THE EFFECTIVENESS OF A CLASSROOM ASSESSMENT TECHNIQUE (CAT) IN THE CALCULUS I CLASSROOM

A Dissertation

by

DAVID A. MURDOCK

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Christine A. Stanley
Committee Members,	Glenda Musoba
	Vicente Lechuga
	Ben Welch
Head of Department,	Mario Torres

August 2018

Major Subject: Educational Administration

ABSTRACT

This study examined the effectiveness of a classroom assessment technique (CAT) in the Calculus I classroom using a quasi-quantitative research design methodology. A single university, single discipline, single professor and single class setting was utilized to minimize confounding variables and locate all data under a single institutional culture. Purposive homogenous sampling was utilized to select Calculus I students for the treatment and control groups. The data were analyzed using statistical methods to test the hypothesized relationships between the treatment group which was the Calculus I students who used the classroom assessment technique, and the control group which was the Calculus I students who did not use the classroom assessment technique.

The problem of too many college freshman that major in science, technology, engineering or mathematics (STEM) and fail to succeed, is linked to the lack of mathematical skills. More specifically, the low retention rate of freshmen STEM majors corresponds to the lack of student success in Calculus I, which is typically the first math course for the freshmen STEM major. If calculus is the linchpin to success in the STEM disciplines, then how it is typically taught and assessed has significance in understanding and mitigating this problem.

There is a dearth of studies in the literature regarding the effectiveness of certain classroom assessment techniques posited by past researchers, and previous studies underscore the need for additional research. At present, there are no known studies of these specific techniques in the Calculus I classroom. This study intended to add to the literature of higher education by analyzing the effectiveness of a Classroom Assessment Technique (CAT) in the Calculus I classroom and comparing the findings to previous studies. The ultimate intention of

this study was to contribute to the national chronicle that seeks to improve student learning and student success in the STEM disciplines. Results of the analysis were null findings. The final results indicated that the Classroom Assessment Technique (CAT) tested did not make a significant difference.

DEDICATION

I love to learn. I am indebted to many who sustained me through this long journey of learning. My wife Chesnee made the most sacrifices; for many nights after a long day of work herself, she would call my office and say: “Please stay and work as long as you need to on your dissertation; me and the children are fine.” Thank you Chesnee for sustaining me through this journey. This dissertation is dedicated to you and our daughters Wemberly and Rebekah. My prayer is that our daughters will love learning as much as I do.

ACKNOWLEDGEMENTS

There are far too many people to thank and acknowledge in the space that I have here. First, I must thank my family for their support and encouragement. My wife Chesnee made many sacrifices without complaining, and my daughters encouraged me to write even when they wanted me to spend time with them. Chesnee, Wemberly and Rebekah, I am so grateful for your encouragement, love and support. I would have never reached the finish line without you. I would also like to thank Mark and JoAnne Batson for their prayers and financial assistance for my family during this long journey.

I also owe a tremendous amount of gratitude to my supervisor Bethany McCraw. Bethany not only said yes to every day I asked to be off work to write, but she also encouraged me and celebrated the victories with me along the way. I cannot thank Bethany enough for her enthusiasm and support for this dissertation.

Many other individuals helped me by offering their expertise and sound advice along the way: Dr. Jack Tubbs, Divya Lakshminarayanan, Melinda Sanson, Deborah Holland, Joyce Nelson, Dr. Don Gehring and Dr. Yvonna Lincoln just to name a few. These individuals were invaluable and a source of comfort when I struggled with some of the more difficult questions of this study. Dr. Jack Tubbs was particularly instrumental in helping me understand the limitations of this study.

I would like to thank my committee chair, Dr. Christine A. Stanley, and my committee members, Dr. Glenda Musoba, Dr. Vicente Lechuga, and Dr. Ben Welch for their constant

support and encouragement. Dr. Stanley and Dr. Musoba acted as if I was the only doctoral student they had; both were so quick to read and respond to e-mails and chapter drafts.

And finally, I want to thank all of my professors in the program. I benefited from every class and never felt than any class assignment was a waste of time. Dr. Homer Tolson for example is such a gifted teacher. I looked forward to every Statistics class because I knew Dr. Tolson would keep the classes fun and interesting. Dr. Candace Schaefer's Literature Review class and Dr. Yvonna Lincoln's Proposal class were beyond invaluable to me at the beginning of this study. With respect to all of my classes and professors, I cannot imagine having a better educational experience anywhere else.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This study was supported by a dissertation committee consisting of Professor Christine A. Stanley [advisor] and Associate Professor Glenda Musoba of the Department of Educational Administration and Human Resource Development, Associate Professor Vicente Lechuga of the Department of Educational Administration and Human Resource Development, and Assistant Dean for Executive Education and Clinical Professor Ben Welch of the Department of Management.

All work for the dissertation was completed independently by the student.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Effective instructional design and assessment are critical variables for higher education to succeed in its mission of preparing students for the 21st century. Massive tertiary enrollment increases compounded with decreases in public resources underscore the need for greater efficiency and effective accountability measures at every institution of higher education. When individual students are successful in their chosen majors, then the institution is successful on this particular mission critical output. There are many variables that factor into the equation of student success and some of these variables are beyond the control of the institution such as individual student talent, giftedness and internal motivation to succeed. Effective instructional design and classroom assessment techniques are within the control of the institution and this responsibility lies primarily with the individual faculty members.

So what constitutes effective instructional design and effective classroom assessment? The higher education literature is replete with articles and research studies that are intended to illuminate and answer this question. Faculty seek the answer to this question in order to provide optimal student learning and students, whether they realize it or not, are also interested in this question because they also desire optimal learning and their own success. Although there is agreement that effective instructional design and assessment are critical for higher education to succeed, there are also different perspectives and variation with respect to the application of the instructional design (classroom teaching) and how to measure the effectiveness of the design (classroom assessment).

The concept of effective instructional design and assessment has not and does not change; however, the language and different approaches to illuminate and explicate this concept does vary. Moreover, teaching methodologies and the assessment and evaluation of these different methodologies, is a dynamic process with respect to the context. When the context and the learning participants change, the instructional design and how it is assessed must also be reexamined and modified to fit the new context. Effective instructional design and assessment cannot be a set routine and the only way to remain on the cutting edge and continuously improve is to constantly assess and evaluate each time the context changes. The context is usually multidimensional; for example, the context may include the specific setting, the specific learning needs and learning styles of the students being taught and finally, the professor's needs assessment and sometimes personal preferences and skill set.

One crucial nexus between effective instructional design and assessment is the alignment of the two. When the intended learning goals, along with the instructional methods, are properly aligned with the assessment techniques, the measurement and evaluation of the goals are more structured and objective which equals greater internal validity for the process. Moreover, Kaplan, Meizlish, and O'Neal (2007) found that aligning the intended learning outcomes (goals), instructional methods, and assessments (testing) can lead to significant gains in student learning. Of course, there are many other aspects of effective instructional design and assessment. For example, Angelo and Cross (1993) stress assessment techniques that are continual and based on collecting feedback regarding student learning in the local classroom context.

It is well understood that for higher education to survive and prepare students for the 21st century, every institution of higher education needs to achieve the highest possible quality of student learning. Effective instructional design and effective assessment are two key variables

for every faculty member. The teaching profession is not only a high calling but one of grave importance for creating and maintaining an educated citizenry. Both faculty and students have equally important roles to play in the learning process and effective instructional design and assessment can empower both faculty and students to improve the quality of learning in the classroom (Angelo & Cross, 1993).

Statement of the Problem

Currently, one area of much needed improvement in higher education is student success in the Science, Technology, Engineering and Mathematics (STEM) disciplines. Nationally, only 40% of incoming freshman STEM majors persist and obtain a STEM degree (Young et al, 2011). Moreover, additional research suggests that student interest in the STEM disciplines is either declining or remains constant while the economic demand for STEM workers is increasing. According to Carnevale, Smith, and Melton (2011), a meta-analysis of studies demonstrated that the work-force demand for STEM majors has been increasing from 1971 to 2009 while the number of students pursuing STEM majors in the US has remained constant at around 30%. Even more recently, a report from the US President's Council of Advisors on Science and Technology (PCAST 2012), indicated that the US economy will need an increase in STEM workers for continued economic growth (Ellis, Kelton, and Rasmussen, 2014). Therefore, the problem of low STEM major success in higher education is compounded by declining student interest in the STEM disciplines and high attrition rates.

The reasons undergraduate students fail to persist in the STEM disciplines has been a popular topic of interest for past researchers. Seymour and Hewitt (1997) posited the following general reasons from their work in a book *Talking About Leaving: Why Undergraduates Leave the Sciences*: 1) students experience difficulty in the introductory courses and lose confidence in

their ability to complete the program due to low grades, 2) the STEM culture is highly competitive and some students do not respond well or strive in a highly competitive environment, 3) some students find the curriculum and degree plan overwhelming and too fast paced, 4) ineffective teaching by STEM faculty members, 5) poor advising and lack of support for students who need extra help.

Other researchers have narrowed down this list of reasons posited by Seymour and Hewitt (1997) and determined that mathematics is the key skill for success in the STEM disciplines and even more specifically, Calculus I and II is the basic foundation. Therefore, student experiences in Calculus I and II can be the linchpin to success for STEM majors. The pedagogical strategies employed in introductory mathematics courses can make or break a student's STEM career. Seymour and Hewitt (1997) also posited that unsuccessful STEM students reported uninspiring classroom instruction that lacked an emphasis of conceptual understanding and authentic application as one of the major reasons for changing their majors. Similarly, Pampaka et al., (2012) found that students who opted not to continue their math education beyond the minimum requirements, was intensified by the transmission teaching perspective. The transmission teaching perspective (more commonly referred to as lecturing) is probably the most common and well-known in primary, secondary and post-secondary education. According to Pratt, (2002), "the purpose of teaching, from a transmission perspective, is to move knowledge or skills from a content expert (teacher or text) to the learner" (p. 65) and teacher expertise and efficient content delivery are the main defining attributes of the transmission perspective of teaching.

Background of the Study

The problem of too many freshman STEM majors failing to succeed and obtain a STEM degree is not just a problem for STEM faculty members or higher education. This problem is multi-faceted and has large implications. One important implication is the US economy and American global competiveness. According to several studies, the economic demand for STEM workers has been consistently increasing while the number of freshman students pursuing a STEM career has remained steady or declined and this is exacerbated by the attrition rate for freshman STEM students which is far too high to keep up with the current economic demand.

Some of the causes of this problem are also multi-faceted and beyond the reach of STEM faculty members and higher education administrators; however, effective instructional design and assessment practices are two variables that are within the reach of faculty members and higher education decision makers. Seymour and Hewitt (1997) asserted that mathematical skills are crucial for student success in the STEM disciplines. More specifically, Calculus I and Calculus II are the foundational building blocks for success in the STEM disciplines, and Brown (1991) found that Calculus I has been determined to be a major roadblock for freshman STEM majors.

One disquieting aspect of this problem is that researchers in the US and other countries found that first year university mathematics courses (which would be Calculus I for most STEM majors) are often used as a filter to prevent underprepared students from pursuing a STEM career. Whether this is good or bad practice for higher education is debatable but it is clear that this cooling out function (Clark, 1960) increases the overall STEM attrition rate with respect to the needs of the US economy and with respect to the desire for institutions of higher education to successfully graduate as many students as possible. On the one hand, it is true that some

students who enter a STEM discipline may not have the necessary talent and motivation to succeed, and they should pursue a different major. On the other hand, some of these initial underprepared students may have the necessary talent and motivation to succeed and they just need extra help or perhaps a modified instructional design in order to succeed.

The fact that most introductory first year university mathematics courses are taught via large lecture due to practical and economic reasons does not necessarily have to be viewed as problematic. In his book, *How to Teach Mathematics,* Krantz (1999) asserted there is no data to support the claim that small classes are better than large classes for teaching mathematics; however, studies suggest that students in the smaller classes feel better about the class, they also seem to enjoy the smaller classes more and finally, they perform better and persist at higher rates than students in larger classes. Additionally, the increasing number of students enrolling in higher education and decreasing resources for higher education will most likely continue to make the large lecture a popular choice for institutions grappling with efficiency and maintaining quality with increasing quantity. Moreover, the large lecture appears to be deeply engrained in the higher education landscape; researchers have found that some students are fond of lecturing and even prefer it over other pedagogical strategies.

The use of the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) appear to be limited in the STEM disciplines. Many faculty members design the lecture for the maximum amount of content to be transmitted to the students during the 50 or 75 minute time period. The desire to maintain control of the class, so that the prescribed content can be transmitted efficiently, are salient characteristics of the large lecture and although large lecture in some disciplines seems to work well, the STEM classroom where the attrition rate is approximately 60% (Young, et al, 2011), is an area for potential improvement. Moreover, the

assessments used for calculus classes are usually summative which occur at the end of the instruction as opposed to formative assessments which take more time but allow the students the opportunity to correct and learn from their mistakes without a grade penalty. If formative assessments are truly less common in the STEM disciplines and more specifically, in Calculus I and Calculus II, this could be an area for potential improvement and may positively impact the problem of low success rates for freshman STEM majors.

There is clearly tension between the primary pedagogical strategy (large lecture) in the calculus classroom and the low success rate of freshman STEM majors nationally. Given this tension, additional studies are needed to further explore possible solutions to this very important problem that has national implications. The use of the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) is worth exploring in the Calculus classroom.

Purpose of the Study

With respect to the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993), the broad and central question of prior studies found in the literature is: Do Classroom Assessment Techniques (CATs) improve student learning? The purpose of this dissertation study is to build on this prior research by analyzing the effectiveness of one CAT (*Documented Problem Solutions*) in the Calculus I classroom. If this CAT is shown to be effective with respect to student learning, then perhaps STEM Faculty members might take note and incorporate *Documented Problem Solutions* in their instructional designs. The ultimate desired outcome is to positively impact student success in calculus and therefore, improve freshman STEM major success and positively impact retention in the STEM disciplines. The research objective of this study was to test the effectiveness of the Classroom Assessment Technique, *Documented Problem Solutions*, in the Calculus I classroom.

Significance of the Study

Scholars and educators seek to better understand and improve the classroom on a continual basis, and this concept is generally referred to as classroom research. Classroom research is a subset of the larger action research movement of the 1940s. Action research is intended to produce results that have clear implications for action. In other words, the goal of action research are results that are likely to produce meaningful change and improvement. In 1983, President Ronald Reagan's National Commission on Excellence in Education published a report called: *A Nation at Risk* which is considered to be watershed moment in American education because this report spurred reform efforts at the local, state and federal level.

In general, in the higher education context, classroom research is seen as more pragmatic and less scholarly than other strands of research. For example, classroom research focuses on the local context and this research seeks to illuminate topics of imminent interest to faculty in order to improve teaching and student learning. Classroom research usually does not consist of highly formal studies with large sample sizes across many institutions of higher education with the goal of a large inference space and generalizability back to the population of higher education or a particular sector of higher education; for example, community colleges or public institutions. Moreover, classroom research is not likely to win research dollars, and it is also not likely to help faculty members with their publishing requirements. Rather, the focus on classroom research is classroom assessment and evaluation with the limited goal of better teaching and learning.

According to Cross (1998), classroom assessment and classroom research play different but complementary roles in improving student learning in the classroom. Classroom assessment has two facets: 1) assessment for accountability and 2) assessment for improvement. Assessment

for accountability purposes usually involves reporting assessment information to external parties such as accreditation agencies or a state government or perhaps the taxpayers. Assessment for accountability is usually about satisfying predetermined requirements from an external party. Assessment for improvement is more closely linked with classroom research because the goal is very different; to increase the quality and effectiveness of student learning. Moreover, the classroom research component usually involves students in the assessment process as opposed to making students the object of the assessment process. "In the final analysis, it is teachers and students, working together, who are ultimately responsible for the quality of learning" (p. 6).

Classroom research is usually focused on two types of questions regarding student learning; "why" and "how" questions such as, why did students miss the main point of the instruction, or how can the instruction be revised to better hold the students' attention. Cross (1998) asserts that if faculty members are serious about their responsibilities to maximize student learning, then they must know more about how students learn. Moreover, Cross posited there is an urgent need for more research on teaching and learning in the disciplines because there is great variety across the disciplines in higher education. "Teachers of a given discipline—whether male or female, full time or part time, experienced or inexperienced, teaching in a public community college or a private four-year college—share a value system with respect to teaching goals that is distinctively discipline-related and significantly different from that of colleagues in other disciplines" (Cross, 1998, p. 8).

Classroom research also stands for the proposition that no one is better qualified to do this research than faculty members who know their disciplines and are familiar with the problems students have in their classrooms (Cross, 1998). Additionally, classroom research is different from traditional research because it is not an add-on activity; rather, it is usually

imbedded in the classroom activities. Faculty members and their students have the mutual goal of improving learning in the real world of the classroom. Cross (1998) also posited that "at its best, Classroom Research involves students as collaborators rather than subjects in the research. Knowledge about human learning, especially their own, is of high value and high interest to students" (p. 9).

Additional characteristics of classroom research include pertinence and relevancy. In classroom research, the faculty member decides the research questions that are relevant with respect to the students, class, and discipline. Traditional research may not reach the same degree of relevancy for a number of reasons. Cross (1998) asserted that traditional research has been ineffective in improving student learning because faculty are too busy to read research articles that seem to result in equivocal findings that may or may not apply to their teaching and classroom goals. Moreover, traditional research may only indicate that certain things seem to be related or are correlational, but classroom research may yield more powerful findings from less participants due to higher relevancy. And finally, traditional research is often about fixed characteristics such as age or ethnicity, but classroom researchers are more interested in malleable characteristics such as pedagogical strategies and study techniques for students (Cross, 1998).

The Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) are an even narrower strand of classroom research. On the one hand, some of the arguments for studying CATs include positive student responses, especially when the students believe their input is considered in the course design. Additionally, CATs tend to generate intellectual excitement and better collegiality between faculty members and their students. On the other hand, some of the most common arguments against studying CATs include costs and time

constraints; faculty members are keenly aware of the need to cover the course content within the time constraints of each class period. Idiosyncratic findings and informal methodologies are additional arguments typically leveled against studying CATs. Most teaching faculty are not formal learning researchers, nor are they social scientists, and as a result, CATs research is viewed differently than collaborative action research that involves teachers and social science researchers for example. In general, teaching faculty are more likely to participate in departmental or disciplinary associate classroom research programs or attend a Teaching Center seminar designed to help teachers relate their personal theories to formal learning theories.

Although the work of Angelo & Cross in their book *Classroom Assessment Techniques: A Handbook for College Teachers* (1993) is popular and viewed as canonical in the higher education community (Simpson-Beck, 2011), there are not very many studies in the literature that examine the validity of these 50 formative CATs. There are strands of studies that seem to validate some of the CATs in some classroom contexts but there are other studies; for example, Cottell & Hardwood (1998) and Simpson-Beck (2011) that did not validate the efficacy of the CATs studied. Moreover, the majority of the studies appear to be qualitative research designs or only anecdotal observations; there are only two quantitative studies in the literature, so the quantitative research design lens is even more limited. Currently, there are no studies in the literature regarding the use of CATs in a Calculus I or Calculus II classroom. More research regarding the use of CATs in the STEM classroom and more specifically in the calculus classroom may at least shed light on the question of whether or not CATs might be a possible solution or positive influence on freshman STEM major success rates. More quantitative studies regarding Calculus I and Calculus II instruction could provide balance to the current dearth in the literature of the quantitative research design lens and also, if future studies confirm the efficacy

of CATs in the calculus classroom, more STEM faculty members might take note and incorporate CATs in their instructional designs.

This study intended to add to the literature of higher education by analyzing the effectiveness of a Classroom Assessment Technique (CAT) in the Calculus I classroom and comparing the findings to previous studies. Just as the Cottell & Harwood (1998) and Simpson-Beck (2011) studies drew comparisons of their findings to previous studies, this present study was intended to further expand the study of a CAT within the specific context of Calculus I and contribute to the national chronicle that seeks to improve student learning and student success in the STEM disciplines. If the findings are positive, this study at least has the potential to impact the instructional design and assessment of Calculus I. Additionally, an improvement in the instructional design and assessment of Calculus I, has the potential to increase student success and therefore, decrease the number of students who fail Calculus I and drop out of the STEM disciplines.

Rationale and Justification

The gap found in the higher education literature is a dearth of quantitative studies that test the effectiveness of a Classroom Assessment Technique (CATs) posited by Angelo and Cross (1993). After an extensive and nearly exhaustive review of the literature, only two quantitative studies were found and each study called for additional research using the quantitative research design. Therefore, a quantitative research design was selected in order to attempt to fill this gap. Additionally, the two past quantitative studies in the literature indicated that the CATs evaluated were not helpful, yet other research design studies indicated the opposite result. This current study was needed to bring some degree of clarity to this quandary.

The Cottell and Harwood (1998) study tested six different CATS based on one semester of classes taught by two professors at two universities. Additionally, the researchers made general midcourse adjustments, based on feedback from the CATs sections, to both the control sections and the treatment sections which introduced a bias against the CAT treatment sections: "We decided to go ahead and make those changes, even though we realized we introduced a bias against finding the result that CATs improved learning" (p. 41). In addition to midcourse adjustments to both the treatment and control groups, testing multiple CATs simultaneously as Cottell and Harwood did, makes it difficult to isolate which CAT made or did not make a difference with respect to student learning. This feedback from Cottell and Harwood and other researchers influenced the research design of the current study.

Population and Sample

The population is Calculus I students in a single university setting. The sampling of choice in the quantitative research design is pure random sampling so that causation may be inferred and generalized back to the population; however, this is not always possible in the social sciences (Rudestam & Newton, 2001). Given that Calculus I students are not randomly assigned to class sections, this study must be a quasi-experimental research design that compromises some of the rigor of the controlled experiment but maintains the argument and logic of experimental research (Cook & Campbell, 1979). In summary, this research design was a one discipline (Calculus), one professor, one-class setting (Calculus I) model which offered the best research design to measure whether or not one specific Classroom Assessment Technique (*Documented Problem Solutions*), which is one specific instructional method, will improve student learning.

Data Collection

The independent variable (IV) for this study was the Classroom Assessment Technique (CATs) of *Documented Problem Solutions.* This CAT focuses on critical thinking and the steps taken to solve a calculus problem. Students must show every step (document) of their problem solving process in reaching the final answer (solution) to a calculus problem. The treatment section (s) of Calculus I required all students to use *Documented Problem Solutions* in solving their homework problems. The control section (s) of Calculus I did not require the students to utilize *Documented Problem Solutions* in completing the same homework problems.

The dependent variable (DV) for this study was the student outcomes for the course which consisted of overall course grades. The underlying assumption is that grades measure the competencies that students develop over the course of the semester. The summative student outcome measures for the Calculus I course in this study were: Four examinations, multiple quizzes and one standardized departmental final examination (written by the calculus professors in the math department and graded by the math department calculus professors) and the final class averages. A second assumption is that the student characteristics in the Calculus I class sections are similar to a degree sufficient for comparison with respect to the independent variable.

Data Analysis and Testing Procedures

The researcher made use of International Business Machines (IBM) Statistical Package for the Social Sciences (SPSS), version 24 software. The treatment and control groups were compared with respect the American College Test (ACT) Math scores and ACT Composite scores to ensure the participants were similar enough for comparison with respect to the dependent variable. Moreover, an analysis of covariance (ANCOVA) was performed to control

for any significant differences with respect to mathematical competencies for the Final Examination student outcome. And finally, a two group independent sample t-test was utilized to evaluate each research hypothesis ($H_{A1} - H_{A6}$) with respect to statistical significance.

Research Hypotheses

The research hypotheses identify the questions that the researcher wants to test; moreover, the research hypotheses define and operationalize the study's variables (Calabrese, 2006). This study consisted of the following research hypotheses:

Research Hypothesis 1 ($H_{A1:}$): There will be a significant difference between the CAT section and the control section with respect to Examination One.

Research Hypothesis 2 ($H_{A2:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Two.

Research Hypothesis 3 ($H_{A3:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Three.

Research Hypothesis 4 ($H_{A4:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Four.

Research Hypothesis 5 ($H_{A5:}$): There will be a significant difference between the CAT section and the control section with respect to standardized departmental Final Examination.

Research Hypothesis 6 ($H_{A6:}$): There will be a significant difference between the CAT section and the control section with respect to the Final Course Grade.

Research Objective

The objective of this study was to: Test the effectiveness of the Classroom Assessment Technique, *Documented Problem Solutions*, in a Calculus I classroom.

Definition of Terms

Classroom Assessment Techniques (CATs): For this study, this term is limited to the 50 classroom assessment techniques posited by Angelo and Cross (1993):

> Classroom Assessment is a systematic approach to formative evaluation, and Classroom Assessment Techniques (CATs) are simple tools for collecting data on student learning in order to improve it. CATs are "feedback devices," instruments that faculty can use to find out how much, how well, and even how students are learning what they are trying to teach. Each Classroom Assessment Technique is a specific procedure or activity designed to help faculty get immediate and useful answers to very focused questions about student learning (p. 25).

Documented Problem Solutions: The Documented Problem Solutions classroom assessment technique prompts students to keep track of the steps they take in solving a problem. Students are required demonstrate how they worked the problem by showing each solution step, and by analyzing these detailed protocols—in which each solution step is documented, teachers can gain valuable information on their students' problem-solving skills (Angelo and Cross, 1993).

STEM Disciplines: STEM (an acronym for science, technology, engineering and mathematics) has its origins in the 1990s at the National Science Foundation (NSF) and has been used as a generic label for any event, policy, program, or practice that involves one or several of the STEM disciplines (Bybee, 2010). In general, STEM disciplines include: life sciences, technology related disciplines, all engineering disciplines, and mathematic disciplines including statistics and computer programming.

Calculus I: Course Description: Differential calculus of a single variable. Introduction to the definite integral and the Fundamental Theorem of Calculus (Nichols, 2017).

Classroom Research: Broadly defined, Classroom Research focuses on the "why" and "how" questions with respect to student learning; it differs in many ways from traditional research in or on classrooms because it is not viewed as an add-on activity, rather Classroom Research is embedded in the regular ongoing work of the class (Cross, 1998). Classroom Research encourages teachers to use their classrooms as laboratories for the study of learning (Cross and Steadman, 1996).

Quasi-Experimental Research: In this design, the researcher does not utilize experimental manipulation or random assignment of participants to conditions because the conditions have already taken place or they are not able to be manipulated (Kerlinger and Lee, 1999). This design is also referred to as "ex post facto research" (Rudestam, and Newton, 2001). Moreover, this approach is more popular in social science research because pure randomization is often not feasible in the social sciences. Cook and Campbell (1979), asserted that this design compromises some of the rigor of the pure experimental design but maintains the argument and logic of experimental research.

Experimental Research: In a pure experimental design, comparison is used to determine differences between two or more groups (or classes of participants). In this design, the researcher "sets the table" and attempts to isolate all known variables (independent, dependent, and extraneous) so that inferences can be made (based on the laws of probability) regarding the variables of interest. When the comparison groups are drawn from the population of interest via pure random sampling and all known variables are isolated, then the researcher may claim (or infer) that the resultant differences were due to the treatment variable. With the proper controls, the researcher may infer a causal relationship and with random sampling, generalize the results back to the original population (Rudestam, and Newton, 2001).

Independent Variable: For the purpose of this study, the independent variable is the Classroom Assessment Technique, *Documented Problem Solutions.* More generally, the independent variable is also referred to as the "classification variable" or the "categorical variable" and it does not vary as a result of another variable. It is "independent" of other variables. The independent variable is changed or controlled by the researcher to test the effects on the dependent variable (Tolson, 2012).

Dependent Variable: For the purpose of this study, the dependent variable is the summative student outcome measures for the Calculus I class. More generally, the dependent variable is the variable being tested and measured; the dependent variable is "dependent" on the independent variable. In other words, as the independent variable is changed or manipulated, the effects on the dependent variable are observed and recorded (Tolson, 2012).

Null Hypothesis: The null hypothesis (H_0) is a statement of no difference (thus Null) for a parameter (s) (Tolson, 2012).

Research Hypothesis: The research hypothesis or alternative hypothesis (H_A) is a statement of anticipated difference or some other relationship between variables. The research hypothesis is presented as $\neq$, $<$ or $>$. The $\neq$ option indicates that both ends of the distribution are of interest (thus, a two-tailed test). The $>$ or $<$ options indicate a one-tailed test is of interest (Tolson, 2012).

Summative Student Outcomes: For the purpose of this study, the summative student outcome measures were: The Calculus I Examinations 1 – 4, the standardized departmental Final Examination and finally, the Calculus I Final Class Grades. More generally, summative assessments are one-time assessment exercises based on a criterion. Summative assessments are also referred to as "Traditional Assessments," and Cross (1998) posited that these assessments

"tell students----often too little and too late—how they have done on a test in a course, but not how they are doing as learners" (p. 6).

Formative Assessments: Formative assessments increase student knowledge and learning by providing on-going corrective information throughout the learning process either with or without grade penalty (Goldstein, 2007). For the purpose of this study, the CAT, *Documented Problem Solutions,* is a formative assessment technique.

CHAPTER II

LITERATURE REVIEW

The purpose of this literature review is to provide an overview of the pertinent literature regarding how Calculus is typically taught and assessed in higher education. This review is organized into three sections. The first section provides a very brief overview of calculus instruction and assessment in the higher education setting. The second section addresses some of the major pedagogical strategies employed in the STEM disciplines. The third section focuses on classroom assessment; more specifically, the Classroom Assessment Techniques (CATs) posited by Angelo and Cross, (1993) and how some of these techniques might be considered to improve success in the calculus classroom.

How Calculus Is Typically Taught and Assessed

One pedagogical strategy often used with confidence in undergraduate mathematics courses is lecturing (Kensington-Miller, Sneddon, Yoon & Stewart, 2013). According to Bergsten (2007), the typical undergraduate mathematics course is taught via lecture as lectures have a long history in American higher education as being both practical and economical. Lectures consist of a subject matter expert usually standing behind a podium and wearing a microphone delivering an oral presentation to a large group of students with the students sitting in rows listening to the teacher describe mathematical ideas and methods while taking notes (Beswick, 2005). This description from Beswick matches the description of other researchers such as Phillips (2005) who asserted that lecturing has been the dominant mode of teaching in not only mathematics but in most university disciplines. Many researchers cite tradition, practicality, and economics as the main reasons for the popularity of lecturing in higher

education; however, other reasons are prevalent in the literature. Yoon et al., (2011), found that students prefer lecturing even while simultaneously acknowledging that they do not expect to learn much from this pedagogical strategy.

The rich history and strong popularity of lecturing in higher education has not thwarted researchers from pointing out the shortcomings of this pedagogical approach. In the large lecture, the range of individual student characteristics and learning preferences can be quite broad, which creates challenges for the professor seeking to reach every student with effective instruction. Biggs (1999) found that the practical problems faced by the teacher and the students are directly proportional to class size. In other words, as the class size increases, the problems faced by both teacher and students increase and change in nature. Much of the criticism of lecturing falls into the category of problems that generally have negative effects on student learning (Mulryan-Kyne, 2010). Additional research regarding the challenges and limitations of the large lecture in higher education is plentiful and well known. More specifically, some researchers have examined introductory mathematic large lectures. For example, a study conducted by Hourigan & O'Donoghue (2007) suggested that in undergraduate mathematics lectures, students often become passive listeners and they experience difficulty in comprehending the content and paying attention throughout the lecture.

The literature regarding how calculus classes are taught is plentiful; however, the literature regarding how calculus students are assessed is sparse. More generally, Goldstein (2007) found that college teachers often utilize and provide students with summative evaluations of their work as opposed to formative evaluations. Examinations either periodic, mid-term or final that can be graded using a Scan Tron form are a popular summative assessment technique for large lecture classes. Cross (1998) asserted that traditional assessment techniques "tell

students----often too little and too late---how they have done on a test in a course, but not how they are doing as learners" (p. 6). In contrast to summative assessments, formative assessments are rarer in higher education (Goldstein, 2007). Goubeaud (2010) found that "assessment practices that could be considered formative with potential to promote student learning appear to be underutilized by all science faculty" (p. 1). Formative assessments can be graded or not graded but they are fundamentally different from summative in that they provide students with an ongoing evaluation of their learning and mastery of the content. One of the biggest advantages of formative evaluations is timing; formative evaluations are typically completed during the instruction (as opposed to the end) and therefore, both the student and the teacher can make adjustments designed to improve student learning.

Instructional Pedagogy in the STEM Disciplines

The pedagogy of university teaching and subsequent student learning is a popular topic in the higher education literature. Pedagogy is generally defined as the art, science, or profession of teaching which is where the old adage of "teaching is both an art and science" stems from. All pedagogical strategies are based on learning theories and all good educators seek to maximize learning in their classrooms. The literature is quite broad regarding the topic of learning theory but in general there are two broad categories of learning found in the literature: deep learning and surface learning (Walker, 2012).

Most higher education faculty seek to achieve deep learning, which involves critical thought about new ideas and concepts and how these ideas connect to prior knowledge. Deep learning connotes wisdom and the ability to discern the nature of knowledge, the limits of knowledge and finally, how to apply new knowledge in different or unfamiliar contexts to solve

real world problems. In contrast to deep learning, surface learning involves less critical thought and the mere acceptance of facts and memorization.

Although there are only two broad categories of approaches to learning (Deep vs Surface), there are other descriptors and theories of learning found in the literature regarding how learning is achieved. For example, on the one hand, the term passive learning usually connotes students sitting quietly in a classroom and receiving information from the instructor. On the other hand, the term active learning connotes students being more involved in the instruction by asking questions, making mistakes, correcting mistakes, and learning from each other. Passive learning and active learning correlate with the pedagogical strategies of teacher-centered learning and student-centered learning respectively. Although the terminology varies over time (sometimes according to the latest fad) the general ideas and concepts are the same regarding student learning.

The research in this area is important because there is a clear need to increase student learning and persistence in the STEM disciplines (Ellis, Kelton, and Rasmussen, 2014). Again, nationally, only 40 percent of the incoming freshman STEM majors are successful in earning a STEM degree (Young et al, 2011). Moreover, the subject matter and the type of learning sought are not the only variables in selecting a pedagogical strategy; there are other variables such as feasibility, economic considerations, instructor preferences, student expectations, and institutional culture. Therefore, the intent of the next section is to examine the current pedagogical strategies associated with the STEM disciplines.

Major Categories of STEM Pedagogical Strategies

The literature regarding university pedagogy is voluminous to say the least and expands over 100 years in time. As a starting point, most educators view pedagogical strategies through

either the passive learning or the active learning lens but teaching and learning is never this simple. There is a substantial amount of research regarding the traditional lecture which is teacher centered and active learning which is student centered. Moreover, regarding calculus, there seems to be more research on how calculus is taught and less on how calculus is typically assessed. In order to examine the questions of how calculus is typically taught and assessed more thoroughly, it is important to take an in depth look at some of these prior studies.

For example, researchers at the University of Maine found that faculty members teaching STEM courses cannot simply be classified into two groups: (1) traditional lecturers or (2) instructors who teach in a high interactive manner (Smith, Vinson, Smith, Lewin, & Stetzer (2014). According to prior research over many decades, active learning approaches are very effective in the STEM disciplines. For example, a meta-analysis of studies from 1942 to 2009 by Freeman et al. (2014) indicated that students learn more and are more likely to persist in science, technology, engineering, and mathematics (STEM) courses that use active-engagement instructional approaches. This is a significant period of time (67 years) and these studies seem to validate the notion that active learning pedagogical approaches have the potential to improve student learning in the calculus classroom.

In light of this research regarding the effectiveness of pedagogical strategies based on active learning techniques, researchers at the University of Maine sought to examine and quantify how many STEM faculty members were utilizing active learning in their classrooms. The researchers cited the lack of systematic data being collected by institutions of higher education regarding how many faculty members were using active learning (Wieman and Gilbert, 2014) and secondly, the need to establish baseline data to inform policy decisions regarding STEM instruction as justification for this particular study (Smith et al., 2014).

Therefore, the ultimate goal was to improve professional development opportunities for STEM faculty members at the University of Maine and elsewhere.

The researchers used purposive and convenience sampling to recruit STEM faculty members in 13 different departments for this study. The faculty members were asked to allow middle and high school STEM teachers to come in and observe their teaching. The researchers also appeared to use some degree of deception because they told the faculty members via e-mail: "the teachers were helping to capture a snapshot of STEM instruction at UM; we did not describe the observation protocol being used or what instructional practices the teachers were capturing until the study was complete" (Smith et al., 2014, p. 626). The middle school and high school observer teachers utilized *The Classroom Observation Protocol for Undergraduate STEM* (COPUS) (as posited by Smith et al., 2013) to observe 51 STEM courses, 44, of which were at the introductory level. In addition to the classroom observations, all faculty member participants were surveyed using a *Teaching Practices Inventory* developed by Wieman and Gilbert (2014) and all statistical analyses were conducted using IBM, SPSS software (Smith, et al., 2014)

The hypothesis for this study was that STEM faculty members at the University of Maine would fit into two pedagogical categories: (1) those who primarily lecture their students (passive learning) and, (2) those who use a variety of active learning techniques and lecture very little (active learning). The results of the collapsed COPUS instructor codes across all 51 STEM courses observed did not support this hypothesis. Instead, the researchers observed a continuum of teaching techniques. More specifically, the researchers observed a continuum from 2 to 98% with respect to the collapsed code of Presenting (P) defined as: Lec: "Lecturing or presenting information and RtW: Real-time writing and D/V Showing or conducting a demo, experiment, or

simulation" (Smith, et al., 2014, p. 627). Therefore, faculty members were not able to be classified into the two pedagogical categories in the hypothesis.

A second hypothesis was to explore whether class size correlated with certain pedagogical strategies; more specifically, whether or not the larger classes would correlate with higher percentages of lecturing. From this sample, the researchers did find that instructors of larger classes tended to lecture more; however, the correlation was a weak positive correlation and the overall findings were mixed. For example, some larger classes had a wide range of pedagogical strategies and some smaller classes had less variety and more lecturing.

The researchers were interested in quantifying how many STEM faculty members were actually using active learning pedagogical strategies in light of previous research demonstrating the effectiveness of these strategies in the STEM disciplines. Moreover, can faculty members be distinctly classified based on passive learning techniques and active learning techniques, and whether or not traditional teaching practices dominate in the STEM disciplines at the University of Maine? The results of this study suggest that the passive learning / active learning classifications are a false dichotomy. The findings demonstrated a broad continuum of pedagogical strategies across a number of different classification variables. Further research is needed to develop a more comprehensive system for classifying pedagogical strategies in the STEM disciplines.

A second implication is that the popular sentiment (and prior research findings) of instructors with larger classes tending to lecture more may not always hold true. It is clear from the literature that some researchers have demonstrated that class size is a reliable predictor of time spent lecturing students (Murray & Macdonald, 1997; Ebert-May et al., 2011); however, this current study found STEM faculty who taught large classes using active learning

pedagogical strategies and also the opposite, STEM faculty who taught smaller classes, presenting and lecturing more. A final implication from this study is that faculty members who have had some success with active learning pedagogical strategies in larger classes would be ideal in leading teaching seminars for junior faculty or new faculty who hold the popular view of larger classes dictating more lecturing (Smith et al., 2014).

Other Studies of STEM Pedagogical Strategies

In US institutions of higher education, introductory STEM classes tend to be large and taught primarily via large lecture. Due to economics, practicality and other reasons, lecturing is the most popular pedagogical approach when teaching a large number of students. Some reasons are fairly simply to understand; for example, lecturing has a long history and is a rich tradition in American higher education. Many students come to the university expecting to be lectured and many faculty (who were lectured as students) perpetuate the cycle by teaching the way they were taught. Another reason that is simple to comprehend is the economic benefits of lecturing. After World War II, college and university enrollments significantly increased as a result of federal programs that made college more affordable and accessible. Generally, higher enrollments will equal larger classes and with large lecture, more students can be taught with only one instructor.

Other reasons for the popular pedagogical approach of large lecture are not as obvious and some appear to be menacing if not borderline nefarious. According to Steen (1988) and Wake (2011), in the US and elsewhere, first year university mathematics courses often function as a filter, preventing large numbers of students from pursuing a career in the STEM disciplines. Clark (1960) has researched and written about a phenomenon he referred to as “cooling-out” at community colleges. The cooling-out function is necessary to point out gaps between career aspirations and the scholastic ability and talents needed to be successful in a particular career.

The efficacy of filtering students or "cooling them out" to use Clark's term via an introductory mathematics course taught via large lecture is a debatable matter.

As a result of the persistence and popularity of the large lecture in higher education, other researchers have conducted studies that focus on individual teacher beliefs about pedagogical strategies and how to best challenge and persuade teachers to consider new strategies or combine a new strategy with their traditional lecture. For example, researchers at the University of Auckland (using data from a research project that investigated the social norms of lecturer-posed questions in large undergraduate mathematics lectures) convinced one teacher to deviate from his normal practice of primarily large lecture and allow class time for students to work interactively on a few questions (Kensington-Miller, Sneddon, Yoon, & Stewart, 2013). The researchers cited prior research that was critical of large lectures as justification for their study.

This study involved only one instructor (Chris) who agreed to examine his beliefs regarding lecturing with a research group of mathematics educators. Chris was teaching a large undergraduate mathematics class and he agreed to change part of the lecture to allow students to work interactively on a mathematical question once during each lecture. The lectures were altered using the questioning technique in calculus and linear algebra on three occasions in 2009, 2010 and 2011 and Chris kept a kept a reflexive journal regarding how his thoughts, goals, beliefs, knowledge, and identity had been affected throughout the study (Kensington-Miller, et al., 2013).

Before this study, the lone study participant, Chris, was a typical traditional mathematics lecturer who taught by transmitting expertise to his students; Chris stated:

> Early on in my teaching career, I would have been reluctant to spend more than a small amount of time making my lectures more interactive. I think I gave good lectures, but for the most part they were 'sage on the stage' style" (Kensington-Miller, et al., p. 8).

Chris also spoke of other common attributes associated with lecturing such as teacher control, efficiency, and the need to cover a certain amount of content. After the study, Chris stated that his personal beliefs regarding lecturing had been challenged, and he began to see the value in introducing new pedagogical strategies in his teaching. More specifically, Chris stated:

> They've got an opportunity to think about what you've just taught them, what you've been talking about at least, to see whether they've learnt anything. So I think that it kind of, yeah, it gets their brains working on a different level. It gets them thinking about the material rather than just sitting and passively listening to it" (Kensington-Miller, 2013, et al., p. 11).

In this study, one mathematics lecturer (Chris) who was a successful and popular teacher, agreed to examine his beliefs about lecturing with a team of other mathematics educators. Chris expressed a desire to be open minded regarding his pedagogical strategy of lecturing, and he stated he was open to more innovative thinking and teaching practices. Although beliefs, habits and deeply engrained preferences regarding university pedagogy are powerful influences, the willingness to engage with other teachers and critically reflect can be helpful in examining teaching practices. The researchers also mentioned that there is often a gap between espoused beliefs and enacted beliefs regarding teaching. The most salient implication is that this framework of working collectively with similar educators "provides a robust pathway for reflecting and testing out lecturers' beliefs about teaching practice" (Kensington-Miller, et al., p. 13).

Other researchers examined ways in which the large lecture may be modified but not fundamentally changed to increase student success. Kajander (2006) studied the pedagogical strategy of using smaller student groups in large mathematics classrooms. Small groups within the structure of a large lecture gives students the opportunity to interact more personally with their peers and provides more hands-on learning and collective problem solving as opposed to all passive listening and note taking. This study was informed by social constructivist theory and the hands-on learning and personal problem solving within the large lecture was based on active learning theory. The author noted that recent reforms in math education prompted the inclusion of more active learning strategies in the classroom: "Engagement, problem solving and communication are often viewed as integral parts of recent reform movements in mathematics education, which promote students actively working together on problems in the classroom in order to develop deeper conceptual understanding" (Kajander, 2006, p. 234).

The premise of organizing students into small groups was based the author's prior observation that students entering post-secondary introductory mathematics courses tend to be insecure regarding their mathematical ability and they are unlikely to be naturally engaged and independent learners (Kajander, 2006). Moreover, in the large lecture setting with the instructor covering the content with little hesitation, insecure students are often hesitant to ask questions in front of their peers. The author found that putting students in small groups was helpful for their mindset and in learning the content: "working with other students with similar concerns has proven to be helpful to many students both psychologically as well as mathematically" (p. 235). Based on these prior observations, the students were organized into learning groups on the second day of classes and each group consisted of six students. The course was taught via large lecture twice a week.

This was a qualitative study and the author primarily used student journals to judge the efficacy of the learning groups. Most students indicated they enjoyed their learning groups and the learning groups were given some degree of autonomy with the course assignments. Individual student assignments were still required, which were problem sets, but group assignments and group presentations were also required and counted toward the final grade of each student. A comparison of initial student journal entries with the final student journal entries revealed that most students had a positive learning experience in the class. The student journals suggested that sometimes students find working in groups an initial challenge, but usually worth it in the long run. One student wrote:

> We struggled to get started on this problem and it was a great frustration. I think my concern over starting the problem was because we were working in groups. I am sometimes better if I am left to work on my own, but in this case I was wrong. It was because of our group that we found the pattern… This took many ways of viewing the results to achieve success (Kajander, 2006, p. 240).

The main implication of this study is that the large lecture pedagogical strategy need not be completely changed or shelved to improve student learning in the mathematics classroom. The challenges that are inherent to the large lecture structure may be mitigated by incorporating other pedagogical strategies. In this study, the small student learning groups appeared to be effective in mitigating student insecurities and lower levels of engagement in the introductory mathematics classroom. This study underscores the relationship between good teaching and effective pedagogy. Good teaching is more than an eloquent, clear and well-delivered lecture that captivates students causing them to admire and praise the instructor. This can be a valid definition of good teaching, but this study suggests that good teaching and learning includes an

effective combination of pedagogical strategies (whatever this looks like for a particular group of students) and "that content and process are intrinsically related" (Kajander, 2006, p. 240).

Other scholars have examined methods to make the large lecture pedagogical strategy more effective in terms of student achievement and success rates. For example, Millett (2002) sought to systematically improve student achievement in a Pre-calculus course called Mathematics 15 at the University of California, Santa Barbara (UCSA). Millett was disquieted by the fact that the final grade distribution in this course indicated that few students actually mastered the content, and many students were failing to pass the class with the minimum grade of C. Moreover, some majors at UCSA require students to take and pass a "gateway examination" even if they were successful in Mathematics 15 and obtained the minimum grade to pass the course.

Millett (2002) asserted that his teaching perspective is informed by the maxim: "*if students haven't learned, you've not taught,"* and he explicitly rejects the assumption of many lecturers who carp that the cause of any failure with traditional lecturing lies primarily with the students. According to Millett's viewpoint, the widespread failure of large numbers of students to become mathematically proficient cannot be blamed solely on students, and lecturers must make student achievement the principal measure of success in their courses. Therefore, the question he sought to answer is what changes can be made to the large lecture structure to facilitate better results in terms of individual student achievement?

Millett (2002) noted the following observations of his Pre-calculus students based on prior experience: (1) students come with a history of limited mathematical success, (2) students have limited appreciation of their ability to learn and do mathematics, (3) students lack successful strategies for learning math and simply want to "pass the course" (4) students do not

seek an understanding of math sufficient to allow them to apply it in new settings (5) few students recognize the need to change their approach to learning, even with a record of previous failure and, (6) students tend to be passive listeners and they come to the lecture unprepared to learn or unprepared to participate actively in the class. Therefore, the goal of this study was to modify the traditional large lecture structure to address these student deficiencies and increase student learning and success in Mathematics 15 (Millett, 2002).

The following changes to the large lecture structure were implemented: (1) graded homework problems were assigned on a weekly basis to encourage student preparation for the lecture (2) homework grades were combined with "surprise quizzes" grades to count for 25% of the final course grade (3) students who received a 'physical calculus qualifying grade' on the gateway exam before the end of the quarter received a bonus award of 15 points which was added to their final grade (4) students were informed prior to the final comprehensive exam that their final course grade would not be lower than their grade on the final exam. The idea behind this fourth change was to encourage a sustained effort through the end of the final exam, and the beauty of this approach is that it gave students, who may have done poorly on prior class assessments, an opportunity to redeem themselves and their final grades (Millett, 2002).

Moreover, a new textbook was selected and mandatory recitation sessions were added to the large lecture. The recitation sessions were 50 minutes long and they were facilitated by graduate teaching assistants (TAs). The TAs focused on graded homework and answering any questions regarding the homework or material from the large lecture (Millett, 2002). And finally, the seventy-five minute lecture format was modified to be more interactive in style as opposed to the traditional format of pure lecture and content delivery from the instructor. More specifically, students were asked to find a partner to discuss the relevant mathematical concept

for that day, and Millett then selected students to introduce their partner and then summarize their conversation regarding the mathematical concept:

> Finally, I ask about five students to describe their partner's example and to compare it with their own. This exercise requires a fair bit of time but it is used to lead into an explanation of what I am seeking from them by way of interaction with me during the class meeting (Millett, 2002, p. 144).

To assess the results of this study, statistical analysis were conducted on all student assessments in the course. The two exam scores, homework and quiz grades and the final exam scores were listed by class rank. Surprisingly, the graded homework did not make a statistically significant difference with respect to student performance on the final examination. However, the offer of bonus credit for passing the calculus gateway examination was statistically significant with respect to student success on the final examination. Some students in the class were taking the calculus gateway examination due to being required to for their particular major which allowed for a subset and the incentive if 15 extra points for passing the calculus gateway exam appeared to be effective (Millett, 2002).

Millett (2002) posited several consistent student themes based on the interactions with students during this Pre-calculus course: (1) students preferred that the course to be graded on a curve (2) students did not like being asked unfamiliar math questions (3) students failed to make the distinction between memorization of problem solving steps and actually "understanding the concepts" (4) students believed that grades should be based on their efforts rather than their actual performances on the course assessments. Furthermore, "In interviews with students the goal of working problems and studying the course in order to make it possible to solve problems without 'having to think' arose frequently" (p. 150). Millett also noted that: "In looking at

student work on examinations, it seems likely that a careful reading of the problem statement and the question is not occurring" (p. 150).

The implications of this study did not bode well for the students of this particular study. Millett (2002) wrote: "Despite an intensive effort to actively engage all students in productive work on the curriculum, many of the students remained on the sidelines unconcerned gaining the expected understanding of the mathematics" (p. 150). One implication is that although many studies indicate that active learning approaches in the large lecture setting can produce better results, there are always outliers and cross-current studies that suggest the opposite results. Another implication is that regardless of the pedagogical strategy or combination of pedagogical strategies which may include active learning for students, the student has the primary responsibility to participate actively in the course (Millett, 2002). It seems clear from the findings of this study that the level of student participation most likely at every point (class lectures, recitation sessions, homework…etc.) was inadequate. However, the author's maxim at the beginning of the study was: "*if students haven't learned, you've not taught"* and therefore, the best combination of effective curriculum design and assessment for this Pre-calculus class at UCSA remains elusive.

Another study that sought to examine a method of improving the large lecture pedagogical strategy with respect to student achievement and success was conducted by F. Eugene Brown at Northern Virginia Community College in Alexandria, Virginia. Brown (1991) studied the effect of structured notetaking on student success in Calculus I. In this study, the author noted that the typical teaching perspective in mathematics classes is lecturing or the transmission teaching perspective. For example, in most mathematics classes the communication is one-way with the professor writing problems on the board while students listen, take notes and

imitate the correct steps to solve the problems. The class discussion is usually limited to the students who are brave enough to ask questions, and the author also noted that it is rare for the professor to stop and assess student learning. As a result of these typical classroom characteristics of a lecture in mathematic classes from the author's perspective, he desired to test the effectiveness of structured notetaking in the Calculus I classroom.

The main assumption of this study was that taking accurate notes during the lecture was essential for student success in the Calculus I. Additionally, the author noted that many students, by their own admissions, do not look at their notes until it is time to study for the examination, and moreover, some students will not bother to ask clarifying questions even when they know their class notes are incomplete (Brown 1991). Therefore, the idea was to provide more structure for student notetaking by providing a specific notetaking format and more feedback by evaluating and responding to the student's notes on a daily basis. When the author discovered incomplete or incorrect notes, he wrote comments to the students, and if he discovered trends that indicated a more general problem, he addressed it with the entire class (Brown 1991).

The justification for this study is the same justification for this dissertation study; the author cited that mathematics is the key skill for success in the STEM disciplines: "Calculus I has been determined to be a major roadblock for students pursuing programs in science and engineering" (Brown, 1991, p. 261). This study explored just two questions: "1) Will a structured notetaking procedure improve a student's performance in Calculus I? 2) Do students perceive the structured notetaking procedure as being beneficial to their study of calculus?" (p. 261). The research design utilized was mixed method. The mixed method research design combines the quantitative and qualitative approaches into one single inquiry. More specifically,

the author employed the Sequential Study Mixed Method design which is the quantitative and qualitative data are generated in two distinct phases within the single study.

The author sought to answer the first question (Will a structured notetaking procedure improve a student's performance in Calculus I?) by using a previous section of Calculus I (Fall 1988) as the control group, and he used the upcoming section of Calculus I (Fall 1989) as the experimental group (Brown, 1991). Moreover, since the author taught both sections himself, he was intentional with respect to teaching the sections the same way:

> That is, I began each class by answering their questions and then lectured the rest of the period. The course started with a review of algebra and trigonometry, followed by an introduction to differential calculus. A short quiz was given each Friday. The quizzes counted 10% of the final grade. At the end of the four weeks the first major test was given (Brown 1991, p. 263).

In order to establish a baseline for comparison purposes, the author waited until after the first major test (at the end of the first four weeks) to introduce the specific notetaking format in the experimental section. His rationale was that late registration had ended and therefore no new students would be entering the class. Conversely, most of the students who were going to drop the class did so by this point in the semester, and finally, the remaining students in the class who were struggling would be receptive to a change in their notetaking strategy (Brown 1991). By establishing an internal baseline and judging success in terms of movement from the baseline, the author minimized extraneous variables, and he also accounted for potential differences between the two sections of Calculus I that were also threats to the internal validity of the study.

The specific notetaking format was introduced in the experimental section and class time was utilized to discuss the directions and what was expected on each section of the notetaking

form. In order to make room in the class schedule and incentivize student participation, the author dropped the weekly quizzes and graded the notetaking forms. The specific notetaking forms would count for ten percent of the final grade (Brown, 1991). The specific instructions given to the students were as follows:

> Their first assignment was to organize their notes from class according to their required format. This was to be done as a homework assignment. I emphasized that they were to use their own words to restate the main concepts in the lesson. The conclusion section of their notes could be used to let me know how they felt they were doing in the class, to ask questions, and/or make comments about the class in general (p. 263).

Some of the initial student feedback regarding the formalized notetaking form was less than stellar. For example, students felt there was either not enough time to re-write their notes on the notetaking form, or this time would be better spent by completing homework problems. One student remarked that taking notes did not fit his/her learning style and therefore did not write anything on the notetaking form. Due to student feedback concerning the amount of time they were spending on the form, the original notetaking form was shortened after only one week after the original form was introduced (Brown, 1991).

Regarding question two (Do students perceive the structured notetaking procedure as being beneficial to their study?) which was the qualitative piece of this study, the author asked for a written evaluation of the project twice during the study. The results from the twenty respondents were the following: 1) 60 percent were in favor of the structured notetaking approach. 2) 10 percent thought the approach was helpful with reservations. 3) 30 percent of the respondents indicated the structured notetaking approach had no benefit (Brown, 1991). Brown also noted that throughout the semester, the students were submitting comments that he had

never received in class before. Furthermore, Brown stated he believed his students perceived that he was personally really interested in their success which is a documented characteristic of classroom research.

On the last day of class, Brown (1991) administered the second survey which yielded the following results: 1) 62.5 percent were in favor of the project. 2) 25 percent had reservations. 3) 12.5 percent of the 16 respondents indicated the structured notetaking approach had little to no value. In summary, the first survey was administered after one month into the study, and the second study was administered on the last day of class. There was a slight increase in the students who were in favor of the structured notetaking approach, and an increase in the number of students who reported the approach was helpful with reservations, and finally, a decrease in the number of students who reported the approach was not helpful. Therefore, the overall results were a positive increase between survey one and survey two on every survey question.

Regarding question one (Will a structured notetaking procedure improve a student's performance in Calculus I?) which was the quantitative piece of this study, the author compared two sections of Calculus I with respect to the final class average which was the dependent variable. In order to determine if the student entry characteristics were similar enough for comparison, the first author analyzed student grade point averages (GPA) in both the treatment and control sections of Calculus I. Brown (1991) originally intended to calculate all student participant GPAs in mathematics courses only; however, not all students in the sections had taken a math class at Northern Virginia Community College (NOVA). Therefore, Brown chose to calculate the student participant GPAs in all courses taken at NOVA. Additionally, Brown calculated the class GPAs in both sections prior to the first test since the scores on the first test were used as a base-line for the two sections.

The final class average for the treatment section was 69.2 and 66.4 for the control section; however, the student scores in the treatment group on the first test were higher than the student scores on the first test in the control group which indicated the experimental section may have scored higher without the intervention of structured notetaking (Brown, 1991). When this was taken into account and controlled for statistically, the overall result with respect to question one was insignificant findings: "When this fact is taken into account and a statistical adjustment is made, the final averages are 69.2 for the 1989 class and 66.4 for the 1988 class. Using a t-test it was determined that this difference is not statistically significant" (Brown, 1991, p. 268).

Although the statistical analysis yielded insignificant findings, (Brown, 1991) reported several perceived benefits of structured notetaking. For example, every student who participated in the project on a regular basis earned a grade of C or better. Moreover, "the number of students who indicated that the notetaking procedure was not helpful decreased from 30% to 12.5% by the end of the course" (p. 269). Brown also noted that the process of reviewing and re-writing class notes gave students an opportunity to ask more substantive questions as opposed to the vague questions he normally gets during the class lecture. The student questions on the notetaking form were more specific and directly related to problem areas with the content, and since Brown had the student questions beforehand, he had more time to prepare the answers as opposed to answering after hearing the question for the first time during the lecture. A final observation regarding student questions from Brown was that students will ask questions on paper that they were too embarrassed to ask in class during the lecture.

Just as other researchers who experienced non-significant findings after testing a CAT, Brown (1991) reported benefits which indicated that the CAT had some value and was not a complete waste of time. An additional benefit mentioned by Brown was when he reviewed the

student notes on the specific notetaking form, this provided insight into the individual student learning styles, and he also mentioned the fact that not all students learn best by writing notes. According to Brown, the most important benefit from his perspective, was his feedback to each student: "I reviewed their papers and returned them the next class day so that the students could see corrections or additions that I made to their papers" (p. 269). Also, the privacy of the feedback seemed to be well received from the student perspective (Brown, 1991).

This study by Brown (1991) is not a study about one of the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) because "structured notetaking" is not one of classroom assessment techniques in the Angelo and Cross handbook. Therefore, this study is in the "other studies of STEM pedagogical strategies" section of this literature review. However, this study by Brown was conducted in the setting of Calculus I, and there are hints of the *Documented Problem Solutions* classroom assessment technique (posited by Angelo and Cross, 1993) in this study by Brown. For example, the original notetaking form contained the following statement: "Examples from the board with written explanation of each step" (p. 264). This requirement is essentially equivalent to *Documented Problem Solutions* because this classroom assessment technique requires the same thing with less detail. When utilizing *Documented Problem Solutions* students must show each step of the problem solving process. Brown required students to not only document each step in writing, but he also required a written explanation of each problem solving step.

Unfortunately, this statement was removed when the original notetaking form was revised and shortened just one week into the study. Brown's (1991) objective was not to test whether requiring students to document each step of the problem solving process would improve student learning, but rather his objective was to test whether requiring students to take notes in a

specific and prescribed way would increase student success in Calculus I. This study is a good example of classroom research conducted by teaching faculty on their own students. Brown recognized that fewer than 60 percent of the students who enrolled in his Calculus I course successfully completed the class with the grade of C or better and attempted to improve his student success rate by introducing a specific student notetaking format (Brown, 1991).

In summary to this section regarding other studies pedagogical strategies in the STEM disciplines, it is clear from the literature that pedagogical strategies are of great interest to higher education scholars and researchers. The studies examined in this section focused mainly on the pedagogical strategy of the large lecture in introductory undergraduate mathematics courses because Calculus I is the first math course for most freshman STEM majors. Moreover, the large lecture appears to be the most popular pedagogical approach for Calculus I and Calculus II, at least at most large research institutions. The challenges, limitations, and perhaps opportunities of the large lecture need to be analyzed and understood in order to improve freshman STEM major success. In the next session regarding classroom assessment techniques in higher education, an explanation of the Classroom Assessment Techniques posited by Angleo & Cross (1993) will be given and some additional studies that demonstrate how Calculus is typically assessed will be examined.

Classroom Assessment Techniques in Higher Education

The higher education literature is replete with slightly different definitions of classroom assessment or just "assessment" and some scholars offer more nuanced definitions depending on the teaching context. For example, Falchikov (2005) makes a distinction between performance assessments and authentic assessments. Performance assessments stress skill demonstration in real world environments and authentic assessments measure knowledge of ways to solve actual

problems. Angelo & Cross (1993) described classroom assessment as an approach designed to help teachers find out what students are learning in the classroom and how well they are learning it; "Classroom assessment helps individual college teachers obtain useful feedback on what, how much, and how well their students are learning. Faculty can then use this information to refocus their teaching to help students make their learning more efficient and more effective" (p. 3).

Although the definitions of classroom assessment vary and exist in different teaching contexts, there is general agreement in the literature regarding the purpose of classroom assessment and that is to improve teaching and learning. Effective classroom assessment techniques empower both teachers and their students with information to improve the quality of learning in the classroom. Angelo & Cross (1993) asserted that teaching without learning is not really teaching at all but just talking, and unless learning takes place, effective teaching has not occurred.

For the past two and one half decades, compliance and accountability concerns in higher education have been rising which created more public scrutiny and calls for a re-examination of the quality of teaching and learning in colleges and universities. The classroom assessment techniques (CATs) as posited by Angelo & Cross (1993) and other assessment measures are at the heart of this national conversation because all assessment measures can be used to directly address these concerns and identify deficiencies for improvement or provide evidence that higher education meets agreed upon standards and is worth public investments. Educators and administrators must prove to their stakeholders that their institutions can deliver on the promise of quality education at a competitive price. Goubeaud (2010) asserted that "assessments have potential not only to enhance the learning process and inform instruction but also to provide evidence to stakeholders that students are competent" (p. 237).

Three Categories of Assessment

In the literature, there are three categories of assessment: 1) diagnostic assessment, 2) formative assessment and 3) summative assessment. Diagnostic assessment involves the teacher's learning about student needs prior to instruction and is not very common in higher education given the prerequisite system. Formative assessment provides prescriptive feedback to assist students in reaching their overall class goals and summative assessment is a comprehensive or cumulative measure and helps to establish whether students have attained the overall class goals. (Joosten-ten Brinke et al., 2007). In general, the timeline of these categories of assessment with respect to the instruction is diagnostic first, before the instruction, formative second, during the instruction, and summative assessment occurs at the end of the instruction. All 50 of the CATs posited by Angelo and Cross (1993) fall into the formative assessment category.

Although the higher education literature regarding the categories of assessment is rather voluminous, two authors found that there is very little common understanding of the terms often used to describe forms of assessment in policy documents and other literature at one university: Scaife & Wellington (2010) asserted that the university documents they reviewed had the correct assessment terminology (diagnostic, formative and summative), but there was no evidence of formal or working definitions for the reader in the documents. These authors recommended that institutions of higher education define these terms in their published policies consistent with previous literature on assessment in higher education. Moreover the authors offered their own definitions for the three assessment categories: "Formative assessment principally concerns students learning from teachers' feedback or from self or peer assessment; summative assessment principally concerns the categorizing of students' assessed work; diagnostic

assessment principally concerns the teacher learning about the students' learning needs" (Scaife and Wellington, 2010, p. 146).

While there is general agreement in the literature regarding the efficacy of all three major types of assessment (diagnostic, formative, and summative), scholars disagree at times regarding who should be primarily responsible for assessment in different teaching contexts. Inoue (2004) asserted that students should be directly involved in the assessment of their work, and they can be trusted to assess themselves; however, this assertion was limited to formative assessments in the development of writing skills and Inoue acknowledged that self and community-based assessment may not be a viable option in other disciplines. The rationale behind formative assessments is that often times the process of making mistakes and then correcting them without a grade penalty has more value with students than a one-time summative assessment and a final grade. Formative assessments that are not graded typically removes the grade anxiety and places a greater emphasis on learning. Inoue asserted that allowing students to assess their own work has instructive value and students respond positively to this approach because they get to share the authority inherent in the assessment process. Scaife & Wellington (2010) asserted essentially the same data point, that students value assessments that are not graded as long as the quality of the feedback is meaningful and helpful to their learning.

Classroom Assessment Techniques (CATs) in the Literature

Since the publication of *A Nation at Risk* (1980) which was a nationwide study on the condition of education in the United States, scholars began an intense reexamination of the quality of teaching and learning at all levels of education (Angelo and Cross, 1993). Since the 1980s, education reform has been largely driven by the setting of academic standards which subsequently prompted what has been called the "assessment movement" in higher education.

More recently, federal policies continue to reflect the desire for clear and measurable standards in education. Reig and Wilson (2013) stated: "No Child Left Behind and many professional accreditation associations, such as the National Council for Accreditation of Teacher Education, mandate student performance outcomes as indicators of teaching and program effectiveness" (p. 1). As a result, much has been written regarding assessment in education and in higher education. Moreover, in the last two decades, educators have begun implementing a wider variety of assessment types including alternative and student-centered assessment practices (Goubeaud, 2010).

Consequently, the assessment movement is still a powerful influence in higher education today. Scholars continue to develop, implement and write about classroom assessment techniques. Simpson-Beck (2011) posited that the work of Angelo & Cross in their book *Classroom Assessment Techniques: A Handbook for College Teachers* (1993) has come to be viewed as canonical in the higher education community. However, other researchers point out a dearth in the literature regarding studies that examine the validity of the CATs as posited by Angelo & Cross. For example, Adams (2004) stated there is a notable lack of experimentation and research evidencing quantitative empirical support that classroom assessment techniques improve overall learning. Instead, much of the writing on classroom assessment techniques has been anecdotal and the research studies have been qualitative and generally limited in scope to:

- The conceptualization and application of classroom assessment techniques (CATs)
- The documentation of the various types of CATs
- Surveys of student and / or faculty opinions assessing the effectiveness of CATs
- Surveys on how CATs improved student course satisfaction

- And finally, surveys of faculty attitudes toward the use of CATs in the classroom (Simpson-Beck, 2011).

One study by Cottell and Harwood (1998) utilized a quantitative quasi-experimental research design to test the effectiveness of the following CATs: (1) *background knowledge probe,* (2) *minute paper,* (3) *feedback form,* (4) *directed paraphrasing,* (5) *pro-and-con grid,* (6) *what did I learn from the exam,* (7) *classroom assessment quality circle,* (8) *group instructional feedback technique*, and finally, (9) *group-work evaluation form*. Cottell and Harwood were accounting professors at the time of this study, and as accountants often do, they begin the think about the cost benefit analysis of classroom assessment. Classroom assessment techniques are formative assessments, and in general, the costs are related to the class time that is consumed in implementing the techniques. Therefore, the central question of this study was do the benefits of assessment exceed the costs of assessment? Is there a positive bottom line and if the benefits of CATs exceed the costs, then students should learn more in classes where CATs are utilized. The overall class objective was to minimize the difference between the accounting competencies that employers demand with the accounting competencies that the students possessed.

The authors began by formulating a model of the learning process and selecting CATs that were congruent with the class goals and content. This learning model and the instructional design process involves key assumptions about the students and their learning needs, and these assumptions were tested by using CATs throughout the semester. The learning model contained two primary variables: 1) student entry competencies and individual ability. 2) student attitude and level of motivation for learning. (Cottell and Harwood, 1998).

The research design was one semester of accounting classes taught by two professors and each professor taught two sections of the same accounting class. The second class immediately

followed after the first class for both professors, and the treatment group was the first class and the control group was each professor's second class. The researcher's deliberately selected the first class as treatment and the second class as control to minimize what they called a "second class effect." "Professors often think there's a second class effect, meaning the second class tends to go better than the first" (Cottell and Harwood, 1998, p. 40-41). The idea behind the second class effect is that professors get valuable student feedback from the first class and therefore, make slight adjustments for the second class. In order to avoid a potential second class affect biasing the study in favor of significant findings, the researchers selected the first class as the treatment group and the second class as the control group.

A common research ethical dilemma surfaced in this study with respect to how to handle midcourse adjustments. The researchers knew that the treatment group, which utilized the CATs being tested, would yield student feedback that could be used to improve teaching and learning in both sections given that the classes were the same. The dilemma was how could this feedback be withheld from the control group students? (Cottell and Harwood, 1998). The researchers decided to bifurcate the student feedback into two categories, specific prompts and general prompts. "…Specific prompts the professor or students to make detailed, discrete, and often temporary changes in behavior…general prompts the professor or student to make broad, sweeping, and typically lasting changes in behavior" (p. 41). The researchers decided to implement the latter, the general changes, to both the treatment and control groups while acknowledging this decision introduced a bias against finding significant results that CATs improved student learning. Simon-Beck (2011) asserted the following regarding this decision by Cottell and Harwood: "However, Cottell and Harwood (1998) may have biased their results by

implementing course changes during the study, based on information obtained from the use of CATs, to both the treatment and control groups (i.e. contamination bias)" (p, 127).

Cottell and Harwood (1998) defended this study and their results by noting that even without the general changes made by the professors, both treatment and control groups benefited from the specific changes the professors made and the general and specific changes that the students made in both treatment and control groups. Additionally, Cottell and Harwood noted that they did not make a lot of general changes: "As it turned out, we did not make many general changes to both classes." (p. 41). Yet they did not define or quantify the term "many," in this study, nor did Cottell and Harwood specify which general changes were implemented at the midpoint of the semester. This reality indicates the results of this study should be interpreted with some degree of caution.

A second aspect of this study that cast a degree of caution over the final results were the use of multiple CATs. Moreover, "we selected CATs from many of their assessment categories. We used each CAT at least twice, generally once a week" (p. 41). Cottell and Harwood (1998) tested a total of six CATs, and when multiple CATs are tested in just one study, it is difficult to pin point caution and determine which particular CAT may or may not have made a difference with respect to student learning. However, the focus of this study was centered on the sum effect of the selected CATs on student learning with respect to a cost benefit analysis and not pinpointing or isolating the benefit or lack of benefit of one particular CAT.

Cottell and Harwood (1998) found no significant differences between the treatment group and the control group with respect to the dependent variable which consisted of student grades, student participation and student perceptions:

> Interestingly, at each university one class seemed to dominate the other, typically earning slightly higher grades on nearly every course element. Because the dominant class was the CATs class at University A and the Control Class at University B, these results do not suggest that CATs systematically played a role in student learning as measured by graded course elements (p. 42-43).

Student participation was tracked by both professors and graphed to show the trends for the treatment and control group. When these trends were compared, the patterns between the CATs and control classes were highly variable, and the authors concluded that if changes in student participation were connected to the CATs, the effects were not enduring. "Participation in the last quarter of the class was highly variable in both CATs and control classes. Given this data, we could not conclude that using CATs had a systematic effect on the number of students who participated in class." (Cottell and Harwood, 1998, p. 43). Moreover, the students were surveyed regarding their perceptions of whether or not the CATs improved their learning. The survey results indicated the opposite anticipated conclusion. The control students, not the CATs students, tended to feel that they learned slightly more. (Cottell and Harwood, 1998).

Finally, the authors ran a regression model in order to isolate the CATs studied and account for other factors. The results were the same in that the CATs were not shown to be a positive influence on student's perceptions of their learning:

> We did this by running a regression model that included a variable to capture the difference between CATs and control class students along with variables that controlled for students beginning competencies, ability, attitude and motivation, effort, and other aspects of the learning environment. Given the results in Table 4.3, we were not too

surprised that we did not find CATs to be an important factor in explaining students' perceptions of their learning (p. 44).

The results of this study did not find the six CATs studied to be statistically significant with respect to student grades in the context of the accounting classroom at two universities. Moreover, the results also did not indicate the six CATs increased student classroom participation or improved student perceptions of their own learning in the accounting classroom at two different universities. Other researchers have asserted that although the results of their CATs studies did not have significant findings, this should not be interpreted to mean the CATs tested have no value. Cottell and Harwood (1998) posited "we found a bottom line that suggested no gain or loss from using CATs" (p. 44).

Victoria Simpson-Beck found in her review of the literature regarding the CATs posited by Angelo and Cross (1993) a dearth of studies that examined CATs using the quantitative research design. "Although the various types of CATs have been extensively documented and qualitatively studied, there appears to be little quantitative research assessing the effectiveness of these techniques in improving student learning" (Simpson-Beck, 2011). As stated previously, other researchers have made this same observation. For example, Adams (2004) found a notable lack of experimentation and research on CATs. As a result, Simpson-Beck decided to empirically test one of the oldest CATs, the *Muddiest Point,* by using a pre-test / post- test, quasi-experimental design comparing student learning outcomes across two sections of a fall 2009 introductory undergraduate criminal justice course at a mid-sized northern university in the U.S.

Simpson-Beck (2011) sought to avoid the contamination bias that Cottell and Harwood (1998) had in their study. "The purpose of this study was to fill a void in the CATs literature by

empirically and quantitatively exploring the effectiveness of CATs, while controlling for contamination bias, which prior research has failed to do" (p. 127). At this point in 2011, the Cottell and Harwood study was the only study in the literature that utilized an empirical quantitative analysis to measure the relationship between the use of CATs in the classroom and increased student learning. Simpson-Beck also decided to differentiate her study from the Cottell and Harwood study by testing only one CAT to avoid confounding results. If two CATs are studied simultaneously in the same study, it is possible that one technique may be effective while the other technique is not.

Utilizing a convenience sample of 64 students from her Monday, Wednesday and Friday section of Introductory Criminal Justice, this section served as the experimental group. The Tuesday and Thursday section of the same course consisted of 62 students and served as the control group. The research questions were as follows: RQ1: Does the use of the Muddiest Point CAT improve chapter test scores? RQ2: Does the use of the Muddiest Point CAT improve cumulative final examination scores? RQ3: Does the use of the Muddiest Point CAT improve overall course grades? (Simpson-Beck, 2011, p. 127). The Muddiest Point CAT was selected because it matched the Introductory Criminal Justice course goal of building foundational knowledge. According to Angelo and Cross (1993), the Muddiest Point CAT promotes declarative learning or memory learning and allows for the ongoing assessment of recall and understanding.

During this study, Simpson-Beck (2011) taught both sections using the same course requirements and grading rubrics except for the Muddiest Point assignment which was given only to the experimental section. Both sections were given the same chapter quizzes at the beginning of each week, and the Muddiest Point assignment was given to the experimental

section at the end of each week (Friday) for the following reasons: 1) to increase participation in the assignment. 2) to prevent one potential form of bias (p. 128). More specifically, with respect to number one above, graded in-class assignments were given to both sections each Friday, therefore, class attendance was higher on Fridays due to the grade incentive. With respect to number two above, by waiting to give the Muddiest Point assignment on Fridays, this decreased the potential for this assignment to bias information presented to the control group. In other words, the researcher did not receive the student feedback from the Muddiest Point assignment until the end of the week when that particular body of course content was over; this feedback could not bias the researcher in the control section because again at the end of the week, this particular course content was history and the new week introduced new course content.

> After the Muddiest Point assignment was completed in the experimental group on Fridays, the instructor did not meet with the control group until subsequent Tuesdays, at which time a new weekly topic chapter was begun for the control group and that topic was unrelated to the previous Muddiest Point assignment" (Simpson-Beck, 2011, p. 128).

Formative assessments may be graded or not graded, and formative assessments may be optional or required depending on instructor preferences. In this study, Simpson-Beck (2011) elected not to grade the Muddiest Point assignment, and she made the assignment anonymous to encourage class participation. Although the assignment was not graded and anonymous, she told the students they were required to participate to further increase class participation. Students were told they could write about something else if they did not have a Muddiest Point for that week. The researcher acknowledged that this may have prompted some students to write about a Muddiest Point: "Telling students that they were required to participate and that they must write

something regardless of whether they had a muddiest point, may have resulted in prompting more students to write about a muddiest point" (p. 128).

Students were given approximately one to two minutes to complete the Muddiest Point assignment each Friday morning in the treatment section, and after the papers were collected, the researcher read the muddiest point remarks while the students worked on a different in class assignment. The instructor used the last 15 minutes of class time to share the most prevalent muddiest points with the class, and because many of the muddiest points were similar, the majority of the remarks were covered with the students. On some weeks, all of the muddiest point were shared with the class (Simpson-Beck, 2011). Moreover, students were told if they did not hear their muddiest points discussed, they could reach out to the instructor.

> To prevent students from potentially feeling excluded when class time prevented the instructor from addressing all muddiest points, students were told that the instructor would try to cover all of the muddiest points submitted, but would begin with the most prevalent ones and continue, as time allowed. Then if a particular muddiest point was not covered during class time, the student was invited to speak with the instructor after class, by email, or during office hours (p. 129).

The research design utilized in this study was a quasi-experimental pre-test / post-test design. The independent variable was the Muddiest Point assignment, and the dependent variable was student outcome measures (student grades). More specifically, this introductory criminal justice course had the following student outcome measures: 1) nine chapter quizzes and 2) one final examination. The pre-test, which was given to all students at the beginning class, establishes a base-line for comparison, and the post-test (which was the same instrument) is designed to detect any differences (above and beyond differences expected due to chance) that

may have occurred after the pre-test. To use the researchers words, "The quasi-experimental pre-/post-test research design allows for a straightforward assessment of an intervention by detecting the difference between two points in time" (Simpson-Beck, 2011, p. 129).

Although not a specific research question in this study, but an important question nonetheless, was whether or not any learning took place after the pre-test measurement point. Again, the central question of this study was would the students utilizing the Muddiest Point CAT learn more and retain more, than the students who did not utilize the Muddiest Point CAT. However, before this question could be answered, the question of whether new learning took place in either group (treatment and control) had to be answered first. To answer this question, the author conducted a paired-samples t-test analysis and significant differences between the pre-test and post-test scores were observed: "Students in both groups tested statistically significantly higher, on average, on the post-tests compared with the pre-tests" (Simpson-Beck, 2011, p. 129).

Since new learning did occur in both treatment and control, the next question was if a difference in learning could be detected between the two groups over the period of time between the pre-test and post-test. To answer this question, which is essentially the three research questions of this study, the author utilized an independent-samples t-tests to compare group means between the two sections of students. No significant differences were detected with respect to the dependent variable, chapter test scores (RQ1), final exam scores (RQ2) and overall course grades (RQ3). Therefore, using the Muddiest Point CAT did not result in increased student learning in the experimental group as compared to the control group.

In summary, the results of this study were that the *Muddiest Point* CAT did not result in statistically significant differences with respect to chapter test scores, cumulative final examination scores, and final grades for the course. The author stated that decades of qualitative

studies and anecdotal observations regarding CATs, which tend to validate the effectiveness of these techniques in improving student learning, should be tested further, and the empirical research should be expanded by employing quantitative research designs to examine the validity of CATs. Both Cottell and Hardwood (1998) and Simpson-Beck (2011) were careful to point out that although their studies did not confirm the validity of the CATs studied, more research is needed and their results should not be interpreted to mean that the CATs studied have no value at all.

Goubeaud (2010) found a void in the literature regarding research studies that examined the extent college science faculty members use a variety of assessment practices, such as a combination of both formative and summative assessments: "Assessment practices that could be considered formative with potential to promote student learning appear to be underutilized by all science faculty" (p. 217). Most of the published studies regarding different assessment practices in the science classroom, were from single institutions using either small or convenience samples. Goubeaud sought to expand the literature by conducting a large scale study to examine how college faculty members assess science learning and to specifically measure the extent that a variety of alternative as well as traditional assessment practices were being used (Goubeaud, 2010).

Using a sample from the National Study of Postsecondary Faculty (NSOPF) which is a National Center for Education Statistics (NCES) dataset sponsored by the U.S. Department of Education, Goubeaud (2010) focused on faculty members who taught biology, chemistry and physics at undergraduate institutions. The assessment techniques examined were: (1) *Multiple-Choice Exams,* (2) *Short-Answer Exams,* (3) *Essay Exams,* (4) *Research and Term Papers.* Moreover, the grading practices for formative purposes examined were: (1) *Students' Evaluation*

of Each Others' Work, (2) *Multiple Drafts of Written Work*. And finally, the grading practices for summative purposes were: (1) *Grading on a Curve* and (2) *Competency-Based Grading.*

The results of this study indicated there were statistically significant differences with respect to the assessment and grading between the three science disciplines (biology, chemistry and physics). More specifically, a greater proportion of biology faculty members used a wider repertoire of assessment types than the physics or chemistry faculty members. There were other differences observed in this study, for example, less than half of physics and chemistry faculty used assessments that required students to express their ideas in writing such as essay answers or term papers (Goubeaud, 2010). Another result from this study is that formative assessment tools might be underutilized by faculty of all science disciplines which indicated that helping faculty understand the usefulness of various types of assessment for their disciplines might be helpful for faculty development efforts (Goubeaud, 2010).

There are also studies in the literature where scholars slightly modified a CAT as posited by Angelo & Cross (1993) for their specific purpose. For example, the Angelo & Cross *Minute Paper* (which is also commonly referred to as the *One Minute Paper*) was modified to a *Five-Minute Essay* by Isaksson (2008) at Stockholm University. The *Five-Minute Essay* was utilized as a continuous classroom assessment technique in a course called *Scientific Methods in Archaeology ---Applications and Problems*. This course is taught primarily via traditional lecture with only one final examination. The researcher stated that the primarily lecture and only one final exam structure was problematic because students tend to sit through the lecture and not study until it is time to take the final exam which reinforces a superficial or surface approach to learning.

By using graded *Five-minute* essays, Isaksson (2008) hoped to increase student preparedness and commitment to the lectures by enticing students with an exercise that could increase their final grades. In other words, the students would be more attentive during the lectures because a *Five-minute* essay must be written at the end of each lecture and it would be graded and counted toward their final grade. Each student received his or her graded essay with feedback at the following lecture so the students were able to adjust their efforts in the lectures immediately. Therefore, this particular structure and classroom technique would be primarily formative; however, since it is graded and added to the final grade, it is also a summative assessment. The grading scale in Sweden is a three-level scale Fail (F), Pass (P) and Pass with Distinction (PwD) and in this study, there was a steady increase in the number of essays receiving the grade of PwD (Isaksson, 2008). The analysis showed a strong and statistically significant correlation between the number of essays with the grade of PwD and the number of weeks in the course. More specifically, the coefficient of determination (r^2) showed that approximately 76% of the variation in the data was accounted for by the model (Isaksson, 2008).

Although the strong correlation between the *Five-minute essay* grades and time in the course indicated the technique was effective as a student learning activity, the researcher acknowledged several limitations in this study. For example, this study lacked a control group and was not a true experimental research design. Moreover, correlation does not equal causation; correlation is only the mathematically computed degree of linear relationship between two attributes (Tolson, 2012). However, even with these limitations, the author maintained that the *Five-minute essay* prompted students to think more critically about each lecture and sort out their thoughts in writing while the impact of the lecture was still fresh in their minds. On the front end of the lecture, the students knew that each essay would affect their final grade which

made it more likely they put in some effort to understand the content before each lecture (Isaksson, 2008).

In summary, the author concluded that the greater number of essays receiving the highest grade (PwD) was indicative that a greater number of students reached a higher level of understanding. Moreover, students rated the *Five-minute essay* favorably (although some students indicated it was too stressful), and there were no significant differences in the degree of appreciation of this technique based on gender. And finally, the authored observed that the formative aspect of the *Five-minute essay* was effective because it gave each instructor a means to give students direct feedback early in the course.

Gary S. Goldstein an associate professor of psychology at the University of New Hampshire at Manchester also noted the rarity of CATs being tested in the higher education literature, and he decided to test several CATs in an undergraduate introductory statistics in psychology course (Goldstein, 2007). This was a classroom research study where an instructor uses his or her own students to test a pedagogical strategy with the goal of improving teaching and learning. However, this study was not a pure CATs study with respect to the CATs posited by Angelo and Cross (1993) because Goldstein modified some of the CATs tested or used a variation of a CAT published by Angelo and Cross. "…I describe several CATs (some from Angelo and Cross's 1993 text and some that I have created) used in SIP, the rationale for their use, the impact that CATs have on class time, and grading students' performance on CATs" (p. 78). The acronym, SIP, in this direct quote stands for Statistics in Psychology.

When Goldstein (2007) first introduced CATs in his introductory SIP class at the University of New Hampshire, he did not grade the CATs and students did not sign their names on the techniques. This was based on the recommendation from Angelo and Cross (1993) that

student responses to CATs should be anonymous and not graded. However, when the CATs were optional, only 15 to 20 percent of the students elected to participate. In order to increase student participation, Goldstein made the CATs mandatory, but did not grade individual assignments (Goldstein, 2007).

After reviewing the benefits of CATs from his own personal experience in the classroom, Goldstein (2007) decided to test the following hypothesis: Is it possible that requiring CATs (as opposed to making them optional) and having them contribute to the final grade affects students' perceptions of their usefulness? (p. 80). In order to test this hypothesis, he compared students' ratings of inventory items in two sections where CATs were optional with the same ratings in a section in which CATs were required. A comparison of the means of the two combined sections (CATs optional) with the one section (CATs mandatory) using a two-tailed test for each inventory question, indicated that the students in the section where CATs were mandatory rated the techniques more positively. However, *t* tests for eight of the nine items on the inventory were not statistically significant (Goldstein, 2007).

Goldstein (2007) stated that his results should be interpreted with caution given the lack of statistical control for covariates such as different levels of student motivation between the three sections of Statistics in Psychology (SIP). Moreover, question seven on the inventory ("Completing the learning assessment instruments helped me earn a better grade in the course than I would have had we not completed them") reached statistical significance but this finding is most likely explained by the fact that the CATs mandatory section could earn additional points in the course.

A study by Walker (1991) indicated that CATs increased student learning in an introductory psychology course (Psychology 101); however, Walker also modified the CATs

(posited by Angelo and Cross, 1993) based on a learning model that he created. Walker noticed that over time, his introductory psychology course had become less satisfying to him as a teacher, and he was also concerned about low student test scores: "Negative indicators such as decreased class attendance, lower participation levels, and most significant of all, lower student test scores, suggested that I had problems to solve in Psychology 101" (p. 67). Walker explicitly refers to his study as a Classroom Research effort designed to improve his teaching and ameliorate these negative indicators he had observed.

The goal of this study was very simple: To increase the number of students who master the course subject matter. Moreover, the researcher had two additional research questions: "Which topics are more difficult to learn?" and "What effect does my teaching have on student evaluations?" (Walker 1991, p. 68). To answer these questions and select the CATs for this study, the researcher created a theoretical lens based on psychology theories and models, and he developed a set of guidelines called the MORE guidelines. The MORE acronym stands for Motivation, Organization, Rehearsal and Elaboration. The CATs selected for this study were: 1) Memory Matrix and 2) The Concept Map (Walker, 1991).

The researcher used his own students over a period of two academic years: 1988-89 and 1989-90. The 1988-89 students were designated as the control group and were taught without using any CATs. The 1989-90 students were designated as the treatment group and were taught using the modified CATs by the researcher. The researcher taught both groups over the two year period of time and noted that "because the course content, texts, and examinations were similar across these two academic years, I felt a comparison could be useful" (Walker, 1991, p. 74).

As with any traditional parallel statistical design, other variables present threats to the internal validity of the study. Walker (1991) discussed two confounding variables that presented

challenges to his findings: 1) retesting and 2) student entry characteristics. The treatment group (1989-90) was given the opportunity to take "recovery quizzes" and this may have contributed to their higher examination scores. In this study, the concept of recovery quizzes equaled each student having the opportunity to answer five additional questions, on a subsequent date, with respect to the exam material that was the most problematic for the student. Additionally, the recovery quiz points were retroactively added to the students' exam scores. The control group did not have the opportunity to take recovery quizzes.

Student entry characteristics can differ and therefore possibly account for differences observed between treatment and control. Researchers typically try to measure and account for student entry characteristics by comparing college admissions test scores (such as the Scholastic Assessment Test (SAT) or the American College Testing (ACT) standardized tests) and also, High School grade point averages (GPA). Moreover, if the researcher is conducting a study in the context of a particular discipline, such as mathematics for example, the researcher may choose to compare GPA's in all former mathematics classes as an additional means to measure and account for differences in academic ability. In this study, Walker (1991) compared cumulative grade-point averages of both groups and found that the control group (1988-89) had a slightly higher average, mean GPA = 2.79, than the treatment group (1989-90) which had a mean GPA of 2.71.

A comparison of the examination means of the treatment and control student groups using tests of statistical significance yielded a statistically significant difference. The treatment group scored significantly higher on the exams at the .005 level of statistical significance. "The total test means of these two groups were 86.3 percent and 76.5 percent…only 18 percent of the 1988-89 students received A or B grades on tests, 59 percent of the 1989-90 students received A

or B test grades" (Walker, 1991, p. 74-75). More specifically, the course consisted of five units, and the CATs section scored statistically significantly higher on each of the five units. This held true for the 1989 section and the 1990 section.

The second question of this study was what effect the teaching might have on student evaluations and would there be a noticeable difference between the CATs section and the control section. Past research indicates a positive correlation between student evaluations of a course and their overall learning in the course (Howard, 1984). Due to a change in the student evaluation form, the researcher was not able to adequately answer this question; however, one of the questions remained unchanged and a comparison was made. The question that remained unchanged on the student evaluation form was the following:

> "Overall, as it was presented by the instructor, a student can learn a lot from this course," the 1989-90 students gave higher ratings than the 1988-89 students; the means of these two groups were 4.58 and 4.00, respectively. This difference was statistically significant at the .001 level" (p. 76).

The result of this comparison with respect to this one question suggests that the CATs students may have learned more and therefore rated the class higher than the control students. The researcher indicated that this comparison combined with other measures such as, feedback from students in the form of "written comments, thank-you notes, letters sent to the dean of arts & sciences, and increased class attendance, strongly indicate a difference in overall satisfaction between the CATs section and the control section (Walker, 1991).

Although this study indicated that the use of CATs had a positive effect on both student outcome measures (grades) and instructor outcome measures (course evaluations), the results should be interpreted with some degree of caution. The fact that the treatment sections were

given recovery quizzes and the control sections did not have recovery quizzes presents at least some degree of bias in finding higher test scores in the treatment sections. The author presented the following data point to counter this observation: "Overall, approximately half of the 9.8 percent improvement in mean test score totals of the CAT-instructed group can be attributed to the effect of recovery quizzes" (Walker, 1991, p. 75). A second observation with respect to the recovery quizzes is that with only five additional questions being asked, some students may have marked the correct answers by guessing. The author suggested this was not the case based on his overall results. "Even assuming that students earned one out of a possible five points per quiz by chance, the mean difference between the 1988-89 and 1989-90 grades was still statistically significant at the .05 level" (p. 75).

Walker (1991) also made an anecdotal observation in this study with respect to the popular notion that disciplinary research equals better teaching. The students suggested that Walker cover less material during the class lectures and therefore, he reduced material covered in all five units except for the unit on social psychology which was his area of research expertise. The decision to cover more material on the social psychology unit due to Walker's expertise did not produce a noticeable result.

> This was the only unit in which I observed little or no difference in student learning between the classes taught with and without assessment techniques. This result does not support the often made assumption that doing disciplinary research is the best way to prepare for teaching. As other Classroom Researchers have discovered, I found that there is no direct relationship between my research expertise in the discipline and my teaching effectiveness (Feldman, 1987) (p. 77).

A second anecdotal observation was that CATs seemed to give new life to the class. When presenting the two CATs to the experimental sections, the students were told that each classroom assessment technique was designed to improve the professor's teaching and help the students learn the information for the unit examinations. This fact, along with the recovery quizzes (which was requested by students), seemed to provide a greater sense of empowerment over each student's own learning. However, other changes were made during study that may also have contributed to the greater sense of empowerment. For example, active learning techniques were used instead of traditional lecturing: "..instead of simply lecturing on Freud's theory of personality structure, I changed many of my lectures to respond to their needs…I had students role play the id, ego, and superego in a dating-game simulation" (Walker, 1991, p. 77).

This study by Walker (1991) is another study where a faculty member employed classroom research to attempt to improve his own teaching and subsequent student learning. Additionally, this study fits into the strand of studies where the researcher modified one or more of the CATs posited by Angelo and Cross (1993), and multiple CATs were tested in one study which makes it nearly impossible to attribute any observed differences back to a single CAT. In looking through the classroom research lens with the primary goal of classroom research in mind, better teaching and learning, this study has value. The goal of this study was not to precisely test whether or not one specific CAT would improve student learning; rather, the goal was to improve student learning without regard to the typical goals of other strands of research such as, a large inference space or generalizability of the results. Moreover, the reproducibility of this study was also not a primary goal. Walker selected two CATs the Memory Matrix and the Concept Map based on his desire to improve his introductory psychology class, and he adapted each CAT using a learning model he created, the MORE instructional design.

Although it has been 25 years since CATs were introduced by Angelo & Cross (1993), the higher education literature is rather sparse regarding the study of CATs. There have been several qualitative studies and studies where an original CAT was modified; there are also some mixed method studies and several anecdotal pieces, but only two quasi-experimental studies (Cottell and Harwood, 1998 and Simpson-Beck, 2011). Additionally, both of these studies indicated that the CATs tested did not make a significant difference with respect to student learning. Simpson-Beck called for additional research utilizing the quantitative research design to further test whether CATs can be shown empirically to be effective in the classroom. The current study seeks to further expand the higher education literature by testing the *Documented Problem Solutions* CAT in the context of Calculus I; moreover, if this CAT is shown to be effective in the Calculus I classroom, it may have the potential to improve student retention in Calculus I and the STEM disciplines.

CHAPTER III

METHODOLOGY

Chapter III Introduction

The previous chapter contained a review of the relevant literature, and this chapter outlines the methodology utilized in the present study. This chapter consists of the following sections: The research design, theoretical perspective, research questions, research context, an overview of the participants in the study (population and sample), unit of analysis, dependent and independent variables, **data collection** procedures, a brief explanation of the statistical methods used for the data analysis, and finally, a summary of the study limitations.

Research Design

This study utilized a quasi-experimental quantitative research design to test the effectiveness of a classroom assessment technique (CAT) in the Calculus I classroom. This design is also referred to as "ex post facto research" (Rudestam, and Newton, 2001) and is more popular in social science research because pure randomization is often not feasible in the social sciences. True experimental research designs are characterized by the random assignment of study participants to groups by the researcher over the intervention; however, for this study, the random assignment of study participants was not possible. Cook and Campbell (1979) asserted that quasi-experimental designs compromise some of the rigor of the pure experimental design but maintains the argument and logic of experimental research.

The participants for this study were selected based on enrollment in pre-identified sections of Calculus I. The researcher, with the help of the Math Department at the institution, identified a faculty member who was already using the classroom assessment technique of

interest in some sections of Calculus I but not in other sections. In other words, the faculty member was already using two different formats for the homework problems in Calculus I, and the institution welcomed this study because it would test the effectiveness of one homework problem format over the second homework problem format with respect to the student outcome measures for the course. In summary, the treatment group sections of Calculus I utilized the classroom assessment technique (CAT) to complete their homework problems, and the control group sections did not utilize the classroom assessment technique to complete the same homework problems.

Maxim (1999) noted that in the absence of random assignment to groups, self-selection of study participants into groups can threaten internal validity. For this study, random assignment was not possible but self-selection bias was mitigated by the nature of the enrollment process for Calculus I. The two different homework formats were predetermined by the faculty member but not advertised to the students beforehand. Therefore, the study participants enrolled with no knowledge of the difference between treatment and control sections.

Theoretical Perspective

A theoretical or descriptive framework provides a conceptual lens through which a particular problem or phenomenon may be viewed. For this study, the theoretical framework was based on formative assessment theory. All 50 of the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) are considered to be formative assessments. Formative assessments increase student knowledge and learning by providing on-going corrective information throughout the learning process either with or without grade penalty. Usually the corrective information is shared more than once and continual improvement is the goal up to the point of a minimal satisfactory criterion. The efficacy of formative assessments are generally

acknowledged in the higher education literature, and Yorke (2003) noted there is agreement that the central purpose of formative assessment is to contribute to student learning through the provision of information about performance.

Black and William (1998) completed a substantial review of formative assessment studies and found that it was helpful in promoting student learning across a wide range of educational settings including different disciplines, different learning outcomes and varying levels; however, the majority of this research was conducted in the K-12 setting. In contrast, the majority of studies in the higher education setting focus on summative assessment which is determining the extent to which a student has achieved the learning objectives. The distinction between formative assessment and summative assessment is not always clear and some assessment activities have characteristics of both.

Research Questions

With respect to the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993), the broad and central question of prior studies found in the literature is: Do these Classroom Assessment Techniques (CATs) improve student learning? This study sought to answer the same question but within the context of the Calculus I setting and for only one CAT at a mid-sized private university in the southwestern United States. The following research questions were central in the study:

1) Does the Classroom Assessment Technique (CAT), *Documented Problem Solutions,* posited by Angelo and Cross (1993) improve student learning?

2) Are there significant differences with respect to student grades after students participate in a treatment condition (CAT, *Documented Problem Solutions*, homework format) and a control condition (On-line homework format)?

3) Are there significant differences with respect to student grades between students who utilized the CAT, *Documented Problem Solutions*, homework format and those who do not, after controlling for preexisting levels of mathematical ability?

The Research Context

This study examined the effectiveness of the Classroom Assessment Technique (CAT), *Documented Problem Solutions,* in the Calculus I classroom at a large, private not-for-profit, more selective, four-year research institution in the southwestern United States with a primarily full-time and primarily residential student population. This university's Carnegie classification is "Research University – Higher Research Activity." The student body consists of approximately 17,059 students of whom 84 percent are undergraduate students. In the interest of maintaining the confidentiality of the student data, the institution is not identified in this study.

Methodology of Population and Sample

The theoretical population for this study was Calculus I undergraduate students in a single university setting. The sample utilized for this study were student participants who were full-time (enrolled in 12 credit hours or more) and enrolled in a section of Calculus I. Additionally, the study participants enrolled in their Calculus I sections without knowledge of the difference between treatment and control. The two different homework formats were set by the faculty member prior to the students enrolling in the sections, and the homework formats were not disclosed to the students beforehand. To assess the impact of the independent variable, the sample for this study were students enrolled in two sections of Calculus I in the spring 2017 semester and two sections of Calculus I in the fall 2017 semester. Of the purposive sample selected, 84 ($N = 84$) students completed Calculus I and yielded usable data to address the

research questions through statistical analysis. The treatment group had 40 participants ($n = 40$), while the control group had 44 participants ($n = 44$).

Unit of Analysis

One of the most important concepts in any research design is the unit of analysis. The unit of analysis is the entity being studied and analyzed in the study. In social science research, the typical unit of analysis is an individual person. In this study, student participants were the unit of analysis because each student had a numerical score with respect to the student outcome measures being compared. For example, data was collected and statistical analysis were conducted on each individual student, and this study's conclusions were based this analysis.

Dependent and Independent Variables

There was only one dependent variable in this study: academic achievement in Calculus I and academic achievement was judged by the student outcome measures for the course. As stated previously in chapter one, a key assumption in this study is that grades accurately reflect the competencies that students develop over the course of the semester. The summative student outcome measures were equivalent for each section of Calculus I and consisted of the following: Four Examinations, one standardized departmental Final Examination (written by the calculus professors in the math department and also graded by the calculus professors) and the Final Course Grade.

There was only one independent variable in this study: The Classroom Assessment Technique (CAT), *Documented Problem Solutions.* This technique focuses on the explicit steps and methods used in problem solving. In mathematics and the sciences, it is not uncommon to find that students can often solve the neat textbook problems, but they cannot transfer their problem solving ability to new contexts or solve similar but messy real-world problems (Angelo

and Cross, 1993). In large lecture settings where a Scan-Tron form is used for student examinations, the steps of the problem solving process are not considered and only the final answer is graded. This makes for more efficient and easier grading, but the value of learning the steps of problem solving is not taught and reinforced. Therefore, many students fail to develop an explicit awareness of the proper steps in problem solving, and the result is they cannot easily adjust to new real world problems in a different context.

Documented Problem Solutions focuses on the process of solving a problem and not just the final answer. Additionally, students who simply memorize the steps of a particular problem without giving any thought to the underlying concepts or fundamental question of the problem will be equally lost when it comes to solving a similar but different context real world problem. When students can articulate and explain the "why" and "how" of each step and also realize "where" and "why" they are stuck on a problem, they learn the fundamental concepts the problems are intended to teach. The problems change and the steps vary, but the fundamental mathematical concepts remain the same. Unfortunately, many math students solve problems by plugging the problem into a formula or equation they have memorized, which can result in limited success in some areas of mathematics, but this strategy is far from sufficient in developing authentic mathematical proficiency in the STEM disciplines. *Documented Problem Solutions* is intended to promote metacognitive skills which is ways of thinking about thinking and how the mind processes information (Angelo & Cross, 1993).

Data Collection Procedures

This study utilized student data from four sections of Calculus I over a period of two semesters (spring 2017 and fall 2017). The data collection period spanned two semesters with 84 students ($N = 84$) producing usable data. The same faculty member taught all four sections, and

the course designs were equivalent with the exception of the format for the homework problems; moreover, the homework problems were nearly identical in all four sections. The CAT sections were required to submit their homework problems utilizing the traditional paper format, and the students were required to show each step of the problem solving process. The homework problems were graded by a Teaching Assistant (TA) and additionally, students were encouraged to review their mistakes and re-submit. The control sections completed the same homework problems via an On-line format and were not required to document or show the problem solving steps. Similar to the CAT sections, students utilizing the On-line homework format were able re-submit problems that were incorrect; however, the problem solving steps were not part of the feedback with the On-line system. With the On-line format, students received either a green checkmark for a correct answer or a red "x" for an incorrect answer. The On-line web application provided no feedback whatsoever regarding the problem solving steps.

On-line application forms outlining the components of the proposed study were completed and sent to two university Institutional Review Boards (IRB) for approval: The university of enrollment for the researcher and the university were the study was conducted were required to approve the study. Upon approval from both Institutional Review Boards, student participant course evaluation information was shared with the researcher by the Calculus I instructor. Each participant's university issued student identification number served as a unique identifying number to match the classroom data with institutional data. After the match was confirmed to be accurate, the identifiers were no longer utilized, and the student identification numbers were removed and replaced with a code assigned by the researcher so that the electronic data no longer contained personally identifiable information.

The course evaluation information was entered by the researcher and stored on the researcher's university-issued computer. This computer is protected with Pretty Good Privacy (PGP) encryption and with the researcher's username and password. Additionally, this computer was locked when not in use and stored in a locked office suite with a security system. To the extent permitted by the Family Educational Rights and Privacy Act (FERPA), personally identifiable information obtained in this study was kept confidential and not shared with anyone not associated with this study. The data will be kept for at least three years after completion of the research.

Data Analyses

The research questions of this study were analyzed using statistical methods by converting each question to a specific research hypothesis and comparing the results of each statistical analyses to the null hypothesis. Prior to evaluating each research hypothesis, the treatment and control groups were examined for any significant differences with respect to beginning mathematical competencies. This was completed using independent samples *t*-tests and an analysis of covariance (ANCOVA) test. The analysis of covariance analysis was limited to the Final Examination student outcome measure. The groups were compared with respect to American College Test (ACT) Math scores and ACT Composite scores. Results of these analyses yielded no significant differences between the treatment and control group. Additionally, the original intention was to calculate each student participant's grade point average (GPA) in mathematics courses taken at the institution; however not every student had taken a mathematics course prior to enrolling in Calculus I. Moreover, in most of the STEM disciplines, Calculus I is the first math course. As a result, the cumulative GPAs were calculated;

however, the fall 2017 study participants did not have an institutional GPA at the beginning of the Calculus I course which made this technique unfeasible for this study.

An independent sample *t*-test was performed for each hypothesized relationship in this study (H_{A1} - H_{A6}) in order to determine any statistically significant differences. The following hypothesized relationships were evaluated:

Research Hypothesis 1 ($H_{A1:}$): There will be a significant difference between the CAT section and the control section with respect to Examination One.

Research Hypothesis 2 ($H_{A2:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Two.

Research Hypothesis 3 ($H_{A3:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Three.

Hypothesis 4 ($H_{A4:}$): There will be a significant difference between the CAT section and the control section with respect to Examination Four.

Research Hypothesis 5 ($H_{A5:}$): There will be a significant difference between the CAT section and the control section with respect to standardized departmental Final Examination.

Research Hypothesis 6 ($H_{A6:}$): There will be a significant difference between the CAT section and the control section with respect to the Final Course Grade.

Prior to conducting these independent sample *t*-tests, statistical testing was performed to ensure the data were normally distributed which is essential for statistical inference based on probability. These specific tests included quantile-quantile (Q-Q) plots, Stem and Leaf plots, and Histograms. In addition to the independent sample *t*-tests which compares the means for each hypothesized relationship, further testing was conducted to confirm or not confirm the initial results of the *t*-tests if needed. For instance, if the assumptions are violated for an independent

sample *t*-test, other statistical tests, such as non-parametric tests, are available to further evaluate the hypothesized relationship.

The researcher utilized International Business Machines (IBM) statistical software; Statistical Package for the Social Sciences (SPSS), version 24 to conduct each statistical analyses. A two group independent sample *t*-test was utilized to evaluate each research hypothesis with respect to potential significant differences. The results of each analyses are shared in chapter 4.

Limitations of the Study

The design of this study resulted in several limitations. First, potential differences between treatment and control study participants may exist beyond beginning mathematical competency. For instance, varying student work ethic and internal motivation to succeed. It is possible that student attitudes toward the Calculus I course varied, and some students were more committed to success than other students. Internal motivation to succeed in a class is difficult to measure and account for. Likewise, the potential for variation with respect to the number of credit hours the study participants were enrolled was also present. This potential variation would result in some students having more or less time for the Calculus I course during the study. Furthermore, not only the number of credit hours may have varied, but also the level of difficulty of the other courses enrolled in could equal additional variation between treatment and control. A true experimental design with random assignment of study participants to groups would have been better with respect to these potential differences beyond mathematical competency but again, for this study, random assignment was not feasible and self-selection bias was mitigated by the enrollment process at the institution.

Second, the time exposure to the treatment condition was limited to the homework problems. Students were only required to utilize the classroom assessment technique (CAT) while completing the homework problems. A greater time exposure may have produced different results. For example, had students been required to utilized, *Documented Problem Solutions,* for all classroom activities such as notetaking, quizzes, examinations including the final examination, there would have been a significantly greater time exposure to the CAT being tested. However, in comparison, the time exposure for the treatment intervention in this study was greater than the exposure in the Simpson-Beck (2011) study.

Third, the design of this study limits generalizability to other contexts. Due to the fact that possible pre-existing conditions were not accounted for and the lack of random assignment to groups, equals a smaller inference space for this study. Additionally, this design also limits conclusions regarding causality. Quasi-experimental designs are not as powerful as true experimental designs in establishing a causal link or causal relationship between variables. Therefore, causality may be inferred but it is less definitive than true experimental research designs (Rudestam & Newton, 2001).

Finally, this study was conducted at a single institution, in a single discipline, with a single faculty member, and only one Classroom Assessment Technique was tested with a limited time exposure to the study participants. Moreover, the sample size for the study was relatively small ($N = 84$) though sufficient for the statistical methods employed. These facts further limit the inference space of this study. However, the results may still be generalizable to similar institutions with similar student populations.

Chapter III Summary

This chapter outlined the methodology utilized for the present study. This study sought to test the effectiveness of a particular classroom assessment technique in the Calculus I setting by comparing student outcome measures between treatment and control conditions. Six hypothesized relationships were evaluated with respect to the student outcome measures for the Calculus I course. The following chapter presents the results of the data analyses.

CHAPTER IV

RESULTS

Chapter IV Introduction

The fourth chapter contains the data analysis and results of the study. This study utilized a quasi-experimental design to test the effectiveness of the Classroom Assessment Technique (CAT), *Documented Problem Solutions,* in the Calculus I classroom. Moreover, the overall desired outcome of this study was to positively impact student success in calculus and therefore, improve freshman STEM major success and positively impact retention in the STEM disciplines. The treatment variable in this study was the CAT, *Documented Problem Solutions,* and the treatment group utilized this CAT to complete all homework problems. The treatment group had a Teaching Assistant (TA) to grade their homework problems, and the TA evaluated and provided feedback regarding the documentation of each step of the problem solving process. The control group did not utilize this CAT and completed the same homework problems on-line using software that did not require students to show and document their problem solving steps.

Research Question Results

This chapter presents the results of the statistical analyses that were employed to answer the research questions of this study:

1) Does the Classroom Assessment Technique (CAT), *Documented Problem Solutions,* posited by Angelo and Cross (1993) improve student learning?

2) Are there significant differences with respect to student grades after students participate in a treatment condition (CAT, *Documented Problem Solutions*, homework format) and a control condition (On-line homework format)?

3) Are there significant differences with respect to student grades between students who utilized the CAT, *Documented Problem Solutions*, homework format and those who do not, after controlling for preexisting levels of mathematical ability?

To address these questions, the following null and research hypotheses were constructed:

Null Hypothesis ($H_{0:}$): There will be no significant differences between the CAT group and the control group with respect to the student learning outcomes of this study.

Research Hypothesis 1 ($H_{A1:}$): There will be a significant difference between the CAT group and the control group with respect to Examination One.

Research Hypothesis 2 ($H_{A2:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Two.

Research Hypothesis 3 ($H_{A3:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Three.

Research Hypothesis 4 ($H_{A4:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Four.

Research Hypothesis 5 ($H_{A5:}$): There will be a significant difference between the CAT group and the control group with respect to the standardized departmental Final Examination.

Research Hypothesis 6 ($H_{A6:}$): There will be a significant difference between the CAT group and the control group with respect to the Final Course Grade.

Statistical analyses were performed to evaluate treatment fidelity and to compare the treatment and control groups with respect to the dependent variable of student outcomes (student grades for the Calculus I course). An independent samples *t*-test and Analyses of Variance (ANCOVA) were utilized to detect any possible significant differences. An ANCOVA was

utilized to control for preexisting levels of mathematical ability with respect to the Final Examination student outcome measure.

Data Collection

The first semester of data collection was spring 2017 and a total of 48 students enrolled in Calculus I but only 41 students completed the course. More specifically, 19 students utilized the Classroom Assessment Technique, *Documented Problem Solutions,* by completing the homework problems via traditional paper format with a Teaching Assistant (TA) grading the homework problems and requiring that each problem solving step be documented. The remaining 22 students did not utilize the Classroom Assessment Technique, *Documented Problem Solutions,* and completed the same homework problems via on-line which did not require that each problem solving step be documented. Given that 7 students dropped the course and the treatment group only contained 19 students and the control group only contained 22 students, the decision was made to collect additional data the next semester, fall 2017.

The second semester of data collection was fall 2017, and a total of 54 students enrolled in Calculus I; 11 students dropped the course, and a total of 43 students completed the course. More specifically, 21 students utilized the Classroom Assessment Technique, *Documented Problem Solutions,* by completing the homework problems via traditional paper format with a Teaching Assistant (TA) grading the homework problems and requiring that each problem solving step be documented. On average, the TA had the homework problems graded and returned to the students within two days. The remaining 22 students did not utilize the Classroom Assessment Technique, *Documented Problem Solutions,* and completed the same homework problems via on-line which did not require that each problem solving step be documented. The on-line system provided immediate feedback to the students, but the feedback was limited to

whether or not the submitted answer was correct. Therefore, combining the two semesters yielded a total of 84 study participants ($N = 84$) with 40 ($n = 40$) students in the treatment group and 44 ($n = 44$) students in the control group.

Pre-Analysis Data Screening

After determining that 84 students completed the Calculus I course, and to ensure sufficient quality of data before final analysis, the data were filtered to remove any student who did not complete the student outcome associated with the respective final analysis. This analysis of missing data values revealed minimal missing data for the treatment and control data sets. In the treatment group, three students had missing data: (a) Student number 12 did not have an Exam Four score and a Final Exam score; (b) Student number 15 had a score of three (3) on Exam Four and did not have a Final Exam score; (c) Student number 32 did not have a Final Exam score. All three of these treatment group students were removed from the Exam Four and Final Exam independent samples *t*-test analysis in order to prevent these missing data from possibly skewing the results. These three treatment group students did not drop the course, and they each received a final grade of "F" for the course.

In the control group, two students had missing data: (a) Student number 9 did not have an Exam Four score and a Final Exam score; (b) Student number 20 did not have an Exam Four score and a Final Exam score. Both of these control group students were removed from the Exam Four and Final Exam independent samples *t*-test analysis in order to prevent these missing data from possibly skewing the results. These two control group students did not drop the course, and they each received a final grade of "D+" and "F" for the course respectively.

In order to compare two groups of study participants with respect to the treatment variable, the two groups must be similar enough for the statistical comparison to be valid and

meaningful. One method commonly used, especially in classroom research, to determine if the two groups (treatment and control) are similar enough for comparison is to compare the study participants grade point averages (GPA) at the beginning of the study. If the GPA's are similar, which does not mean identical, but similar to the degree that the GPA means between the two groups are not so different that they are statistically significant, then the researcher may assume the groups are similar or stated differently, homogenous enough with respect to other variables, such as academic ability or more specifically for this study, mathematical competency. For example, if the mathematical competency in one group is significantly higher than the other, any observed difference may be due to this difference and not the treatment variable of interest.

For this study, 41 of the study participants were enrolled and completed Calculus I in the spring 2017 semester. All 41 students had an institutional grade point average (GPA) at the beginning of the Calculus I class, and transfer credit hours were not included in these GPA calculations. However, the majority of the 43 study participants who were enrolled and completed Calculus I in the fall 2017 semester did not have an institutional grade point average (GPA) at the beginning of the Calculus I class because fall 2017 was their first semester at the institution. As a result of not having institutional GPAs for the fall 2017 student participants, the method of using pre-study GPAs to assess the homogeneity of the two groups was not possible.

A second method commonly used to determine if two groups (treatment and control) are similar enough for comparison is to compare the means of the study participants standardized college admissions test scores. This method was utilized for this study. Of the 40 students in the treatment group, 30 students completed the American College Test (ACT), four students completed the Scholastic Assessment Test (SAT), and two students completed both the ACT and SAT; moreover, four students completed the institutions Math Placement Examination. Of the

44 students in the control group, 28 students completed (ACT), 12 students completed the (SAT), and one student completed both the ACT and SAT; moreover, three students completed the institutions Math Placement Examination.

The majority of students in both treatment and control had ACT Math scores. Therefore, an ACT/SAT conversion chart was used to convert the minority of students who had SAT Math scores to ACT Math scores. Given that no conversion was available for the institutions Mathematics Placement Examination to ACT or SAT, the four students in treatment, and the three students in control were removed and not included in the independent samples *t*-test to compare treatment and control means with respect to ACT Math scores. Using Statistical Package for the Social Sciences (SPSS) version 24, an independent samples t-*test* was conducted to compare the means of the treatment group and control group with respect to their ACT Math scores. The groups did not differ significantly, $t\ (73) = .114$, $p = .910$, $d = .024$. The mean for the treatment group ($M = 26.77$, $SD = 3.858$) was not significantly different from the control group ($M = 26.68$, $SD = 3.489$). These findings indicate that the treatment group and the control group are very similar with respect to mathematical competency as measured by ACT Math score.

Although ACT Math scores are probably the best indicator of mathematical competency for this study, ACT Composite scores were also analyzed for additional comparison between treatment and control. Of the 40 students in the treatment group, 30 students had an ACT composite score and three students had an SAT composite score; moreover, two students had both ACT and SAT composite scores; one student had only an SAT composite score, and the remaining four students had the institutions Math Placement Exam scores. Of the 44 students in the control group, 28 students had an ACT composite score and 11 students had an SAT

composite score; additionally, one student had both ACT and SAT composite scores; one student had only an SAT composite score, and the remaining three students had the institutions Math Placement Exam scores. Using the same methodology as before, converting the SAT composite scores to ACT composite scores using an ACT / SAT conversion chart and removing the Math Placement Exam students from the analysis, an independent samples t-*test* was conducted to compare the means of the treatment group and control group with respect to their ACT composite scores. The groups did not differ significantly, t (75) = .653, p = .516, d = 0.149. The mean for the treatment group (M = 26.97, SD = 2.762) was not significantly different from the control group (M = 26.49, SD = 3.620). These findings also indicate that the treatment group and the control group are very similar with respect to mathematical competency as measured by ACT Composite score.

In addition to the missing data analysis and comparison of the treatment group and control group with respect to over variables, such as mathematical competency at the beginning of the study, each data set was tested to determine if the data was normally distributed. Statistical inferences are made based the laws of probability and the assumption of normally distributed data. To test the assumption of normal distribution, each data set was compared with respect to the independent variable (which is simply the treatment and control groups) using three different representations of the data: 1) Normal quantile-quantile Q-Q plots (Expected Normal vs Observed Values). 2) Stem and Leaf Plots (Frequency of certain classes of values). 3) Histograms (Frequency distribution regarding how often each different value in a data set occurred). For each between-group tests conducted using independent samples *t*-test, a visual inspection of normal Q-Q plots suggested that the dependent variable was normally distributed in each data set. The data sets were not perfectly normally distributed, and there was some

skewness in the data sets, but no obvious violations of normality were detected. Stevens (2002) posited that deviations from multivariate normality had a minimal effect on the possibility of increasing Type I error, and that even considerably skewed distributions do not diminish power. The final result of this pre-analysis data screening indicated that pre-levels of mathematical competency, missing data, and departures from normality did not pose a significant threat to the internal validity of this study. The demographics of the study participants are shown below (Table 1).

Table 1

Demographic Characteristics of Study Participants (N = 84)

	Treatment (n = 40)		Control (n = 44)	
Characteristics	n	%	n	%
Gender				
Female	22	55.0	27	61.4
Male	18	45.0	17	38.6
Race/Ethnicity				
African American	1	2.5	7	15.9
American Indian			1	2.3
Central South American	5	12.5	2	4.6
Chinese	2	5		
Filipino			2	4.6
Hispanic	5	12.5	3	6.8
Indian	1	2.5	1	2.3
Korean	1	2.5		
Other	3	7.5	3	6.8
Vietnamese	1	2.5	2	4.6
White	21	52.5	23	52.3

Note: Some students selected up to three descriptors and in these cases, the researcher listed only the first descriptor in this Table.

The results of the ANCOVA analysis controlling for beginning mathematical competencies with respect to the Final Examination are shown below in (Table 2).

Table 2

ANCOVA Dependent Variable: Final Examination

Source	***Type III Sum of Squares***	***df***	**Mean Square**	***F***	***Sig.***
Corrected Model	1510.074[a]	2	755.037	1.573	.215
Intercept	2426.246	1	2426.246	5.054	.028
ACT Math	1276.314	1	1276.314	2.658	.107
Grouping	246.324	1	246.324	.513	.476
Error	33606.173	70	480.088		
Total	451159.00	73			
Corrected Total	35116.247	72			

a. R Squared = .043 (Adjusted R Squared = .016)

Analysis of Data

Each research hypothesis (H_{A1} - $H_{A6)}$ was tested utilizing an independent samples *t*-test to determine whether statistically significant differences existed between the treatment and control group with respect to student grades. The student grades consisted of four Examinations, one standardized departmental Final Examination and the Final Grade for the course. An alpha of .05 was used for each independent sample *t*-test and each test was a two sided (2-tailed) test.

Research Hypothesis 1 ($H_{A1:}$): There will be a significant difference between the CAT group and the control group with respect to Examination One (Table 3).

Table 3

Independent Samples t-test for Equality of Means, Examination One

	Treatment (Paper Homework)		Control (On-line Homework)					
Examination 1	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed	*d*
	73.805	13.401	70.931	18.443	.822	78.322	.414	.178

The results indicated null findings; however, Levene's Test for Equality of Variances was violated for the Exam one independent samples *t*-test ($p = .011$). Stated differently, there was a statistically significant difference between the variance of the treatment group and the variance of the control group. When the equal assumption of variances test is not met, it becomes more difficult to detect a significant with respect to the dependent variable which was Exam one in this case. Due to Levene's Test for Equality of Variances not being met, a second statistical test was employed to test for significant differences. The researcher chose an independent samples median test, *Mann-Whitney U Test*, to further test for significant differences by comparing medians. The *Mann-Whitney U Test* is a nonparametric statistical test, and the result was the same a non-significant relationship amongst the variables (Table 4).

Table 4

Independent Samples Median Test - Independent Samples Mann Whitney U Test

	Treatment (Paper Homework)	Control (On-line Homework)	
Examination 1	*Null Hypothesis*	*Test*	*Sig.*
	The medians of Exam 1 are the same across categories of Grouping	Independent Samples Median Test	.827
	The distribution of Exam 1 is the same across categories of Grouping	Independent Samples Mann Whitney U Test	.747

Research Hypothesis 2 ($H_{A2:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Two. Results of this independent samples *t*-test indicated there was not a statistically significant difference with respect to Examination Two (Table 5).

Table 5

Independent Samples t-test for Equality of Means, Examination Two

	Treatment (Paper Homework)		Control (On-line Homework)					
Examination 2	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig. (2-tailed*	*d*
	66.200	20.527	66.932	17.866	-.175	82	.862	-.039

Research Hypothesis 3 ($H_{A3:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Three. Results of this independent samples *t*-test indicated there was not a statistically significant difference with respect to Examination Three (Table 6).

Table 6

Independent Samples t-test for Equality of Means, Examination Three

	Treatment (Paper Homework)		Control (On-line Homework)					
Examination 3	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed	*d*
	62.400	19.541	60.674	20.100	.388	81	.699	.086

Research Hypothesis 4 ($H_{A4:}$): There will be a significant difference between the CAT group and the control group with respect to Examination Four. Results of this independent samples *t*-test indicated there was not a statistically significant difference with respect to Examination Four (Table 7).

Table 7

Independent Samples t-test for Equality of Means, Examination Four

	Treatment (Paper Homework)		Control (On-line Homework)					
Examination 4	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed	*d*
	73.395	19.660	72.595	22.368	.169	78	.866	.038

Research Hypothesis 5 ($H_{A5:}$): There will be a significant difference between the CAT group and the control group with respect to the standardized departmental Final Examination. Results of this independent samples *t*-test indicated there was not a statistically significant difference with respect to standardized departmental Final Examination (Table 8).

Table 8

Independent Samples t-test for Equality of Means, Final Examination

	Treatment (Paper Homework)		Control (On-line Homework)					
Final Exam	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed	*d*
	79.757	14.223	77.429	17.270	.649	77	.518	.148

Research Hypothesis 6 ($H_{A6:}$): There will be a significant difference between the CAT group and the control group with respect to the Final Course Grade. Results of this independent

samples *t*-test indicated there was not a statistically significant difference with respect to the Final Course Grade. (Table 9).

Table 9

Independent Samples t-test for Equality of Means, Final Course Grade

	Treatment (Paper Homework)		Control (On-line Homework)					
Final Course Grade	*M*	*SD*	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed	*d*
	71.581	18.031	71.687	18.059	-.027	82	.979	-.006

Chapter IV Summary

This chapter presented the findings of the data analysis conducted to address the study's research questions. The results were null findings. Each analysis indicated that there was no significant difference between the treatment and control group with respect to student grades in the Calculus I course. Moreover, the researcher conducted an additional analysis to test whether or not the treatment condition (the Classroom Assessment Technique (CAT), *Documented Problem Solutions,)* made a significant difference for the non-white students in the class with respect to the standardized final examination and the final course grade. Although the sample size was not sufficient for the independent samples *t*-test to be valid and reliable (treatment $n = 18$; control $n = 19$), these findings were also null. The primary question of this study was would the Classroom Assessment Technique (CAT) of *Documented Problem Solutions* significantly

increase student learning as measured by student grades. The final results of this study did not indicate that the CAT treatment condition made a significant difference.

CHAPTER V

SUMMARY

Chapter V Introduction

Effective instructional design and assessment are key variables for student success and retention in higher education. Researchers have previously established that student success and retention in the Science, Technology, Engineering, and Mathematics (STEM) disciplines is particularly problematic on many fronts. For example, according to a meta-analysis of studies covering a period of 38 years, the economic demand for STEM expertise has been increasing while student interest in the STEM disciplines has either remained steady or slightly decreased. Moreover, Young, et al, (2011) has determined that the retention rate for freshman STEM majors is approximately 40 percent, which corresponds to an attrition rate of approximately 60 percent. The research in this area seems clear and unequivocal regarding the need to increase student learning and persistence in the STEM disciplines (Ellis, Kelton, & Rasmussen, 2014).

A second dimension to this problem is the more specific question of why students are unsuccessful and leave the STEM disciplines. Seymour and Hewitt (1997) studied this question and posited the following general findings: 1) students experience difficulty in the introductory courses and lose confidence in their ability to complete the program due to low grades, 2) the STEM culture is highly competitive and some students do not respond well or strive in a highly competitive environment, 3) some students find the curriculum and degree plan overwhelming and too fast paced, 4) ineffective teaching by STEM faculty members, 5) poor advising and lack of support for students who need extra help. Additional researchers focused on and studied Seymour and Hewitt's findings and zeroed in on number one above "Students experience

difficulty in the introductory courses and lose confidence in their ability to complete the program due to low grades." And even more specific to introductory courses, mathematics has been determined by this prior research to be a key skill for success in the STEM disciplines. Mathematics is both sequential and hierarchical (Tyre, 2016), and a student's math skill set is critical in succeeding in introductory mathematics courses for the STEM disciplines. Matt Larson, the president of the National Council of Teachers of Mathematics (NCTM) posited last year, that "Math Education *Is* STEM Education" (Larson, 2017). Moreover, Brown (1991) found that Calculus I, which is the introductory or first math course for many of the STEM disciplines, is a major roadblock for freshman STEM majors.

Summary: Purpose of the Study

The present study sought to build on this prior research by continuing to focus on introductory mathematics courses, and also a second general finding from Seymour and Hewitt (1997) which was their fourth data point: 4) ineffective teaching by STEM faculty members. This particular perspective or lens in viewing the low STEM retention rate problem lies within the instructional design and assessment category. Of course, there are other perspectives and lenses to look through at this same problem, but this study sought to focus on the instructional design and assessment lens by combining another general finding in the higher education literature with respect to classroom assessment.

Turning now to the classroom assessment aspect of this study, prior researchers discovered a paucity of studies in the higher education literature regarding the testing of the Classroom Assessment Techniques (CATs) posited by Angelo and Cross (1993) 25 years ago. Adams (2004) stated there is a notable lack of studies that confirm these CATs actually improve student learning. He also noted that most of the work that has been completed with respect to

CATs has been qualitative studies and anecdotal pieces. Simpson-Beck (2011) posited essentially the same data point seven years later; "although these CATS posited by Angelo and Cross are very popular and viewed as canonical in the higher education community, there is a notable lack of empirical studies that effectively test these techniques in the higher education literature" (p. 9).

By putting these two aspects of the STEM problem together, this study tested a CAT in the Calculus I classroom utilizing a quantitative research design, and additionally, this study is the first known study to test the effectiveness of a CAT, posited by Angelo and Cross (1993), in the Calculus I classroom. There are two quantitative design studies in the literature that tested CATs, posited by Angelo and Cross, and there are several studies where a CAT posited by Angelo and Cross was slightly modified by the researcher and then tested, but no known study has tested a CAT in the Calculus I or Calculus II classroom. The purpose of this study was not only to answer the call from prior researchers to test these CATs experimentally, but to test one specific CAT, *Documented Problem Solutions,* in the Calculus I classroom empirically to see if this instructional design would increase student success in Calculus I.

Design of the Study

This study utilized a quasi-experimental design to compare a classroom assessment technique intervention with a control condition in a Calculus I course. The study included data from 84 students ($N = 84$) who were enrolled in Calculus I at a private university in the Southwest. The data was collected over a period of two semesters and consisted of four sections of Calculus I students taught by the same faculty member. The rationale behind this rather simple design was to test only one classroom assessment technique (CAT) in only one discipline (Calculus I) with only one professor in order to minimize extraneous variables. Of the 84 study

participants, the treatment group consisted of 40 students ($n = 40$) and the control group consisted of 44 students ($n = 44$).

The classroom assessment technique (CAT) and independent variable for this study was a homework format that required students to show and document each step of the problem solving process. This CAT, posited by Angelo and Cross (1993), is called *Documented Problem Solutions*. The Calculus I students in the treatment group were required to use the CAT in completing all homework problems in the course. The treatment group submitted their homework problems using paper and had a teaching assistant (TA) who graded the homework problems and provided feedback. The TA provided feedback by writing comments on the homework assignments each week throughout the course. This feedback was more than whether or not the student arrived at the correct answer but involved each step of the problem solving process from start to finish. Students were also given partial credit if some of the problem solving steps were completed correctly even though the final answer was incorrect. This CAT focuses more on the problem solving process and less on whether or not the final answer is correct. And finally, the treatment group students were given an incentive to correct problems and re-submit them for additional credit.

The control condition was a completely different homework format that did not require students to document and hand in the problem solving steps. Students in the control group completed their homework problems on-line and the problems were also graded and submitted to the professor by the on-line system. With the on-line homework format, students received immediate feedback regarding whether or not the answer they submitted was correct. If the submitted answer was incorrect, the system gave the student an opportunity to re-submit a different answer up to 15 tries per problem; however, after six attempts at the problem, a 10

percent reduction occurred with each additional attempt. In summary, six problem attempts were free from grade penalty and beginning with the seventh attempt there was a 10 percent grade penalty (per attempt) up to a total of 15 allowed attempts per homework problem.

The critical difference between the two homework formats in this study was that the feedback from the on-line homework system was "answer driven" feedback. When the student submitted an answer to a homework problem, they received either a green check mark for a correct answer or a red "X" for an incorrect answer. The on-line system provided no feedback with respect to the problem solving steps or "why" the submitted answer was incorrect. For example, the magnitude of the answer could be correct, but if the sign (positive or negative) of the answer was incorrect, the on-line system would respond with a red "X" because the sign is wrong and therefore, the answer is incorrect. One possible improvement for the on-line system is software that can provide expanded feedback to students beyond the correct or incorrect answer.

In contrast, the feedback from the TA in the treatment group was "problem solving steps" driven feedback. With the traditional paper format and the requirement to document each problem solving step, the TA could examine the problem solving process from start to finish. Using the same example from the above paragraph, if a student submitted an answer that was correct in magnitude but incorrect with respect to the sign (positive or negative), then the TA could determine where the student erred in the problem solving process and give partial credit for getting the magnitude of the problem correct. This type of feedback is more thorough in that the student can see precisely where the error was made and therefore, in theory at least, learn from the mistakes made with respect to the specific steps taken to solve the problem.

It is important to note that the same Calculus I problem solving steps were taught to all students in this study. A second rationale for having the same professor teach all students was to

standardize the problem solving process as much as possible. There are different approaches to problem solving but the professor in this study taught all four sections using the same curriculum even though the homework formats were different. Therefore, each student in this study, again in theory at least, would be applying the problem solving approaches and steps they learned in the lectures to the homework problems. The difference is that the traditional paper homework format placed an emphasis on the steps taken to solve the problems, and the on-line homework format placed no emphasis on the problem solving steps.

The dependent variable in this study was the student outcome measures (student grades). All four sections were given the same examinations, and one standardized departmental final examination, which was written by the calculus professors in the department, and each calculus professor graded the problems they wrote for every student exam. The standardized departmental final examination (The Math Department at the institution) was perceived by the researcher to be the most objective student outcome measure due to the fact that the same professor graded the problems he or she wrote for every Calculus I final exam for the semester. Stated differently, the calculus professors wrote and graded the department final exam collectively to avoid differences and variation in the grading metrics between professors.

The study was designed to address the following research questions:

1) Does the Classroom Assessment Technique (CAT), *Documented Problem Solutions,* posited by Angelo and Cross (1993) improve student learning?

2) Are there significant differences with respect to student grades after students participate in a treatment condition (CAT, *Documented Problem Solutions*, homework format) and a control condition (On-line homework format)?

3) Are there significant differences with respect to student grades between students who utilized the CAT, *Documented Problem Solutions*, homework format and those who do not, after controlling for preexisting levels of mathematical ability?

Summary of Findings

The results of the analysis for all three research questions indicated there was not a statistically significant difference between the treatment and control group with respect to the student outcome measures. For each data set (Examination 1 – 4, the standardized departmental Final Examination and the Final Course Grade) no significant differences were observed and therefore, the null hypothesis could not be rejected. Potential differences were examined using an independent samples *t*-test and a one-way ANCOVA which controlled for pre-existing levels of mathematical competency; the ANCOVA was utilized only for the Final Examination student outcome measure. These findings reveal that the treatment intervention did not lead to an observed difference above and beyond what might be observed by chance alone.

Moreover, additional testing was conducted on the Examination One data set due to an observed statistically significant difference in the variation between the treatment and control group with respect to the means. More specifically, Levene's Test for Equality of Variances detected this significant difference at $p = .011$ on the independent samples *t*-test which indicated it would be more difficult to detect a significant difference with respect to the independent variable. An independent samples *t*-test compares the means (the arithmetic average) of the treatment and control group, but it is more difficult to detect a significant difference in the means when the variation with respect to each mean is not approximately equal between the two means of the two groups being compared. As a result of Levene's Test being violated for the Examination One data set, a different statistical test was conducted to examine the medians

between treatment and control to further check for a significant difference. The result of this test which was the *Mann-Whitney U Test,* was the same as the independent samples *t*-test for this data set which was null findings. The null hypothesis could not be rejected because the medians were not significantly different from each other. The fact that two different statistical tests, one parametric and the other non-parametric, indicated the same result lends greater credibility to this particular finding.

Interpretation of Findings

The findings of this study matched the findings of the two quantitative research design studies on CATs in the higher education literature. Even though these two prior studies are also classroom research studies, this study was somewhat different in its design. The first study by Cottell and Harwood (1998) was also a quasi-experimental design but this study tested the effectiveness of multiple CATs simultaneously. When multiple CATs are tested simultaneously, it is more difficult to attribute any observed difference to a particular CAT. However, the focus of this study was not precise causation, but rather would there be a positive cumulative effect of the CATs being tested. In addition, Cottell and Harwood decided to implement changes and make midcourse adjustments in both the treatment and control group based on feedback from the treatment group (the sections utilizing CATs) which introduced a bias against finding significant results that the CATs being tested improved student learning. Cottell and Harwood acknowledged this potential bias and stated they could not withhold what they perceived to be helpful changes and adjustments from the control sections. They also defended the results of the study by stating they did not make very many changes to both sections. The results of this study were null findings which means the null hypothesis could not be rejected. Stated differently, each research hypothesis was not supported due to insignificant findings.

The second quantitative research design study in the literature by Simpson-Beck (2011) was more similar to the current study in that only one CAT was tested and the sample size was similar ($N = 64$). Simpson-Beck had read the Cottell and Harwood (1988) study and she sought to avoid any contamination bias in her research design: "The purpose of this study was to fill a void in the CATs literature by empirically and quantitatively exploring the effectiveness of CATs, while controlling for contamination bias, which prior research has failed to do" (p. 127). The intervention for this study was the CAT, *The Muddiest Point,* which is a short writing assignment where students are asked to describe the class content they did not understand or something that needed clarification from the professor. The Muddiest Point exercise was completed each Friday, and the feedback was shared with the entire class. The researcher noted that many of the muddiest points were similar, and the majority of the muddiest point feedback was shared; however, students were told if they did not hear their muddiest points discussed, they could contact the professor and follow up individually.

This study utilized the quasi-experimental pre-test / post-test design and the dependent variable was the student outcome measures. The advantage of the pre-test / post-test design is that the pre-test scores establish a baseline at a specific point in time from which success can be measured and judged at a later point in time. Paired samples *t*-test were used to compare pre-test and post-test scores and both groups had significant differences between their pre-test and post-test scores. But the main research question of this study was would the muddiest point CAT treatment group have higher scores on the post-test scores as compared to the control group that did not utilized the muddiest point CAT. To answer this question independent samples *t*-test were used to compare group means between treatment and control with respect to chapter test scores (RQ1), final examination scores (RQ2) and final course grades (RQ3). The results were

null findings; students in the both treatment and control group showed no statistically significant differences between pre-test and post-test scores.

These two prior studies, Cottell and Harwood (1998) and Simpson-Beck (2011), are the only known studies in the literature that tested the effectiveness of a pure CAT posited by Angelo and Cross (1993) using a quantitative research design. A nearly exhaustive examination of the higher education literature, with the assistance of a research librarian, yielded no additional quantitative research design studies. However, additional studies may exist; Pascarella (2006) noted that given the volume of higher education research, "it is nearly impossible for any review of such a large body of evidence to be absolutely encyclopedic" (p. 508). In light of these two studies, the results of this study are not that surprising; however, the Cottell Hardwood study examined multiple CATs simultaneously, and the Simpson-Beck study tested only one CAT but it seemed to be a minimal intervention at least with respect to time. The inherent differences in the design of these studies and time exposure to the CAT being tested were not insignificant.

To further explain, in the Simpson-Beck study, the muddiest point CAT was administered to the treatment group each Friday, which is obviously only one day each week, and moreover, students were given only one to two minutes to complete the Muddiest Point assignment. The professor then used the last 15 minutes of the class time to share the most prevalent muddiest points with the class. Additionally, on some weeks, 15 minutes was not enough time to cover all of the muddiest points with the class, but students were told they could reach out to the professor if they did not hear their specific muddiest points discussed. In summary, the student exposure to the *Muddiest Point* CAT in this study was only one day a week for approximately 17 to 20 minutes; again, students had one to two minutes to write their muddiest points and the instructor used approximately the last 15 minutes of class time to discuss the muddiest points submitted.

In contrast, the present study yielded a much greater time exposure to the CAT being tested. In the present study, the paper homework format was required every day of the week for the entire semester. The traditional paper homework format and teaching assistant (TA) was part of the instructional design for these particular sections of Calculus I and students had no choice in whether or not they would complete the homework problems using this method. For these reasons, exposure to the CAT, *Documented Problem Solutions*, via the traditional paper homework format was mandatory and it was sustained throughout the entire semester. Due to this much greater exposure to the CAT being tested, the researcher thought this study's intervention might be more powerful and yield a statistically significant change with respect to the student outcome measures. This turned out not to be the case, and the findings of this study were somewhat surprising to the researcher especially when looking at the results from the time exposure to the CAT being tested point of view.

Another key difference between this study and the Simpson-Beck (2011) study was one of the student outcome measures. The outcome measures in this study were the Examinations, Final Examination and Final Course Grades. Similarly, the outcome measures in the Simpson-Beck study were Chapter Tests, Final Exam, and overall Course Grades. However, the Calculus I Final Examinations in this study were not graded by the Calculus I professor. As stated previously, the Calculus I Final Examination was a standardized departmental test written and graded by the departmental calculus professors. The researcher perceived this difference in the design and grading of the Final Examination to be a more objective measure of student competencies at the end of the course. Due to this perceived greater objectivity, the researcher believed it might be more likely to observe a significant difference with respect to the Final

Examination than on the other student outcome measures. Yet, this turned out not to be the case as well.

Implications of Findings

This present study is now the third known study to experience null findings when testing one or more of the CATs posited by Angelo and Cross (1993) with respect to increased student learning as measured by student grades and exam scores. On the one hand, this study does not bode well for the contention that these CATs are beneficial for student learning, yet this study aligns with the Simpson-Beck (2011) study in that the results were the same. On the other hand, non-significant findings do not necessary equal no value. Cottell and Harwood (1998) stated: "We found a bottom line that suggested no gain or loss from using CATs" (p. 44). Simpson-Beck (2011) was also careful to point out that although her study did not confirm the validity of the *Muddiest Point* CAT, her results should not be interpreted to mean the muddiest point exercise had no value at all for her students. Simpson-Beck also called for further testing by stating the empirical research of these CATs should be expanded by employing quantitative research designs.

There is clearly tension between the findings of decades of qualitative studies and anecdotal observations that tend to validate the effectiveness of these CATs and the findings of quantitative studies that tend not to validate these same CATs in the higher education literature. However, it is important to note here that the favorable findings of the qualitative and anecdotal studies significantly outnumber the comparatively small number of unmodified or pure CAT quantitative research design studies. Additionally, this observation is not uncommon and exists with other research topics. On just about any research topic there can be a strand of mainstream studies or meta-studies that seem to have the same or similar findings but there can also be a

strand of cross-current studies or meta-studies that tend to contradict the mainstream strand of studies with respect to findings on the same topic. In addition, on some research topics the volume of studies is still evolving and there is no clear trend to indicate a general finding over time. In these cases, more peer reviewed research is needed to develop a trend that is meaningful and tenable.

On the other hand, with respect to CAT studies seen in the higher education literature, the key difference seems to be the research designs with the qualitative studies and quantitative studies seemingly contradicting each other. Some of this phenomenon may be accounted for and explained by the fundamental nature of what is being measured and assessed in these very different research designs. For instance, study participant perceptions can usually best be assessed qualitatively via participant interviews. The value of this approach is that open ended questions may yield data that was not foreseen by the researcher. Study participant perceptions can be also be measured with a survey instrument and later quantified an analyzed using statistics. Pascarella (2006) noted: "that the many powerful quantitative tools we might bring to bear in college impact research are probably more suited to establishing the existence of potential causal relationships than they are to understanding or explaining why those causal relationships exist" (p. 515).

Had the current study been a mixed-method research design and the 40 students (n = 40) in the treatment group been interviewed regarding the efficacy of the paper homework format and especially the assistance of the TA, it is likely that at least some of the students if not the majority, would have reported positive perceptions about the paper homework format and therefore, the researcher could state the *Documented Problems Solutions* CAT was viewed by the students to be helpful for their learning. This possibility of a mixed result occurred in a study of a

researcher modified CAT (s) by Goldstein (2007). Goldstein, used modified CATs in his Statistics in Psychology class on an optional basis, but then decided to test the following hypothesis: "Is it possible that requiring CATs (as opposed to making them optional) and having them contribute to the final grade affects students' perceptions of their usefulness? (p. 80). To test this hypothesis, Goldstein used an inventory instrument to compare the student ratings of inventory items in three sections of his Statistics in Psychology class. In two class sections, the CATs were optional and the CATs were required for the third section. A comparison of the means using a two-tailed test for each inventory question, indicated that the students in the mandatory CATs section rated the CAT techniques more positively. In contrast however, two-tailed independent sample *t*-test for eight of the nine other items on the inventory were not statistically significant. An important limitation to this study was the fact that the mandatory CATs section could earn additional points by completing the CATs which introduced a bias toward the student participants viewing the CATs more favorably.

Context of Findings

The context of this study was limited to one institution and further limited to only one faculty member who taught only four sections of Calculus I. The sample size was 84 study participants ($N = 84$) with 40 students in the treatment group ($n = 40$) and 44 students in the control group ($n = 44$) which is a sufficient size for independent sample *t*-tests; the general rule for conducting independent sample *t*-test when the data is with normally distributed is ≥ 30 observations in each group being compared. And finally, only one CAT was tested in this study. This study's design and context was intentionally selected due to the goal of determining whether or not one CAT would make a significant difference with respect to student learning. The goal was not to test the cumulative effect of more than one CAT, rather the goal was to

determine if the one CAT, *Documented Problem Solutions,* could be attributed to higher scores on the student outcome measures. The choice of only one professor and one institution was also intentional and designed to minimize confounding variables. Given this specific design, only limited comparisons can be made with similar student populations at similar institutions.

Limitations of the Study

This study is a classroom research study that sought to improve teaching and student learning by exploring the efficacy of a classroom assessment technique. Classroom research is heavily grounded in the local classroom context which limits generalizability. Good (1983) posited that arguably, the complexities and uniqueness of each classroom make it impossible to follow a simple research-into-practice model (p. 127). At a minimum, classroom research can yield findings and illuminate concepts that will help both the teacher and the students. The major limitation though is the changing classroom context. For example, simple reductionistic models cannot account for the varied and complex variables in the classroom (Good, 1983).

Had the results of this study been different; for example, a statistical difference was observed on several of the student outcome measures, these findings would be fairly limited to this specific classroom context. The variation in student characteristics in classes can be large or small over time, and an instructional method or assessment technique that seemed effective for one group of students, may not necessarily be effective for a different group of students. However, this is not to say that classroom research cannot produce predictive data or trends that may hold true for the majority of college students most of the time in approximately the same classroom contexts.

The rationale behind the CAT, *Documented Problem Solutions,* is that when students understand and are required to show each problem-solving step, they are more likely to arrive at

the correct answer. On its face, this rationale makes sense and one would think it would hold true for most students. But, on the other side of the ledger, a talented group of math students who are accustomed to skipping the simple steps and working at a faster pace, may find this requirement tedious and frustrating which may result in less problems being completed in a fixed period of time. Moreover, frustration during the problem solving process can lead to an increased error rate. In this study, the Documented Problem Solutions exercise was only required for the homework problems where in theory at least, time would not be an issue. Although students were highly encouraged to show their work in this study, they were not required to document their problem solving steps for the examinations and final examination, which are timed tests. Students were also not limited to the problem solving approaches and steps taught in the class. If the correct answer was obtained via a different problem solving approach, this was perfectly acceptable.

In summary, the main limitation of this study was a small inference space due to the study's context. At best, positive findings may have sparked additional studies in Calculus I and perhaps Calculus II at other institutions to test whether *Documented Problem Solutions* could help additional students be successful. Had similar studies been conducted with favorable results, then it would have been plausible over time to view the increasing number of studies as a positive trend line further demonstrating the efficacy of this particular CAT in the calculus classroom. Likewise, the other possibility is that this study could be the start of a negative trend due to null findings.

Suggestions for Future Research

The results and limitations of this study lead to several thoughts regarding additional inquiry. First and foremost, it is important to keep in mind that this is the first known study to

test a CAT posited by Angelo and Cross (1993) in the calculus classroom. This study found that the CAT tested did not affect academic success over a period of two semesters. But this is only one study that was nuanced quite narrowly which has limited implications, and additional research is needed to further explicate the proposition that these CATs, posited by Angelo and Cross, actually improve student learning.

Future researchers may consider a different research design with more rigorous statistical controls. As an example, the first research design considered for this study was a cross-over design which is a repeated measures design that is popular in medical research. In this design, the study participants serve as their own matched statistical control by switching (crossing over) from the treatment group to the control group halfway through the experiment. Had this design been employed for this study, all student participants would have utilized the paper homework format for the first seven weeks of the semester. Then, at the midpoint of the semester, all student participants would have been switched over to the on-line homework format. At the completion of the study, the student outcome measures would have been compared on the basis of the first and second half of the semester. The major advantage of the crossover design is that it obviates the need to analyze and compare student entry characteristics between the two groups due to the fact that each student participant serves in both groups and therefore, serves as her or his own matched control.

As with any research design there are advantages and disadvantages. For this study, the main concern about the cross-over design was the possibility of students doing well with one homework format but not the other. A student who earns good grades for the first seven weeks but is then switched to a different homework format and then experiences a drop in grades was an untenable risk for this study. A second disadvantage of the cross-over design for this study

was the student outcome measures being disproportional between the first and second half of the semester; the Final Exam for example, covers content from the entire semester making it indistinguishable between the two homework formats had students experienced both formats. The same would be true for the Final Course Grade student outcome measure. Despite these disadvantages, future researchers may be able to utilize a cross-over design with the proper student consent and student outcome measures that are similar for the first and second half of the semester.

Furthermore, additional inquiries would broaden the scope of this study. This study took place at a private university in the southwest that tends to attract academically high achieving students. For example, the average ACT Math scores for the treatment and control group were 26.77 and 26.68 respectively. These average scores for both groups are approximately six to seven points higher than the national ACT Math average score which is approximately 20. The students in this study as a whole may have demonstrated a higher level of mathematical competency than students would at other institutions. Similar research at other institutions, such as a community college with remedial mathematics students may yield a different result.

In addition, future studies that have larger sample sizes could possibly help to further determine the overall effect of the *Documented Problem Solutions* CAT on student learning in calculus. Larger sample sizes at other institutions should have students who differ from the participants of this particular study. For instance, students at an institution that is less selective should equal a greater range of mathematical competencies. Another example is students at institutions that have a more diverse study body. Larger and more diverse samples may provide greater clarity and trigger more nuanced inquires based on race, gender and other relevant subcategories that embody social identity. Pascarella (2006) posited that it is a mistake to

assume that an intervention will have the same impact on all students; that is, the general effects on all can be vastly different from a conditional effect on a particular subset of students. "The same intervention or experience might not have the same impact for all students, but rather might differ in the magnitude or even the direction of its impact for students with different characteristics or traits" (p. 512).

Applying Pascarella's (2006) observations above to the present study, no general effects or conditional effects were observed with respect to the student outcome measures. The independent samples *t*-test for the non-white students in this sample was the same result for all students in the sample. However, the sample of non-white students was insufficient and future studies with larger sample sizes may bear different results such as a conditional effect that does not hold true for all students in the larger sample. This logic certainly applies to other student characteristics beyond the traditional categories of diversity based on ethnicity or race, such as, background diversity, affluence and disability to name just a few.

Conclusion

Although the results of this study were null findings, this study is the first known study to test an unmodified or pure classroom assessment technique (CAT) posited by Angelo and Cross (1993) in the calculus classroom. The fact that there are no other known studies in calculus is somewhat surprising given the well-established link between calculus and success in the STEM disciplines. Brown (1991) noted 27 years ago that calculus can either be a major stumbling block or the linchpin to success in the STEM disciplines. Interestingly, the Brown study was just two years after Angelo and Cross published their second edition of *Classroom Assessment Techniques: A Handbook for College Teachers* (1993) which list the CAT, *Documented Problem Solutions,* as a classroom assessment technique specifically for mathematics. More recently, Tyre

(2016) noted essentially the same data point as Brown: "Between 2003 and 2009, 48 percent of students pursuing a bachelor's degree in a STEM field switched to another major or dropped out---many found they simply didn't have the quantitative background they needed to succeed" (p. 5).

The findings of this study need not be disheartening but rather a call for more work to be done on the problem of low STEM major success and how calculus is being taught and assessed. If calculus is truly the linchpin to success in the STEM disciplines, more inquiries are needed to broaden the scope and further explicate how it is being taught, and how it might be improved. There are other lenses to look through and no one study or even strand of studies can give the full picture of the true complexity of student success in the calculus classroom. The problem solving aspect is just one variable in a much larger picture.

This study sought to test whether requiring the completion of homework problems a certain way, based on a classroom assessment technique, would lead to better student learning. Future researchers, in conjunction with faculty, might expand this approach well beyond the homework problems to all classroom assignments. It is already common for calculus instructors to teach different problem solving approaches and encourage students to show their problem solving steps for partial credit on short summative assessments. But this can be uncommon at large institutions where introductory mathematics classes are taught via large lecture. In these large classes, scantron testing is popular because these tests can be graded very quickly. Scantron testing is both practical and economical for the institution but does not allow for more thorough grading beyond the correct answer. Documented Problems Solutions stands for the proposition that students benefit by solving problems one step at a time and showing each so that the problem solving process may be graded holistically rather than correct answer driven grading

only. One way to further test this proposition, for example, is to look at the testing methodologies in calculus at large research institutions.

To help more students succeed in calculus and in the STEM disciplines, effective instructional design and assessment are key variables but not the only variables. For educators to fully grasp the nettle of student success, more work is necessary. This work should not be viewed as a burden but rather an opportunity to effect meaningful solutions to problems that can be solved. We owe our students nothing less than our best efforts to help them succeed.

REFERENCES

Adams, P. (2004). Classroom assessment and social welfare policy: Addressing challenges to teaching and learning. *Journal of Social Work Education, 40*(1), 121-142.

Agrawal, D., & Khan, Q. (2008). A quantitative assessment of classroom teaching and learning in engineering education. *European Journal of Engineering Education, 33*(1), 85-103.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for faculty.* Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning.

Bergsten, C. (2007). Investigating quality of undergraduate mathematics lectures. *Mathematics Education Research Journal, 19*(3), 48-72.

Beswick, K. (2005). The beliefs/practice connection in broadly defined contexts. *Mathematics Education Research Journal, 17*(2), 39-68.

Biggs, J. B. (2011). *Teaching for quality learning at university: What the student does.* UK: McGraw-Hill Education.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Boud, D., & Falchikov, N. (2005). Redesigning assessment for learning beyond higher education. *Research and Development in Higher Education, 28*, 34-41.

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics, 112*(1), 3-11.

Brown, F. E. (1991). The effect of structured notetaking on student success in Calculus I. *Community/Junior College Quarterly of Research and Practice, 15*(3), 261-270.

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher, 70*(1), 30-35.

Bybee, R. W. (2010). What is STEM education? *Science (New York, N.Y.), 329*(5995), 996. doi: 10.1126/science.1194998

Carnevale, A. P., Smith, N., & Melton, M. (2011). *STEM: Science technology engineering mathematics.* Washington, DC: Georgetown University Center on Education and the Workforce.

Clark, B. R. (1960). The "cooling-out" function in higher education. *American Journal of Sociology,* 569-576.

Cook, T. D., & Campbell, D. T. (1979). *The design and conduct of true experiments and quasi-experiments in field settings* [Reproduced in part in *Research in Organizations: Issues and controversies*]. Pacific Palisades, CA: Goodyear Publishing Company.

Cottell, P., & Harwood, E. (1998). Do classroom assessment techniques (CATs) improve student learning? *New Directions for Teaching and Learning, 1998*(75), 37-46.

Cross, K. P. (1998). Classroom research: Implementing the scholarship of teaching. *New Directions for Teaching and Learning, 1998*(75), 5-12.

Cross, K. P., & Steadman, M. H. (1996). *Classroom research: Implementing the scholarship of teaching*. San Francisco: Jossey-Bass.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *Bioscience, 61*(7), 550-558.

Ellis, J., Kelton, M. L., & Rasmussen, C. (2014). Student perceptions of pedagogy and associated persistence in calculus. *Zdm, 46*(4), 661-673.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education, 26*(3), 227-298.

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 of the United States of America, 111*(23), 8410-8415. doi: 10.1073/pnas.1319030111

Goldstein, G. S. (2007). Using classroom assessment techniques in an introductory statistics class. *College Teaching, 55*(2), 77-82.
Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=25127245&site=ehost-live

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* Minneapolis, MN: University of Minnesota Press.

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics. *Journal of Science Education and Technology, 19*(3), 237-245.

Hourigan, M., & O'Donoghue, J. (2007). Mathematical under-preparedness: The influence of the pre-tertiary mathematics experience on students' ability to make a successful transition to tertiary level mathematics courses in Ireland. *International Journal of Mathematical Education in Science and Technology, 38*(4), 461-476.

Inoue, A. B. (2004). Community-based assessment pedagogy. *Assessing Writing, 9*(3), 208-238.

Isaksson, S. (2008). Assess as you go: The effect of continuous assessment on student learning during a short course in archaeology. *Assessment & Evaluation in Higher Education, 33*(1), 1-7.

Joosten-ten Brinke, D., Van Bruggen, J., Hermans, H., Burgers, J., Giesbers, B., Koper, R., & Latour, I. (2007). Modeling assessment for re-use of traditional and new types of assessment. *Computers in Human Behavior, 23*(6), 2721-2741.

Kajander, A. (2006). Striving for reform based practice in university settings: Using groups in large mathematics classes. *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 16*(3), 233-242.

Kaplan, M., Meizlish, D., O'Neal, C., & Wright, M. (2007). A research-based rubric for developing statements of teaching philosophy. In D. R. Robertson and L. B. Nilson (Eds.), *To Improve the Academy* (pp. 26, 242-262). San Francisco: Jossey-Bass.

Kensington-Miller, B., Sneddon, J., Yoon, C., & Stewart, S. (2013). Changing beliefs about teaching in large undergraduate mathematics classes. *Mathematics Teacher Education and Development, 15*(2). Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1018710&site=ehost-live

Krantz, S. G. (1999). *How to teach mathematics.* St. Louis, MO: American Mathematical Soc.

Larson, M. (2017, May). National Council of Teachers of Mathematics. *Math Education Is STEM Education!* Retrieved from: https://www.nctm.org/News-and-Calendar/Messages-from-the-President/Archive/Matt-Larson/Math-Education-Is-STEM-Education!/

Maxim, P. S. (1999). *Quantitative research methods in the social sciences.* UK: Oxford University Press.

Millett, K. C. (2002). Making large lectures effective: An effort to increase student success. In D. Holton et al. (Eds.), *The Teaching and Learning of Mathematics at University Level* (pp. 137-152). The Netherlands: Kluwer Academic Publishers.

Mulryan-Kyne, C. (2010). Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education, 15*(2), 175-185.

Murray, K., & Macdonald, R. (1997). The disjunction between lecturers' conceptions of teaching and their claimed educational practice. *Higher Education, 33*(3), 331-349.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal,* 113-130.

Nichols, J. (2017). *MTH 1321: Differential calculus of a single variable—Introduction to the definite integral and the Fundamental Theorem of Calculus.* Unpublished syllabus for Department of Mathematics. Waco, TX: Baylor University.

Pampaka, M., Williams, J., Hutcheson, G., Wake, G., Black, L., Davis, P., & Hernandez-Martinez, P. (2012). The association between mathematics pedagogy and learners' dispositions for university study. *British Educational Research Journal, 38*(3), 473-496.

Pascarella, E. T. (2006). How college affects students: Ten directions for future research. *Journal of College Student Development, 47*(5), 508-520.

Phillips, R. (2005). Challenging the primacy of lectures: The dissonance between theory and practice in university teaching. *Journal of University Teaching & Learning Practice, 2*(1), 2.

Pratt, D. D. (1998). *Five perspectives on teaching in adult and higher education.* ERIC. Retrieved from https://eric.ed.gov/?id=ED461013

Rieg, S. A., & Wilson, B. A. (2009). An investigation of the instructional pedagogy and assessment strategies used by teacher educators in two universities within a state system of higher education. *Education, 130*(2), 277.

Rudestam, K. E., & Newton, R.R. (2001). Surviving your dissertation: A comprehensive guide to content and process (2nd ed.). Thousand Oaks, CA: Sage Publications.

Scaife, J., & Wellington, J. (2010). Varying perspectives and practices in formative and diagnostic assessment: A case study. *Journal of Education for Teaching, 36*(2), 137-151.

Seymour, E. H., & Hewett, N. M. N. (1997). *Talking about Leaving: Why Undergraduates Leave the Sciences.* Boulder, CO: Westview Press.

Simpson-Beck, V. (2011). Assessing classroom assessment techniques. *Active Learning in Higher Education, 12*(2), 125-132.

Sin, L. Y., Cheung, G. W., & Lee, R. (1999). Methodology in cross-cultural consumer research: A review and critical assessment. *Journal of International Consumer Marketing, 11*(4), 75-96.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sciences Education, 12*(4), 618-627. doi: 10.1187/cbe.13-08-0154

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE Life Sciences Education, 13*(4), 624-635. doi: 10.1187/cbe.14-06-0108

Steen, L. A. (1988). *Calculus for a new century: A pump, not a filter.* Papers presented at a colloquium (Washington, DC, October 28-29, 1987). MAA notes number 8. ERIC. Retrieved from https://eric.ed.gov/?id=ED300252

Tolson, H. (2012). *EDAD/EHRD 690A: Theory of Educational Research*. Unpublished syllabus for Department of Educational Administration and Human Resource Development. College Station, TX: Texas A&M University.

Tyre, P. (2016, March). The Math Revolution. *The Atlantic.* Retrieved from https://www.theatlantic.com/magazine/archive/2016/03/the-math-revolution/426855/

Wake, G. (2011). Introduction to the special issue: Deepening engagement in mathematics in pre-university education. *Research in Mathematics Education, 13*(2), 109-118.

Walker, C. J. (1991). Classroom research in psychology: Assessment techniques to enhance teaching and learning. *New Directions for Teaching and Learning, 1991*(46), 67-78.

Walker, D. (2012). Classroom assessment techniques: An assessment and student evaluation method. *Creative Education, 3*(06), 903.

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE Life Sciences Education, 13*(3), 552-569. doi: 10.1187/cbe.14-02-0023

Yoon, C., Kensington-Miller, B., Sneddon, J., & Bartholomew, H. (2011). It's not the done thing: Social norms governing students' passive behaviour in undergraduate mathematics lectures. *International Journal of Mathematical Education in Science and Technology, 42*(8), 1107-1122.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*(4), 477-501.

Young, C., Georgiopoulos, M., Hagen, S., Geiger, C., Dagley-Falls, M., Islas, A., Forde, D. (2011). Improving student learning in calculus through applications. *International Journal of Mathematical Education in Science and Technology, 42*(5), 591-604.