University of Nevada, Reno

An Examination of Flipped Instructional Method on Sixth Graders’ Mathematics Learning: Utilizing Propensity Score Matching

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Counseling and Educational Psychology

by

Darren G. Ripley

Dr. Leping Liu/Dissertation Advisor

May 2015
We recommend that the dissertation prepared under our supervision by

**DARREN G. RIPLEY**

Entitled

**An Examination Of Flipped Instructional Method On Sixth Graders’ Mathematics Learning: Utilizing Propensity Score Matching**

be accepted in partial fulfillment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

Leping Liu, PhD, Advisor

Cleborne Maddux, PhD, Committee Member

Livia D’Andrea, PhD, Committee Member

Robert Quinn, PhD, Committee Member

Jerry Johnson, PhD, Graduate School Representative

David W. Zeh, Ph. D., Dean, Graduate School

May, 2015
Abstract

There is a widely held belief among stakeholders in the field of mathematics education that we as a nation are losing ground when it comes to educating our students. In the past, technology has been used by educators to augment student learning. However, as we move deeper into the twenty-first century, the role of technology is beginning to change from that of supporting instruction to actively teaching students. Classroom flipping is an example of how technology can be used in this manner, and it has been posited that it could change the educational landscape forever. Classroom Flipping is the practice of taking direct instruction and moving it from the group learning environment to the individual learning environment (Mussalam, 2012). The concept of classroom flipping is a relatively new idea in the field of education, but is becoming increasing prevalent in the educational lexicon, as well as the research literature. Recent surveys in the field demonstrate an increase in the number of teachers, administrators and stakeholders who are interested in the practice and believe it is a valid teaching method that will, for many content areas, become the preferred method of content delivery. Research in the field at the secondary and post-secondary level is becoming more readily available; however there has been very little published research at the elementary level. Research at this level is fraught with ethical, legal, and logistical difficulties.

As traditional experimental methods aren’t always practical to educators and researchers, this study explored the use of a statistical method known as Propensity Score Matching (PSM). PSM enables researchers to use data from observational studies to create a “quasi-experimental” setting that mimics a randomized controlled trial in order to determine treatment effects (Rosenbaum and Rubin, 1983). PSM also has been shown to reduce the biases known to plague observational studies when attempting to use them to determine treatment effects. PSM relies
heavily on the ability of users to establish covariate balance through the use of propensity scores, and this dissertation will provide readers with the criteria by which researchers can ensure covariate balance. For this study, statistical and graphical tools were used to determine that a 1:5 treatment to control group ratio, without replacement of control subjects, and with a .1 caliper distance used to match control units was optimal for the purposes of matching subjects.

Second, utilizing PSM, this study determined that there were no statistically significant differences between the learning outcomes of sixth graders who have received a flipped learning experience and those who haven’t on a standardized assessment. This study utilized the data from a teacher in the Washoe County School District (WCSD) known to have used the flipped learning method with her sixth grade math students and compared their learning outcomes on the Math Criterion Reference Test (CRT) to other sixth grade students who didn’t receive flipped instruction in their sixth mathematics classes.

Third, this study used survey responses from WCSD teachers who have flipped their math and science classes to explore their perceptions of (a) what constitutes a flipped classroom, (b) how student performance has changed as a result of flipping their classes, and (c) how their roles as educators has changed as a consequence of flipping their classes. The survey results showed that teachers’ beliefs about what constitutes a flipped classroom is consistent with the literature. They also believe that student performance has improved as a result of flipping their classes. Lastly, they believe their roles as educators have changed into more collaborative roles, where they are able to spend more time and explore deeper concepts with their students.

This study will add to the growing body of literature around classroom flipping and Propensity Score Matching in educational research. Ideally, educational researchers will use this study as a starting point to continue and expand upon the ideas introduced in this study, and
conclusively determine a concrete set of best practices both for educators choosing to flip their classes and educational researchers wanting to use PSM in their work.
For my son, August, in hopes that you will see that there are some things that can never be taken away from you. You finally get your dad back! For my mother, who has been the ultimate source of support and guidance throughout the process. For my father, whom I wish could have been here to see the journey finished. And, lastly, to my students who have been my inspiration and who have endured my frustrations and struggles from beginning to end.
Acknowledgements

This dissertation would not have been possible without the undying support of the people in my life who love and believe in me. I will be forever grateful to my colleagues and coworkers who have encouraged me and had faith even as mine was flagging. I would also like to take the time to acknowledge my esteemed committee members who have worked patiently to ensure I completed the process. In particular, I want to thank my chair, Dr. Leping Liu, for inspiring me to reach for something I never would have conceived of without her. Your patience and diligence in working with me was humbling and I hope I can live up to the high standard you have set. I look forward to collaborating with you on future endeavors. Thank you all so much.
ABSTRACT .................................................................................................................. i
DEDICATION ........................................................................................................ iv
ACKNOWLEDGMENTS .......................................................................................... v
TABLE OF CONTENTS ......................................................................................... vi
LIST OF TABLES ................................................................................................... xii
LIST OF FIGURES ................................................................................................ xiii

CHAPTER ONE - INTRODUCTION

Background ............................................................................................................. 1

   The State of Mathematics Education ................................................................. 1
   Trying to Change the Tide: Laws and Policies ................................................... 2

Statement of Problem ............................................................................................ 4

Practice in the Field ............................................................................................... 6

   Education Technology ....................................................................................... 6
   A Dearth of Technology Standards .................................................................. 7

   The 21st Century Classroom: Making Technology Part of Standardized
   Practice in the Classroom ................................................................................. 8
   How Learning Theories Inform our Decisions About Technology
   in the Classroom ............................................................................................... 11

Classroom Flipping ............................................................................................... 13

   A Practical Use of Technology that Actually
   Frees Up More Time ......................................................................................... 13
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Research</td>
<td>14</td>
</tr>
<tr>
<td>Research Methods for Robust Examination of Classroom Flipping</td>
<td>15</td>
</tr>
<tr>
<td>Propensity Score Matching</td>
<td>15</td>
</tr>
<tr>
<td>Rationale of Study</td>
<td>17</td>
</tr>
<tr>
<td>Purposes and Research Questions</td>
<td>18</td>
</tr>
<tr>
<td>CHAPTER TWO – LITERATURE REVIEW</td>
<td></td>
</tr>
<tr>
<td>Part I-Classroom Flipping</td>
<td>20</td>
</tr>
<tr>
<td>Introduction</td>
<td>20</td>
</tr>
<tr>
<td>Definition and History of the Flipped Classroom</td>
<td>22</td>
</tr>
<tr>
<td>Flipped, Blended and Online Learning</td>
<td>25</td>
</tr>
<tr>
<td>Research and Theoretical Underpinnings</td>
<td>28</td>
</tr>
<tr>
<td>Students Actively Learning</td>
<td>28</td>
</tr>
<tr>
<td>Peer Instruction</td>
<td>29</td>
</tr>
<tr>
<td>Pre-Exposure to Material and Cognitive Load</td>
<td>30</td>
</tr>
<tr>
<td>Special Needs Students and Culturally Diverse Learners</td>
<td>30</td>
</tr>
<tr>
<td>Instructional Design in Classroom Flipping</td>
<td>32</td>
</tr>
<tr>
<td>ADDIE</td>
<td>32</td>
</tr>
<tr>
<td>Implementation of Classroom Flipping Within the Context of Learning Models</td>
<td>34</td>
</tr>
<tr>
<td>Implementation of Classroom Flipping: Examples from Around the Country</td>
<td>36</td>
</tr>
<tr>
<td>Research in the Field</td>
<td>37</td>
</tr>
<tr>
<td>Byron High School, Minnesota</td>
<td>38</td>
</tr>
<tr>
<td>Clintondale High School, Michigan</td>
<td>38</td>
</tr>
<tr>
<td>University of Puerto Rico, Mayaguez</td>
<td>39</td>
</tr>
</tbody>
</table>
Part I- Using Propensity Score Matching ................................................................. 57

Where Propensity Scores Come From:

The Logistics Regression Equation ................................................................. 57

Creating the Propensity Scores Using the Logit Function:

Selecting the Covariates .................................................................................. 58

The Logit Model ................................................................................................. 59

Counterfactuals and the Quasi-Experimental Setting .................................... 60

Matching Subjects Based on Propensity Scores ............................................. 61

Three Choices a Researcher Must Make Prior to Matching .......................... 62

Three Matching Algorithms ............................................................................ 65

Ensuring Covariate Balance ........................................................................... 68

Part II- Current Study ......................................................................................... 70

Participants ........................................................................................................ 70

Experimental Group ......................................................................................... 71

Control Group .................................................................................................... 72

Measures ............................................................................................................ 73

Covariates for Propensity Score Matching .................................................... 74

The Criterion Reference Test (CRT) .............................................................. 75

Variables for $t$-tests ........................................................................................ 77

Teacher Surveys ................................................................................................. 78

Research Design and Procedures ..................................................................... 81

Data Analysis .................................................................................................... 83
LIST OF TABLES

Table 1. Descriptive Statistics for Omnibus t-test .................................................. 99
Table 2. Omnibus t-test .................................................................................................. 99
Table 3. Descriptive Statistics for Paired t-test for Means of Control Group .......... 100
Table 4. Paired t-test for Means of Control ................................................................. 100
Table 5. Descriptive Statistics for Paired t-test for Medians of Control Group ....... 100
Table 6. Paired t-test for Means of Control Group ...................................................... 100
Table 7. Descriptive Statistics for Independent Samples t-test ................................. 101
Table 8. Independent Samples t-test ........................................................................... 101
LIST OF FIGURES

Figure 1. Common Core State Standards for Math, Content Standards .................. 10
Figure 2. Concept Map of Major Themes in Flipped Learning Literature ............... 21
Figure 3. ADDIE Model ....................................................................................... 33
Figure 4. Side-by-Side Histograms Demonstrating Common Support .................... 64
Figure 5. Side-by-Side histograms C. from Previous Figure ................................ 66
Figure 6. List of Covariates ................................................................................ 76
Figure 7. Independent and Dependent Variables Used for t-tests ....................... 78
Figure 8. PSM Procedures .................................................................................. 86
Figure 9. Hansen and Bowers Overall Balance Test for all Eight Matching Methods .............................................................. 90
Figure 10. Multivariate Imbalance, $L_1$ .............................................................. 90
Figure 11. Summary of Unbalanced Covariates ($|d| > .25$) ............................... 91
Figure 12. Dot Plots for 1:1 Matches Displaying $d$ Before and After Matching ...... 92
Figure 13. Dot Plots for 1:5 Matches Displaying $d$ Before and After Matching ...... 93
Figure 14. Common Support Histograms for all 1:1 Matching Methods ............... 95
Figure 15. Common Support Histograms for all 1:5 Matching Methods ............... 96
Figure 16. Concept Map from Teacher Surveys Using all Questions .................... 103
Figure 17. Concept Map for Questions Four and Five from Teacher Surveys .......... 106
Figure 18. Concept Map for Questions Eleven Through Fifteen from Teacher Surveys .......................................................................................... 110
Chapter One

Introduction

Background

The State of Mathematics Education

There is a widely held belief among stakeholders in the field of mathematics education that we as a nation are losing ground when it comes to educating our students. If the 2012 Programme for International Student Assessment (PISA) results are any indication of the direction of mathematics education in the United States, drastic measures need to be taken in the field of math education, and traditional means of delivering mathematics content may need to be greatly altered or reformed. In 2012, 15-year-olds from the U.S., when compared with their international counterparts rated 30th in mathematics, down from 24th in 2009 Organization for Economic Co-operation and Development (OECD, 2014 and OECD, 2010), as measured by PISA. Started in 2000, the PISA exam is given to approximately 470,000 randomly selected 15-year olds from the 34 member nations and 31 partnership states, although sample size can vary greatly from testing period to testing period.

The Trends in International Mathematics and Science Studies (TIMSS), in their published findings from the 2011 testing, reported that 4th graders from the United Stated ranked eleventh in the world in mathematics ability and 8th graders ranked ninth out of a total of sixty-three countries (Mullis, et al., 2012). While the TIMSS assessment doesn't include as many countries as the PISA, and critics argue that the countries it does include tend to be more biased against countries with larger poverty rates (Tienken, 2013), its conclusions are recognized and often quoted as valid measures of comparative data in international math and science student performance.
The aforementioned statistics demonstrate a lack of mathematical growth in students in the U.S. as they matriculate through the educational system, particularly toward the end of their secondary education careers. What happens to our students between the eighth- and tenth-grades to create such an enormous decline in international rankings?

**Trying to Change the Tide: Laws and Policies**

In the face of such daunting results, the leaders of the U.S. scrambled to stem the tide of what appeared to be a relegation of United States mathematics education into second-class status and begin serious reform. Something had to be done if the United States was going to maintain its competitive edge in a global economy. Education reform is nothing new and comprehensive laws such as The Elementary and Secondary Education Act (ESEA, 1965) and No Child Left Behind (NCLB, 2001) are evidence of the federal government's authority in regulating education in America. However, it must be noted that it was under these pieces of regulatory legislation that the decline initially happened.

The first federal attempt to address the specific problem of an achievement gap between United States math students and the rest of the world came in the form of the Common Core State Standards (CCSS). Announced on June 1st, 2009, but officially rolled out on June 2nd 2010, the CCSS were created to address many issues related to shortcomings of students that were graduating from American high schools who were neither prepared to enter the workplace nor the University (Achieve Inc., 2005). While adoption of the CCSS by states isn’t compulsory, adoption is incentivized by eligibility for the Obama Administration's "Race to the Top Initiative" (USDOE, 2009), wherein states compete for $4.35 billion in federal grant moneys to improve education. The initiative requires that states implement "internationally benchmarked standards and assessments that prepare students for success in college and the work place" (U.S.
Department of Education). Again, though states don't have to adopt the standards, creating a set of standards that meet the criteria is prohibitive, and thus many states - 44 out of 50 - simply opted to use the CCSS to streamline eligibility. Recently, Indiana, Oklahoma, and South Carolina repealed the implementation of the CCSS and, thus withdrew from the member states that had adopted them.

These standards presented an alternative to the "mile wide - inch deep" teaching philosophy that had permeated mathematics education for decades, where students were given large amounts of mathematical topics to cover at each grade level, but weren't able to delve deeply into the concepts. Educators believe that this approach hobbled student understanding and created a culture of memorization of facts rather than genuine learning. The CCSS, rather, focuses on a limited number of general topics - usually three or four at each level - and allows educators to focus mathematical learning around the higher levels of Blooms Taxonomy, most specifically critical thinking and real-word problem solving. The model for the CCSS, not coincidentally, is based on learning done in Taiwan, Singapore and Scandinavian countries, all of which scored in the top tier of the PISA Exams (CCSS Initiative, 2012), thus validating the theoretical underpinnings of the new standards.

Challenges confronting implementation of the CCSS are copious and multifaceted. The CCSS aren't a federal mandate and seven states have chosen not to implement them at all, claiming that their own high standards are sufficient for educating their students, whether this eliminates them for RTT funds or not. States also aren't mandated in how they choose to implement the actual standards themselves. Issues such as order of presentation, educational materials used, and assessment materials used are left entirely up to the states, and sometimes - particularly in Nevada - left to the individual districts to decide. Critics rightfully claim that this undermines one of the
major tenets of the CCSS, which is to provide consistency and uniformity for any student moving through the system (CCSS Initiative, 2012). What has become a major obstacle to implementation is a lack of pedagogical training for educators, specifically at the primary and middle grades. In particular, the level of mathematics teachers themselves experienced during their undergraduate and teacher training is daunting.

It has been well documented that the majority of mathematics teachers in the public school have not earned either a major or minor in mathematics (Seastrom, M. M., et al, 2004) and many of the concepts the CCSS require teachers to communicate to their students are at a depth at which the teachers themselves have never had to master. The amount of supplemental education and training that elementary and middle grade teachers might need to effectively teach the standards with fidelity is daunting at best, and unrealistic at worst. There is potential for a mass migration of talented teachers choosing to change careers rather than endure what could be considered an onerous process akin to earning an entirely new teacher certification requirement. Teachers, therefore, need more than just guidance from a federally suggested set of standards, but instead need a new set of resources to work with when trying to teach more complex and demanding standards.

Statement of Problem

As students of the United States fall farther behind in mathematics, stake holders are seeking ways to reverse the slide. In the absence of a drastic change in the field, the U. S. runs the risk of losing its position as a global leader, both economically and in its role in education and industry. A sea change in the way we think about mathematics education is necessary. Anachronistic practices such as lecture-based models with little to no student interaction are still being used in mathematics classes today. Some mathematics students experience classes in
which they never use a graphing calculator, either because school budgets aren't able to provide them or because individual instructors forbid their use.

As mentioned previously, many instructors aren’t properly schooled in mathematics instruction (Seastrom, M. M., et al, 2004), and don't have the theoretical background to provide students with the requisite knowledge to fully understand individual topics in the context of a grander scheme. As a result, students see and understand mathematics to be a series of tasks with no interrelation with other concepts. They simply move from one set of skills to the next without any sense of urgency for learning the task as a means of making subsequent tasks more easily understood and providing a greater sense of understanding of mathematics as an aggregation and interaction of ideas. Traditional methods of teaching provide no alternative from this model. Our textbooks provide information in chunks (chapters) which are seemingly unrelated to other chapters, at least to the untrained, and teachers are in the habit of assessing after each chapter, with little or no chance to incorporate meaningful review of topics as a prologue to new material.

This problem in modern education isn't a function of incompetence or ignorance, but rather a lack of time available for direct interaction with students. Teachers want to incorporate group- and project-based learning but feel compelled to race through the material to ensure students see (if not learn) it before the next standardized exam, for which they will be held accountable. Administrators at the site level are subjected to the same scrutiny of accountability from their superiors, and the cycle simply perpetuates itself. Any reasonable solution to the problem is going to have to address the lack of time that instructors have in the classroom to interact with - not instruct - their students. Students need more time to assimilate material before they can be expected to perform at the higher levels of Bloom’s taxonomy.
Practice in the Field

**Education Technology**

Technology in education (Ed Tech) and educational systems has become as ubiquitous as the pencil and notebook (Hoyles & Lagrange, 2010), and mathematics is no exception. Educational Technology influences the way in which lessons are taught, planned and assessed, but do we really know how effective technology is in creating better learners, thinkers, and problem solvers? To complicate matters further, the term “Ed Tech” is ambiguous and inchoate, and if a number of math teachers were asked for a definition, one could expect many different responses.

Slide rules, scientific calculators, graphing calculators and computer algebra systems (CAS) were all considered technological leaps forward at their inception, but their impact was instantaneous and easily identified. However, as new technologies emerge, teachers may struggle to determine effectiveness, and may feel overwhelmed by the sheer number available. New learning technologies appear faster and information about them is disseminated more quickly than sound research about their effectiveness can be published.

As early as the mid-1980’s, researchers began grappling with how effective new technologies were performing in the mathematics classroom. Heid (1988) published what is considered by many to be the first study to measure the efficacy of technology in the classroom, wherein it was determined that the use of the CAS improved overall performance by students in a university-level Calculus course when compared to students who weren’t using the technology. Further, Doerr and Zangor (2000) explored how handheld technologies such as graphing calculators enhanced learners’ understanding of mathematics numerically, analytically and graphically. Drijvers (2013) asserts these studies laid the foundation for instructional design that
incorporates personalized delivery of content through the use of computers. However, as technology in the classroom becomes more diverse, nebulous and readily available - either through free apps from the Internet or paid technologies from publishers - stakeholders are forced to make decisions that won’t necessarily have the optimal impact on students in the classroom. Liu, Maddux and Johnson (2008) warn that many new “fad” technologies have permeated education without being properly vetted by research. Moreover, robust research and field trials should be implemented prior to any new technology being institutionalized as part of any curriculum.

Not only would more research provide educators with political cover for their decisions, but it would comfort classroom teachers who may be wary of other such technologies that were discarded after being found ineffective, expensive, or difficult to implement in the classroom setting. Policy makers and curriculum designers at the highest level should be well informed as to the efficacy of the new technologies that are available to classroom teachers, and that information should be passed down to superintendents, technology and curriculum coordinators, and principals and teachers.

A Dearth of Technology Standards

While the CCSS do not mandate specific standards for using technology to teach the mathematics standards - in fact, there are no standards that relate directly to technology in the document – there is an understanding by all stakeholders that technology will be used in the classroom:

“Mathematically proficient students at various grade levels are able to identify relevant external mathematical resources, such as digital content located on a website, and use
them to pose or solve problems. They are able to use technological tools to explore and deepen their understanding of concepts” (CCSS Initiative, MP5, 2012).

Exactly how should classroom teachers integrate technology into their classrooms to better augment the learning of their students?

The International Society for Technology in Education (ISTE) fully endorses the implementation of the CCSS as a means to improve student performances and states: “technology, used effectively, can help all students meet and exceed the rigorous learning goals embedded in the CCSS” (www.ISTE.org). The ISTE makes no formal recommendations for how teachers should apply best practices for implementing technology in their classrooms, although it does have a recommended book list, published and endorsed by them, which provide teachers with resources for doing so.

The 21st Century Classroom: Making Technology Part of Standardized Practice in the Classroom

The Partnership for 21st Century Skills (P21, 2014), founded in 2002, is an organization of business, education and government leaders whose members include, but aren’t limited to the National Education Administration, Apple, Ford, AP, and Intel. The organization's purpose is to: “provide tools and resources to help the U.S. education system keep up by fusing the 3Rs and 4Cs (critical thinking and problem solving, communication, collaboration, and creativity and innovation). While leading districts and schools are already doing this, P21 advocates for local, state and federal policies that support this approach for every school (http://www.p21.org/overview/p21-faq).”

P21 provides a framework, including specific skills students should have in order to be prepared for the 21st century workplace, for all stakeholders but focuses on educators and
policy-makers. P21 emphasizes that for students to be college- and career-ready in the twenty-first century, they must demonstrate literacy in media, information and technology; however it makes no distinct recommendations as to how the use of technology in classrooms should be implemented.

The National Council of Teachers of Mathematics (NCTM), perhaps the most respected source of guidance and leadership for K-12 mathematics teachers, created The Technology Principle, which advocates for the use of technology to enhance and strengthen learning. Specifically, The Technology Principle states that: “Technology is essential in teaching and learning mathematics; it influences the mathematics that is taught and enhances students’ learning” (www.NCTM.org). The NCTM also recommends that our mathematical classrooms reflect the changes that technologies such as graphing calculators and computer software are bringing to mathematics, lauds technology for providing access for special needs students, and emphasizes that technology shouldn’t replace the well-educated and well-trained mathematics teacher in the classroom. However, the NCTM provides educators with no specific guidelines for implementing the use of technology in the classroom. The webpage is replete with lessons and lesson plans educators can use to enrich and engage students for specific standards, however there is no omnibus set of technology standards to which educators can refer in order to ensure their students are technologically prepared to move on, both in the educational and career realms.

It appears that while administrators, educators and policy makers all agree that technology needs to be an integral part of education in general and mathematics in particular, there isn’t a clear path for teachers to take to ensure they are properly preparing their students for the future in technology. Put frankly, unlike the content standards, there are no technology standards in K-12 mathematics education, only recommendations and resources specific to
individual mathematics lessons. If mathematics teachers need to determine a specific Algebra Standard needing to be covered for class, they need only look as far as the CCSS (CCSS Initiative, 2009). See Figure 1 below:

<table>
<thead>
<tr>
<th>Understand the relationship between zeros and factors of polynomials.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CCSS.Math.Content.HSA-APR.B.2</strong> Know and apply the Remainder Theorem: For a polynomial $p(x)$ and a number $a$, the remainder on division by $x - a$ is $p(a)$, so $p(a) = 0$ if and only if $(x - a)$ is a factor of $p(x)$.</td>
</tr>
<tr>
<td><strong>CCSS.Math.Content.HSA-APR.B.3</strong> Identify zeros of polynomials when suitable factorizations are available, and use the zeros to construct a rough graph of the function defined by the polynomial.</td>
</tr>
</tbody>
</table>


However, if a math teacher wants to use technology to help teach this very standard, the teacher would have to pull from previously produced lessons or take the time to come up with one that, hopefully, effectively supplements instruction. An example of a technology standard that could be useful to an educator might be something like:

“Students will use an individual graphing utility (graphing calculator) to identify the zeros of a polynomial and determine that the zeroes obtained graphically correspond to the zeros obtained algebraically. Students will also use the TRACE function on the utility to determine that $p(a) = 0$ if $(x-a)$ is a factor of the polynomial.”

Presently, there are no such standards available to provide guidance to mathematics educators. This would imply that use of technology is a piecemeal process where students can expect different techniques to be taught with different devices using different software; that is, of course, if an instructor chooses to implement technology in the classroom in the first place. In absence of institutional technological standards, there is no way to guarantee uniformity of
practice across mathematics classrooms. If what the policy makers posit is true, that technology isn’t only important for learning mathematics, but important for success in the 21st century, our students aren’t receiving an optimal mathematics education and their futures are at risk.

**How Learning Theories Inform our Decisions About Technology in the Classroom**

Benjamin Bloom’s Taxonomy of Educational Objectives (Bloom, et al., 1956) is well known as the foundation on which most modern learning theories are built (Shane, 1981). One would be hard pressed to find a teacher who had gone through a certification program who couldn’t identify the iconic pyramid and its component structure: Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation. Although there have been modifications to the pyramid – Lorin Anderson proposed the pyramid should be inverted with the most important and complex objectives at the top, and the highest of the objectives be Evaluate and Create (Anderson, 2005) – Bloom’s seminal work is still considered by most educators to be the foundation for all models of good teaching.

Given this fact, how do mathematics educators employ technology to guarantee students ascend through all of the levels of the taxonomy? Most instructors would probably say they would be happy if their students could successfully navigate through Knowledge and Comprehension, particularly because these are the levels at which most standardized assessments measure. With the increased pressure of being held accountable for student performance on these exams, educators may have a valid concern about whether there is enough time to learn how to incorporate new technologies into their methods.

Van Eck (2006) asserts that game-based learning (GBL) can be an important use of technology to help student learning, but he emphasizes creation of software, either by teachers, students, or third party providers, as part of the learning process and requires a buy-in from
educators for the process to be effective. Teachers’ lack of perceived efficacy, Van Eck claims, is an obvious impediment to the successful implementation of this type of direct use of technology as part of instructional design. Mathematics teachers, in particular, may feel that coming up with their own educational programs from scratch may be too onerous and time consuming to commit to the process with fidelity. Third party products are a viable alternative, so long as they aren’t cost prohibitive and address the needs of the classroom directly, assuming the classroom teacher is comfortable using a video game as students’ source of content.

Liu and Cummings (1997) employed the programming language Logo to reinforce geometric thinking and guide students through a learning process the authors referred to as Concrete-Abstract-Concrete thinking. The authors found the algorithmic thinking required to program computers to perform certain geometric tasks helped build schema (concrete to abstract), and assisted students in analyzing and generalizing specific information to broader categories of information. Once the students had achieved this level of abstraction, they could use what they had learned to solve specific problems (abstract to concrete) that required a variety of problem solving skills. This technique, while effective and exciting for students, required a high degree of comfort and efficacy on the part of the instructor.

The limiting factor for the use of any of these theories appears to be time. As mentioned before, if an educator wants to implement technology to help enrich student understanding, regardless of what learning theory is used, that educator must find the time during the school day to do so. Time is needed not just to learn how to use a nascent technology but also there must be time within the instructional day to actually allow the students to interact with and explore the technology and how it augments what they are presently learning. Many teachers would argue that there aren’t enough hours in a day, considering large class sizes and pressure to complete all
of the content, and thus cover all standards. Bearing in mind that educators are held accountable for the performance of their students on standardized exams, and the fact that not all technologies are well researched, this hesitation to fully embrace educational technologies is understandable.

**Classroom Flipping**

**A Practical Use of Technology that Actually Frees Up More Time**

Any use of technology that can actually create more time for the instructor to spend with students is most desirable, and a recent phenomenon in teaching that is beginning to take hold in many classes appears to do just that. The phenomenon is referred to as Classroom Flipping. The Flipped Learning Model, commonly known as Classroom Flipping, was first clearly defined by Ramsey Mussalam. “Classroom Flipping is the practice of taking direct instruction and moving it from the group learning environment to the individual learning environment” (Education Week Webinar, 2012). In 2011, Salman Khan gave a TED talk titled “Salman Khan: Let’s Use Video to Reinvent Education”, which introduced much of the educational world to the idea of “Flipping the Classroom.”

Mr. Khan delivered a twenty minute TED Talk discussing the merits of his own Khan Academy (http://www.khanacademy.org). However, he briefly mentioned the idea of classroom “flipping”, wherein the traditional paradigm, particularly in mathematics education, of a one-size-fits-all, didactic lecture followed by a homework assignment that is done primarily at home by students that may or may not have understood the material, needed to be “flipped.” He claimed that technology now gives educators the ability to deliver material to the students as homework, and use the classroom as a place where students can practice material with the guidance of the instructor, work in groups, and explore deeper meaning within the topics.
Further, educators can more fully explore learning models that, hitherto, had been too time intensive and thus deemed unreasonable.

Flipping gives instructors the much-needed time to teach to all of Bloom’s levels, try new and exciting learning models such as game-based learning or refine models such as group-based learning so that students and teachers alike don’t feel as though they’re going through a perfunctory routine. Rather, classes can work as cohesive units where students and instructors work side-by-side instead of students listening to a teacher deliver content at a pace completely controlled by the teacher. Flipping, if implemented thoughtfully, can empower students to feel they have control over their learning and consider themselves to be as integral a part of the classroom as the teacher (Kim, et al., 2014; Davies, et al., 2013).

**Current Research**

Heretofore, research in the field of classroom flipping has focused mainly on secondary and post-secondary institutions. A query for research in the field will yield studies from University of Miami, University of North Carolina, Chapel Hill, and California State University at Los Angeles, to name a few, and these will be covered in greater detail in Chapter Two. Similarly, secondary schools in Michigan, Colorado and Minnesota as well as others throughout the nation have been studied. Per the literature, there has been little to no published research done at the elementary school level, using younger students’ performances on valid, reliable assessments to determine if classroom flipping is a viable option for elementary educators wishing to change their mathematics classes.

Educators may haplessly use classroom flipping without knowing if it is an effective means of teaching younger mathematics students, or, worse, they may be using a technique that may actually be harmful to younger students’ understanding of the material. There simply has
been no research to determine this one way or the other. A thoughtful study that determines whether classroom flipping, at the very least, does as well as the current instructional techniques can give educators a reasonable case for moving forward with the model. The study presented in this dissertation hopes to accomplish this very task.

**Research Methods for Robust Examination of Classroom Flipping**

**Propensity Score Matching**

Stakeholders in the educational process are more likely to embrace, and thus implement a sweeping change like Classroom Flipping if the model is well researched and significant differences in student performance is shown. Researching flipped classrooms can prove difficult for many reasons, and traditional research techniques have to be rejected for the sake of practicality. The Randomized Control Trial (RCT) is considered by most researchers to be the “gold standard” for research design (D’Agostino, 1998). However, performing an RCT on subjects who are under age students trying to receive an education can prove logistically, legally, and ethically difficult.

An observational study would be a far more convenient technique for a researcher to gather information and gain insight into how students perform in traditionally taught courses versus a course in which content is acquired online. The researcher could simply follow the performance of the students in both the experimental (flipped) class and the control (traditional) class and then measure and compare their performances on assessments throughout the course. Observational studies tend to contain more bias because students don’t have the same probability of placement in the control and experimental groups, thus researchers have been loath to use observational studies for legitimate research in which causal inferences are required to determine treatment effects.
Rosenbaum and Rubin’s seminal paper, *The Central Role of the Propensity Score in Observational Studies for Causal Effects* (1983), created an entirely new tool in statistical analyses, Propensity Score Matching (PSM), to be used by researchers in all fields of interest. Replete with formulae to substantiate their claims, as well as the rigorous proofs of said claims, they discovered a mathematical technique to accomplish what was considered research heresy at the time: use data from an observational study to closely resemble an RCT. This setting is often referred to as a *Quasi-Experimental Design* (Rosenbaum & Rubin, 1983). Rosenbaum and Rubin showed the process so closely mimicked an RCT that the conclusions made as a result of using propensity score analysis (PSM) correctly in research were tantamount to identifying causation, and reliably determining unbiased estimates of treatment effect both at the sample- and population-level (Austin, 2011).

PSM provides researchers with the statistical tools necessary to determine causal relationships within observational studies (D’Agostino, 1998). The premise of PSM is that, through a series of mathematical calculations, any inherent bias that exists in the subjects is removed by creating a "quasi-experimental" setting by virtue of the process. The propensity scores, in themselves, are simply numerical values that are assigned to subjects that are then used to balance data based on common characteristics, known as covariates. Propensity scores are generated using a logistic regression model wherein the set of covariates serve as the independent variable vector and the numerical score - in this case, the propensity score - is the dependent variable. After the propensity scores for each subject within both the experimental and control group have been calculated, subjects from both groups are matched and traditional statistical methods are used to determine whether significant differences exist between or among the groups. Matching is based on the hypothesis that the propensity scores accurately balance
subjects (Love, et al. 2008) and there are a number or matching techniques available to researchers, depending on how the data presents itself after the matching procedure has been completed.

Propensity Score Matching is a relatively new statistical technique in the field of educational research and pivotal to this study. There are a number of statistical platforms that researchers can use to implement PSM. SPSS provides researchers with a convenient way in which to perform PSM, however, there are a number of options that a researcher must choose when using SPSS to create the best possible matches. Among these are whether to match experimental to control subjects with a 1:1 or 1:N ratio, with or without replacement of control subjects, and with or without caliper distances when matching. Greater exploration of the PSM process will be discussed in subsequent chapters.

This dissertation will describe the implementation process for flipping a math class, why the math teachers chosen for the study decided to flip their classes, and whether students’ performances in these flipped classes, by virtue of scores on standardized mathematics tests, were significantly different from those of their peers who did not participate in a flipped mathematics class. Propensity score matching will be employed as the primary statistical method and will be utilized to match experimental and control group subjects prior to the use of more traditional statistical analyses.

**Rationale of Study**

If the United States is to maintain its position among the top countries as having the best economy, educational system and well educated talent pool, decisive action must be taken to ensure its students in general, and math students in particular, are keeping pace with the rest of the world. Policy changes are under way to address this issue, but more needs to happen at the
educator level. Students of mathematics need better methods to receive and assimilate new content. Mathematics teachers are required to be well-versed in technology if they want their students to remain competitive in the 21st Century global education and labor forces. However, they also need time with their students to implement best practices for learning mathematics. Techniques such as classroom flipping may be a good start to a possible solution to the challenges facing mathematics education in the 21st Century.

**Purpose and Research Questions**

The purpose of this study is to examine whether students who participate in "flipped" mathematics classes perform significantly different from their peers who participate in traditional mathematics classes. However, the study will also be an exploration of utilizing Propensity Score Matching as a viable research method. This study will also explore teacher perceptions about the practice of flipping, changes in student performance as a result of flipping, and changes in teacher roles as a result of flipping. The main research questions that will be addressed in the study are as follows:

1. When performing propensity score matching on a particular data set, what are the analytical criteria to determine the optimal covariate balance to decide whether to use (a) 1:1 or 1:N matching ratio, (b) with or without control data replacement, and (c) and with or without caliper distances?

2. Are there significant differences in the mean scores of the Math CRT exam between students who received the "flipped" instructional method and those who received the traditional instructional method?
3. What are the perspectives of instructors who are utilizing flipped learning in their classrooms and, anecdotally, what are their perceptions of the differences in their students' learning?
Chapter Two

Literature Review

This chapter will be divided into two parts. The first part will be solely dedicated to a review of the literature pertaining to classroom flipping. The second part will be an introduction to the statistical method known as Propensity Score Matching (PSM), and recent applications of PSM in the educational field. Because PSM is a relatively new statistical technique in the social sciences and educational research, it warrants a separate examination within this chapter.

Part I. Classroom Flipping

Introduction

Classroom flipping is a recent phenomenon that has gained prominence among educators and policy makers alike. While there are many emerging resources available to educators wishing to flip their classes, there is a decided dearth of research in the field. As a concrete set of best practices materializes, guided by solid research, the field of education has found itself on the precipice of an entirely new and innovative way to engage students in learning. To provide a graphical understanding of the main themes present in the literature pertaining to classroom flipping, a content map, created by the product Leximancer and generated by inputting all articles used in this literature review is presented below in Figure 2. Further explanation of Leximancer and its utility in qualitative research will be explored in Chapters Three and Four, however the figure is intended to provide an understanding of major themes that surround the field of classroom flipping in recent years.
Figure 2. Leximancer output of major themes in flipped learning literature.

For the review of literature for this study, tens of publications and thousands of words were input into Leximancer and the fifteen themes were produced. The “classroom”, “solving”, and “knowledge” themes (spheres) occurred most frequently in the publications, however the “learning”, “students”, and “solving” themes (spheres) had the greatest connectivity within and between documents.
A Definition and History of the Flipped Classroom

The Flipped Learning Model, commonly known as Classroom Flipping, was first clearly defined by Ramsey Mussalam. “Classroom Flipping is the practice of taking direct instruction and moving it from the group learning environment to the individual learning environment” (Education Week Webinar, 2012). The Flipped Learning Network (2014) expanded on the definition by adding “… and the resulting group space is transformed into a dynamic, interactive learning environment where the educator guides students as they apply concepts and engage creatively in the subject matter.”

In 2011, Salman Khan gave a TED talk (http://www.youtube.com/) titled “Salman Khan: Let’s Use Video to Reinvent Education”, which introduced much of the educational world to the idea of “Flipping the Classroom.” A working definition for educators of flipped learning, as posited by Mr. Kahn, among others including the author, is simply using technology to allow content acquisition by students when and where it is convenient for them and, thus, freeing up time during class to create a more dynamic, active space for students to augment their learning. Mr. Khan delivered a twenty minute TED Talk discussing the merits of his own Khan Academy (http://www.khanacademy.org). However, he briefly mentioned the idea of classroom “flipping”, wherein the traditional paradigm, particularly in mathematics education, needed to be “flipped.” I. E, a one-size-fits-all lecture, followed by a homework assignment that is done primarily at home by students who may or may not have understood the material, is an antiquated approach to mathematics education. He claimed that technology now gives educators the ability to deliver material to the students as homework, and use the classroom as a place where students can practice material with the guidance of the instructor, work in groups, and explore deeper meaning within the topics.
The benefits of the technique, Mr. Khan claimed, are multi-faceted for students and instructors alike. Students can choose to acquire information at their own convenience, rather than when they are tired, hungry, or not emotionally prepared to listen to a teacher. They can watch the videos from the comfort of their own homes as many times as they like and whenever their understanding of certain material seems to flag. Students would miss no instruction due to illness or other absence from class. Also, they receive direct, one-on-one instruction from the teacher when they need it, in class, rather than when the teacher has time for it. For instructors, the bane of their existence – time management – becomes a thing of the past. Rather than racing through material to insure students are “taught” concepts, teachers have the ability to engage students individually and customize learning to the needs of their charges. Advanced students can be given rich, insightful projects and problem solving activities to challenge them, while struggling students receive more focused and task-oriented assignments to bring them up to speed and reinforce concepts.

The educational landscape has been fertile ground for ideas like classroom flipping for decades, where technology is used not just as a peripheral tool to augment learning but as a main player in teaching students. Eric Mazur’s 1991 paper: “Can We Teach Computers to Teach?” proposed that educators hadn’t yet utilized technology to its full potential in the field of education. In a moment of insight, Mazur asserted, “I don’t think computers will replace teachers, but I am confident that computers will play an important role in improving teaching.” (1991, p33).

Originally coined as an “inverted classroom” (Lage, et al., 2000), the idea was proposed as a way to mitigate the effects of differences between the learning styles of students and the teaching styles of instructors. In their 1996 landmark study, Lage, et al., used videotaped
lectures and Power Point slides to deliver Economics material to University of Miami students outside of class, and utilized the additional class time to explore projects and labs pertaining to the material covered. Rather than determining whether the “inverted classroom” produced significantly higher grades, the researchers were interested to see if students’ preference for this type of class was significantly higher than the traditional, “chalk and talk” classroom, which it was. Furthermore, they found that students believed they both learned more and received higher grades as a result of the method of teaching (Lage, et al., 2000).

This exact model, although updated with more current technology, has recently been used effectively by Jonathan Bergmann and Aaron Sams at Woodland Park High School in Colorado to teach their students chemistry. Flipping seems like a perfect match for any course in which laboratory instruction is being used to augment understanding, i.e., the hard sciences. Sams and Bergmann have published “Flip Your Classroom: Reach Every Student in Every Class Every Day” (ISTE, 2012), lavishing their praises on the flipped model and providing interested educators with resources, ideas and a theoretical framework in which to operate.

The turn of the 21st century appears to be a pivotal time for ideas about flipped teaching to foment. At the 11th International Conference on College Teaching and Learning (2000), Jack Chambers’ contribution to the field, “The Classroom Flip: Using Web Course Management Tools to Become the Guide on the Side” not only gave the new movement its moniker, but it gave proponents a mantra mission statement: “Become the Guide on the Side, not the Sage on the Stage”. Incidentally, while Chambers is often credited with coining the Phrase “Guide on the Side”, it was Allison King who originally came up with the phrase (King, 1993).

Unbeknownst to most educators, the field of classroom flipping took a leap forward in 2004, when an ex-hedge fund manager by the name of Salman Khan used technology such as
screen capture and tablet software to upload YouTube videos of tutorial lessons he had created
for his cousins, who were struggling in their math classes. Khan chose to leave the availability
options on his videos open, which allowed other students to view his videos and provide him
with feedback, which was resoundingly positive. Khan was so moved by this experience and the
effect it had on others’ lives he chose to invest much of his substantial resources to create the
Khan Academy (www.khanacademy.org), a completely self-contained, comprehensive, free
learning management system (LMS) that allows students to learn not just mathematics, but
thousands of topics across hundreds of fields. The site is meant to be used by students, educators
and coaches (tutors) to supplement student learning and has received mostly positive reviews, as
well as a great deal of publicity in the years following Khan’s aforementioned TED talk,
including a segment on 60 Minutes titled “Khan Academy: The Future of Education?” (60
Minutes, September 2, 2012).

Flipped, Blended and Online Learning

It is important to delineate the differences between flipped, blended and online learning.
Flipped learning is often erroneously confused with the other two models, although there is
overlap between the three. Online learning is widely understood to occur exclusively remotely,
with no physical face-to-face student-teacher interaction (Oblinger & Oblinger, 2005). Most, but
not all interactions between students and teachers and students and other students are
asynchronous and are accomplished through the use of learning management systems (LMS).
Chat rooms, forums and discussion groups are contained within the LMS while the instructor
manages the content, activities and interactions. Clearly, online learning doesn’t lend itself to a
truly flipped experience since there is no “group learning” space, i.e., the classroom.
Blended learning is a combination of both the traditional, brick-and-mortar educational experience wherein students interact directly with an instructor, and some component of online instruction. The use of videos, applets, websites, etc. to obtain content or the use of online tools for assessment and feedback are all considered to fall under the purview of “blended” learning (Allen, et al., 2007). There is an obvious spectrum of what is considered a blended learning experience; however the main attribute of blended learning is the student enjoys the benefits of both face-to-face interaction with a professional educator and the benefits and possibilities of online resources.

Skeptics may say that classrooms where instructors merely sit students down in class and have student watch online videos – presumably of the content which the instructors themselves should be providing – and thus supplanting their own roles in the class with that of the internet, could claim that blended learning is occurring. Those skeptics would be correct. Blended learning prescribes no set of best practices to be used by educators. It is merely the definition of a term.

Classroom flipping is an obvious choice for educators wishing to blend online instruction with in-class interaction, within the context of a more clearly defined pedagogy. Hamdan and colleagues’ (2013) A Review of Flipped Learning, produced for the Flipped Learning Network (FLN, 2014), attempts to create a set of best practices for educators and administrators wishing to flip all or part of their curriculum. They use the predictable acronym, “FLIP”, to highlight guidelines for the practice of flipping:

**Flexible Learning Environments (F).** Flipping is, by definition, a flexible learning environment by virtue of the fact that learners can choose when and where to acquire content. Additionally, teachers should foster this ethos in learners to alter the classroom environment to
supplement content acquisition. The traditional rows-and-columns classroom can hinder group learning, research and project-based learning. Hamdan et al. (2013) assert that “flipped classrooms allow for a variety of modes”, and the classroom environment should be flexible enough to accommodate those modes.

**Shift in Learning Culture (L).** Educational environments, under the flipped paradigm, will shift from the “sage on the stage” to the “guide on the side” (King, 1993). Teachers will need to relinquish some control of their classes and assume more collaborative roles with students. In a well-run flipped classroom, the environment may feel more chaotic as students rely less on the instructor to control the class and more on each other for assistance.

**Intentional Content (I).** Educators must evaluate which content is appropriate for flipping and which must be delivered by direct instruction. Some material isn’t appropriate for flipping (e.g. Socratic discovery) and teachers must critically examine this aspect before deciding to flip content. This said, the purpose of intentional content is to afford teachers the luxury of time, and hopefully abandon the teacher-centered model of learning, in which students are passive recipients of information and the instructor is both the primary source of information and evaluator of whether the student is successful (Huba & Freed, 2000).

**Flipping Requires Professional Educators (P).** Critics of flipping decry it as a way to replace classroom teachers. Quite the contrary, flipping requires hard-working, skilled, well-educated craftsmen who are sensitive to the needs of their students and dedicated to the field of education to create learning environments where students reach their full potential. Teachers must be skilled, flexible critical thinkers and communicators in order to transition students from more antiquated models of direct instruction to newer models in which students have far more control over their learning environments, both in and out of class. Instructors will experience a
degree of disequilibrium as they relinquish the control they once demanded in the traditional paradigm to make way for a more supportive role in what can feel like a chaotic learning environment. Gojak (2012) asserts that teachers must ask themselves not whether or not to flip, but how they will adjust their techniques to fit the new, more flexible environments to ensure learning is occurring.

In summary, flipped learning is an exciting new instructional model that utilizes technology in an intuitive manner, and directly addresses implementing the use of technology in 21st century classrooms. Classroom flipping, while still inchoate, has the potential to demonstrate how traditional educational systems can change to accommodate a new model.

**Research and Theoretical Underpinnings of Flipped Learning**

Because of the nascent nature of classroom flipping, there exists a dearth of quantitative and qualitative research in the field. However, well-researched topics of best practices in education reinforce how flipping, when implemented well, has the potential to create classrooms in which genuine learning occurs.

**Students Actively Learning**

Student-centered learning where participants are actively involved, rather than sit as passive recipients of information, is shown in the research to be far more effective in increasing both learning and achievement (Prince, 2004; Michael, 2006, et al.). As importantly, particularly in the context of the modern assessment paradigm, student academic performance is significantly improved in classrooms where students – not instructors – are at the center of learning and take an active role (Hake, 1998; Knight & Wood, 2005; Freeman, 2007; Chaplin, 2009). As an added benefit, O’Dowd and Aguilar-Roca (2009), found there was an increase in student engagement, critical thinking and positive overall attitudes toward learning in classrooms where students are
actively involved in the learning. In stark opposition to the traditional, “one-size-fits-all”, lecture-based classroom, student-centered learning provides instructors with the opportunity to work alongside students, building collaboration in assessing learning. Michael (2006) asserts that students who have the ability to actively participate in their learning more effectively create, test, and adjust hypotheses that lead to the formation of schema.

Flipped learning provides teachers with the instructional foundation to implement student-centered, activity-based learning in their classrooms. While flipped learning itself isn’t a mechanism for these attributes, it allows the instructor the in-class time to explore them, rather than feeling pressured to “get through” material and stay on schedule with the curriculum.

**Peer Instruction**

Considered by many to be the leading researcher in peer instruction, Harvard’s Eric Mazur has been publishing articles in the field for almost two decades. His seminal 1996 work on the topic, *Peer Instruction: A User’s Manual* emphasized in-class interaction between students and encouraged instructors to engage students and assist them in examining their logic and addressing misconceptions. Mazur posits that once a student’s mind is engaged by the instructor the learning dynamic changes to a collegial relationship in which the student is equally invested in his or her learning. Benjamin Bloom himself (1984) observed that learning and performance are significantly improved when instructors and peers are able to give real-time feedback on a one-to-one basis.

Flipped learning enables teachers to create classrooms in which peer instruction becomes the norm rather than the exception by virtue of the freedom of time it gives them. Because the most time consuming of all classroom activities, the lecture, is transferred to homework, the
teacher is free to explore activities in which students and teacher alike are able to interact openly and in a way that lends itself to a culture of collaborative learning.

**Pre-Exposure to Material and Cognitive Load**

Research shows that previous exposure to material increases memorization and recall (Bodie, et al., 2006). While the researcher's study focused on blended learning and not specifically classroom flipping, the results showed compelling evidence for what the researchers referred to as "priming", or receiving direct instruction outside of the classroom.

John Sweller's Cognitive Load Theory (Sweller, et al., 1998) proposes that there are information processing limits which learners experience when inundated with too much information. Obviously, these limits differ between individuals depending on cognitive processing abilities, but students inevitably "shut down" when the amount of information required for them to process exceeds their limit, and learning is adversely affected. Ramsey Musallam, in a 2010 study, used what he called "pre-training", or exposing students to material prior to the class in which the material is covered, to determine that students required fewer cognitive resources to understand concepts when they received pre-training. Thus, it appears that pre-exposure to material reduces cognitive load for students and allows them to learn more complex content without feeling overwhelmed by the material. The entire foundation of classroom flipping is for students to acquire content prior to any demands of their understanding being placed on them.

**Special Needs Students and Culturally Diverse Learners**

In their 2008 report, the National Mathematics Advisory Panel concluded that instructional methods that can benefit students with learning disabilities include systematic instruction, peer tutoring and self-instruction. Further, while many instructors are already using
technology to supplement instruction, its use in the special needs population, and students with Autism Spectrum Disorder (ASD) in particular, has remained largely unexplored (Ramdoss, et al., 2012). That said, a paper published by Kagohara and colleagues (2013) reviewed 15 studies dealing with the use of technology and special needs, and concluded that the use of mobile devices to supplement instruction can positively impact academic, communication and transitioning skills. Furthermore, Burgstahler (2003) concluded that students with disabilities who are transitioning into post-secondary institutions and/or employment can use technology to assist with transitions, enhance academics, and optimize independence. While it is apparent that technology greatly assists special needs learners, there has been no substantive research in the literature, either qualitative or quantitative, in the field of classroom flipping for these students.

The advantages for a student, regardless of ability level or disability, being able to pause, rewind and repeat a video are invaluable. Positive anecdotes abound, but this field of research appears to be fertile ground for additional work.

Research demonstrates that second language learners benefit from using technology to augment traditional instruction. Marshall and DeCapua (2013) found that a flipped learning model enables second language learners to take as much time as needed with content delivery, again by virtue of the ability to pause, rewind and repeat lectures. Unlike in a traditional classroom where a student struggling with a non-native language must not only process difficult content but also overcome the inability to fully understand the language of delivery, with all of its slang, idioms, and nuance. In the same study, Marshall and DeCapua observe that second language learners in traditional classrooms often wallow at the lower levels of Bloom's Taxonomy (memorization and knowledge) because of their language challenges. Given the opportunity to process the information at their own pace, they can, with the guidance of a skilled
instructor, begin to learn at the higher levels of the taxonomy. An adroit observer may take note that all learners could benefit from a setting where they have the freedom to process content at their own pace, outside of class, and experience learning at the highest levels of the taxonomy in class.

In conclusion, classroom flipping is an excellent tool for implementation of previously researched methods of effective teaching. Students of all types experiencing a flipped classroom, by dint of the decreased amount of time taken away from the class for lecture, can benefit from a more empowered role in the classroom, more face-to-face interaction with the instructor, more peer interaction, and more directed and supported practice with challenging new material.

**Instructional Design in Classroom Flipping**

An important component of any new method of instruction is design. It is incumbent on any teacher wishing to change an instructional method to decide specifically how the new method will be designed, implemented, and assessed. This is particularly true when the method utilizes technology. That said, it can be very difficult to design a course that utilizes technology, which classroom flipping very clearly does, because best practices and the concept of technology integration in the classroom is so difficult to define (Liu & Maddux, 2008). Teachers who wish to use classroom flipping may feel overwhelmed by the idea of creating a well-researched set of principles to guide them when creating lessons, units, and courses. Much like available technology itself, the design needs to be flexible and nimble, reacting quickly to fix any problems that may arise.

**ADDIE**

First developed by Florida State University in 1975 for the US Army, (Branson, et al., 1975) the Instructional System Design (ISD) model, ADDIE, was created to provide instructors
with a model that evolved as feedback from participants within the system was analyzed and improvements were deemed necessary. ADDIE is an acronym for analysis, development, design, implementation, and evaluation, and is presented as a closed loop with “evaluation” at the nexus. Figure 3, below, is one of the many ways to represent the model:

![ADDIE Model Diagram](image)

*Figure 3. ADDIE model.*

The diagram clearly shows that implementation of the ADDIE model is a thoughtful process wherein the designer is constantly evaluating the program through feedback and data provided by stakeholders. This design model lends itself nicely to classroom flipping, particularly for the inexperienced teacher. Flipping is so new to the field of education there are limited resources available to interested educators.

For example, a math teacher may want to do a “full flip”, which implies that 100% of content is available online for students to view outside of class, and no formal lecturing will take place during class time. The teacher can evaluate existing video content (Khan Academy, TeacherTube, e.g.) to ensure it is correctly aligned with the content for a class. If content isn’t available or properly aligned, the teacher might want to create their own video content using
available technology, or collaborate with colleagues to create content. Once the content is created and students are viewing the videos, the teacher might receive feedback from the students or other stakeholders that point out strengths or weaknesses in the videos, which the teacher would then redress. After fully implementing the flip, the teacher might find that test scores are lower than in previous years. The teacher can then scale back the amount of content that is flipped or modify the way in which information is given to students. Fully invested teachers, administrators and curriculum developers can track data longitudinally to determine if student performance is improving and create long-term curriculum strategies, depending on the outcome. Researchers can compare flipping methodologies across subjects and content to determine which methods produce the best outcomes, and the constant feedback structure the ADDIE model provides can ensure a responsive, ever-evolving body of knowledge on which educators can draw.

**Implementation of Classroom Flipping Within the Context of Learning Models**

Project-Based Learning (PBL) is a model built on the constructivist assumptions of Piaget, in which students, working as a team, create solutions to nontrivial questions and present these solutions (artifacts) to demonstrate their knowledge (Blumenfeld, et al, 1991). Constructivist educators can use flipping to allow for more in-class time for students to work on such “nontrivial” projects and free up the environment for more collaborative interactions between students. Because the teacher needn’t control the class so rigidly, they can spend more time with the students and their groups to help tease out ideas and concepts that might have had to be overtly communicated to students in a lecture format.

For example, if a history instructor wishes for his or her students to understand the impact of the fall of the French at Diem Bien Phu and its impact on the United States and the war in
Vietnam, students could be assigned to watch documentaries on colonialism, read primary source documents about the Truman Doctrine and research Vietnam using available resources. They would then work in class to create a treatise to “solve” the problem of the fall of Diem Bien Phu. The treatise would serve as the artifact demonstrating knowledge which the students had constructed from their research which had been done outside of class.

The National Center on Universal Design for Learning (UDL) defines this model as “a set of principles for curriculum development that give all individuals equal opportunities to learn.” UDL defines the three guiding brain networks that are involved with learning as: The Recognition Networks: how learners gather facts and categorize what is perceived, the Strategic Networks: how learners plan and perform tasks, and the Affective Networks: how learners remain engaged and motivated while learning (http://www.udlcenter.org). Proponents of UDL seek to develop curriculum that presents information using a variety of delivery methods, differentiates the manner in which students can demonstrate knowledge and constantly stimulates student interest.

Classroom flipping lends itself to the UDL model readily. Flexible instructors allow students to obtain content in whatever fashion the student learns best, whether it’s from the video casts, from the textbook, or from one-on-one tutorial sessions with the instructor during class time. Assessments could be in the form of exams, both oral and written, as well as written assignments or projects, wherein the students create their own content videos to demonstrate their knowledge. Because class time isn’t hindered by the need of the instructor to lecture constantly, projects, writing assignments, and group work can be used to keep the material novel and engaging.
Proponents of Mayer and Morenos’ Multimedia Learning Theory (1998, 1999) believe that learners learn information more deeply if they are presented the material with both words and pictures than they do with just words (Mayer, 2005). Because flipping creates a paradigm in which teachers must present information using visual digital technology, and not just words from their lectures or Power Point presentations, the flipped classroom, by default, subscribes to the Multimedia Learning Theory. However, any educator interested in using Multimedia Learning Theory to inform their curriculum would need to insure that the 12 Principles of Multimedia Learning are addressed when creating content.

In summary, while only a handful of learning models are mentioned in this dissertation for which flipped learning is ideally suited, it could be argued that any model that requires increased face-to-face time with students, and student-student and teacher-student rapport building is well suited to a flipped class. Learning requires time and attention by an expert in the field and classroom flipping allows for an increase in both.

**Implementation of Classroom Flipping: Examples from Around the Country**

Once an educator has made a decision to flip their class, there appears, predictably, to be as many ways to implement the flip as there are ways to teach a course. Ramsey Musallam, a high school chemistry teacher from San Francisco, flips any material that used to be given in a lecture format. However, Musallam takes a day of class to explore and introduce students to new material before he has them watch videos. Afterwards, students complete laboratory assignments in class. This method gives Musallam’s classes significantly more time to work on labs rather than receiving information, unlike traditional chemistry courses (Saltman, 2011). A similar model was used by Jonathon Bergmann and Aaron Sams in their landmark work in Woodland Park High School chemistry classes. Jay Hooper, an AP Calculus teacher from
Champagne, IL, records all of his direct instruction lectures using a screen capture program named Camtasia (www.techsmith.com) and uploads them to YouTube.com for his students to watch prior to giving them assignments. He then augments the material with projects and group assignments (Saltman, 2011).

Shelley Wright of Cornerstone Christian School in Saskatchewan, Canada takes a day in class to introduce materials through discovery activities and then assigns students to watch Khan Academy or TED Talk videos for supplemental information. This method utilizes more of a blended learning technique, rather than a full flip, but it is important to remember that any method of instruction where content acquisition is obtained outside of the classroom is considered to be, in some degree, a flip (Saltman, 2011).

Bergmann and Sams, considered by many to be leaders in the field of classroom flipping, have written a book, *Flip Your Classroom: Reach Every Student in Every Class, Every Day* (ISTE, 2012), a primer for teachers wishing to flip their classes. Additional publications that directly address flipped learning, both its benefits and challenges that have been published in the last year are *Time for Learning: Top 10 Reasons Why Flipping the Classroom Can Change Education*, (Corwin, 2014), and *Implementation and Critical Assessments of the Flipped Classroom Experience* (Scheg, 2015). Musallam has also become a leader in the field and has given a TED Talk (3 Rules to Spark Learning, 2013) and maintains a website named “Cycles of Learning (http://www.cyclesoflearning.com) to provide a resource for educators and researchers in the field of classroom flipping.

**Research in the Field**

As is expected in an emerging field such as classroom flipping, there is a noticeable lack of empirical research. That said, there are a number of case studies at the K-12 level and several
studies, both qualitative and quantitative at the post-secondary levels, which will be explored below.

**Byron High School, Minnesota**

In 2006, 29.9% of Byron High’s passed the Minnesota Comprehensive Assessment (MCT) for mathematics, their version of the standardized test used for state and federal reporting. The school’s composite score average for the ACT was a 21.2, right at the national average for the year ([www.act.org](http://www.act.org)). In 2009, in an attempt to overhaul the existing system which was creating mediocre outcomes, but faced with the financial challenges of the times which limited available resources, the math department, led by Troy Faulkner, rewrote the department curriculum and decided to flip all of the math classes (Fulton, 2012). As a result, by 2011 73.8% of Byron High’s students passed the MCT and the average composite score on the ACT had increased to 24.5, 3.1 points above the national average of 21.2. These achievements earned the school the National Blue Ribbon School Award in 2010 and the Intel Schools of Distinction award for High School Mathematics in 2011.

**Clintondale High School, Michigan**

After hearing about the flipped learning model in 2010, Principal Greg Green of Clintondale High decided to flip all of his freshman courses. Clintondale High is a school located in a suburb of Detroit, with three quarters of its population being either minority or low-income or both. His school, rife with discipline issues, flagging test scores and unacceptable dropout rates, forced Green was to re-evaluate his approach to instruction, particularly for his freshmen. After hearing about the flipped classroom model, Green decided it was worth experimenting, as the status quo wasn't producing satisfactory results. Green (2012) began to see results almost immediately. From 2010 to 2012, failure rates among freshmen dropped by 33%
in some classes, and discipline issues fell from 736 cases in 2009 to 249 cases in 2010 to 187 cases in 2011. Similarly, parent complaints dropped from 200 cases in 2009 to seven in 2012. Green was so pleased with the results that he chose to implement classroom flipping for all students at all grades. The data from the case study of the entire school hasn't yet been published and it is undetermined if the experiment was successful.

University of Puerto Rico, Mayaguez

Papadopoulos and Roman (2010) used an “inverted” learning model, a synonym for flipped learning, in an electrical engineering class. They discovered that students covered more material faster, which enabled instructors to extend the scope of the course and had a greater depth of understanding of the materials covered. They also found significant differences between student performance on exams between students coming from the flipped environment and their peers in traditional, lecture-based classes. Using a pretest as a baseline, students in the experimental (flipped) group answered 18.3% of the questions correctly while the control answered 17.1% correctly, with no significant difference between the two. After receiving both forms of instruction, the experimental group significantly outperformed the control in percentage of correctly answered questions, 31.2% to 21.4% respectively.

California State University, Los Angeles

Researchers Warter-Perez and Dong (2012) wanted to measure learning outcomes and perceptions of students experiencing a flipped learning model used at CSULA in an introductory-level electrical engineering course. The course’s instructors used flipped instruction to capitalize on in-class time, which they repurposed from lecture to collaborative, project-based learning (CPBL). While the study was primarily concerned with how CPBL affected learning outcomes, the fact that flipped instruction was employed gave the researchers
added value. Quantitatively, the researchers determined that the overall model significantly increased conceptual understanding. Qualitatively, students found that the new model contributed to positive peer interaction, increased understanding, increased note taking, and design skills.

**University of North Carolina-Chapel Hill, North Carolina**

Classroom flipping has expanded into the realm of graduate course methods as well. In an introductory graduate course on pharmaceuticals, which all first year students in the UNC School of Pharmacy must take, Dr. Russell Mumper obtained data from two separate courses: his 2011 course taught using the traditional, lecture-based method, and his 2012 course taught using the flipped method. McLaughlin, et al. (2014) discovered significant differences on the final exam scores. The data also indicated that student attendance increased between the two years and students reported that flipped methods had a greater effect on their overall learning.

**Flipping Isn't a Panacea for all Types of Courses**

Jeremy Strayer, in his 2012 article, *How Learning in an Inverted Classroom Influences Cooperation, Innovation and Task Orientation*, found that students in an introductory statistics course found flipped learning less satisfactory than traditional instruction methods. However, Strayer acknowledges, this may in part be due to the nature of an introductory course, where students have little or no advance knowledge to build on, and thus may feel a heightened sense of cognitive disequilibrium. It may be unreasonable to expect students to learn entirely new material online, without the guidance of a lead instructor, assuaging anxieties over new and often confusing. Johnson and Renner (2012) found no significant differences between assessment scores for students in both an introductory and secondary computer applications course, although the authors report the instructor of the course was hesitant to use the flipped learning model, and
believed there was no need to do so. Similarly, Lape, Levy, and Yong (2014) found no significant differences between students from Harvey Mudd College participating in flipped Science, Technology, Engineering, and Math (STEM) courses there and students receiving a traditional, lecture-based instruction.

**Stakeholders’ Perceptions**

While there is a distinct lack of empirical research in the field of classroom flipping, both qualitative and quantitative, there is an emerging group of surveys that can provide insight into how stakeholders feel about the possibility of flipped learning. The three largest and best known among flipped learning advocates are the 2012 Flipped Learning Network/Classroom Window survey, The Flipped Learning and Democratic Education survey conducted by Tom Driscoll from Columbia University (2012), and the Speak Up online surveys (2012, 2013), conducted by Project Tomorrow, a nonprofit organization dedicated to improving education in the U.S.. What follows is a brief summary of the results of these surveys from the standpoint of the primary stakeholders in the field: students, teachers, administrators and parents.

**Students**

Driscoll's survey of 203 students from flipped classrooms reported that 80% claimed to have had more constant and positive interactions with teachers and peers during class time, as well as better access to content and course materials. 80% also reported the flipped classrooms allowed them the ability to proceed through material at their own pace, exercise greater freedom to choose how to demonstrate their knowledge, and viewed learning as a more active process. Of the students surveyed, 70% believed that flipped learning gave them more choice in how they engage in learning, made them more likely to work collaboratively with peers, and enabled them to participate in activities that fostered problem solving and critical thinking. Driscoll used the
results of the survey to claim that flipped learning is a key mechanism to help democratize the classroom.

Of the 180,000 students from grades 6 - 12 who participated in the 2013 Speak Up online survey, 36% reported using videos in some way to support learning, up from 30% in 2012. Almost two thirds of these students viewed videos created by their instructors, and presumably published on the internet. This is up from 2012, where just over half of the students viewed videos created by their teachers. The other students who were using videos to augment or replace direct instruction were using video resources created by someone other than their teachers, culled for use by either the teacher or the student.

**Classroom Teachers**

The Flipped Learning Network conducted a survey (2012) of 450 classroom teachers using flipped instruction with their students. Of the teachers surveyed, 66% reported improved performance by students on standardized exams, 80% believed students had improved attitudes toward learning as a result of the flipped classroom, and close to 90% reported improved job satisfaction as a result of flipping their classrooms.

Driscoll (2012) reported that all of the 26 educators who had flipped their classes believed their students had become more active participants in the classroom; more than 90% claimed positive interactions with students had increased; students had better access to course materials and content; students could work at their own pace; students had an increased level of critical thinking and problem solving; and learning became more differentiated and personalized. Over 80% of the instructors claimed positive interactions between students increased, collaborative learning increased and students had an increased likelihood of selecting how they demonstrated their knowledge.
The Speak Up online survey was wider in scope than either Driscoll’s or The Flipped Learning Network's surveys. Over 466,000 K-12 teachers, parents, students and administrators participated in 2012 and 403,000 in 2013, and it helped push flipped learning to the forefront of a national dialogue about improving teaching methods. Unlike the heretofore mentioned surveys, the Speak Up survey wasn't directed at stakeholders who were directly involved in flipped classrooms. Its purpose was to report on perceptions of educators and students on broad topics, including but not limited to classroom flipping. More than 56,000 teachers responded to the 2012 survey and, of them, 18% reported interest in flipped learning in their classrooms. 20% percent wanted to learn how to make content-specific videos for use in their classrooms, and 15% were interested specifically in implementation of a flipped learning model. The numbers are a bit more muddled for 2013, as the data is disaggregated differently for the year. A total of 40,000 teachers responded to the survey, and Speak Up reported that 16% of math and science teachers are already implementing flipped instruction in their classrooms, a difference in reporting from the previous year, where there was no delineation based on subject. Of the teachers that didn't teach math or science, an additional 15% expressed interest in "trying flipped learning", and 16% list "learning how to flip my classroom" as one of their desired professional development topics.

It appears there is a ground swell building in the educational ranks around the idea of classroom flipping. As more teachers hear about the concept and see their peers enjoying successes by virtue of their students' test scores, improved job satisfaction, and a greater sense of learning occurring in the classroom, the movement will gain more momentum. As more teachers get involved, opportunities for research will grow, and new forms of classroom flipping will evolve. As of the writing of this dissertation, there are an increasing number of resources for
teachers wishing to flip their classes, cohorts of veteran flipping teachers are beginning to create
professional development around flipping, and barriers to entry are eroding. Teachers no longer
need to strike out on their own when flipping their classes as a culture of collaboration and
assistance is beginning to emerge.

Administrators

The only existing data concerning flipped classrooms vis-à-vis school-, district- and state-
level administrators was collected during the Speak Up surveys (2012, 2013), where over 6,000
and 4,500 administrators responded, respectively. Of those responding in 2012, 23% claimed
their teachers were using videos to replace direct instruction and 19% responded that they had
teachers who were fully flipping their classes. While almost 1,200 administrators claiming that
their teachers are flipping classes may seem impressive, exactly to what extent the teachers
flipped their instruction isn't clearly defined in the survey. It is not clearly understood, for
example, whether an administrator would perceive a teacher using a video to supplant a lecture
constitutes a fully flipped classroom.

The 2013 survey also verified that 24% of principals claimed their teachers were using
online videos to supplement or replace direct instruction, and a quarter acknowledged that
flipped learning was having an impact on changing classroom instruction, eclipsing other
popular technology-based topics such as game-based learning and mobile apps (Speak Up
Survey, 2013). However, while these statistics are useful to determine what administrators saw
in their schools and districts, there is a far more telling set of statistics that came from the survey.
Only five percent of the principals surveyed provided flipped learning as part of their
professional development for their teachers during the previous year. That said, 41% believe
teacher candidates for hire should know how to implement a flipped classroom, and fully two
thirds of principals believed that pre-service teachers need to know how to create and use instructional videos as part of their teacher certification program. It appears that school-level administrators understand the impact flipping might have in their schools, and want their teachers to be prepared for the changes to come.

Parents

Karen Cator, the former director of the Office of Education Technology for the United States Department of Education has fully acknowledged the growth of the flipped classroom and praises it for what she perceives as a great potential for parents to be more involved in their children's education. However, she is cautious in fully supporting the movement until further research is done in the field (Baker, 2012). Families experiencing dramatic changes in the way homework is traditionally assigned may initially feel uncomfortable with the idea of their children participating in a flipped classroom. It is therefore paramount that teachers and principals alike act deliberately to keep parents fully informed of not only what it means to have a child in a flipped classroom, but also why the teacher has chosen to flip the classroom.

In conclusion, the combination of flagging performances by students when compared to their peers and a peaked interest by educators wishing to take real and substantive steps toward utilizing technology to augment education creates an ideal environment for a concept such as classroom flipping to occur. It is incumbent upon the research community to provide all stakeholders with valid recommendations such that they may either continue to develop the field of flipping with confidence, or abandon it and begin their search anew for a solution to the problems our educational system faces.
Best Practices

Before discussing the implementation of flipping through the paradigm of contemporary learning theories, a bit more on best practices will inform the discussion more aptly. Brian Bennett’s quote, “Flipping isn’t a methodology, it’s an ideology” (Flipped Classroom Resources, 2012) is an interesting place to start for educators and administrators who are deciding not only how to flip, but why to flip in the first place. John Bergmann’s website, flipped-learning.com provides useful theoretical ideas pertaining to flipping. Bergmann posted the article “What IS Flipped Learning?” (December, 2012), where he sets out 5 clear theories teachers can aim for when designing lessons using the flipped model. He states that flipped learning should:

- Transfer the ownership of the learning to the students
- Personalize learning for all students
- Give teachers time to explore deeper learning opportunities and pedagogies with their students
- Make learning (not teaching) the center of the classroom
- Maximize face to face time in the classroom

Andrew Miller’s February 2012 blog at Edutopia (www.edutopia.org) outlines his Five Best Practices for the Flipped Classroom. Miller asserts that the successfully flipped classroom teacher:

- Creates a “need to know” for the students. Students are motivated to learn because they feel there is a purpose behind the learning, not just for the test or to pass the class.
- Uses engaging pedagogical models such as project-based learning (PBL), game-based learning (GBL), etc. and uses that model as a foundation for the classroom. Miller
asserts that the instructor needs to master the chosen model and then “use the flipped model to support the learning” within the context of the model,

- Knows and understands which technologies are needed for a successful flip. If the technology is lacking, is there a work-around that can fill the gap?
- Creates activities for students to reflect on the content they have watched. This will help them solidify what they do and don’t understand about the material.
- Creates a time and place for learning. Miller doesn’t advocate a true flip and believes it unfair to demand students watch videos outside of class time. Regardless of whether a teacher agrees with Miller, it is important for teachers to help students create a time and place for learning.

In summary, with these best practices in mind, an educator can use time-tested, well-researched learning theories in combination with a flipped ideology to create meaningful classes where measurable learning takes place.

**Criticisms of the Flipped Learning Model**

**Technology – The Lynch Pin to Flipping**

Unequal access to technology should be the primary concern for any parent, teacher, or administrator desiring to flip a class. Child Trends (2012) found that 85% of students had access to computers at home, but that only two thirds had access to the internet. Both of these percentages are considerably lower in Hispanic, African-American and low income households. It is important to note, however, that the amount of access is increasing continuously as devices other than computers that can access the internet (smart phones, tablets, etc.) are becoming ubiquitous across all populations.
The issue of access should be addressed early and students and parents should be surveyed to determine how many students have the requisite technologies available. Teachers and administrators should collaborate to solve the problems of any students lacking the requisite hardware, software, and internet access (Benjamin, et al., 2012). Furthermore, workarounds must be available for those students who don’t have the required resources, such as access to computer labs during study halls or before and after school is an example of one such workaround. Although this violates one of the guiding principles of Flipping - students should have access to content delivery when and where it is most convenient for them – exceptions must be made when the situation requires. Checking out laptops, tablets, or IPods as though they are textbooks is another example of accommodations that can be made if necessary. However, making such devices available for students may unduly overtax a school’s technology budget. This simply underscores the need for educators as a whole to implement flipping in an intentional, well-reasoned manner so as to address and mitigate any such issues in advance.

**Flipping is Simply an Extension of Best Practices in Teaching**

Because of the amorphous nature of the flipped learning model, many have criticized it as simply giving a name to what teachers should be or are already doing in their classrooms; gathering available materials to augment learning, project-based learning, etc. (Stumpenhorst, 2012). Calling flipped learning a model might be a bit premature, particularly since there has been a lack of research in the field. It is simply a method of content delivery, the process of which is chosen entirely by individual teachers, guided by the one simple principal that content acquisition is removed from the group space and designated to the individual space (Masallam, 2012). Likewise, some express concerns that students who flourish in the traditional, lecture-based model might suffer under the flipped model. It is important to note that the flipped model
does not preclude the use of lectures and direct instruction, and many teachers prefer to use a combination of traditional content delivery and flipping, depending on what the needs of the class and the material dictate (Bergmann & Sams, 2012).

A foreign language course, for example, may be a setting where a fully flipped classroom isn’t the optimal method for teaching. The process of learning in such courses occurs when students speak or write the language, make mistakes and are immediately corrected. That said, rote materials such as vocabulary, grammar and syntax might be components of such courses where an instructor might choose to flip some or all of the materials, which would lead to greater opportunity for peer interaction and actual speaking of the language during class time, with the expert guidance of the instructor.

**Poor Teaching and Lack of Spontaneity**

Concerns have been expressed that flipping will create an opportunity for teachers to simply find videos on the internet that closely address content in their class, force their students to view them and then give them worksheets to work on while the teachers ignore them. This brings up a pivotal idea that must be addressed: bad teachers teach badly regardless of the method being used. It is within the purview of the administrator to evaluate, provide criticism and guidance, and, if necessary, terminate bad teachers. It should be assumed by all stakeholders that flipped learning is a tool that will be used by professional educators that wish to improve the educational experience for all involved.

A different and more valid concern expressed about the topic is the canned nature of the lectures, particularly for teachers not creating their own materials. An extension of this concern is that policy makers might take this opportunity to increase class sizes, or replace teachers entirely with videos of content. While these points may seem entirely different from each other,
they can both be addressed in the same way. There is no replacement for a skilled, experienced and professional teacher who knows how to use resources to optimize learning in the classroom (Hamdan, et al., 2013). Teachers should constantly observe students, solicit their feedback and adjust materials to suit their needs and accommodations. Spontaneity and teachable moments should still abound in the flipped classroom. They simply won’t happen while the teacher is standing at the board or lecture podium.

**Confounding Bias between Flipping and Great Teaching**

The researcher must be leery to address any and all issues of bias when drawing any causal relationships between classroom flipping and improved performance by students. Proper balancing of subjects must be paramount in the researcher’s mind to ensure that factors such as the types of teachers who are flipping their classrooms aren’t confounding results. If influential, valid research is going to be accomplished in the field of flipped learning, judicious researchers must be sure to address demographic issues such as the characteristics of teachers that are doing the flipping, the types of students participating in the flipping, and the cultures of the institutions in which the flipping is occurring. Without such due diligence, important research may be discarded as irrelevant and opportunities may be lost at a time when solid research is most necessary to either validate the model and establish that this new paradigm of teaching can make a difference, or repudiate the claims made by flipping’s proponents and seek out another method for improving education.

**Conclusion**

Regardless of whether an educator or administrator subscribes in part or in whole to the flipped classroom, one thing appears to be certain; technology’s role in the educational realm is moving from the peripheral to center stage. The ubiquity of educational websites that can be
used by students to acquire content puts teachers in an interesting quandary where they must choose to either compete with technology for the attention of their students or embrace technology as an effective tool and learn to redefine their role in the classroom.

That said, the lack of rigorous research in the field is potentially prohibitive for educators interested in implementing a flipped model. This study determines whether there are statistical differences on the performance of 6th grade students on an end-of-year standardized mathematics exam called the Criterion Reference Test (CRT) between students who received a “flipped” learning experience for their 6th grade math class and those that didn’t. This study seeks to enrich the literature and provide educators with substantive conclusions as to the efficacy of classroom flipping. Its intent is also to inspire further research in the field to create a more complete body of work from which policy makers, administrators, etc. can draw on when creating new curriculum and training high quality teachers.

Part II. Propensity Score Matching

Introduction

This part of the literature review will be divided into three sections. The first section will provide the reader with a working definition of propensity scores and propensity score analysis. The second section will provide a brief description of PSM processes. The last section will provide the reader with a number of recent studies, which can be used as a literary review if one is so inclined as to validate using PSA in one’s educational research. In the interest of completeness, a glossary of pertinent terms will be provided in the appendices of this dissertation.

Propensity Scores and Propensity Score Matching (PSM): A Definition

A propensity score is the conditional probability that a person will be in one group (e.g.
experimental or control), given a specific set of observable covariates, sometimes referred to as the covariate vector (Rosenbaum & Rubin, 1983). The propensity score, \( \hat{e}(x) \), is meant to balance participants in each group. Subjects in each group with similar covariate values will have similar propensity scores. The propensity score can be derived in a number of ways, however, for convenience and ease of use, the logistic regression equation will be employed for this study, and this choice will be further explored in this chapter.

Propensity Score Matching is, simply put, the use of propensity scores to match individual or groups of subjects within a quasi-experimental group with those in the control group by balancing them based on their covariates (Rubin & Rosenbaum, 1983), and then employing traditional statistical methods to determine treatment effects and causal relationships. A subsequent section in this paper will further describe the mathematical and statistical methods by which this is done, but the reader must understand the underlying reason for using PSM. Because observational studies aren’t controlled by the researcher through randomization of placement, they inherently contain selection bias. Researchers are currently using propensity scores to reduce bias, increase precision and draw causal conclusions in their research (D’Agostino, 1998). PSM accomplishes this by creating quasi-experimental conditions in which researchers are able to compare treatment effects of subjects who received treatment – in a Randomized Controlled Trial (RCT), these subjects would be members of the experimental group – versus those who received no treatment, by comparing subjects based on their propensity scores.

Rosenbaum and Rubin’s paper, *The Central Role of the Propensity Score in Observational Studies for Causal Effects* (1983), created an entirely new tool in statistical analyses to be used by researchers in all fields of interest. Replete with requisite formulae to
substantiate their claims, as well as the rigorous proofs of said claims, they discovered a mathematical way to utilize data from an observational study to closely resemble a Randomized Controlled Trial (RCT), the “gold standard” of inferential statistics (D’Agostino, 1998). Rosenbaum and Rubin showed the process so closely mimicked an RCT that the conclusions made as a result of using PSM correctly in research were tantamount to inferring causation, and reliably determining unbiased estimates of treatment effect both at the sample- and population-level.

Initially used only by researchers in the medical and econometric fields, PSM has crossed over into other fields such as the social sciences (Caliendo & Kopeinig, 2008). However, difficulties can arise for researchers wishing to utilize PSM insofar as the process is prohibitively complicated. This chapter will attempt to clarify much of the process and reasons behind the process, employing examples, some elementary theory and algorithms for guidance, such that researchers who might want to delve into PSM will not feel intimidated and, thus, gain access to a useful research tool.

A Brief Overview of the Process of Implementing Propensity Score Analysis

This section will provide the reader with a quick primer of the procedures a researcher would utilize when using PSM in a study. It also lays out the structure which is explained in greater detail in both Chapter Three and in the appendices, which delves more deeply into each of the topics covered here.

According to the literature (Love, 2008, Shadish, et al., 2008, Steiner, et al., 2011) the following nine procedures are implemented in Propensity Score analysis:

1. Covariates are identified and data is collected.
2. Covariate data is divided into experimental and control groups.
3. Covariate matrices are input in SPSS to create a logit model. The IV will be either a 1 if the student participated in the flipped class or a 0. else.

4. The logit model generates sample propensity scores, \( \hat{e}(x) \), for each matrix.

5. Cohen’s \( d \) is used to identify any covariate imbalance prior to matching.

6. In the event of serious imbalance, adjustments to the model are made.

7. A matching algorithm is chosen; in this study, the algorithm is Nearest Neighbor matching.

8. After matching, analytical and graphical tools are used to determine if the criteria for properly balanced covariates are met.

9. If no serious imbalances exist, a matched-pairs \( t \)-test is calculated to determine if significant differences exist between the groups.

The above list provides a process template for potential users of PSM to follow in their research. While common research processes are intuitive for most readers, it is important to fully understand the ideas and concepts presented in this template by virtue of the relatively unique methods propensity score analysis requires. Again, interested readers are invited to read the appendices to gain a more rigorous understanding of the PSM process.

**Examples of PSM from the Literature**

Because the learning curve for PSM is prohibitively steep, many researchers may shy away from using it in their work. However, as properly designed RCT’s in an educational context are either too expensive or illegal to conduct, and as enormous data bases filled with information about student demographics and performance become ubiquitous because of mandated accountability, educational researchers are beginning to find PSM to be a viable tool for their researcher.
Hong and Yu (2008) utilized PSM to assess the effects of kindergarten retention on children’s social-emotion development. In their paper they sight the deficiencies of previous work, on account of selection bias and measurement, and use PSM to overcome these challenges. Incidentally, their research concluded that there is no significant difference between the social-emotional development of students who are retained in kindergarten and those who aren’t.

Wyse, Keesler, and Schneider (2008) used propensity score models to evaluate the effects of small school size on mathematics achievement. Hong and Raudenbush (2006, 2008) investigated both instructional design and kindergarten reading levels using PSM. Staff, Patrick, Loken, and Maggs (2008) discovered (rather obviously) that heavy drinking adolescents experience reduced educational attainment. Morgan, Frisco, Farkas, and Hibel (2008) used propensity score matching to determine the effectiveness of special education services provided to students in the United States. More recently, Melguizo, Kienzl, and Alfonso (2011) studied educational attainment at the junior college level using propensity score matching. Fan and Nowell (2011) produced a brief to introduce educators to PSM, including step-by-step procedures for researchers desiring to use PSM in their work.

**Conclusion**

It is quickly becoming obvious that PSM is a valid alternative to more traditional statistical methods, particularly in educational research, where large amounts of data are becoming readily available (Rubin, 1997). The onus on institutions and researchers who understand how to properly implement PSM is to educate prospective researchers who are interested in using PSM as a tool. It is the author’s belief that PSM is at a pivotal stage in educational research and, if applied thoughtfully and carefully, will emerge as a useful tool for discovering what works in education.
This study will provide future researchers with practical tools when considering using PSM in their research. PSM is a difficult and, at times, counterintuitive approach to statistical analysis, and has a number of pitfalls that an ill-informed researcher can easily fall into. By expanding the body of literature on PSM in general, and its uses in educational research in particular, the author hopes to empower hitherto hesitant researchers to feel confident employing PSM methods.
Chapter Three

Methods

Introduction

As was the case with Chapter Two, this chapter consists of two parts. The first part will introduce the methodological underpinnings for the use of PSM in this particular study, and the second will take a more traditional approach to describe participants, variables, measurements, and statistical methods after propensity score matching is applied to the experimental and control groups for the current study. Because of the relative obscurity of PSM in educational research, it is incumbent upon the author to provide sufficient background information to confidently determine whether the research methods used in this study and the conclusions obtained are sound.

Part I. Using Propensity Score Matching

Where Propensity Scores Come From: The Logistics Regression Equation

In the interest of completeness, it is important to delve into the more theoretical aspects of Propensity Score Matching. This examination will begin with how propensity scores are created and expand into how the scores are employed to match subjects, thus creating a quasi-experimental setting, and finish by exploring how the matched subjects can be analyzed to determine treatment effects. This section, combined with the information provided in Chapter Two, will provide the reader with a step-by-step process for using PSM as well as an overall understanding of how and why propensity scores and their uses are important in determining treatment effects from observational studies.

Logistics Regression is an extension of multiple regression wherein the dependent variable (DV) is not a continuous, quantitative variable, but rather a categorical variable that may
have as few as two outcomes. Logistics regression is most often used in dichotomous, or binary outcome studies because complexity increases as the DV adds more layers (Mertler & Vannatta, 2010). Moreover, in logistics regression (LR), the independent variables (IV’s) can be either qualitative or categorical data, unlike Discriminant Analysis where all IV’s must be quantitative, which makes (LR) more flexible for the purposes of research (George & Mallery, 2000). It is important to acknowledge that multiple regression involves regressing more than one IV onto a single DV using several selection processes, and LR is no exception. It is the dichotomous nature of the Logistics Regression Model that makes it useful to the researcher in that the DV, when input as a value between 0 and 1, outputs a probability of membership into one of two groups. These two groups, for the purposes of PSM, represent experimental and control groups and are treated as such during subsequent statistical examination.

The Logistics Model isn’t a unique tool for creating propensity scores, or probabilities of group membership given a set of covariates. Other techniques using probit models, and semi- and non-parametric techniques are deeply explored in the literature. However, in the interest of saving time, it is reasonable to say that the logistics (or logit) model is most often used in the initial stages of PSM because of its familiarity to researchers and ease of use, particularly when using statistical software (Steiner, 2013).

**Creating the Propensity Scores Using the Logit Function: Selecting the Covariates**

The first and primary assumption that must be met before valid PSM can occur is the absence of nonignorable covariates, otherwise known as the “no unmeasured confounders” assumption that states that all variables (covariates) that can affect either treatment effect or membership in one of the groups measured by the DV (Austin, 2011) and is the most crucial step in creating a valid model for use in PSM (Brookhart, et al., 2006). While it is impossible to
collect a complete set of covariates, and thus ensure this assumption is met, researchers are compelled to explore the literature prior to any research endeavor involving PSM to determine as many covariates as possible that might influence treatment effect or group membership.

The most profound effect that this assumption has in PSM is it allows the researcher to completely disregard all assumptions pertaining to Logistics Regression. The logit model is used exclusively to create the propensity scores, and not to determine treatment effects of any kind. The fact that no conclusions are being drawn from the model allows the researcher to ignore Logistics Regression assumptions. Dr Thomas Love (2008) states that the covariates introduced into the model are meant to “sop up” the signal that represent differences between treatment and control groups, and the researcher must include as many relevant covariates into the model as is possible, else run the risk of creating a model that has bias without ever even knowing it.

The Logit Model

After the data is collected, a statistical software package such as SPSS would then create a logistics (logit) regression equation:

$$\ln \left( \frac{\hat{Y}}{1 - \hat{Y}} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

Algebraically, the logit function can be solved to produce the following, equivalent equations:

$$\frac{\hat{Y}}{1 - \hat{Y}} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i}$$

$$\hat{Y} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i}}$$

$$\hat{Y} = \frac{1}{1 + e^{-\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i}}$$

There is a litany of other practical research applications using logistics regression, however the most important concept to understand when using this form of regression is that the DV will provide the researcher with a probability of a subject being in a...
particular group, given a specific set of characteristics. This set of characteristics creates what is known mathematically as a vector:

$$X = x_1 + x_2 + x_3 + \cdots + x_n$$

or

$$X = [x_1, x_2, x_3, \ldots x_n]$$

The probability model produced by logistic regression isn’t the only useful result. Since the model produces (beta) coefficients, referred to as beta weights, for each characteristic, researchers can examine each beta weight and determine which characteristic is most predictive of group membership. Similarly, the beta weights can be used to determine odds ratios for each characteristic, which can be useful in establishing causation for group membership. However, for the initial purposes of creating propensity score, the beta weights are deemed unimportant to PSM, particularly because traditional assumptions about Logistics Regression are disregarded when creating the model for PSM (Love, 2008).

**Counterfactuals and the Quasi-Experimental Setting**

The use of a propensity score as a statistical tool in general and the propensity score produced by the logit model in particular has entirely to do with its ability to balance subjects based on their covariate matrices. In an ideal situation, all unignorable covariates would be measured and the propensity score would represent a parameter for each possible covariate matrix, $X$. This implies subjects could be perfectly matched based on their respective, identical propensity scores (Austin, 2011). Under this perfect setting, treatment assignment is entirely independent, thus two perfectly matched individuals or groups of individuals could be compared and treatment effect could be determined exactly. I.E., if a member of the control group is perfectly matched with a member of the experimental group, researchers could precisely
determine what would have happened had the experimental subject *not* received treatment and the control subject *had* received treatment, and the counterfactual question could be answered with exact precision. Another way of saying this is if the propensity score can be calculated exactly, and the ignorability assumption isn't violated, then all hidden bias can be removed from a study. As Steiner (2013) states, "hidden bias results when the strong ignorability assumption is not met."

Because it is impossible to determine all unignorable covariates, a researcher has no idea if there is some unmeasured covariate acting as a source of bias in the model (Rubin, 1997). Therefore, the propensity score created is an estimated statistic, \( \hat{e}(x) \), which is used to match subjects. Researchers must perform due diligence to ensure that as many unignorable covariates are measured as possible prior to creating the model that determines propensity scores, and they also understand that it is highly improbable that their models will contain perfect information. It is for this reason that PSM creates what is referred to as a *quasi-experimental* setting. Since the ultimate goal of any study design, be it an RCT or a quasi-experimental study, is to reduce bias, a valid question for researchers to ask is whether PSM indeed achieves this goal. From its inception PSM, when the nonignorability assumption is addressed diligently, has been proven to eliminate a minimum of 90% of selection bias present in an observational study (Rubin and Rosenbaum, 1983). Initial study design informs whether the nonignorability assumption is being addressed exhaustively by the researcher.

**Matching Subjects Based on Propensity Scores**

The crucial step in Propensity Score Matching, which requires the greatest amount of understanding of the overall process of PSM, is subject matching. Matching is the procedure whereby the researcher assigns one or more of the control subjects to a single experimental
subject. Hitherto, there is no discernable difference between PSM and a simple Logistics Regression analysis, other than the fact that the researcher is allowed to ignore assumptions that need to be met before performing a valid Logistics Regression analysis. However, the matching of subjects from the experimental and control group based on the closeness of their propensity scores is the cornerstone of PSM and the technique most often used by researchers wishing to use propensity scores in their research (Steiner, 2013). It should be noted that propensity scores themselves, much like many of the tools from statistics can be utilized in a host of other ways, however this discussion is beyond the scope of the study.

**Three Choices a Researcher Must Make Prior to Matching**

In the literature it is recommended that prior to matching subjects based on propensity scores, it behooves the researcher to make three separate choices concerning the technical nature of the process. Steiner (2013) asserts the researcher must choose:

- A distance metric for determining “acceptable” distances between matched subjects from the control and experimental groups
- The number of subjects from the control group which should be matched to each experimental subject, and whether it should be done with or without replacement.
- The choice of the algorithm used to create the matching.

Because of the relative newness and esoteric nature of PSM, most statistical packages do not include matching applications as part of their default programming. Unlike most statistical applications where researchers need only use drop down menus and radio buttons to achieve their goals, PSM matching requires plug-ins such as R or SAS, which aren’t as user friendly or intuitive as other applications. A hapless user may inadvertently use default settings and end up
with results that appear erroneous or invalid, and as such needs to take the time to be certain the program is indeed running with the above choices being confirmed by the user.

First, the most common distance metric used by PSM researchers is the Mahalanobis Distance (Krzanowski, 2000), created by the mathematical equation, 
\[ d_{ij} = (x_i - x_j)'S_X^{-1}(x_i - x_j), \]
which uses the inverse variance-covariance matrix \( S_X \) and its correlation structure to define distances between subject matrices (Mahalanobis, 1936). This distance metric is preferable to other distance metrics, such as Euclidean distance, because of the use of the matrices. However, Rosenbaum (2009), expressed concern over the behavior of the metric in the case of large outliers in interval and ratio data, and recommends using rank scores in the event of their presence.

Second, the issue of how many control subjects should be matched to each treatment subject can have subtle effects on the outcomes of a study utilizing PSM. If matches are 1:1, where each treatment unit is matched to only one control unit, while bias is certainly reduced, the study will lose efficiency since so many control data values are discarded and any potential treatment effect that these data can provide the study are discarded as well (Caliendo and Kopeinig, 2008). If each experimental unit is matched to many control units (1:N), efficiency is gained, however certain biases may unintentionally be introduced by virtue of the fact that it is improbable that each of the matched control units’ propensity scores will be arbitrarily close to the score of the experimental unit. A reasonable compromise between efficiency and potential bias is struck through the use of calipers which creates an interval of values, centered around the experimental units’ propensity score (Rubin & Thomas, 2000). If a control units’ propensity score falls within the interval, it is considered a match, else it is either matched with a different experimental unit (again, if it falls within the caliper interval of the new experimental unit) or
discarded as unmatched. The caliper distance is established by the researcher and the computer program establishes matched units automatically. If the researcher defines the caliper distance as .3, then the algorithm would only look for potential matches for experimental units with control units that fall within .3 standard deviations of the linear propensity score.

A researcher must also decide whether to allow replacement of previously matched control units, which would potentially enable them to be re-matched with other control units. While loss of efficiency is a possibility in the replacement case, research has shown that matching with or without replacement typically has a negligible effect on both bias and efficiency vis-à-vis treatment effect (Ho, et al., 2007).

Another step a researcher must take after generating propensity scores, but prior to matching is to analyze the distributions of the propensity scores of both the control and treatment groups. This analysis is best done side-by-side, and Figure 4, provided by Dr. Thomas Love (2008) demonstrates how to use side-by-side histograms to this end:

![Figure 4. Side-by-side histograms demonstrating common support (Love 2008).](image)

Notice that in graph A, there is no overlap, or “common support” (Caliendo & Kopeinig 2008) between the propensity scores of the control and treatment groups. In this case, the
researcher will need to reconsider the use of PSM, as any matching method is going to mismatch experimental and control units because of the large distances between propensity scores, and thus inject an inherent bias into any conclusions. In graph C there is a large amount of overlap, however the researcher must contend with how to approach the non-overlapping tails of the distributions. The distributions of the control and experimental propensity scores will inform the choice of matching method, in terms or whether to discard experimental units, control units or both. However to simply allow the algorithm to run without understanding potential creation of bias in the results is statistically feckless and should be avoided whenever possible. Ideally, the distributions should completely overlap, as in graph B, as this guarantees that combinations of characteristics present in the experimental group are also present in the control group (Bryson, et al., 2002). It should be obvious to the reader that each treatment until will have a close match to a control unit and vice versa. This rarely happens and rigorous explanation of choices surrounding deletion and/or mismatching of data should be included in any study.

**Three Matching Algorithms**

The choice of matching algorithm must be made after the researcher has determined a distance metric, a matching caliper distance (if calipers are going to be used) and whether to match with or without replacement, the final decision before matching is which algorithm to implement. There are a number of different algorithms that have been invented to match subjects based on propensity score since its inception in 1983. The following three matching algorithms have been widely used in PSM across many disciplines, as well as in educational research.

1. Stratification/Subclassification
2. Optimal Matching
3. Nearest Neighbor Matching

The first and simplest posited by Rosenbaum and Rubin (1983) was stratification, or subclassification. Stratification involves the researcher breaking both groups into equal strata based on propensity score. Rosenbaum and Rubin (1984) recommend a minimum of 5 strata (quintiles) and have shown that this method removes 90% of bias in a study. After each group is stratified, standard statistical techniques such as t-tests, ANOVA, and regression modeling can be used to determine treatment effects. The caveat with stratification is propensity scores for both groups need to have similar distributions, i.e., if the experimental group has most of its propensity scores close to 1 and the control group has most of its scores close to 0 the researcher will arrive at biased results. Figure 5, below, displays possible distributions of propensity scores for experimental (treatment) and control groups:

![Figure 5](image)

*Figure 5.* Side-by-side histograms C. from previous figure (Love, 2008).

In this case, the top 20% of experimental units (those with a propensity score closest to 1) would be matched to the top 20% of control units and so on until all of the units are matched by quintile. Clearly, this method creates a bias, as each experimental group created will
systematically vary from its control counterpart.

A second matching method is Optimal matching. Optimal matching was invented by way of a computer algorithm by Dimitri Bertsekas (Bertsekas, 1991) and seeks to minimize the overall pairwise Mahalanobis distances between experimental units and their control counterparts by using a field of research referred to as flow theory. One benefit of Optimal matching is that it creates a unique matching environment which is repeatable every time the algorithm is run, which Nearest Neighbor matching cannot always guarantee. Unfortunately, as of the writing of this dissertation, the only way to run Optimal matching is using the R statistical software package (http://www.r-project.org/). While the R programming language has many plug-ins available for user-friendly statistical platforms such as SPSS, there are no such plug-ins available for Optimal matching, and the amount of coding a researcher must learn to effectively use the algorithm is prohibitive.

The last matching method is referred to as Nearest Neighbor (NN). NN is the matching method used in this study because of its ease of use with the SPSS statistical platform. In nearest neighbor matching, the researcher/algorithm selects a treatment unit, which can either be selected at random or by simply starting with the treated unit with the highest or lowest propensity score. The control group unit with the closest propensity score, sometimes referred to as the "nearest neighbor" is matched to the treatment unit. The researcher can choose whether to match exactly one experimental unit with one control unit (1:1) or with N control units (1:N). When matching 1:1 the researcher trades off an increase in covariate variance for a decrease in overall bias (Caliendo & Kopeinig, 2008). The process can also be done with or without replacement, however, when matching with replacement overall bias decreases while variance among the covariates increases. Inversely, matching without replacement results in an increase in overall
bias and a decrease in covariate variances.

The researcher must also choose whether to use a caliper, or propensity score interval in which to allow control subjects to be matched with treated subjects. The bias/efficiency trade-off is identical to matching with replacement, wherein using calipers decreases overall bias but increases variance (Caliendo & Kopeinig, 2008). Caliper matching also can have the unfortunate consequence of experimental units being unmatched to control units, which has a direct effect on the power of the study. This is referred to as an “incomplete matching” (Rosenbaum & Rubin, 1985). Another limitation of Nearest Neighbor matching is the matches can depend entirely on the order in which experimental subjects are chosen, particularly when matching without replacement (Austin, 2011). A control subject can inadvertently be matched to an experimental subject that doesn’t have the closest propensity score match by virtue of the order of matching, which can completely change the overall control group after matching. Processes for trouble shooting these limitations are explored in greater detail later in this chapter.

Ensuring Covariate Balance

As the entire purpose of propensity scores is to balance the covariates (Rosenbaum and Rubin, 1983) after matching, the final step the researcher utilizing PSM must do is to mathematically check for covariate balance. Meeting the criteria that ensures covariate balance is the second of two assumptions in PSM, and any imbalance of even a single covariate can potentially inject bias into a study. While there are a number of algorithms and tests to do so (Steiner, 2013), parsimony dictates that the simplest and most straight forward should be employed. Love (2008), among others recommends comparison of univariate distributions of each covariate separately utilizing Cohen’s $d$, given by $\frac{(\bar{x}_e - \bar{x}_c)}{\sqrt{s_{pooled}^2}}$, where $\bar{x}_e$ and $\bar{x}_c$ are the means of individual experimental and control covariates, and $s_{pooled}^2$ is the pooled variance for
both groups. This value represents the standardized differences between the means of each of the covariates, i.e., the imbalance for each covariate (Rubin, 2001). It is best to measure the amount of covariate imbalance both before and after matching on propensity scores. Large differences prior to matching ($|d| > 1$) imply heterogeneity of distributions of subjects for a single covariate and the researcher might consider either deleting the covariate from the study or combining the covariate with another using multivariate techniques. After matching, imbalance is again determined by how large Cohen’s $d$ is. Stuart and Rubin (2007) recommend $|d| < .25$, however more conservative benchmarks of $|d| < .1$ have been used (Love, 2008, Shadish et al., 2008). If, after matching, Cohen’s $d$ exceeds these benchmarks for any individual covariates it is recommended the researcher try multivariate techniques or combining covariates to mitigate the imbalance. However, if covariate balance is ensured, the PSM procedure is complete and the researcher may use the two data sets as they would any experimental and control group.

Researchers using SPSS for analysis of covariate balance have a number of other tests available to them as well. A test for overall imbalance developed by Hansen and Bowers (2008) is akin to the well-known Hotelling’s $T^2$, and is the first balance test provided in the output when matching in SPSS. It simultaneously determines if any covariates or linear combination of covariates are significantly imbalanced after matching (Thoemmes, 2012). Significant p-values below .05 imply serious imbalances in the covariates and the researcher must use analytical techniques to address such imbalances. This test is only available for analysis when matching is 1:1.

Yet another omnibus test for overall balance of covariates that is provided by the SPSS output, $L_1$, developed by Iacus, King and Porro (2009) uses a technique referred to as automatic coarsening and binning to determine imbalances between variables. If $L_1 = 0$, then the subjects
are perfectly matched on all covariates and if $L_1 = 1$ the variables are perfectly separated (Thoemmes, 2012). Further reading on the topic, including specific algorithms and mathematics, is available for the interested researcher (Iacus, et al., 2011). While there is no cutoff for whether covariates are balanced, researchers should be looking for a reduction of $L_1$ from before matching to after matching. Researchers will use this test in combination with the other information presented in the output to determine if the covariates are adequately balanced.

Finally, SPSS produces graphical outputs for researchers to analyze covariate balance graphically. Jitter plots, line plots for individual differences, histograms for standardized mean differences, and dot plots for standardized mean differences are all graphical output options available to researchers in order to scrutinize data to ensure the criteria for covariate balance is met. These graphical options will be explored in greater detail in Chapter Four.

Part II. Current Study

Participants

For the purposes of the study, data on human subjects were used and IRB approval was granted to acquire the data. All data and any and all information about participants in the study were completely de-identified and, based on the de-identification of the data, IRB approval was expedited. Washoe County School District (WCSD) Department of Accountability supplied all data for the study from their extensive data bases. The subjects are all sixth graders from schools in WCSD for 2011-2012 and 2012-2013 academic years, which amounted to a large amount of data ($N$=9425). However, because the experimental subjects, elaborated on in the next section, are taken exclusively from a Title I school, all subjects not attending a Title I school were eliminated. In the WCSD there are middle schools that have sixth grade classes as well as the traditional seventh and eighth grade classes. Such sixth grade classes were eliminated from the
study for the purpose of reducing bias due to different social and scholastic experiences. Similarly, any Title I students with missing covariate data pertinent to the study were eliminated as well. A Title I school is defined as a school that receives federal funding via the Elementary and Secondary Education Act (ESEA) by virtue of having a large percentage of low income students. Schools whose low income population exceeds 40% of the student population are eligible to receive Title I funding. This cleaning of the data created a new data set ($N = 2370$) to be used for the purposes of the study.

**Experimental Group**

The experimental subject data were defined as any data that was mined from one particular elementary school, henceforth referred to as flipped school, from within the WCSD. Experimental subjects were defined as students who took their mathematics course from one particular teacher, henceforth referred to as flipped teacher, during the 2011-2012 and 2012-2013 academic years, and had no missing covariate values during the year of the study. Flipped Teacher self-identified as a teacher who employed flipped learning methods in her math classes.

Flipped teacher utilized math videos created by Kahn Academy (www.khanacademy.org) during the 2011-2012 and 2012-2013 school years for the purpose of content acquisition for the students. Students were then able to work in groups, individually, or with direct supervision from flipped teacher during the in-class portion of their mathematics course. Students could either acquire content at home using the internet and their families' devices or they could use the computers provided by flipped school during after school hours. Flipped school is a one-to-one school, meaning the ratio of students to computers in the school is 1 to 1; implying access to devices with internet access is always available to students.
Control Group

Control subject data were defined as data acquired from all remaining students from the remaining schools in the study. Principals from those schools were contacted to ensure that none of their students received a flipped learning experience, and all of the principals contacted from control school data maintained that none of their sixth grade teachers used the flipped method in their classes during the years of the study. Because of the high need for access to technology as well as buy-in from administrators and parents alike, principals would necessarily need to be aware if any teachers in their building were implementing flipped instruction.

Identical data was collected from both the experimental and control subjects, and consisted of both qualitative and quantitative data about the students, which was used to form covariate matrices for each student. The covariates used to create the covariate matrix are as follows: gender, ethnic background, socio-economic status (SES) as measured by participation in the free and reduced lunch program (FRL), participation in special education programs offered by the district as measured by the student having an Individualized Educational Program (IEP), English language proficient (LEP), mathematics grade from the year of the study, Criterion Referenced Test (CRT) score for mathematics for the two years prior to the study, and CRT English score for the year of the study. It is important for the author to acknowledge that, while there may be other unignorable data points that weren’t accessible, after much consultation and collaboration with educators and administrators alike, the data collected were deemed to be as complete in terms of ignorability as was feasible. This highlights one of the primary challenges faced by researchers using PSM; all unignorable covariate data must be collected, however there is no way for a researcher to know if any unignorable covariates have been overlooked (Rosenbaum & Rubin, 1983).
Students from the control group with missing data values were eliminated from the study. Missing data from the experimental group were a more serious issue, as the size of the experimental group needed to remain relatively high to insure power of the study. All such students were investigated in greater detail to determine if missing data could be recovered, or data values could be replaced with analogous data. For example, a student transferring from California into the WCSD the year of the study wouldn't have a math CRT score for the year prior, as California uses the Standardized Testing and Reporting (STAR) exam. That student's STAR exam score for mathematics would need to be translated into a CRT score. This could easily be done since STAR and CRT are very similar and are both criterion referenced tests used to measure mathematical content knowledge. However, after the data were collected, it was decided that in all cases there were no analogous data available to replace the scores. An overall experimental groups size of 55 was determined to be large enough to both preserve a high enough power and create a large enough pool of experimental units with which to match control units.

Data were delivered via excel spreadsheet by Holly Mercer, a data specialist for the WCSD. Issues of attrition and non-response weren't an issue as this is an observational study and no surveys were used in the collection of the quantitative data. A survey meant to analyze the qualitative aspects of flipped learning, and answer research question number three, was created for Flipped Teacher. Nine other teachers in the district who have experimented with flipped learning and their responses as well as qualitative analysis are published as part of this dissertation.

**Measures**

Propensity Score Matching (PSM), as described previously was used to match
experimental subjects with control subjects to create the final control and experimental groups used to determine treatment effect. Nearest Neighbor matching was the matching algorithm of choice and was chosen because it simplified the issue of available statistical software to execute the matching. SPSS has a Propensity Score Matching plug-in available through a statistical program called R ([http://www.r-project.org/](http://www.r-project.org/)). For researchers interested in using Propensity Score Matching through the SPSS statistical platform, Thommes (2012) provides step-by-step instructions for doing so. Using Nearest Neighbor matching presents researchers with other additional challenges such as determining how many control subjects will be matched to each experimental subject, whether matching will be executed with or without replacement of control subjects, and whether or not caliper distances will be used to determine if control units are eligible for matching. However, the ease of use of SPSS in combination with PSM, combined with the fact that Nearest Neighbor Matching is the only available method of matching in SPSS, make Nearest Neighbor Matching the most parsimonious choice for matching.

While the literature claims there is very little difference between measured treatment effects between matching algorithms (Gu & Rosenbaum, 1993), using a reliable matching instrument is an important aspect to the study. As a tool for reducing bias, PSM has been shown to have a high degree of validity, so long as the researcher doesn't violate the assumptions of both nonignorable covariates and covariate balance (Rubin & Rosenbaum 1983, 1985), and as long as there is acceptable balancing of control and experimental subject covariates after matching. The matching algorithm choice is the researcher’s prerogative.

**Covariates for Propensity Score Matching**

This study utilizes existing data provided by the WCSD to create an observational study. Data from the observational study created both control and experimental subjects based on
whether the subjects received a flipped learning experience for their sixth grade mathematics (experimental subjects) course or the traditional, lecture-based experience (control). It was assumed by the researcher that the data set used in the study contained all unignorable covariates which might have affected treatment outcome. Propensity scores for each of the subjects within both groups were generated using a logit function. For the logit function, the independent variables (IV's) were defined by the covariates mentioned in the "participant" section of this chapter and the Dependent Variable (DV) was defined as a 1 if the subject participated in a flipped learning experience and a 0, else. All qualitative covariates were assigned dummy variables to represent differences within each covariate and quantitative data were defined using the measures given by the WCSD/State of Nevada. All quantitative data were interval- and ratio- level data. Figure 6, on the following page, provides a list of independent and dependent variables and how they were coded for the purposes of the study.

**The Criterion Reference Test (CRT)**

Per the regulations mandated by the No Child Left Behind Act (NCLB, 2001), all states must assess their students' abilities in mathematics, as well as other subjects, for Annual Yearly Progress (AYP). AYP is meant to demonstrate growth both at individual grade levels as well as for students as they matriculate between grades. In the state of Nevada, the CRT is used to satisfy the AYP mandate ([http://www.doe.nv.gov/Assessments_CRT/](http://www.doe.nv.gov/Assessments_CRT/)). Given to students from grades three through eight, the CRT is meant to be a summative assessment - an assessment meant to provide a single data point, or "snap shot" for every individual student in the public school system in the state of Nevada - which is standardized and linked to state standards. The exam is written by Nevada grade-level content experts, typically educators, reviewed and revised by West Ed, a national assessment consulting group, and then resubmitted to the State of
Nevada's Department of Education for final approval.

<table>
<thead>
<tr>
<th>Covariates in This Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariates</strong></td>
</tr>
<tr>
<td><strong>DV</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.* List of covariates. Independent and dependent variables used in the study.
While exam details vary from grade to grade, the sixth grade exam consists of sixty multiple choice and Constructed-Response questions. Each Constructed-Response question is worth a maximum of three points, and each multiple choice question is worth either one point for a correct answer or zero for an incorrect answer, for a maximum of 500 points cumulative for the exam. Partial credit can be given for Constructed-Response questions on a score from 0 to 3, based on a rubric provided by Measured Progress, the vendor subsidiary of West Ed. Exams are also scored by Measured Progress and results as well as statistical analysis is provided shortly after the exams are received from the state of Nevada Department of Education.

The purpose of the exams is to provide state-, district-, and site-level information as to the performance of individual students and teachers, grade levels within schools and schools as a whole. The "cut scores" for sixth grade students are as follows: a student scoring between 100-249 is considered "emerging/developing", a student scoring between 250-299 is considered "approaching standard", 300-405 "meets standard", and 406-500 "exceeds standard" (http://www.doe.nv.gov/Assessment_Resources/). Cut scores used to determine achievement level differ between grades and can be adjusted by the state depending on the performance of students in any given year in attempts to reduce any unidentified bias present when the exam was created.

**Variables for t-tests**

After utilizing PSM to match both control and experimental groups, both a matched-pair t-test and independent samples t-test was performed to determine if there were significant differences in the performances on the Nevada State Sixth Grade CRT between students who had received math instruction via the flipped experience and those who hadn't. A greater explanation
of why both types of \( t \)-tests were performed will be provided in Chapter Four. Figure 7, below, provides the variables used for the \( t \)-tests:

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Level A - Flipped Learning</td>
<td>CRT Scores for Flipped Learners</td>
</tr>
<tr>
<td>Treatment Level B - Traditional Learning</td>
<td>CRT Scores for Traditional Learners</td>
</tr>
</tbody>
</table>

*Figure 7*. Independent and dependent variables used for \( t \)-tests.

**Teacher Surveys**

A teacher survey was created to determine qualitative understanding around the practice of classroom flipping by educators who are actually flipping their classes. The surveys were distributed to WCSD secondary teachers, as well as the flipped teacher whose data was used in this study. The surveys were created in Word and distributed via email. Completed surveys were returned via email and participants were paid twenty dollars in return for their participation. Teachers were compensated in hopes of obtaining complete, well thought out responses, rather than short answers that would have no value. There were a total of seventeen questions on the survey and the entire survey is available in the appendix. The questions were grouped to tease out overarching themes about how local teachers are specifically flipping their classes, what types of technological obstacles needed to be overcome in order for them to flip their classes, and what results were they seeing from their students as a result of flipping their classes.
Qualitative analysis of this study focused on two separate sets of questions, questions four and five and questions eleven through fifteen. Further research into the qualitative aspects of flipping, both from the standpoint of the teachers and the students, is a compelling project, but one left out of this study in the interest of time. Questions four and five focused primarily on what educators believe constitutes a flipped classroom and how they specifically flipped their classes. Below are the actual questions from the survey:

4. What about your method of instruction constitutes a flipped classroom?

5. Specifically, how do you implement flipped learning in your classroom? Be as specific as possible, mentioning how students acquire content, homework, etc.

Questions eleven through fifteen focused on educators’ beliefs about how flipped learning, changed their classes, changed their teaching, and changed student performance. Below are the actual questions from the survey:

11. What was the initial reaction of your students upon hearing that their class was going to become a flipped class?

12. Did student attitudes toward classroom flipping change after having gotten used to the model? Was this change for the better or worse?

13. How did your role as a classroom teacher change as a result of flipping your classes?

14. What changes, if any, did you notice in your classes after implementation of flipping?

15. Do you feel that learning has improved, stayed the same, or worsened as a result of flipping your classes?

Excerpts from some of these responses will be included in Chapter Four.
Returned surveys were graphically analyzed using the *Leximancer* program to determine overall themes surrounding educators’ beliefs about classroom flipping. *Leximancer* proved to be a useful tool in determining major themes in flipped classrooms, both in the literature and by instructors using the flipped learning model. The *Leximancer* product (info.Leximancer.com) was used for the purposes of qualitative and graphical analysis of the extant literature surrounding classroom flipping, as well as of the survey responses from teachers who were using the flipped classroom model. *Leximancer* claims the software “automatically analyzes text documents to identify the high level concepts in your text documents, delivering the key ideas and actionable insights you need with powerful interactive visualisations and data exports,” (info.leximancer.com) and has been used in numerous qualitative studies across many fields of research. Examples of the diversity of the product for the research purposes include: psychology (Cretchley, et al., 2010), supply chain management (Yu, 2011), and the health sciences (Travaglia, et al., 2009) to name a few. Penn-Edwards (2010) determined that *Leximancer* is a valid research tool for phenomonography, or the different ways in which people think about the world.

*Leximancer* employs two algorithmic techniques to create outputs, semantic and relational, as the software determines not only how often words appear in text, but also the relationships between these words and others within and between publications (Smith & Humphreys, 2006). Both the concept map, which is a multi-dimensional Venn Diagram, and the Thematic summary visually reflect these relationships, however, the concept map demonstrates overlap and direct connectivity between themes while the thematic summary provides analytics to numerically determine to what extent themes are related to one another. Themes are written over colored spheres in the same color as the sphere they are describing. Inside of each sphere
can be found additional words, written in black and in a smaller font. These words can be thought of as minor themes within each major theme, and are generated by the algorithm in the same manner as the major themes.

The themes in the thematic summary are the same as the spheres in the concept map. Connectivity between themes is determined by the two algorithms and each theme is given a relevance value and corresponding bar on the Pareto chart. The relevance value can be misleading when compared to the concept map because small relevance values in the thematic summary don’t necessarily correspond to small spheres in the concept map. The methods by which connectivity and relativity are determined is quite complex and beyond the scope of this study. A simplified way of interpreting the output and theme values is by thinking of relativity/connectivity values and how often a particular theme relates to other themes, both within and between publications, and their corresponding sphere size as a measure of how often the theme appears in text, regardless of its relationship to other themes. To prevent confusion, thematic summaries will not be presented in this dissertation; however, they are available in the appendices.

**Research Design and Procedures**

Using Propensity Score Matching, control and experimental subjects were matched by way of the Nearest Neighbor algorithm. Nearest Neighbor matching was accomplished using the R-Essentials plug-in for SPSS ([http://www.r-project.org/](http://www.r-project.org/)) in combination with SPSS. R is a free, open source statistical programming language that has an available plug-in for SPSS users. SPSS, while capable of doing some matching, doesn't have the available language and commands to execute certain matching methods, which was the reason the Nearest Neighbor algorithm was chosen, as SPSS is capable of executing this algorithm.
Prior to matching, covariate balance between experimental and control groups was examined using Cohen's $d$, and any covariates with an imbalance outside the acceptable tolerances ($|d|>1$) would have been re-evaluated and combined with other variables to create a composite score where pre-matching balance assumptions weren't violated. SPSS provides an output with individual balances for each covariate, and none of the covariates had unacceptable imbalance prior to matching. Post-matching assessment of covariate balance, again using Cohen's $d$ as well as the other tests and graphs provided by SPSS mentioned earlier, checked for covariate imbalances in the matched groups. Any such imbalances could have been addressed by either combining variables to create composite scores (as mentioned in pre-matching imbalances) or utilizing nonlinear techniques to create a new logit model. However, after analysis, there were no serious post matching, covariate imbalances.

After subjects were matched and the criteria for covariate balance was satisfied, math CRT scores were introduced into SPSS for each of the subjects. Nearest Neighbor matching resulted in experimental subjects being matched to more than one control subject, 1:5, and differences between experimental and control subjects’ CRT scores were calculated for both the mean and median of the five control subjects to each of the experimental units, thus producing fifty-five differences. Once differences in CRT scores were calculated for all experimental subjects and their respective matched control subject estimated value, SPSS determined the $t$-value for the differences, as well as the associated p-value and significance was determined. A more detailed exploration of both the matching methods used in the study and the statistical analysis of the CRT mathematics scores is forthcoming in Chapter Four.

Any threats to internal validity in the study would have come from data collection techniques used by WCSD, as the researcher had no control over how these data are collected. It
is the legal obligation of the WCSD to collect, compile, and present data in an accurate manner, and it was the assumption of the researcher that the data received met this high standard. If another researcher wished to replicate the study using different data, reliability would be reliant on said researcher using the same covariates, as well as the identical PSM techniques. Any variation from these two major factors could result in different conclusions. Further exploration of possible criticism of the study will be discussed in Chapter Five.

Data Analysis

Question 1. When performing propensity score matching on a particular data set, what are the analytical criteria to determine the optimal covariate balance to decide whether to use (a) 1:1 or 1:N matching ratio, (b) with or without control data replacement, and (c) and with or without caliper distances.

Data analysis procedures for Research Question 1 were as follows: a logistic regression was performed using the aforementioned covariates, and the regression equation produced was used to calculate propensity scores for individual students. Nearest Neighbor matching was used to match experimental and control subjects. The assumption of nonignorability of covariates was implied in the study. Similarly, covariate balance was addressed prior to any statistical procedures.

Question 2. Are there significant differences in the mean scores of the Math CRT exam between students who received the "flipped" instructional method and those who received the traditional instructional method?

Data analysis procedures for Research Question 2 were as follows: both a matched-samples t-test and an independent samples t-test were used to determine significant differences between the groups. The independent variable was the instructional method used with two levels,
the control group and the experimental group. The dependent variable was students’ mathematical comprehension, as measured by their CRT scores. The null hypothesis tested in the $t$-test was $H_0: \mu_T = \mu_C$, or there was no difference in performance on the math CRT scores for students who received flipped mathematic instruction versus those who received traditional mathematics instruction. For both $t$-tests, all assumptions were checked to insure valid results. The alpha level was chosen at $\alpha = .05$ for determining significant differences between the treatment and control groups.

*Question 3.* What are the perspectives of instructors who are utilizing flipped learning in their classrooms and, anecdotally, what are their perceptions of the differences in their students' learning?

Data analysis procedures for Research Question 3 wer as follows: The *Leximancer* product (info.Leximancer.com) was used for the purposes of qualitative and graphical analysis of the extant literature surrounding classroom flipping, as well as of the survey responses from teachers who were using the flipped classroom model. The *Leximancer* outputs from the survey responses were analyzed to determine if distinct thematic trends were present about what educators believe were the most important attributes of the flipped learning model.

Provided in the appendices is the survey filled out by instructors in the Washoe County School District known to have used the flipped method in their mathematical instruction. The qualitative data from the survey responses was analyzed and summarized to provide insight as to the teachers’ experiences, and will be included in the "Results" chapter.

**Rationale for Statistical Methods Used**

To create a truly Randomized Controlled Study (RCT) in hopes of answering the research question for this study would be impossible logistically, economically, and ethically. PSM
presents a new tool that helps create a “quasi-experimental” setting using pre-existing descriptive statistics wherein treatment effects can be determined without the prohibitive aspects of the RCT. PSM is an exciting technique in the eyes of the research and one which could potentially transform the educational research landscape, particularly if educational researchers can decode what appears from the outside to be an enormously complex statistical procedure. The possibility of playing a part in educating other researchers and distilling the process down to simpler processes is the determining factor for this researcher’s choice of statistical tool.

Measuring treatment effects was the purpose of the matched pair $t$-test and the independent samples $t$-test, and was the reason they were chosen for post-matching statistical analysis. The $t$-tests’ parsimony and ease of use with existing statistical software made it a logical choice for the task. Clearly, as researchers become more confident with using PSM, the panoply of statistical tools available to analyze data at the more granular level can be employed. However, in the interest of simplicity, clarity and brevity the $t$-test sufficed to determine whether significant differences existed.

Summary

A brief summary of the processes used in this study are as follows:

1. Covariate data from similar sixth grade students in Washoe County School District were collected and broken up into groups who received a flipped learning experience (experimental) and those who didn’t (control).
2. A logistics regression model was created by the covariate data and used to generate propensity scores for each student.
3. The Nearest Neighbor matching algorithm was used to match students based on their propensity scores.
4. Covariate balance was checked before and after matching, and because of the choice of matching algorithm, no covariate imbalances existed.

5. After matching, a matched-pairs $t$-test was used to determine if any significant differences between performances by each group on the math CRT exist.

Figure 8 below graphically shows the PSM procedures for this study:

Figure 8. PSM Procedures

As parsimony and simplicity are greatly valued in educational research, the intent of the study was to create a simple, easy to use template for future researchers wishing to use PSM to reference. The study will also add to the existing, but rather sparse body of research in the field of classroom flipping, especially at the elementary school level, in hopes of providing information to educators, administrators and parents to help them decide if flipping is a viable option to aid in educational reform.
Chapter Four

Results

Introduction

This chapter will present overall results from the study, with answers to the research questions presented in Chapter One. Because Propensity Score Matching (PSM) is a statistical method which has been hitherto relatively unused in the field of educational research, a more extensive analysis of the methods used to match the experimental and control group will be explored in answering Research Question 1 than would typically appear in the Results chapter of a dissertation. The author believes it is important to report in this manner to provide clarity to readers who have no experience with what can be considered a complicated and confusing statistical approach. Results from Research Question 2 and Research Question 3 will be presented in a more traditional manner, with quantitative analysis provided in the results from Research Question 2 and qualitative analysis provided in the results from Research Question 3.

Results by Question

Research Question 1 Results

Question 1. When performing propensity score matching on a particular data set, what are the analytical criteria to determine the optimal covariate balance to decide whether to use (a) 1:1 or 1:N matching ratio, (b) with or without control data replacement, and (c) and with or without caliper distances?

Data from 9,425 6th grade students from the Washoe County School District (WCSD) were de-identified and obtained from the district. The data was then cleaned to include only students from Title 1 elementary schools and subjects with any missing covariate values were removed from the study. The final number of students to participate in the study was 2,370. Of
those, 55 were formally defined as the experimental group and the remaining 2,315 subjects were considered part of the control group pool to be matched with experimental subjects.

Nearest Neighbor matching algorithm was chosen to match subjects because of its ease of use when combined with SPSS. While SPSS didn’t provide the logistics regression (logit) model used to create the propensity scores for each of the experimental and control subjects, it was easy to produce the logit model in SPSS using the “Regression” and “Binary Logistic…” commands. After the control and experimental data were input into SPSS, the following logit model was produced:

$$\ln \left( \frac{\hat{P}}{1 - \hat{P}} \right) = -7.512 + .164x_1 - .154x_2 + .629x_3 + .393x_4 + .598x_5 - .001x_6 + .008x_7$$

$$+ .009x_8 - .542x_9$$

Where \( \hat{P} \) is the propensity score for each subject. Recall that none of the assumptions about logistics regression were checked or were assumed to be true for this model and therefore no conclusions could be drawn from the model, other than it was used to determine the propensity scores for each of the subjects.

SPSS utilizes the “psmatching” plug-in utilizing code written in R (Ho, et al., 2007) to perform PSM. The following algorithms are used in concert with the SPSS interface to create the graphical outputs and tests for covariate imbalance: “MatchIt” by Ho, Imai, King, and Stewart (2007), “RItools”, by Bowers, Fredrickson, and Hansen (2010), and “cem” by Iaccus, King, and Porro (2009). Other matching algorithms such as optimal matching, full matching, etc. are available for researchers wishing to use PSM; however the user interface for these programs isn’t as easy to use as SPSS, and often requires the researcher to learn a programming language such as R (http://www.r-project.org/) in order to be able to use them, which can be prohibitive.
Experimental and control data were input into SPSS for analysis. Because SPSS was easy to use and NN options were easily customizable using radio buttons and dialogue boxes within the platform, 1:1 and 1:5 matching with and without replacement and with and without calipers set at within .1 standard deviations of the linear propensity score were calculated and outputs were examined. Ming and Rosenbaum (2000) determined that while 1:N matching can be beneficial, particularly when the number of control subjects greatly outnumbers the experimental units; any advantage in balance is usually optimized at around 1:5. A complete examination of the diagnostic processes, with examples, for determining which matching method to use is available in the appendices. For the purposes of answering Research Question #1 tables of pertinent diagnostic information are provided in this section, as well as the graphical evidence that informed the matched model decision made for this study.

Summary of Tests for Covariate Balance with Analysis

For the purpose of selecting the balanced model to be used to determine significant differences between students who received a flipped mathematics experience and those who hadn’t, the following diagnostics tools were used to meet the criteria for adequately balancing matched subjects:

1. The Hansen and Bowers (2008) test for overall balance. This test was only available for analysis when matching is 1:1, and without replacement. Figure 9 provides all of the overall balance values for each of the matching methods. Recall that covariates are considered poorly balanced if the test value is significant, or < .05. Clearly, in the two cases where the Hansen and Bowers Test was valid, there was no evidence of serious imbalance of covariates.
2. Relative Multivariate Imbalance, \( L_1 \), a second test for overall balance of covariates developed by Iacus, King and Porro (2009). There is no cutoff value for whether covariates are balanced, but a reduction of \( L_1 \), from before matching to after matching, will be observed when covariates are sufficiently balanced. See Figure 10 below:

<table>
<thead>
<tr>
<th>Type of matching</th>
<th>Before Matching</th>
<th>After Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1, W/O Replacement, W/O Caliper</td>
<td>.997</td>
<td>.945</td>
</tr>
<tr>
<td>1:1, W/ Replacement, W/O Caliper</td>
<td>.997</td>
<td>.945</td>
</tr>
<tr>
<td>1:1, W/O Replacement, .1 Caliper</td>
<td>.997</td>
<td>.964</td>
</tr>
<tr>
<td>1:1, W/ Replacement, .1 Caliper</td>
<td>.997</td>
<td>.945</td>
</tr>
<tr>
<td>1:5, W/O Replacement, W/O Caliper</td>
<td>.997</td>
<td>.971</td>
</tr>
<tr>
<td>1:5, W/ Replacement, W/O Caliper</td>
<td>.997</td>
<td>.971</td>
</tr>
<tr>
<td>1:5, W/O Replacement, .1 Caliper</td>
<td>.997</td>
<td>.967</td>
</tr>
<tr>
<td>1:1, W/ Replacement, .1 Caliper</td>
<td>.997</td>
<td>.967</td>
</tr>
</tbody>
</table>

*Figure 9.* Hansen and Bowers Overall Balance Test for all eight matching methods.

*Figure 10.* Multivariate Imbalance, \( L_1 \) (Iacus, King and Porro, 2010).
The Relative Multivariate Imbalance Test, $L_1$, (Iacus, et al., 2010) demonstrated that in all eight cases, $L_1$ was reduced from before to after matching, indicating again that no matching technique should have been eliminated based on this diagnostic test. There is no standard in the literature for how big the difference needs to be from before and after matching, only that a reduction is necessary to consider a matching technique to have improved covariate balance (Thommes, 2012).

3. Summary of unbalanced covariates. Figure 11 displays whether individual or combinations of covariates display imbalance ($|d| > .25$) after matching.

<table>
<thead>
<tr>
<th>Type of matching</th>
<th>Unbalanced Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1, W/O Replacement, W/O Caliper</td>
<td>GenderRecodedXLEPRecodedGenderRecodedXIEPRecoded</td>
</tr>
<tr>
<td>1:1, W/ Replacement, W/O Caliper</td>
<td>GenderRecodedXLEPRecodedGenderRecodedXIEPRecodedRaceRecodedXIEPRecodedIERecodedXCRTReadingRecoded</td>
</tr>
<tr>
<td>1:1, W/O Replacement, .1 Caliper</td>
<td>FRLRecodeXMathGradeRecodedIEPRecodedXLEPRecoded</td>
</tr>
<tr>
<td>1:1, W/ Replacement, .1 Caliper</td>
<td>GenderRecodedXLEPRecodedGenderRecodedXIEPRecodedRaceRecodedXIEPRecodedFRLRecodeXMathGradeRecoded</td>
</tr>
<tr>
<td>1:5, W/O Replacement, W/O Caliper</td>
<td>GenderRecodedXLEPRecoded</td>
</tr>
<tr>
<td>1:5, W/ Replacement, W/O Caliper</td>
<td>GenderRecodedXLEPRecoded</td>
</tr>
<tr>
<td>1:5, W/O Replacement, .1 Caliper</td>
<td>No Covariates Exhibit Large Imbalance</td>
</tr>
<tr>
<td>1:5, W/ Replacement, .1 Caliper</td>
<td>GenderRecodedXLEPRecoded</td>
</tr>
</tbody>
</table>

*Figure 11.* Summary of Unbalanced Covariates ($|d| > .25$).

The summary of unbalanced covariates demonstrates that seven of the eight matching techniques had interactive covariate terms that displayed a significant imbalance ($|d| > .25$), and 1:5 matching without replacement and with a .1 caliper had no significant imbalance, interactive or otherwise.
4. Dot Plots to graphically determine whether Cohen’s $d$ was reduced from before to after matching are shown in Figure 12 and Figure 13.

![Dot Plots for 1:1 Matching](image)

**Figure 12.** Dot Plots for 1:1 matches displaying $d$ before and after matching.
Figure 13. Dot Plots for 1:5 matches displaying $d$ before and after matching.

The Dot Plots show that all eight matching techniques had a reduction of imbalance for most of the covariates, however, based on the graphs, the best reduction of imbalance appeared
in the 1:5 experimental to control ratio, with replacement and a .1 caliper distance method (7 of 9 covariates display reduced imbalance) and the 1:5 experimental to control ratio, without replacement and a .1 caliper distance method (8 of 9 covariates display reduced imbalance).

Examination of histograms provided the information that all eight matching techniques provided ample evidence of common support, or propensity score distribution overlap.

5. Histograms to determine if common support exists between the control and experimental subjects’ propensity scores after matching are displayed in Figure 14 and Figure 15 on the following two pages.

These histograms were used to ensure that the criteria of common support between matched treatment and control groups were met. A quick examination of the distributions of the treatment and control group for each matching method demonstrated that ample common support was present.
Common Support Histograms for 1:1 Matching

<table>
<thead>
<tr>
<th>1:1, W/O Replacement, W/O Caliper</th>
<th>1:1, W/Replacement, W/O Caliper</th>
</tr>
</thead>
</table>

**Figure 14.** Common Support Histograms for all 1:1 matching methods.
<table>
<thead>
<tr>
<th>1:5, W/O Replacement, W/O Caliper</th>
<th>1:5, W/ Replacement, W/O Caliper</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="chart1" alt="Matched Treated" /></td>
<td><img src="chart2" alt="Matched Treated" /></td>
</tr>
<tr>
<td><img src="chart3" alt="Matched Control" /></td>
<td><img src="chart4" alt="Matched Control" /></td>
</tr>
<tr>
<td><img src="chart5" alt="Matched Treated" /></td>
<td><img src="chart6" alt="Matched Treated" /></td>
</tr>
<tr>
<td><img src="chart7" alt="Matched Control" /></td>
<td><img src="chart8" alt="Matched Control" /></td>
</tr>
</tbody>
</table>

*Figure 15.* Common Support Histograms for all 1:5 matching methods.
Choice of Matching Method

After reviewing all diagnostic tools, a decision needed to be made as to whether each matching method meets the criteria for balancing covariates. Based on all tests for imbalance in combination with the graphical data, it was the conclusion of the author that the 1:5 treatment to control ratio, without replacement and with a .1 caliper distance was the matching method that best produced covariate balance and decreased sampling bias. This choice of matching technique wasn’t without consequence. In Nearest Neighbor with 1:N matching, SPSS was forced to match five control subjects to each experimental subject, however, when using calipers, it is possible for there not to be enough control subjects that match within the caliper distance to produce the 1:5 ratio. When this occurs, SPSS assigns weights to each of the control subjects within groups where the 1:5 ratio isn’t possible. While this, in theory, fulfills the algorithm’s requirement of producing the correct ratio, the statistical consequence is an imperfect match. This is the trade-off for insuring that control matches are relatively close to their experimental counterparts. This was the case in this study. In the process of matching, SPSS matched 267 control units with 55 experimental units, which was clearly not a 1:5 ratio. A total of 8 control units couldn’t be matched to experimental units by virtue of enforcing a caliper distance on the matching algorithm. However, the weighting of control units within groups that didn’t have a 1:5 match accommodated for the loss of control data matched to experimental data (Ho, et al., 2007).

Once a matching method that doesn’t produce a 1:1 treatment to control ratio was chosen, the researcher had to decide how to proceed with statistical analyses that measured differences between groups. Mainly, when comparing the control and experimental groups, should the two groups have been treated as independent or paired? Elizabeth Stuart (2008, 2010) and Schafer
and Kang (2008) state that accounting for matched pairs isn’t necessary for two reasons. First, conditioning of the two groups will suffice so long as a linear regression, such as the logit model, is used to determine propensity scores. Second, the matching of the treatment and experimental subjects based on propensity scores doesn’t imply that they match equally on all covariates, and thus assuming perfectly matched subjects might produce biased results. However, Imbens (2004) and Austin (2011) argue that, by the very nature of matching subjects based on covariates, treated and control subject data comes from the same multivariate distributions, and thus should be treated as matched pairs. That said, in the interest of completeness, this study examined control and treatment subjects for treatment effects using both an independent samples t-test and a matched pairs t-test. In the case of the matched pairs t-test, two separate analyses were tested: one in which the mean of the five control units’ Math CRT scores were compared with the individual score of the matched experimental unit, and one in which the median of the five control units’ Math CRT scores were compared with the individual score of the matched experimental unit.

**Research Question 2 Results**

*Question 2. Are there significant differences in the mean scores of the Math CRT exam between students who received the "flipped" instructional method and those who received the traditional instructional method?*

Data from 9,425 6th grade students from the Washoe County School District (WCSD) were de-identified and obtained from the district. The data were then cleaned to include only students from Title 1 elementary schools and subjects with any missing covariate values were removed from the study. The final number of students to participate in the study was 2,370. Of
those, 55 were formally defined as the experimental group and the remaining 2,315 subjects were considered part of the control group pool to be matched with experimental subjects.

An independent samples $t$-test was performed prior to matching to determine if significant differences existed between the control group’s Math CRT scores and the experimental group’s Math CRT scores ($N = 2370$). The test failed to reveal statistically significant differences between the Math CRT scores of the students receiving traditional mathematics instruction versus those receiving flipped instruction. The power to detect effect size was determined to be .302. The results are summarized in the following Table 1 and Table 2 below:

Table 1. Descriptive Statistics for Omnibus $t$-test

<table>
<thead>
<tr>
<th></th>
<th>Experimental Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>309.53</td>
<td>295.61</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>64.92</td>
<td>71.23</td>
</tr>
<tr>
<td>N</td>
<td>55</td>
<td>2315</td>
</tr>
</tbody>
</table>

Table 2. Omnibus $t$-test ($\alpha = .05$)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>$t$ (2368)</th>
<th>$p$</th>
<th>Cohen’s $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.434</td>
<td>0.152</td>
<td>0.1965</td>
</tr>
</tbody>
</table>

After matching, two separate matched-pairs $t$-tests were performed using both the means and medians for the groups of five control subjects matched to each experimental subject. For the means, it was determined that no statistically significant differences existed between the control group’s Math CRT scores and experimental group’s Math CRT scores. Using the median scores also failed to reveal statistically significant differences between the control group’s Math CRT scores and the experimental group’s Math CRT scores. In both cases, a
negligible effect size \((d = .028)\) was present and the power to detect effect size was .075. The results are summarized in the Table 3, Table 4, Table 5, and Table 6 below:

Table 3. Descriptive Statistics for Paired \(t\)-test for Means of Control Group

<table>
<thead>
<tr>
<th></th>
<th>Experimental Group</th>
<th></th>
<th>Control Group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>309.53</td>
<td>64.92</td>
<td>307.93</td>
<td>42.38</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Paired \(t\)-test for Means of Control Group\((\alpha = .05)\)

<table>
<thead>
<tr>
<th></th>
<th>Paired Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.592</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>71.16</td>
</tr>
<tr>
<td>(t(54))</td>
<td>0.166</td>
</tr>
<tr>
<td>(p)-value</td>
<td>0.869</td>
</tr>
<tr>
<td>(d)</td>
<td>.055</td>
</tr>
</tbody>
</table>

Table 5. Descriptive Statistics for Paired \(t\)-test for Medians of Control Group

<table>
<thead>
<tr>
<th></th>
<th>Experimental Group</th>
<th></th>
<th>Control Group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>309.53</td>
<td>64.92</td>
<td>311.15</td>
<td>46.82</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Paired \(t\)-test for Medians of Control Group\((\alpha = .05)\)

<table>
<thead>
<tr>
<th></th>
<th>Paired Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-1.62</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>75.00</td>
</tr>
<tr>
<td>(t(54))</td>
<td>-0.16</td>
</tr>
<tr>
<td>(p)-value</td>
<td>0.873</td>
</tr>
<tr>
<td>(d)</td>
<td>.055</td>
</tr>
</tbody>
</table>

After matching, an independent samples \(t\)-test \((N = 322)\) was also performed on the data, and also failed to determine statistically significant differences between the control group’s Math CRT scores and the experimental group’s Math CRT scores. Again, there was a negligible effect
size \((d = .014)\) and the power to detect effect size was \(.055\). The results are summarized in Table 7 and Table 8 below:

Table 7. Descriptive Statistics for Independent Samples \(t\)-test

<table>
<thead>
<tr>
<th></th>
<th>Experimental Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>309.53</td>
<td>308.49</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>64.92</td>
<td>77.15</td>
</tr>
<tr>
<td>N</td>
<td>55</td>
<td>267</td>
</tr>
</tbody>
</table>

Table 8. Independent Samples \(t\)-test \((\alpha = .05)\)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>(t(320))</th>
<th>(p)</th>
<th>Cohen's (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.093</td>
<td>0.926</td>
<td>0.014</td>
</tr>
</tbody>
</table>

There were clearly no significant differences on math CRT’s between students who received flipped instruction and those who received traditional instruction. That said, it is important to notice the reduction in bias that resulted from the PSM procedure. When the omnibus \(t\)-test was performed, a hapless researcher might conclude that there was a small-to-medium effect size from the flipped learning experience. However, after matching, irrespective of whether a matched or independent samples \(t\)-test was performed, any indication of treatment effect disappeared. This implies that the PSM procedure removed the internal bias that could lead to the conclusion that the experimental group outperformed the control group.

**Research Question 3 Results**

*Question 3.* What are the perspectives of instructors who are utilizing flipped learning in their classrooms and, anecdotally, what are their perceptions of the differences in their students’ learning?
The overall proportion of teachers using classroom flipping was unknown, and given the emerging nature of the model, it was very difficult to locate teachers in the Washoe County School District who both used flipping and were willing to fill out surveys about their experience. The author contacted department chairs, curriculum coordinators, principals and colleagues from around the district and received a total of ten responses from teachers willing to complete the survey. All ten teachers had flipped their classes in either a mathematics or science course. There was one other teacher who had flipped her classroom but wouldn’t return request emails. While ten may seem to be a paltry number of responses, when the information provided by the responders was analyzed using the Leximancer software and combined with the Leximancer outputs from the literature, distinct trends emerged about what educators believed were the most important attributes of the flipped learning model.

This section will be split up into three parts. The first part will look at the concept map for teacher responses for the entire survey and provide a brief discussion concerning the major themes that arise from the map. The second part will disaggregate questions four and five from the survey, with a few excerpts of teacher responses from each question. The concept map and major themes will then be presented and discussed. The final part will disaggregate questions eleven through fifteen from the survey and again provide excerpts from the teacher responses, as well as concept maps and major themes.

Concept Map from Entire Survey

Figure 16, below, is the concept map for all ten surveys using all of the questions from the survey. While ten surveys wasn’t a huge number of surveys, and future researchers wishing to conduct similar studies would want a far greater number of respondents, the concept map provided insight as to what teachers who were actually using flipping were saying on the topic.
This concept map, when compared to the concept map created for the entire literature (see Chapter Two, Figure, 2), demonstrates that major themes such as “classroom” and “students” were pervasive, not only in the research, but by educators who were flipping their classes.

*Figure 16. Concept map from teacher surveys using all questions.*

Notice that the “students” theme has the greatest connectivity; however “class” and “classroom” have the greatest occurrence in the concept map. Leximancer isn’t able to discern that “class” and “classroom” are ostensibly the same theme. Notice that “flipped” and “flipping” are disaggregated. As one would hope, “students” appears to be the most important theme, by dint of its relationships with other themes, for teachers who used the flipped model. “Time” also became an important theme to teachers in the field who used the flipped model, whether it was a
bounty of time or a lack thereof, Leximancer isn’t able to make a distinction. However, it is reasonable to assume that teachers believe that flipping saved time.

Survey Questions Four and Five with Excerpts

The following are questions four and five from the teacher survey, with selected excerpts. Again, questions four and five focused primarily on what educators believe constitute a flipped classroom and how they specifically flipped their classes.

**Question 4. What about your method of instruction constitutes a flipped classroom?**

“To put it most succinctly, what my students used to do at school they now do at home, and what they used to do at home they now do at school. I used to teach students skills or concepts in the classroom, and then have them apply that knowledge at home. However, now I have “flipped” that process and do the exact opposite.”

“The classroom is setup where the students watch the lectures at home via video and their homework assignments are completed in class as well as projects and labs.”

“Essentially all of my lectures are delivered through online videos.”

**Question 5. Specifically, how do you implement flipped learning in your classroom? Be as specific as possible, mentioning how students acquire content, homework, etc.**

“I post one to two videos a week on Edmodo, they have to take notes and I quickly check them the next day. They also have to answer a short question on Edmodo to
receive points for watching the video. I assign homework through an online website, MathXL, and they work on this in class on the iPads.”

“I would give my students the Math topic to be covered the following day. They were to watch the Holt Math tutorials and the Khan Academy tutorials that went with the topic. They would watch the videos prior to class which then allowed time for us to delve deeper into the skill. Once the class time started, a very brief overview would take place to refresh their memory of the topic. The remainder of class time would be spent going deeper with the skills, using them in a real-world setting, and developing a better understanding through activities.”

“I post my video lectures to my website, that link to YouTube to have students watch on the computer. For those students that do not have a computer at home, I burn a DVD of the unit for them to watch at home. 100% of my students are using the flipped instruction via the video as I only lecture if the students are struggling with a certain concept in the classroom setting. For homework, students are asked to complete assignments in class that followed the video.”

**Concept Map for Questions Four and Five**

Figure 17, below, is the concept map for questions four and five, questions designed to determine specifically how teachers implemented flipping in their classrooms:


**Figure 17.** Concept map for questions four and five from teacher surveys.

**Major Themes from Questions Four and Five**

Much like the review of the literature, the “students” and “class” themes have the greatest frequency in these questions, with “students” having the greatest amount of connectivity. Strangely, there is no mention of technology, other than “videos”, anywhere in the concept map. Although unexpected, and certainly a topic for future research, it can be hypothesized that teachers who flipped their classes were comfortable enough with technology that it wasn’t a concept worth mentioning. To that end, none of the survey responses contained any mention of technology, other than specific products and applications teachers were using to flip, reinforcing the hypothesis that teachers who endeavored in flipping their classes were technologically savvy.
Another overarching theme for questions four and five was the consistency with which teachers understood the general idea of what flipping was; removing the instructional portion of a lesson from the class, having students view the content on their own time, and working on problems, projects, labs, etc. during class time. Teachers consistently understood the general concept of classroom flipping, however, the spectrum of methods by which educators chose to flip their classes was a compelling aspect of their responses. One teacher simply used Khan Academy videos, while another created their own YouTube videos. In both cases, the teachers had students completing homework in class, as opposed to the first excerpt, where the teacher had students finishing homework using the online platform MathXL (www.mathxlforSchool.com). Future research into which of these methods is most effective is an exciting extension of this study.

Survey Questions Eleven through Fifteen with Excerpts

The following are questions eleven through fifteen from the survey, with selected excerpts. Recall, questions eleven through fifteen focused on educators’ beliefs about how flipped learning changed their classes, changed their teaching, and changed student performance. In the interest of brevity, only one or two excerpts from each question are given.

Question 11. What was the initial reaction of your students upon hearing that their class was going to become a flipped class?

“I polled the students on Edmodo at the beginning of the school year on whether they loved it or hated it. 99 students responded out of 110, with 75% saying they loved it.”

“I think it was very mixed. Some students thought it was a great idea and some heard “regular homework” and didn’t seem very pleased.”
**Question 12.** Did student attitudes toward classroom flipping change after having gotten used to the model? Was this change for the better or worse?

“I think most of my students see the value in the flipped classroom. Also, students that I taught last year, who had already expressed their outright enjoyment of my class, heard what I was doing and expressed their desire to have been a part of my “flipped” curriculum.”

“Their attitudes did change. I commonly elicit feedback concerning which type of homework was more enjoyable and which type helped them learn more. Without fail students support the flipped model.”

**Question 13.** How did your role as a classroom teacher change as a result of flipping your classes?

“I became more of a facilitator than instructor. I still provide students with small doses of direct instruction, especially for my accelerated classes. However, class time is more vibrant and I exist to have conversations with individual students and small groups in a busy, noisy environment.”

“I still deliver content, though now I do so remotely. Now I spend significantly more time directly helping students with problem solving.”
Question 14. What changes, if any, did you notice in your classes after implementation of flipping?

“I see an increase in academic outcomes and far less behavior issues. My students are far more active in the learning process. My curriculum is much better suited for deep and real examination of historical concepts and issues.”

“Much higher test grades and student retention of material. Students are able to understand the material and apply it to real world settings.”

Question 15. Do you feel that learning has improved, stayed the same, or worsened as a result of flipping your classes?

“I truly believe that learning has improved by using the flipped model of instruction. I am able to work with the students who normally struggle and quit and guide them in the right direction.”

“I am much more concerned about students developing the ability to develop an argument using reasoning and evidence to support their claims. Given my aforementioned perspective, the learning has greatly improved for all of my students.”

Concept Map for Questions Eleven through Fifteen

Figure 18, below, is the concept map for questions eleven through fifteen. These questions were written to determine teacher perceptions of how flipped learning had affected their teaching as well as their perceptions of student performance as a result of flipping their classes:
Major Themes from Questions Eleven through Fifteen

Notice again how the primary theme for these questions is “students”. It appears as though the qualitative analysis of flipped learning affirmed what the proponents of the model assert: the flipped model appears to be a student-centered approach to education. That said, the last Leximancer output is the most compelling because the questions it analyzed focused on teachers’ perceptions of their flipped classrooms and how they believe their students perceived the classes. The themes had a distinctly positive tone which leads one to believe that both students and teachers viewed the flipped learning experience in a positive light. Survey responses indicated educators perceived that their students’ performance had “improved” as a
result of their flipped classes, and that students learned content at a much deeper level. They
evaluated the change in their roles as positive, and believed students enjoyed the flipped classes
better than the traditional classes, although it took some students a period of adjustment before
becoming comfortable with the new technique. Lastly, teachers believed that flipping their
classes saved them time that was better spent collaborating with students, allowing students to
work on projects, and working collaboratively. They felt this shift in roles increased their job
satisfaction and made their jobs more rewarding.

**Summary**

In summary, the following are the findings from this study:

1. Graphical and statistical tests were used to determine the best method for matching
control and experimental units. A 1:5 experimental to control ratio, without
replacement of control units, and with a caliper distance within .1 standard deviations
of the linear propensity score was decided upon based on the evidence.

2. An omnibus, independent samples *t*-test was performed using 55 experimental
subjects and all 2315 control subjects from the study to compare performances on the
Math CRT. Per the *t*-test, a small-to-medium effect size appeared to be present (*d* =
.1965), however, there were no significant differences between the two groups.

3. After PSM, three separate *t*-tests were performed using a 1:5 experimental to control
subject ratio. The first *t*-test compared the means of each 5-subject control group to
the matched experimental subject. The second *t*-test compared the medians of each 5-subject control group to the matched experimental subject. Lastly, an independent
samples *t*-test compared the entire experimental group to the entire control group.
There were no significant differences between control and experimental groups,
however, a reduction in bias that over stated the effect size of the flipped group in the omnibus $t$-test occurred as a result of PSM.

4. The major themes from the surveys were consistent with many from the literature; primarily, flipping is a student-centered teaching model that is gaining popularity in both educational and research circles.

5. Primary themes from disaggregated survey responses which focused on how educators’ understanding of classroom flipping and the mechanics of flipping their classes indicated that while educators clearly understood what classroom flipping is, the manner in which they choose to flip their classes existed on a large spectrum.

6. Primary themes from disaggregated survey responses which focused on the educators’ perceptions of how student performance and instructor job satisfaction in flipped courses had changed were consistently positive, with the belief that students performed better and learned material more deeply, while teachers believed their jobs were more collaborative and rewarding.
Chapter Five

Conclusions and Discussions

Introduction

The term “flipping” is becoming ubiquitous in the lexicon of educators at all levels. While the practice of flipping occurs across a spectrum, the underlying theme is the use of technology to remove direct instruction, or content acquisition, from the classroom and transplant that time with dynamic, student-centered activities that augment learning. Content acquisition is now considered homework, which “flips” the traditional paradigm (Davies, et al., 2013). While research is emerging as to the efficacy of flipped learning, there has been little, if any, done at the elementary level. The purpose of this study was to add to the existing body of research on classroom flipping, particularly at the elementary school level. The study will also provide researchers with analytical tools such as Propensity Score Matching (PSM) to help overcome the challenges associated with performing Randomized Controlled Trials (RCT) by using the large amounts of educational data available to the public to create quasi-random settings. It is important to note that, in absence of large amounts of data, PSM won’t necessarily produce valid conclusions by virtue of potential violations of the necessary assumptions for the use of PSM (Rosenbaum & Rubin, 1983).

Summary of Research Questions

1. When performing propensity score matching on a particular data set, what are the analytical criteria to determine the optimal covariate balance to decide whether to use (a) 1:1 or 1:N matching ratio, (b) with or without control data replacement, and (c) with or without caliper distances.
2. Are there significant differences in the mean scores of the Math CRT exam between students who received the "flipped" instructional method and those who received the traditional instructional method?

3. What are the perspectives of instructors who are utilizing flipped learning in their classrooms and, anecdotally, what are their perceptions of the differences in their students' learning?

Conclusions by Research Question

Conclusions from Research Question 1

The primary assumption for Propensity Score Matching (PSM) is the “absence of nonignorable covariates” assumption for covariate vectors entered into the logit model, which calculates the propensity scores. The assumption implies that all variables that have an effect on treatment assignment must be measured and included as part of the covariate matrix. Violation of this assumption can potentially produce biased results of treatment effect (Brookhart, et al., 2006, Rubin & Rosenbaum, 1984). Rubin (2001) states that careful planning of the observational study is crucial when using PSM to determine treatment effects, and such planning should include rigorous examination of variables to insure all nonignorable covariates are included. This study assumed all nonignorable covariates were included in the logit model that determined probability of membership in either the control (traditional) or experimental (flipped) groups.

SPSS was used to create the logit model that determined propensity scores for each of the members of the control and experimental groups. A total of 2,370 subjects, 55 in the experimental and 2,315 in the control group, were input to create the model. After the logit model was created, SPSS, in combination with the R Essentials plug-in, matched control and
experimental subjects using Nearest Neighbor matching to create a total of eight different scenarios as follows; control to experimental subject ratios were 1:1 and 1:5; with and without replacement of control subjects when matching; and with and without caliper distances of .1 standard deviations of the linear propensity score.

After the eight total matching models were generated, analysis of the diagnostics tools provided by SPSS determined which model produced the best covariate balance between control and experimental subjects. The Hansen and Bowers (2008) test for overall balance, $L_1$ (Iacus, et al., 2009), and Cohen’s $d$ were used to determine any significant imbalance between covariates individually as well as overall imbalance in the model. Dot Plots were used to determine best improved covariate balance before and after matching. Lastly, propensity score histograms of both experimental and control subjects insured common support (overlap) after matching. Analysis of diagnostics to ensure that the criteria for well-balanced covariates were met led to the conclusion that the optimal matching method that lead to the greatest balance between control and experimental subjects was the 1:5 control to experimental ratio, without replacement of control units, with a caliper distance of .1 standard deviations of the linear propensity score.

**Conclusions.** PSM is a valid statistical tool for researchers working in the field of education who have access to data procured from observational studies where covariates are thoughtfully chosen to ensure the nonignorability assumption is met. Given this, and if the necessary criteria to demonstrate covariate balance is also met, researchers may assume that any existing bias in the study is reduced. Because bias is a major criticism against the use of observational studies for measuring treatment effect, this reduction lends credibility to research done using PSM.

**Conclusions from Research Question 2**
An independent samples \( t \)-test was performed prior to matching to determine if significant differences existed between the control Math CRT scores and the experimental Math CRT scores \((N = 2370)\). The test failed to reveal statistically significant differences between the Math CRT scores of the students who received traditional mathematics instruction and those who received flipped learning mathematics instruction. After matching, two separate matched-pairs \( t \)-tests \((N=55)\) were performed using both the means and medians for the groups of five control subjects matched to each experimental subject. For the means, it was determined that no statistically significant differences existed between the control group’s Math CRT. Using the median scores also failed to reveal statistically significant differences between the control group’s Math CRT scores and the experimental group’s Math CRT scores. Also after matching, an independent samples \( t \)-test \((N = 322)\) was performed on the data, and also failed to determine statistically significant differences between the control group’s Math CRT and the experimental group’s Math CRT scores.

Separate matched-pairs \( t \)-tests were performed to repudiate any criticisms about whether matched experimental subjects are independent of their control counterparts. Austin (2011) and Imbens (2004) claim that matched subjects should be treated in a manner consistent with paired experimental design. Stuart (2010) and Schafer and Kang (2008) argue that it is statistically acceptable for matched groups to be considered as independent. In the interest of completeness, both analyses are included.

**Conclusions.** While there were no significant differences on Math CRT scores between students receiving the flipped learning experience and those receiving the traditional experience, there are still a number of pertinent conclusions that can be drawn from the study. The first is that the evidence suggests classroom flipping performed at least as well as the traditional method
of instruction. Akin to the medical field, a “do no harm” approach should be observed when considering the implementation of new learning models and pedagogies. This implies that educators wishing to offer a richer technological experience for their students through classroom flipping needn’t worry that they can potentially harm the students’ educational process.

Second, as shown by Liu, Maddux, and Johnson (2008), the presence of no significant differences should in no way lead to the conclusion that students who experienced the flipped learning model didn’t receive a benefit from it. Liu, et al., go on to state that empirical research isn’t designed to disprove anything, but rather to build a body of evidence around theories. This study used one data point, the Math CRT, and one instructor using one technique of flipping to determine differences. The literature has shown that flipped instruction is on a spectrum that uses a variety of methods across a number of content areas. This study shouldn’t be considered a repudiation of classroom flipping, but rather a starting point for future research in the field.

Lastly, while “flipped teacher” was an experienced, professional instructor and well regarded by her supervisors as an exemplary sixth grade teacher, she had little formal training with designing courses using technology. Her response to question 2 of the survey, “What research, if any, did you do prior to flipping your class?” was as follows:

“I spent my summer researching articles about the flipped classroom. I logged many hours on the internet finding information.”

This response indicates that “flipped teacher” wasn’t able to receive any formal guidance to prepare her for flipping her classes. At the time when she first began flipping, there was very little, if any, empirical research done on the field. There were also no extant best practices available to guide her in developing and implementing the flipped math course. This is in no way an indictment of “flipped teacher”. Instead, she should be praised for the hard work and
courage it took to try a new approach to teaching mathematics. She shouldn’t be disheartened by this study since, at the very least, it showed her students performed as well as other students with similar backgrounds using the new instructional method.

**Conclusions from Research Question 3**

A total of ten surveys were collected from WCSD teachers who are using flipped instruction in their math or science courses. The documents from the review of the literature and survey responses were input into Leximancer to determine if there are thematic similarities between teachers’ beliefs about classroom flipping and what the research has shown. Selected questions were disaggregated to more directly address the research question. Teacher responses and Leximancer concept maps were used to determine overarching themes.

**Conclusions.** Qualitative analyses of overall themes in classroom flipping point to a belief by educators that classroom flipping is a student-centered approach to education that affords more time in the classroom to interact with students. Respondents to the survey for this study claimed that classroom flipping provided students with more time in class, a greater number of resources for learning, and better access to the teacher. Teachers found their classroom roles changed, and they became more facilitators of learning, rather than driving the learning as they had when they taught in the traditional manner. Teachers also believed they had more time to interact with students and as a result believed, anecdotally, that students performed better in their classes. Teachers also felt a greater sense of job satisfaction as a result of the changes brought about by flipping their classes. None of the respondents stated they had any institutional training in how to flip their classes or had received any information about classroom flipping other than what they had researched on their own.
Implications

The following section will describe the implications that this study has on the educational research community. A study of this nature seems to create more questions than it answers and this section will attempt to add clarity to the direction future research can take. This section will also provide recommendations for researchers who might be interested in research in a similar field or using similar statistical techniques.

Propensity Score Matching in Educational Research

As data is readily available and public school districts are legally obligated to collect it, PSM should be considered a useful alternative to more traditional statistical methods for researchers wishing to determine the effectiveness of learning models, curricula, differences in learning outcomes, etc. Researchers should seek to form lasting relationships with the data administrators in school districts where they hope to do research. Researchers should also work carefully with their Institutional Review Board (IRB) to ensure all data is collected and research is performed in ways compliant with institutional policies.

Researchers also need to be well trained in creating observational studies that meet the assumptions of PSM. Rubin (2000) and Rosenbaum (2009) assert that the conclusions of any study using PSM are only as good as the observational study created by the researcher. In the case of this study, *ex post facto* data were used, and the collection of said data was not directly under the control of the researcher. In cases such as this, the research must be diligent about ensuring that all nonignorable covariates are collected. Whether the researcher is designing the observational study or using *ex post facto* data, meeting the primary assumption for PSM can be accomplished through extensive study of the literature and consultation with experts in the field.
to determine which covariates are ignorable and which aren’t (Luellen, et al., 2005). For researchers with the ability to create observational studies, a pilot study can be very insightful in determining ignorability of covariates.

Researchers need to be well trained in the use of PSM to avoid specious and misleading conclusions. PSM is a confusing and involved statistical technique that requires a great deal of study on the part of the uninformed researcher prior to attempting to utilize it as part of any study. Researchers who feel they have weaknesses with quantitative methods are advised to avoid using PSM in absence of a mentor who is well schooled in the nuances involved with its use, lest they run the risk of drawing feckless conclusions about their findings. A recommendation for future research would be the creation of a step-by-step, user friendly textbook, along with ancillary computer programs, which researchers in the social sciences can use if interested in utilizing PSM in their research.

PSM can open up fields of research in education that have hitherto been untapped because of the issues of legality, ethics and cost. Researchers in all fields have long been hindered by ethics, legal issues, and cost when attempting to conduct studies, and rightly so. The protection of the research subjects should be paramount in the minds of researchers at all times, however, this code of conduct isn’t always conducive to performing relevant and compelling research. Randomized controlled trials are expensive, time consuming and, at times, difficult to get through IRB. PSM, particularly when used in the manner it was in this study, entirely avoids these issues. *Ex post facto* data are readily available, and all data are often collected by a third party, ostensibly one that behaves in a legal and ethical manner. The only cost in such studies is the labor required to procure the data and any stipend the researcher is able to provide for the data administrator. In the interest of full disclosure, a small fee of less than $500 was provided
the data administrator for the WCSD in return for their assistance with the study.

Observational studies and any accompanying pilot studies may require more energy and incur greater cost to the researcher than using already existing data, but it can be assumed that this would still be substantially less than the cost of an RCT.

**The Field of Mathematics Education**

This study confirms what many prior empirical studies have concluded: provided that students have access to appropriate technology, classroom flipping is an option for educators to use in place of the traditional, lecture-based model. Although this study showed no significant differences in the performances of the flipped class versus the traditional classes, it is important to restate that the flipped subjects performed *at least as well* as the control group on the Math CRT. This study notwithstanding, there is a growing body of research available for teachers interested in flipping their classes across all grades and content areas. In addition to the studies cited in the literature review, there are a number of studies available at the Flipped Learning Network ([www.flippedlearning.org](http://www.flippedlearning.org)) that educators can use to provide evidence to their stakeholders as to the efficacy of flipped learning. Flipped Learning Network also provides teachers with guidelines for how to design and develop their flipped courses such as the “Four Pillars”, books on flipped learning and webinars about flipped learning and the information surrounding it.

There is also a growing body of videos for teachers who aren’t comfortable creating their own. Sites like YouTube ([www.youtube.com](http://www.youtube.com)) and TeacherTube ([www.teachertube.com](http://www.teachertube.com)) have libraries of videos that span large amounts of content areas and topics within those areas, some of which are expressly created to align with individual state standards, as well as the Common Core State Standards (CCSS). It is the author’s belief that flipped learning is most effective when
teachers create their own videos, however there is no extant research addressing this, which leaves the matter open for future research in the field.

**Mathematics Education at the Elementary School Level**

Unlike studies done at the secondary and post-secondary levels, this study focusses on elementary school aged students. An extensive search of the literature produced none others like it, although there may be some as of yet unpublished research on the topic. It is reasonable to assume that, like classroom flipping at the secondary and post-secondary levels, there is a growing number of elementary teachers who are either expressly utilizing classroom flipping or are interested in doing so. It is also highly improbably that “flipped teacher” is the only elementary school teacher in the world who flips her mathematics classes. Lest we allow teachers at the elementary level to move forward with no training, no best practices specific to their grade levels, and no research to bolster their ideas and methods, the research community needs to focus its efforts on insuring that these teachers are well supported with a wealth of empirical research on the subject. All stakeholders should be well informed prior to making broad changes to curriculum to ensure that methodologies and pedagogies *vis-à-vis* classroom flipping are sound and shown to improve, or at least not impede, student outcomes. This is true for all educational levels, but most would agree and the literature verifies that the elementary level is the most important for grabbing student interest and building a foundation for success at subsequent levels.

**Qualitative Research in the Field**

This study focused mainly on quantitative analysis of classroom flipping, however qualitative research will play an important part in the field as it grows and transforms as part of the educational mainstream. How students feel about participating in flipped classrooms, student
perceptions about their success in flipped classrooms, and how well students believed they were prepared for the next course in a sequence after a flipped experience are examples of important questions that a competent qualitative researcher can address. Likewise, teachers’ perceptions about student performance in flipped classes, how the role of the teacher changed after flipping, and how flipping affected job satisfaction are compelling questions the future research can focus on. These are merely a few examples selected from the innumerable possibilities that qualitative methods afford the research community, and it is left to other researchers to begin to address some of the more compelling questions.

Limitations

Observational Data Used in the Study

The researcher had no control over the data that was collected, how it was collected, and which covariates were made available for the study. The data provided by WCSD was exhaustive in its demographic and quantitative aspects, however there was an absence of any qualitative data and informal assessment data used by flipped “flipped teacher” and teachers of the control group during to time of the study. For this reason, the nonignorability assumption, pivotal to PSM, was assumed to be met. It was reasonable to infer that the data provided by WCSD was ample for meeting the nonignorability assumption, but one can’t be certain if an important covariate had been overlooked. Researchers wishing to duplicate similar research should run a pilot study to determine which covariates need to be observed to insure valid conclusions. Optimally, researchers should work with data administrators to create an observational study from scratch where the researcher has control over which data will be collected over the course of the study.
Instructor Used in the Study

At the time of the study, flipped learning was still in its infancy. There were limited resources available for teachers wishing to flip their classes, and no institutional training available. “Flipped teacher” simply heard about it flipping, thought it would be a good idea, obtained permission from her administrator and moved forward. Her total understanding of flipping came from Salman Khan’s TED Talk. Again, this is in no way an indictment of her as a professional educator, but rather she should be commended for her willingness to take a risk and use an instructional technique that seemed intuitively better that the traditional way of teaching mathematics. She felt that her students would receive a greater benefit from a flipped experience and she had the courage to try it. Her lack of training in the use of technologies such as screen capture and video editing software led her to the choice to use videos from Khan Academy which are informative, mathematically vetted, and well produced. Unfortunately, Khan Academy videos aren’t created to address specific standards, but rather individual topics. Future research may determine that if teachers create their own videos, directly aligned with the standards exams are written to address, significant differences between flipped learners and traditional learners may, in fact, exist.

Software Limitations Available in SPSS

Researchers wishing to use PSM in their work, but who aren’t trained in the use of statistical software packages such as R and STATA are forced to use plug-ins with more user-friendly programs such as SPSS. R, for example, is a difficult program to learn, and proficiency with the program is tantamount to mastering a new programming language, C++ (Sekhon, 2011), which the program is written in. R is an incredibly powerful program capable of running whatever statistical analysis a researcher might want. It can easily run PSM algorithms such as
Optimal Matching, Greedy Matching, and Nearest Neighbor Matching. However, because it is so difficult to use, many researchers may find themselves opting for the ease of use of the SPSS plug-in, which was that case for this study.

The problem with this approach is the matching algorithm options are limited in SPSS. The researcher only has access to Nearest Neighbor matching, combined with 1:N experimental to control ratio, matching with or without replacement of control subjects, and with or without the use of calipers. While this may seem like an impressive array of options, the lack of flexibility afforded by other matching algorithms may limit the researcher’s ability to draw substantive conclusions, and thus force the researcher to abandon the use of PSM. Researchers wishing to use PSM on a regular basis in their work should consider becoming formally trained in the use of R. This will open up a variety of options that SPSS doesn’t have the capability to offer.

**Difficulty in Replicating the Study**

WCSD and the State of Nevada Department of Education are in the process of phasing out the use of CRT’s as an assessment to meet the requirements of No Child Left Behind (NCLB). This year, the CRT is being replaced by the Smarter Balanced Assessment (www.smarterbalanced.org) that is aligned to the CCSS, which the CRT was not. It will also be difficult for a researcher to find an elementary-level teacher who had flipped their classes from 2011-2013. As the number of elementary-level teachers employ flipping in their math classes increases, researchers will have to use the Smarter Balanced Assessment as their dependent variable to determine differences, otherwise risk bias by virtue of the difference between assessments.
Recommendations

**Instructional Design in Flipped Learning**

One of the four pillars of flipped learning is Intentional Content. Because flipped learning is so reliant on technology, educators wishing to flip should have at least a rudimentary understanding of Instructional Design and Technology (IDT). Educators wishing to create their own content need to be trained not only with the technology but with how to integrate the technology into their courses so that students feel well supported, both in class and while they are acquiring content online. If the end goal of flipping is to create more independent learners, students must have a way to evaluate their understanding as they acquire knowledge.

All of these issues can be addressed if educators have a strong foundation in Instructional Design. Teachers well-versed in Instructional Design models such as the ADDIE model mentioned in this dissertation are thoughtful enough to create intentional content, flexible enough to change their teaching methods as needed and constantly gathering data and evaluating whether flipping is effective and students are learning.

Future research needs to concentrate around which instructional design techniques work best in the context of flipping. This will better inform educators using the practice of flipping as well as institutions who train educators to insure that a solid set of best practices is available and being utilized by those in the field.

**Continued Research in the Field of Classroom Flipping**

**Elementary education.** Foundations for future success are developed at the elementary level. It is vital that educators are using the best possible practices to educate the young. Classroom flipping should be unequivocally determined to be *at least as effective* at teaching elementary school students. Assuming classroom flipping’s effectiveness, comprehensive
research across the spectrum of techniques that teachers are using to flip their classes should be emphasized in order to inform best practices and allow classroom flipping to become institutionalized in teacher education programs.

**Secondary education.** Classes at the secondary level begin to specialize into periods and, historically, student performance in mathematics in the U.S. begins to drop off at this level (OECD, 2012, TIMMS, 2012). It is again vital for teachers wishing to use flipping to be well informed as to design and implementation, as well as the efficacy of the model.

**Post-Secondary education.** Unencumbered by the legal issues present when researching minors, research at the post-secondary level has the ability to use RCT’s versus less robust experimental methods and thus conclusively determine if classroom flipping improves learning outcomes. If so, the post-secondary level can engender a solid set of best practices, model design methods, and teacher training in the field of classroom flipping.

**Statistical Methods other than PSM**

This study focused primarily on the use of PSM in determining treatment effect. It is important to note the other options available to researchers wishing to do research in the field of classroom flipping. The following are a few of the many more traditional methods researchers can use to determine if significant differences exist between flipped learners and traditional learners.

**RCT: The “gold standard” of experimental design.** After an RCT has been designed and implemented, the data collected can be analyzed using the full array of traditional statistical methods (ANOVA, Factorial Analysis, MANCOVA, MANOVA, etc.) to determine differences and treatment effects between flipped and traditional classrooms.
Multiple regression analysis. A researcher can create a regression model wherein classroom flipping is one of the independent variables, along with other covariates, and the dependent variable is the score on an assessment. Traditional regression techniques such as odds ratios and beta weight analysis can help determine the relationship between classroom flipping and assessment outcomes.

Logistics regression and discriminant analysis: For these statistical methods, a researcher can include participation in a flipped classroom as an independent variable and then determine what should be defined as the dependent variable. Since these methods define dependent variables as membership in a category, examples of dependent variables could be pass versus fail, exceeds standard vs. meets standard vs. below standard, enjoyed course vs. not enjoyed course, etc.

Non-Parametric analysis. In the event that assumptions for the aforementioned techniques can’t be met or adjusted for, the full panoply of non-parametric methods can be employed to help determine differences and treatment effects.

Summary

In summary, this study was an opportunity to perform empirical research on the exciting and emerging field of classroom flipping. There is a great deal of research still to be done in the field of classroom flipping before it can be vetted as a legitimate learning model. This study presented evidence that classroom flipping, at the very least, doesn’t appear to do any harm to students, and should be used as a call for further research in the field, particularly at the elementary school level. The fact that classroom flipping exists on a spectrum and overlaps with other learning methods such as blended and online learning can make the task of research
difficult. However, motivated researchers interested in being on what appears to be the cutting edge of education will find a way.

This study was also an opportunity to introduce an innovative statistical technique to researchers who may never have heard of PSM, and are stymied by issues associated with running Randomized Controlled Trials (RCT). This study demonstrates that PSM is a viable statistical tool in the field of educational research and effective at reducing bias related to observational studies. Educators in search of statistical techniques that can overcome their study challenges may find that PSM is an excellent tool to work with. However, PSM is a difficult statistical tool to work with and easy-to-use training materials need to be created if all but the most technologically savvy researchers are to have access.

In summary, this study adds to the body of knowledge in the fields of both Propensity Score Matching and classroom flipping. Ideally, it will create opportunities for further investigation into classroom flipping and, at the same time provide a resource for researchers in education and the social sciences fields wishing to use Propensity Score Matching.
References


FLORIDA STATE UNIV TALLAHASSEE CENTER FOR EDUCATIONAL TECHNOLOGY.


*PNA*, 8(1), 1-20.


Appendix A

Flipped Instructors’ Survey

Introduction: While the interpretation and implementation of Flipped Learning vary greatly, the best definition that I’ve heard for Flipped Learning was posited by Ramsey Mussalam: “Classroom Flipping is the practice of taking direct instruction and moving it from the group learning environment to the individual learning environment.” If you have flipped even one lesson, you are considered a flipped teacher. You needn’t have flipped your entire class to fill out this survey. Thank you for taking the time to assist me in my research.

Instructions: Please fill out the following survey to the best of your knowledge, by editing this Word document. Upon completion, please submit the survey via email to dripley@davidsonacademy.unr.edu. All answers and your identity will be kept confidential, however, I would be happy to show you my dissertation upon completion to help inform any future decisions you make around classroom flipping.

1. Where/from whom did you first hear about the Flipped Learning Model?

2. What research, if any, did you do prior to flipping your class? Did you watch videos, see an article on the news, etc.?
3. What about the Flipped Learning Model Intrigued you to point that you were interested in its implementation in your classroom?

4. What about your method of instruction constitutes a flipped classroom?

5. Specifically, how do you implement flipped learning in your classroom? Be as specific as possible, mentioning how students acquire content, homework, etc.
6. Prior to implementation, what conversations did you have with your supervisors/administrators?

7. Prior to implementation, what conversations did you have to have with other stakeholders such as colleagues and parents?

8. How did you prepare your students for the flipped classroom?

9. What technological obstacles did you have to overcome prior to implementation?
10. What technological obstacles did you have to overcome after implementation?

11. What was the initial reaction of your students upon hearing that their class was going to become a flipped class?

12. Did student attitudes toward classroom flipping change after having gotten used to the model? Was this change for the better or worse?

13. How did your role as a classroom teacher change as a result of flipping your classes?
14. What changes, if any, did you notice in your classes after implementation of flipping?

15. Do you feel that learning has improved, stayed the same, or worsened as a result of flipping your classes?

16. Do you have any data to back up your claims from question #14? You needn’t provide this data.
17. Would you recommend flipping for any of your peers in either your subject area or another? Are there any subject areas that you feel wouldn’t be conducive to flipping?
Appendix B

A Glossary of Terms and Examples For Use in Propensity Score Analysis

**Randomized Controlled Trial (RCT)**

A specific type of scientific experimental design wherein subjects participating in the experiment are randomly assigned to either a control group, which receives no treatment (but sometimes a placebo), or the experimental group, which receives the treatment is called a Randomized Controlled Trial (RCT). After the experiment is completed, researchers compare the outcomes between the control group and the experimental group to measure for differences (Chalmers, et al, 1981). The differences in outcomes between the control and experimental group are referred to as Treatment Effects. Treatment effects that are generalized by the researcher to the overall population are referred to as Average Treatment Effects (ATE), while treatment effects that are generalized to individuals are referred to as Average Treatment Effects for the Treated (ATT). In an RCT, these two values can be assumed to be the same because subjects are randomly assigned to each group, which implies that the treated group should not systematically differ from the overall population. By virtue of the randomization process, any bias between control and experimental groups should be eliminated, within the limits of sampling error (Austin, 2011).

For Example, a university mathematics professor wishes to measure whether the performance of students who use an online method for content acquisition will differ significantly from students who receive content in a traditional, lecture-based class. The researcher then selects all 500 students taking a college algebra course and flips a coin to determine whether they will be in the online content acquisition course (the experimental group) or the lecture-based group (the control group). After the first mid-term, the researcher takes the
mean and standard deviation from the exam for both groups and determines, using standard statistical analysis, whether there were significant differences in the performance on the midterm. Any differences between the two groups is referred to as the treatment effect.

**Observational Study**

A study wherein researchers measure the differences in outcomes between a control and experimental group is called an observational study. However, unlike an RCT, researchers have no control over the assignment of subjects to either the experimental or control groups (Porta, 2009).

Clearly, the RCT example provided above has both ethical and practical difficulties. It is considered unethical to deprive a math student of the opportunity to take a course that might teach them more effectively, or to force a student to take a course for which they might not be suited. Furthermore, the scheduling logistics alone required to create enough courses at convenient times for 500 students would be daunting. An observational study would be a far more convenient technique for a researcher to gather information and gain insight into how students perform in traditionally taught courses versus a course in which content is acquired online. The researcher could simply follow the performance of the 500 students, some of which would have self-selected to participate in the traditional course and some in the online course, and then measure the differences in their performances on their first midterms. Notice that, in the observational study, the researcher has no control over whether the subject is enlisted as a member of the control or the experimental group. This the key difference between an RCT and an observational study. More importantly, observational studies can’t be used in the same manner as RCT’s to generalize conclusions to a population because of this lack of control over assignment.
Covariates

A covariate is any variable that is possibly predictive of the outcome being studied (Gujarati and Porter, 2009). In the above observational study, suppose that the experimental group (online content acquisition) performed significantly lower on the first midterm than the control group. However, when the researcher looked at the subjects, it was discovered that the experimental group was considerably younger than the control group. Younger college students perform historically poorer in college algebra. In this case, the age of the students in each group could be a predictor of their performance on the midterm, not the treatment (method of instruction) given. This phenomenon is referred to as confounding, since the researcher doesn’t know whether the treatment or the age of the student was the reason for the difference in performance. The age of the students, in this case, would be referred to as a covariate.

Counterfactuals

In statistical parlance, a counterfactual is a thought experiment in which a researcher hypothesizes how a subject would have performed had they not received the treatment they received (Lewis, 1973). In the above observational study, the counterfactual of interest for the researcher would be, “how would a particular student, who self-selected into the experimental group, have performed on the mid-term had they instead been in the control group?” The counterfactual question of how a student, who self-selected into the control group, would have performed had they instead been in the experimental group would be identical.

Propensity Score

A propensity score is the conditional probability that a person will be in one group (e.g. experimental or control), given a specific set of observable covariates, sometimes referred to as the covariate vector, or \( e(x) = p(Z = 1 | X) \) (Rosenbaum and Rubin, 1983). In this equation, \( Z = \)
1 represents the subject being assigned to the treatment group, and \( \mathbf{X} \) is the vector representing the set of covariates for a given subject. The propensity score balances participants in each group. What this means is subjects in each group with similar covariate values will have similar propensity scores. The propensity score can be derived in a number of ways, however, for convenience and ease of use, the logistics regression equation will be employed throughout this paper.
Appendix C

SPPS Procedure

Once the data is ready for analysis, and imported into SPSS the researcher must use the correct commands to execute the appropriate PSM technique. The following is a series of SPSS screen captures, with explanation, used in initiating the PSM procedure:

1. In SPSS, under the “Analyze” menu, click on “PS Matching”

2. This Propensity Score Matching dialogue box will pop up:
3. The researcher must input an ID variable, the independent variable, and the covariates. For the purposes of this paper, “Additional covariates” will not be used, and isn’t typically used for simple PSM.

Notice the default setting, after inputting variables, is “Matching with Replacement” = “False”. This means that once a control subject is matched to a treatment subject, that control subject is removed from eligibility to be matched with any other treatment subject. “Caliper Definition” = “No Caliper” is by default. A researcher may set the caliper to a certain distance, which implies that treatment units can only
be matched to control units if the control unit is within that distance of the treatment unit. The consequence of caliper matching is that some treatment subjects may not be matched to any control units. The last default is the “Match one to many” radio button, which is set to “Match 1: 1”. If a researcher wishes to match more than one control subject to each treatment subject, the “Match 1: many” button must be turned on and the researcher gets to choose the ratio. Caliendo and Kopeinig (2008) warn that matching treatment subjects with multiple control subjects increases bias in the results. Conversely, using calipers reduces bias, as does matching with replacement. This reduction in bias doesn’t come without a price. Any time there is a reduction of bias through a certain procedure, there is a subsequent increase in variance in the variables, which can reduce the researcher’s ability to accurately predict treatment effect (Caliendo and Kopeinig, 2008). It is the onus of the researcher to determine the trade-offs when choosing a matching technique.

4. The researcher must choose the plots and outputs that will be displayed in the SPSS output.
The authors believe the more information a researcher has, the better, which is why all of the plots and the “Detailed balance” option are chosen.

5. The researcher is now ready to press the “Ok” button in the “Propensity Score Matching” dialogue box and interpret the output.

**Interpreting the Outputs**

In the following section, SPSS outputs for the samples in which the unmatched treatment: control ratio was 1: 1, and 1: 5 and for the sake of simplicity, matching was done without replacement and without calipers. These outputs will be discussed and a determination of which of the models would be optimal in determining treatment effects will be reached. The entire output will be examined to inform the reader as to what each value implies for the researcher. The end of the section will explore outputs in data when default matching settings aren’t used.

1. 1: 1

   **Warning**

   No warnings in estimation or matching procedure

   If there are missing data, or any other problems with the data, this window will inform the researcher. Proper data cleaning is crucial to PSM working properly with any statistical software, including SPSS
This window provides basic diagnostics about how many of the control and treatment group were selected. Researchers should use this window to check for missing treatment data and to insure that the parameter inputs in SPSS were correct.

<table>
<thead>
<tr>
<th>Sample Sizes</th>
<th>Control</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2315</td>
<td>55</td>
</tr>
<tr>
<td>Matched</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>Unmatched</td>
<td>2260</td>
<td>0</td>
</tr>
<tr>
<td>Discarded</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

This test for overall imbalance was developed by Hansen and Bowers (2008) and is akin to the well-known Hotelling’s $T^2$, which simultaneously determines if any covariates of linear combination of covariates are significantly imbalanced after matching (Thoemmes, 2012). Significant p-values below .05 imply serious imbalances in the covariates and the researcher must use analytical techniques to address such imbalances. This test is only available for analysis when matching is 1:1.

<table>
<thead>
<tr>
<th>Overall balance test (Hansen &amp; Bowers, 2010)</th>
<th>chisquare</th>
<th>df</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>2.176</td>
<td>9.000</td>
<td>.988</td>
</tr>
</tbody>
</table>

Yet another omnibus test for overall balance of covariates, $L_1$, was developed by Iacus, King and Porro (2009) uses a technique referred to as automatic coarsening and binning to determine imbalances between variables. If $L_1 = 0$, then the subjects are perfectly matched on all covariates and if $L_1 = 1$ the variables are perfectly separated (Thoemmes, 2012). Further reading on the topic,
including specific algorithms and mathematics, is available for the interested researcher (Iacus, King and Porro, 2011). While there is no cutoff for whether covariates are balanced, researchers should be looking for a reduction of $L_1$ from before matching to after matching. Researchers will use this test in combination with the other information presented in the output to determine if the covariates are adequately balanced.

<p>| Summary of unbalanced covariates (|d| &gt; .25) |
|----------------------------------|
|                                   |</p>
<table>
<thead>
<tr>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Std. Mean Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenderRecodedxLEPRecoded</td>
<td>.055</td>
<td>.145</td>
<td>.356</td>
</tr>
<tr>
<td>GenderRecodedxIEPRecoded</td>
<td>.036</td>
<td>.091</td>
<td>.290</td>
</tr>
</tbody>
</table>

This window displays whether any of the individual covariates are imbalanced after matching. Stuart and Rubin (2007) recommend $|d| < .25$, where $d$ is Cohen’s $d$, however more conservative benchmarks of $|d| < .1$ have become standard (Love, 2008, Shadish et al., 2008). In this case, only the interactive terms GenderRecodedXLEPRecoded and GenderRecodedXIEPRecoded are displaying imbalances and because interactive terms weren’t used to create the logit model, the researcher needn’t address any imbalance issues. That said, because GenderRecoded is imbalanced as an interactive term with two separate covariates, the researcher might consider addressing the issue using traditional means of combining GenderRecoded with another variable or removing it as a covariate in the model all together. Before-matching values of Cohen’s $d$ shouldn’t exceed 1, and SPSS doesn’t provide any such warnings in the event that this occurs. The following tables provide researchers the ability to check before- and after-matching imbalances for covariates by way of Cohen’s $d$. 
<table>
<thead>
<tr>
<th></th>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Std. Mean Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>propensity</td>
<td>.035</td>
<td>.023</td>
<td>.016</td>
<td>.510</td>
</tr>
<tr>
<td>RaceRecoded</td>
<td>4.236</td>
<td>4.498</td>
<td>1.372</td>
<td>-.208</td>
</tr>
<tr>
<td>GenderRecoded</td>
<td>.527</td>
<td>.482</td>
<td>.500</td>
<td>.091</td>
</tr>
<tr>
<td>iEPRecoded</td>
<td>.145</td>
<td>.156</td>
<td>.363</td>
<td>-.031</td>
</tr>
<tr>
<td>LEPRecoded</td>
<td>.255</td>
<td>.252</td>
<td>.434</td>
<td>.005</td>
</tr>
<tr>
<td>FRLRecoded</td>
<td>.873</td>
<td>.800</td>
<td>.400</td>
<td>.218</td>
</tr>
<tr>
<td>CRTMath1yearprior</td>
<td>332.545</td>
<td>320.270</td>
<td>82.932</td>
<td>.174</td>
</tr>
<tr>
<td>CRTReadingYOS</td>
<td>336.309</td>
<td>309.363</td>
<td>76.080</td>
<td>.359</td>
</tr>
<tr>
<td>MathGradeRecoded</td>
<td>2.636</td>
<td>2.754</td>
<td>.917</td>
<td>-.136</td>
</tr>
<tr>
<td>CRTMath2yearsprior</td>
<td>319.182</td>
<td>308.559</td>
<td>44.997</td>
<td>.252</td>
</tr>
<tr>
<td>propensityxpropensity</td>
<td>.002</td>
<td>.001</td>
<td>.001</td>
<td>.358</td>
</tr>
<tr>
<td>propensityxRaceRecoded</td>
<td>.140</td>
<td>.097</td>
<td>.068</td>
<td>.458</td>
</tr>
<tr>
<td>propensityxGenderRecoded</td>
<td>.021</td>
<td>.012</td>
<td>.017</td>
<td>.329</td>
</tr>
<tr>
<td>propensityxIEPRecoded</td>
<td>.006</td>
<td>.003</td>
<td>.011</td>
<td>.133</td>
</tr>
<tr>
<td>propensityxLEPRecoded</td>
<td>.008</td>
<td>.006</td>
<td>.013</td>
<td>.105</td>
</tr>
<tr>
<td>propensityxFRLRecoded</td>
<td>.032</td>
<td>.020</td>
<td>.018</td>
<td>.472</td>
</tr>
<tr>
<td>propensityxCRTMath1yearprior</td>
<td>11.760</td>
<td>7.621</td>
<td>6.168</td>
<td>.496</td>
</tr>
<tr>
<td>propensityxCRTReadingYOS</td>
<td>12.591</td>
<td>7.691</td>
<td>6.753</td>
<td>.469</td>
</tr>
<tr>
<td>propensityxMathGradeRecoded</td>
<td>.091</td>
<td>.060</td>
<td>.047</td>
<td>.459</td>
</tr>
<tr>
<td>propensityxCRTMath2yearsprior</td>
<td>11.371</td>
<td>7.313</td>
<td>5.724</td>
<td>.496</td>
</tr>
<tr>
<td>RaceRecodedxRaceRecoded</td>
<td>19.509</td>
<td>22.117</td>
<td>13.719</td>
<td>-.220</td>
</tr>
<tr>
<td>RaceRecodedxGenderRecoded</td>
<td>2.200</td>
<td>2.178</td>
<td>2.445</td>
<td>.009</td>
</tr>
<tr>
<td>RaceRecodedxIEPRecoded</td>
<td>.527</td>
<td>.714</td>
<td>1.752</td>
<td>-.139</td>
</tr>
<tr>
<td>RaceRecodedxLEPRecoded</td>
<td>1.000</td>
<td>1.008</td>
<td>1.743</td>
<td>-.004</td>
</tr>
<tr>
<td>RaceRecodedxFRLRecoded</td>
<td>3.764</td>
<td>3.537</td>
<td>2.126</td>
<td>.123</td>
</tr>
<tr>
<td>RaceRecodedxCRTMath1yearprior</td>
<td>1417.927</td>
<td>1446.226</td>
<td>610.061</td>
<td>-.048</td>
</tr>
<tr>
<td></td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>RaceRecodedxCRTReading YOS</td>
<td>1433.782</td>
<td>1399.009</td>
<td>584.102</td>
<td>0.059</td>
</tr>
<tr>
<td>RaceRecodedxMathGradeRecoded</td>
<td>11.218</td>
<td>12.456</td>
<td>5.887</td>
<td>-0.220</td>
</tr>
<tr>
<td>RaceRecodedxCRTMath2yearsprior</td>
<td>1356.618</td>
<td>1391.108</td>
<td>484.746</td>
<td>-0.071</td>
</tr>
<tr>
<td>GenderRecodedxGenderRecoded</td>
<td>0.527</td>
<td>0.482</td>
<td>0.500</td>
<td>0.091</td>
</tr>
<tr>
<td>GenderRecodedxEIPRecoded</td>
<td>0.036</td>
<td>0.059</td>
<td>0.236</td>
<td>-0.121</td>
</tr>
<tr>
<td>GenderRecodedxLEPRecoded</td>
<td>0.055</td>
<td>0.107</td>
<td>0.309</td>
<td>-0.229</td>
</tr>
<tr>
<td>GenderRecodedxFRLRecoded</td>
<td>0.491</td>
<td>0.384</td>
<td>0.486</td>
<td>0.212</td>
</tr>
<tr>
<td>GenderRecodedxCRTMath1yearprior</td>
<td>183.145</td>
<td>156.133</td>
<td>171.637</td>
<td>0.149</td>
</tr>
<tr>
<td>GenderRecodedxCRTReadingYOS</td>
<td>187.382</td>
<td>154.655</td>
<td>167.919</td>
<td>0.176</td>
</tr>
<tr>
<td>GenderRecodedxMathGradeRecoded</td>
<td>1.509</td>
<td>1.375</td>
<td>1.560</td>
<td>0.085</td>
</tr>
<tr>
<td>GenderRecodedxCRTMath2yearsprior</td>
<td>173.400</td>
<td>149.232</td>
<td>157.784</td>
<td>0.144</td>
</tr>
<tr>
<td>IEPRecodedxEIPRecoded</td>
<td>0.145</td>
<td>0.156</td>
<td>0.363</td>
<td>-0.031</td>
</tr>
<tr>
<td>IEPRecodedxLEPRecoded</td>
<td>0.127</td>
<td>0.087</td>
<td>0.282</td>
<td>0.120</td>
</tr>
<tr>
<td>IEPRecodedxFRLRecoded</td>
<td>0.145</td>
<td>0.134</td>
<td>0.341</td>
<td>0.031</td>
</tr>
<tr>
<td>IEPRecodedxCRTMath1yearprior</td>
<td>37.964</td>
<td>36.136</td>
<td>88.836</td>
<td>0.198</td>
</tr>
<tr>
<td>IEPRecodedxCRTReadingYOS</td>
<td>36.709</td>
<td>33.359</td>
<td>82.051</td>
<td>0.035</td>
</tr>
<tr>
<td>IEPRecodedxMathGradeRecoded</td>
<td>0.327</td>
<td>0.331</td>
<td>0.841</td>
<td>-0.004</td>
</tr>
<tr>
<td>IEPRecodedxCRTMath2yearsprior</td>
<td>40.127</td>
<td>40.739</td>
<td>96.317</td>
<td>-0.006</td>
</tr>
<tr>
<td>LEPRecodedxLEPRecoded</td>
<td>0.255</td>
<td>0.252</td>
<td>0.434</td>
<td>0.005</td>
</tr>
<tr>
<td>LEPRecodedxFRLRecoded</td>
<td>0.236</td>
<td>0.217</td>
<td>0.412</td>
<td>0.046</td>
</tr>
<tr>
<td>LEPRecodedxCRTMath1yearsprior</td>
<td>68.764</td>
<td>66.475</td>
<td>120.255</td>
<td>0.019</td>
</tr>
<tr>
<td>LEPRecodedxCRTReadingYOS</td>
<td>65.982</td>
<td>62.439</td>
<td>112.379</td>
<td>0.030</td>
</tr>
<tr>
<td>Term</td>
<td>Estimate</td>
<td>Std. Error</td>
<td>t value</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>----------</td>
<td>------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>LEPRecodedxMathGradeRecoded</td>
<td>.582</td>
<td>.574</td>
<td>1.076</td>
<td>.007</td>
</tr>
<tr>
<td>LEPRecodedxCRTMath2yearprior</td>
<td>73.109</td>
<td>70.838</td>
<td>123.581</td>
<td>.018</td>
</tr>
<tr>
<td>FRLRecodedxFRLRecoded</td>
<td>.873</td>
<td>.800</td>
<td>.400</td>
<td>.218</td>
</tr>
<tr>
<td>FRLRecodedxCRTMath1yearprior</td>
<td>285.818</td>
<td>254.287</td>
<td>147.218</td>
<td>.245</td>
</tr>
<tr>
<td>FRLRecodedxCRTReadingYOS</td>
<td>288.945</td>
<td>244.909</td>
<td>140.177</td>
<td>.335</td>
</tr>
<tr>
<td>FRLRecodedxMathGradeRecoded</td>
<td>2.236</td>
<td>2.188</td>
<td>1.367</td>
<td>.041</td>
</tr>
<tr>
<td>FRLRecodedxCRTMath2yearprior</td>
<td>275.418</td>
<td>245.340</td>
<td>129.378</td>
<td>.265</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTMath1yearprior</td>
<td>115495.564</td>
<td>109447.559</td>
<td>51737.018</td>
<td>.129</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTReadingYOS</td>
<td>115130.200</td>
<td>103406.117</td>
<td>44995.537</td>
<td>.280</td>
</tr>
<tr>
<td>CRTMath1yearpriorxMathGradeRecoded</td>
<td>906.109</td>
<td>926.230</td>
<td>462.259</td>
<td>-.050</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTMath2yearsprior</td>
<td>107999.218</td>
<td>101697.013</td>
<td>37128.905</td>
<td>.194</td>
</tr>
<tr>
<td>CRTReadingYOSxCRTReadingYOS</td>
<td>118648.127</td>
<td>101491.362</td>
<td>45429.909</td>
<td>.357</td>
</tr>
<tr>
<td>CRTReadingYOSxMathGradeRecoded</td>
<td>921.745</td>
<td>888.550</td>
<td>428.446</td>
<td>.078</td>
</tr>
<tr>
<td>CRTReadingYOSxCRTMath2yearsprior</td>
<td>109351.236</td>
<td>97660.062</td>
<td>33572.442</td>
<td>.349</td>
</tr>
<tr>
<td>MathGradeRecodedxMathGradeRecoded</td>
<td>7.691</td>
<td>8.427</td>
<td>4.773</td>
<td>-.160</td>
</tr>
<tr>
<td>MathGradeRecodedxCRTMath2yearsprior</td>
<td>865.073</td>
<td>872.482</td>
<td>364.523</td>
<td>-.021</td>
</tr>
<tr>
<td>CRTMath2yearspriorxCRTMath2yearsprior</td>
<td>103625.582</td>
<td>97232.228</td>
<td>27226.220</td>
<td>.235</td>
</tr>
</tbody>
</table>
## Detailed balance after matching

<table>
<thead>
<tr>
<th></th>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Std. Mean Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>propensity</td>
<td>.035</td>
<td>.035</td>
<td>.023</td>
<td>-.001</td>
</tr>
<tr>
<td>RaceRecoded</td>
<td>4.236</td>
<td>4.145</td>
<td>1.297</td>
<td>.072</td>
</tr>
<tr>
<td>GenderRecoded</td>
<td>.527</td>
<td>.527</td>
<td>.504</td>
<td>.000</td>
</tr>
<tr>
<td>IEPRecoded</td>
<td>.145</td>
<td>.200</td>
<td>.404</td>
<td>-.153</td>
</tr>
<tr>
<td>LEPRecoded</td>
<td>.255</td>
<td>.273</td>
<td>.449</td>
<td>-.041</td>
</tr>
<tr>
<td>FRLRecoded</td>
<td>.873</td>
<td>.836</td>
<td>.373</td>
<td>.108</td>
</tr>
<tr>
<td>CRTMath1yearprior</td>
<td>332.545</td>
<td>331.200</td>
<td>83.014</td>
<td>.019</td>
</tr>
<tr>
<td>CRTReadingYOS</td>
<td>336.309</td>
<td>337.873</td>
<td>65.491</td>
<td>-.021</td>
</tr>
<tr>
<td>MathGradeRecoded</td>
<td>2.636</td>
<td>2.773</td>
<td>.942</td>
<td>-.157</td>
</tr>
<tr>
<td>CRTMath2years prior</td>
<td>319.182</td>
<td>322.418</td>
<td>45.454</td>
<td>-.077</td>
</tr>
<tr>
<td>propensityxpropensity</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>-.003</td>
</tr>
<tr>
<td>propensityxRaceRecoded</td>
<td>.140</td>
<td>.143</td>
<td>.097</td>
<td>-.032</td>
</tr>
<tr>
<td>propensityxGenderRecoded</td>
<td>.021</td>
<td>.018</td>
<td>.024</td>
<td>.120</td>
</tr>
<tr>
<td>propensityxIEPRecoded</td>
<td>.006</td>
<td>.010</td>
<td>.026</td>
<td>-.234</td>
</tr>
<tr>
<td>propensityxLEPRecoded</td>
<td>.008</td>
<td>.009</td>
<td>.021</td>
<td>-.086</td>
</tr>
<tr>
<td>propensityxFRLRecoded</td>
<td>.032</td>
<td>.030</td>
<td>.026</td>
<td>.066</td>
</tr>
<tr>
<td>propensityxCRTMath1year prior</td>
<td>11.760</td>
<td>11.763</td>
<td>8.164</td>
<td>.000</td>
</tr>
<tr>
<td>propensityxCRTReadingYOS</td>
<td>12.591</td>
<td>12.334</td>
<td>9.262</td>
<td>.025</td>
</tr>
<tr>
<td>propensityxMathGradeRecoded</td>
<td>.091</td>
<td>.090</td>
<td>.061</td>
<td>.016</td>
</tr>
<tr>
<td>propensityxCRTMath2years prior</td>
<td>11.371</td>
<td>11.494</td>
<td>8.033</td>
<td>-.015</td>
</tr>
<tr>
<td>RaceRecodedxGenderRecoded</td>
<td>2.200</td>
<td>2.200</td>
<td>2.313</td>
<td>.000</td>
</tr>
<tr>
<td>RaceRecodedxIEPRecoded</td>
<td>5.27</td>
<td>.818</td>
<td>1.816</td>
<td>-.216</td>
</tr>
<tr>
<td>RaceRecodedxLEPRecoded</td>
<td>1.000</td>
<td>1.145</td>
<td>1.929</td>
<td>-.084</td>
</tr>
<tr>
<td>RaceRecodedxFRLRecoded</td>
<td>3.764</td>
<td>3.582</td>
<td>1.997</td>
<td>.099</td>
</tr>
<tr>
<td>RaceRecodedxCRTMath1year prior</td>
<td>1417.927</td>
<td>1389.909</td>
<td>600.599</td>
<td>.047</td>
</tr>
<tr>
<td>RaceRecodedxCRTReadingYOS</td>
<td>1433.782</td>
<td>1417.764</td>
<td>535.888</td>
<td>.027</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>RaceRecodedxMathGradeRecoded</td>
<td>11.218</td>
<td>11.564</td>
<td>5.719</td>
<td>-.061</td>
</tr>
<tr>
<td>RaceRecodedxCRTMath2years prior</td>
<td>1356.618</td>
<td>1345.382</td>
<td>469.158</td>
<td>.023</td>
</tr>
<tr>
<td>GenderRecodedxGenderRecoded</td>
<td>.527</td>
<td>.527</td>
<td>.504</td>
<td>.000</td>
</tr>
<tr>
<td>GenderRecodedxIEPRecoded</td>
<td>.036</td>
<td>.091</td>
<td>.290</td>
<td>-.289</td>
</tr>
<tr>
<td>GenderRecodedxLEPRecoded</td>
<td>.055</td>
<td>.145</td>
<td>.356</td>
<td>-.397</td>
</tr>
<tr>
<td>GenderRecodedxFRLRecoded</td>
<td>.491</td>
<td>.436</td>
<td>.501</td>
<td>.108</td>
</tr>
<tr>
<td>GenderRecodedxCRTMath1year prior</td>
<td>183.145</td>
<td>165.855</td>
<td>172.254</td>
<td>.095</td>
</tr>
<tr>
<td>GenderRecodedxCRTReadingYOS</td>
<td>187.382</td>
<td>177.727</td>
<td>176.613</td>
<td>.052</td>
</tr>
<tr>
<td>GenderRecodedxMathGraderecoded</td>
<td>1.509</td>
<td>1.573</td>
<td>1.654</td>
<td>-.040</td>
</tr>
<tr>
<td>GenderRecodedxCRTMath2years prior</td>
<td>173.400</td>
<td>170.545</td>
<td>167.134</td>
<td>.017</td>
</tr>
<tr>
<td>IEPRecodedxIEPRecoded</td>
<td>.145</td>
<td>.200</td>
<td>.404</td>
<td>-.153</td>
</tr>
<tr>
<td>IEPRecodedxLEPRecoded</td>
<td>.127</td>
<td>.109</td>
<td>.315</td>
<td>.054</td>
</tr>
<tr>
<td>IEPRecodedxFRLRecoded</td>
<td>.145</td>
<td>.182</td>
<td>.389</td>
<td>-.102</td>
</tr>
<tr>
<td>IEPRecodedxCRTMath1year prior</td>
<td>37.964</td>
<td>54.164</td>
<td>115.919</td>
<td>-.171</td>
</tr>
<tr>
<td>IEPRecodedxCRTReadingYOS</td>
<td>36.709</td>
<td>55.600</td>
<td>118.243</td>
<td>-.199</td>
</tr>
<tr>
<td>IEPRecodedxMathGradeRecoded</td>
<td>.327</td>
<td>.445</td>
<td>.984</td>
<td>-.137</td>
</tr>
<tr>
<td>IEPRecodedxCRTMath2years prior</td>
<td>40.127</td>
<td>57.818</td>
<td>118.934</td>
<td>-.178</td>
</tr>
<tr>
<td>LEPRecodedxLEPRecoded</td>
<td>.255</td>
<td>.273</td>
<td>.449</td>
<td>-.041</td>
</tr>
<tr>
<td>LEPRecodedxFRLRecoded</td>
<td>.236</td>
<td>.255</td>
<td>.440</td>
<td>-.042</td>
</tr>
<tr>
<td>LEPRecodedxCRTMath1year prior</td>
<td>68.764</td>
<td>80.618</td>
<td>138.084</td>
<td>-.098</td>
</tr>
<tr>
<td>LEPRecodedxCRTReadingYOS</td>
<td>65.982</td>
<td>77.927</td>
<td>129.859</td>
<td>-.100</td>
</tr>
<tr>
<td>Covariate</td>
<td>Before PSM</td>
<td>After PSM</td>
<td>Standard Mean Difference (Cohen’s d)</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------------------------------</td>
<td></td>
</tr>
<tr>
<td>LEPRecodedxMathGradeRecoed</td>
<td>.582</td>
<td>.682</td>
<td>1.203</td>
<td>-.094</td>
</tr>
<tr>
<td>LEPRecodedxCRTMath2yearprior</td>
<td>73.109</td>
<td>80.618</td>
<td>133.362</td>
<td>-.059</td>
</tr>
<tr>
<td>FRLRecodedxFRLRecoded</td>
<td>.873</td>
<td>.836</td>
<td>.373</td>
<td>.108</td>
</tr>
<tr>
<td>FRLRecodedxCRTMath1yearprior</td>
<td>285.818</td>
<td>273.818</td>
<td>145.739</td>
<td>.093</td>
</tr>
<tr>
<td>FRLRecodedxCRTReading</td>
<td>288.945</td>
<td>279.473</td>
<td>139.472</td>
<td>.072</td>
</tr>
<tr>
<td>FRLRecodedxMathGradeRecoed</td>
<td>2.236</td>
<td>2.300</td>
<td>1.332</td>
<td>-.054</td>
</tr>
<tr>
<td>FRLRecodedxCRTMath2yearprior</td>
<td>275.418</td>
<td>265.255</td>
<td>125.560</td>
<td>.090</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTMath1yearprior</td>
<td>115495.564</td>
<td>116459.527</td>
<td>53235.443</td>
<td>-.020</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTReading</td>
<td>115130.200</td>
<td>115654.018</td>
<td>44484.368</td>
<td>-.013</td>
</tr>
<tr>
<td>CRTMath1yearpriorxMathGradeRecoed</td>
<td>906.109</td>
<td>959.464</td>
<td>485.142</td>
<td>-.132</td>
</tr>
<tr>
<td>CRTMath1yearpriorxCRTMath2years</td>
<td>107999.218</td>
<td>109769.036</td>
<td>39291.707</td>
<td>-.054</td>
</tr>
<tr>
<td>CRTReadingYOSxCRTReading</td>
<td>118648.127</td>
<td>118369.036</td>
<td>41920.671</td>
<td>.006</td>
</tr>
<tr>
<td>CRTReadingYOSxMathGradeRecoed</td>
<td>921.745</td>
<td>964.927</td>
<td>432.384</td>
<td>-.102</td>
</tr>
<tr>
<td>CRTReadingYOSxCRTMath2years</td>
<td>109351.236</td>
<td>111115.345</td>
<td>32862.936</td>
<td>-.053</td>
</tr>
<tr>
<td>MathGradeRecoedxMathGradeRecoed</td>
<td>7.691</td>
<td>8.559</td>
<td>5.085</td>
<td>-.189</td>
</tr>
<tr>
<td>MathGradeRecoedxCRTMath2years</td>
<td>865.073</td>
<td>915.609</td>
<td>393.608</td>
<td>-.141</td>
</tr>
<tr>
<td>CRTMath2yearsxCRTMath2years</td>
<td>103625.582</td>
<td>105981.945</td>
<td>29306.510</td>
<td>-.086</td>
</tr>
</tbody>
</table>

These two tables provide researchers with univariate diagnostic information for each covariate before and after PSM is used. The right-most column is the most informative because it gives the Standard Mean Difference (Cohen’s $d$) value for each of the covariates and allows
researchers to determine which covariates, if any display unacceptable imbalances. SPSS provides Standard Mean Differences for all quadratic and interaction terms of the covariates, if no imbalances exist in the linear terms, the researcher needn’t look any further. If, however, there are imbalances that exceed the $|d| > .25$ for post-matching and $|d| > 1$ pre-matching, researchers can use the quadratic and interactive values to determine how they want to manipulate the original data to produce a model without these imbalances, which is the reason SPSS provides this information in the table.

**Graphs and Plots**

The first of the graphs SPSS provides is the Jitter Plot, which gives the researcher the opportunity to insure common support, or overlap of distributions of the control and treatment data based on their propensity scores. This graph below clearly demonstrates that matched units have a great amount of distribution overlap, which implies that treated subjects won’t be matched to control subjects who vary greatly per their covariates. This type of distribution is desirable for post-matching statistical analysis.
The graph below is the Line Plot for Individual Differences, and graphically displays the differences between pre- and post-matching changes in covariate balance as measured by $|d|$. It is apparent by this graph that matching subjects greatly reduced any imbalance present in the covariates. The **bold** lines in the graph represent covariates where matching actually *increased* the covariate imbalance. Researchers need to review the detailed balance tables provided above and determine which covariates display an increased $|d|$. In this case, many imbalances for covariates appear to have increased, however, closer inspection reveals that all increases occurred on interactive terms, which weren’t used to create the logit model, or remained well below the .25 value for $|d|$ (Stuart and Rubin, 2007).

Much like the Jitter Plot, the histograms below provide researchers with evidence for or against common support (distribution overlap) before and after matching. In this case, notice that the matched treated and control subjects have very similar distributions and a great deal of overlap, which is highly desirable for post-matching analyses (Caliendo and Kopeinig, 2008).
These histograms graphically represent the distributions of $|d|$ before and after matching. By virtue of forcing both histograms on the same x-axis scale, the researcher can easily view the result of matching on the standardized differences.
This dot plot is yet another way to demonstrate the effects of matching on standardized differences of the covariates. It is also the easiest way for researchers to graphically determine the effects of matching on balance of covariates, and thus assist the researcher in deciding which matching technique is best. Researchers should use this graph together with the tables provided above to corroborate the existence of any imbalanced covariates.
At this point, the researcher needs to determine if the data are sufficiently balanced to move forward with conventional statistical analysis on the variable of interest, in this case the subjects’ performance assessment scores. All of the tables and graphs indicate that analysis can begin, but because of the increase in imbalance for many of the covariates, a researcher may want to increase the pre-matching treatment: control ratio or utilize different options within SPSS. This appendix is meant to provide only a basic introduction to PSM using SPSS, and the authors will simply demonstrate how the outputs differ as the treatment: control ratio changes. It is up to the reader to experiment with different types of matching techniques to determine which technique balances covariates best for their data.
2. 5:1

**Warning**

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2315</td>
<td>55</td>
</tr>
<tr>
<td>Matched</td>
<td>275</td>
<td>55</td>
</tr>
<tr>
<td>Unmatched</td>
<td>2040</td>
<td>0</td>
</tr>
<tr>
<td>Discarded</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Sample Sizes**

Hansen and Bowers (2010) test of global imbalance is currently only implemented for 1:1 matching without replacement.

**Relative multivariate imbalance L1 (Iacus, King, & Porro, 2010)**

<table>
<thead>
<tr>
<th></th>
<th>Before matching</th>
<th>After matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate imbalance measure L1</td>
<td>.997</td>
<td>.971</td>
</tr>
</tbody>
</table>

**Summary of unbalanced covariates (|d| > .25)**

<table>
<thead>
<tr>
<th></th>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Std. Mean Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenderRecodedxLEPRecorded</td>
<td>.055</td>
<td>.131</td>
<td>.338</td>
<td>-.333</td>
</tr>
</tbody>
</table>
Summary of unbalanced covariates (|d| > .25)

No covariate exhibits a large imbalance (|d| > .25).

Notice the initial output tables provide the same information as in the 1:1 example. The change in $L_1$ from before matching to after appears to indicate a greater improvement in balance, and, again, SPSS tells us that there are no individual covariates that display unacceptable imbalance, however GenderRecodedXLEPRecoder is still imbalanced. Overall, the covariates used to determine the logit model are well balanced both before and after matching. Because the “Detailed Balance” tables display a daunting amount of information, and there are no significant imbalances in the data, this section will forego further analysis of the tables, and move straight to the graphs.

![Distribution of Propensity Scores](image)

Again, excellent common support as displayed by the Jitter Plot.
The histograms provide even more evidence for common support.
In this case, we have fifteen covariates and interaction terms, represented by **bold** lines whose balance has worsened as a result of matching. The researcher would need to consult the “Detailed Balance” tables and the final dot plot to determine which of the covariate balance worsened and if that should be considered a problem worth addressing.

![Graph showing covariate balance before and after matching](image)

Notice the dot plot shows that all of the covariate balance improved after matching. This informs a researcher that the covariates that worsened in the line plot above don’t appear to have been part of the model.
Lastly, these graphs reinforce the conclusions that matching resulted in improved balance after matching.
Appendix D

Thematic Summary of Review of the Literature

<table>
<thead>
<tr>
<th>Theme</th>
<th>Connectivity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>students</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>solving</td>
<td>83%</td>
<td></td>
</tr>
<tr>
<td>classroom</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>knowledge</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>studies</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>37%</td>
<td></td>
</tr>
<tr>
<td>questions</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td>data</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>online</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>transfer</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>mathematical</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>technology</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Learning</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>
Appendix E

Thematic Summary for Teacher Responses for Entire Survey

Thematic Summary

<table>
<thead>
<tr>
<th>Theme</th>
<th>Connectivity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>students</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>51%</td>
<td></td>
</tr>
<tr>
<td>classroom</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td>school</td>
<td>35%</td>
<td></td>
</tr>
<tr>
<td>flipped</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>flapping</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>year</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>using</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>learning</td>
<td>09%</td>
<td></td>
</tr>
<tr>
<td>parents</td>
<td>09%</td>
<td></td>
</tr>
<tr>
<td>day</td>
<td>08%</td>
<td></td>
</tr>
</tbody>
</table>
Appendix F

Thematic Summary for Teacher Responses on Questions Four and Five

**Thematic Summary**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Connectivity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>students</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>classroom</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td>work</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>long</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>upper</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>topic</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>student's</td>
<td>09%</td>
<td></td>
</tr>
<tr>
<td>material</td>
<td>03%</td>
<td></td>
</tr>
</tbody>
</table>
Appendix G

Thematic Summary for Teacher Responses on Questions Eleven Through Fifteen

Thematic Summary

<table>
<thead>
<tr>
<th>Theme</th>
<th>Connectivity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>students</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>students</td>
<td>36%</td>
<td></td>
</tr>
<tr>
<td>work</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>classroom</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>better</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>facilitor</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>drastically</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>traditional</td>
<td>04%</td>
<td></td>
</tr>
<tr>
<td>issues</td>
<td>01%</td>
<td></td>
</tr>
</tbody>
</table>