A SELF-REGULATED LEARNING INTERVENTION FOR DEVELOPMENTAL MATHEMATICS STUDENTS AT A COMMUNITY COLLEGE: EFFECTS OF STUDY JOURNALS ON ACHIEVEMENT AND STUDY HABITS

A Dissertation

by

JENNIFER LYNN TRAVIS

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Vincent P. Schielack, Jr.
Committee Members,	Jane F. Schielack
	Carol Stuessy
	Victor Willson
Head of Department,	Yeping Li

May 2014

Major Subject: Curriculum and Instruction

ABSTRACT

Self-regulated learners continually monitor and adjust the learning process through a recursive loop of forethought, learning enactment, and reflection. The literature review for this study used a systematic approach with defined criteria to evaluate the effects of self-regulated learning interventions for college students. The review revealed a shortage of rigorous achievement-based research in authentic settings.

This study evaluated a study-journaling intervention for developmental mathematics students at a large urban community college. Two weekly study journal worksheets were designed, based on self-regulated learning theory. In each of nine pairs of intact classes, one class was randomly assigned to the treatment (study journal) condition and the other to control.

The mixed methods research design had two strands: a confirmatory strand that evaluated the intervention's effect, and an exploratory strand that sought information about the students' study habits. The statistical analysis had two phases: propensity score matching to strategically trim the groups so they had similar distributions of starting characteristics, and logistic regression to estimate the intervention's effect on binary variables representing course success and final exam success. Departing students were counted among the unsuccessful.

Due to implementation shortcomings, the original sample (117 treatment, 140 control) was replaced by a modified sample (60 treatment, 77 control). Propensity score matching trimmed this sample further (54 treatment, 54 control).

Control students experienced significantly higher course success rates and slightly higher final exam success rates. Treatment students were significantly more likely to leave the class than control students (odds ratio 2.94). However, qualitative data from focus groups and surveys indicated the study journals may have positively affected study habits. Taken together, the qualitative and quantitative results suggest the intervention increased students' awareness of study habit inadequacies and time constraints. This position was supported by qualitative analysis of the study journal entries.

This study shows that study journals have potential to improve achievement. However, caution is advised, as the journals may also influence students to leave the class due to increased awareness of problems. Research recommendations include combining study journals with training, feedback or peer support; and collecting subsequent-semester data and data on students' reasons for departure.

DEDICATION

To my Father who planned every element of this journey;

and to my father, Don Lynn Renfro, who started me on the path that led to this point.

Wish you were here, Dad.

ACKNOWLEDGMENTS

I once thought of a dissertation as an independent research project and of the dissertation writing process as a lonely endeavor. Nothing could be further from the truth. I could not have reached this milestone without the direct contributions of many people. I would like to express my heartfelt appreciation to:

First and foremost, my advisor, Dr. Vince Schielack: For his infinite patience, wise guidance, and constant cheerfulness. For valuing quality over speed, and for not being the least bit attached to any of my pet timelines. For not saying “you’ve done enough” when I hoped he would, but instead suggesting I do something else—something that made the paper more defensible. For pointing out logical flaws in my writing; through his extreme care with the meanings of words and sentences, he picks up on subtle problems most readers would miss. Because of the many hours we spent together reading and discussing my drafts, both the dissertation and my writing skill improved tremendously.

My committee members: For their support, input, and fair-mindedness, and for shaping this project into something I could be proud of. Dr. Janie Schielack: For turning routine emails into little gifts of encouragement. For being up-front about expectations and concerns with the article format; through my efforts to forestall those concerns, I reaped the benefits of both formats. Dr. Carol Stuessy: For teaching me to see the literature review as something valuable in itself, rather than as an obstacle to be overcome before the real work begins. Dr. Victor Willson: For teaching me to be

ever-alert for threats to validity, and to be constantly thinking not just about what inferences to make, but also about how I can make those inferences more credible.

My mentors and teammates with the POWER writing support service: For believing in me and bearing with me, even when I was scared, exhausted, and self-absorbed. Dr. Pat Goodson: For helping me fall in love with writing. Through her example and guidance, I have come to see writing as a craft I will hone for the rest of my life. Finishing this dissertation is not the end of my writing journey, but rather the beginning. Dr. Dominique Chlup: For showing me the importance of managing my fear, anxiety, and motivation. Without the accountability system she helped me establish, I might still be analyzing my data. Dr. Maggie Huerta: For reading my daily time use reports. Every day without fail, for over six months, she responded with precious words of encouragement. My other POWER friends: For writing with me, encouraging me, providing feedback, and sharing their own writing so we all could learn.

All those who made this study a reality: the administration of Lone Star College-North Harris, for expressing interest in the project and allowing it to move forward; my mathematics colleagues who were willing to implement the study journals in their classes; and especially, the students who volunteered to participate.

The administration of the Lone Star College System, and of Lone Star College-North Harris: For the sabbatical semester that allowed me to concentrate single-mindedly on writing and research. By immersing myself in the process, I learned and grew in ways I could not have done while dividing my time between writing and teaching. The time and knowledge I gained through the sabbatical greatly improved the quality of the final product.

My Lone Star College-North Harris mathematics family: For their constant support, encouragement, and forgiveness. Every day I walk onto campus, I give thanks for the privilege of working with such a fantastic group of people. Their level of care for students is unparalleled. Their support for one another, both professionally and personally, is also unparalleled. Especially these last two years, they made my teaching schedule and committee work as manageable as possible, so I could finish this dissertation. They have frequently picked up the slack for me, though they are too generous to point it out.

My extended (non-math) family at Lone Star College-North Harris: For discussing my ideas, reading my drafts, and celebrating my milestones. Special thanks to Seth Batiste, Joyce Boatright, Cher Brock, Lashun Griffin, and Bob Lynch, for respecting my writing efforts and believing I can make a contribution to the writing process of others.

My mom, Rebecca Renfro: For devoting hours upon hours to helping me with my dissertation. Many people have provided feedback on small chunks of my dissertation; my mom is the only person outside of my committee who has slogged through the entire thing. Her feedback really tightened up the organization of the paper. She also formatted the tables, figures, and references, and combined my scattered chapters into a single document with a consistent set of headings. By taking the formatting out of my hands, she freed up my time to polish the writing.

My husband, Scott Travis: For sticking with me through the entirety of a long process, and for being a calm voice of reason in a sea of perfectionism-driven fears. For recognizing the benefit of an apartment in College Station, so I could meet regularly with my advisor and write without distractions. He carried all the responsibilities of home without complaint, allowing me to focus on the dissertation. Words cannot express my appreciation for his steadfast love and encouragement.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

Page

Page

CHAPTER I

INTRODUCTION

Background and Setting

In 2012, over 6.7 million students, or about 33% of all postsecondary students, were enrolled in public two-year institutions, generally known as community colleges (National Student Clearinghouse Research Center, 2012). Nationally, the proportion of undergraduates enrolled in community colleges has remained relatively stable over the last two decades, at about 40% of all undergraduates and about 25% of full-time undergraduates (Baum, Little, & Payea, 2011). However, in Texas, the proportion of students attending two-year colleges has been steadily increasing; in 2013, about 53% of all Texas postsecondary students attended two-year colleges (Texas Higher Education Coordinating Board, 2013, p. 15).

Although community colleges have undoubtedly expanded access to higher education, access does not guarantee success. Of the 2003–4 cohort of first-time community college enrollees, 61% expressed an intention to transfer to a four-year college, and 50% expressed an intention to earn an associate's degree. However, three years later, only 16% had completed a degree or certificate and only 50% were still enrolled in college, either at the same community college or at another institution; forty-five percent had left college altogether without earning a degree or certificate (Provasnik & Planty, 2008, pp. 22–24). In Texas, of the 2006 cohort of first-time degree-seeking community college enrollees, only 28% had earned an associate's degree, certificate, or bachelor’s degree within six years, and 60% of the cohort had left college without earning a degree or certificate. Only 36% of the 2002 cohort earned a degree or certificate within 10 years. Of the students starting at a community college and transferring to a four-year college, only 54% graduate within four years of transferring (Texas Higher Education Coordinating Board, 2013, pp. 10–12).

Because of the low cost, convenience, and open access of community colleges, the student bodies of community colleges are very different from those of most four-year universities. On average, community college students are older than students at four-year institutions, are more likely to have children, work more, and are more likely to attend part-time. Only about 40% of community college students fit the profile of the “traditional” college student in that they are under 24 years of age and still dependent on their parents (Provasnik & Planty, 2008, p. 12). While the median age for students in four-year institutions is 21, the median age for community college students is 24, and about 35% are 30 or older. About 35% of community college students have children, and nearly half of these are single parents (Horn, Nevill, & Griffith, 2006, p. 10). About 32% of community college students work full-time, compared to about 16% of students at four-year colleges (Horn et al., 2006, p. 13). In 2012, about 60% of community college students were enrolled part-time, compared with only 27% of students at public four-year colleges and universities (National Student Clearinghouse Research Center, 2012).

Due to academic deficiencies, about 43% of community college students begin in developmental mathematics or English courses (Horn et al., 2006, p. 137). In Texas in 2011, about 54% of first-time two-year college students were classified as "not college ready." Of the 2008 cohort of developmental students entering Texas community colleges, 82% enrolled in developmental mathematics courses. Three years later, only about 16% of these had successfully completed a college-level mathematics course (Texas Higher Education Coordinating Board, 2013, p. 13). Improving developmental course success is a key issue in the research agendas of several well-funded think tanks, including the Community College Research Center at Columbia University's Teachers College, the Bill and Melinda Gates Foundation, and the Lumina Foundation.

Impetus for the Current Study

This project took root in my own experience as a community college mathematics instructor. During my years teaching mathematics, particularly developmental mathematics, I have become convinced that students' study habits have more to do with their mathematics success, or lack thereof, than their mathematical talent or starting skill level. Some students with weak mathematical skills overcome those deficits through persistent effort; other students do not, either because they do not exert effort or because their efforts do not bear fruit. A similar dichotomy exists among students who have good starting skills and who grasp mathematics concepts with ease. Some of these students thrive in their mathematics courses, while others flounder, often because they do not supplement their mathematical talent with effective practice.

Each semester, I ask my students to submit a short introductory essay describing their academic and career goals, their objectives for the mathematics class, and their strategies for reaching those objectives. Most students write about their desire to excel, their intention to attend class every day and ask for help when necessary, and their plans to work every homework problem whether it is graded or not. Though nearly all students begin the class with a vision of studying regularly and effectively, for many of them, that vision does not become reality. Like many educators, I have lamented my inability to motivate my students to invest time in consistent and deliberate mathematics practice, and to seek help from me or the mathematics tutoring center when they are confused.

Often I have wondered whether it was possible to nudge students into a more intentional approach to their mathematics studying, one in which they take more responsibility for their own learning. My interest in helping students become proactive about their learning led me to read extensively about self-regulated learning theory, review existing research on self-regulated learning interventions, and design a study-journaling intervention based on self-regulated learning theory.

Overview of Self-Regulated Learning

The literature contains a variety of self-regulated learning models, based on different theoretical perspectives. Instead of focusing on the distinctions between different theoretical models or on the details of a particular model, this dissertation focuses on the common elements of self-regulated learning models. Self-regulated learning models address learners' thoughts and behaviors before the learning activity (forethought phase), during the learning activity (learning enactment phase), and after the learning activity (reflection phase). In a recursive loop, self-regulated learners incorporate internal and external feedback into their planning for later tasks. Instead of regarding learners as passive recipients of learning activities provided by their teacher, self-regulated learning theory assumes learners are active participants who can exert control over the learning process (Butler & Winne, 1995; Pintrich, 2000, p. 454; Zimmerman, 1990, 2001, 2002; Zimmerman & Schunk, 2001).

To facilitate connections between research efforts based on different theoretical models, Zimmerman (1994, 1998) developed a framework composed of six psychological dimensions of self-regulated learning. Each dimension of the framework is associated with a key word, making it easy to remember. Although the dimensional framework is applicable to many learning settings, both academic and nonacademic, I applied it specifically to academic studying. We can examine students' motivation (*why* students study), their strategies (*how* they study), their time (*when* they study), their performance outcomes (*what* they self-monitor during their studying), their environment (*where* they study) and their social context (with *whom* they study or from *whom* seek help during their studying). For each dimension, task conditions, self-regulatory attributes, and self-regulatory processes can be examined (Zimmerman, 1994, 1998). Zimmerman's dimensional framework plays a pivotal role in this dissertation, serving to organize the results of both the literature review and the empirical study.

Purpose of the Current Study

As previously mentioned, this project arose from my own experience—specifically, from wondering whether I, as a community college mathematics instructor, could help students learn to study more effectively. For that reason, I wanted to design a relatively simple intervention that could be implemented by an individual teacher, without administrative assistance in the form of changes to classroom scheduling, textbooks, or computer software. The intervention needed to be respectful of both instructional time and the teacher's preparation time.

Under these constraints and the guidance of self-regulated learning theory, I designed a simple study-journaling intervention that included goal setting, planning, self-monitoring, and reflection. The intervention was implemented in developmental mathematics classes at a large urban community college in Texas.

The purpose of the current study was to investigate how this study-journaling intervention affected student success in the developmental mathematics course and on the final exam, and what it revealed about the students' study habits. Nine developmental mathematics classes implemented the study journal project; another nine classes served as a comparison group. For ease of discussion, students in classes receiving the intervention will often be referred to as study journal students; their study logs, goal sheets, and reflective writings will be collectively referred to as study journals. Students in control classes did not keep study journals and will be referred to simply as control students.

The study was designed around the following research questions:

1. Are study journal students more likely to pass the course and the final exam than control students?
2. What are the perceptions of the study journal students regarding the effects of the journaling process on their study habits and academic performance?
3. What are the study habits of the study journal students, as shown by their written goals, study logs, and reflective writings?

4. For the study journal students, which of these study habits distinguish successful students from unsuccessful students?

Rationale for the Study Methodology

My interest in the intervention's effect on mathematics success suggested a quantitative approach. My interest in the students' perspectives suggested a qualitative approach. The study journals of the students provided a rich data set from which I could use qualitative methods to glean traces of how students were studying and self-regulating their learning. Therefore, I chose a mixed methods design with two strands: (1) a confirmatory strand in which I evaluated the effectiveness of the study journaling intervention and (2) an exploratory strand in which I sought information about the study habits and strategies of the study journal students. The first research question anchors the confirmatory strand, and the second question supplements it. The third question anchors the exploratory strand. The last research question seeks connections between study habits and success, linking the two strands and thus linking the qualitative and quantitative data, as recommended by Tashakkori and Creswell (2007).

Although a detailed description of the quantitative methodology will be provided in a later chapter, a short overview is warranted here. By providing a brief description of the challenges community college students present to quantitative researchers and my approach to these challenges, I will furnish the reader with the means to understand the chapter divisions in this dissertation.

Community college students do not proceed through college in neat cohorts. In the same classroom, traditional college students mingle with substantial numbers of older students supporting families; students with high school diplomas mingle with students holding general educational development (GED) credentials; students aspiring to earn bachelor's degrees mingle with students pursuing two-year degrees or certificates. In developmental mathematics in Texas, the setting for this study, several different placement tests are used to place students in the appropriate courses in the three-course developmental mathematics sequence. Students repeat courses frequently; sometimes the attempts occur in consecutive semesters, but sometimes they are separated by gaps of several years. In the second and third courses of the developmental mathematics sequence, any given classroom should be expected to contain substantial proportions of students repeating the course, students placed directly into the course via placement test, and students entering directly from the prerequisite course. In the current study, these proportions were 39.7%, 19.1%, and 41.2%, respectively.

The variety in student backgrounds and placement mechanisms presents problems for researchers attempting to use regression or analysis of covariance with traditional success predictors, such as high school grade point average (GPA), college GPA, Scholastic Aptitude Test (SAT) score, or previous mathematics course grade. In developmental mathematics classes, nonexistent data on these predictors is the norm, not an aberration. In many cases of nonexistent predictor values, "not applicable" is a more apt descriptor than "missing."

I handled these difficulties through a statistical technique called propensity score matching. For propensity scores to be relevant, participants need to be divided into two groups (treatment and control) and have scores for several background variables—the covariates for which we would like to control. The propensity score of an individual is defined to be the conditional probability that the person will be in the treatment group given that particular person's vector of values on the covariates (Rosenbaum & Rubin, 1983). By matching on the propensity score, we can trim the initial sample into groups that have similar distributions on the covariates, even if we are not sure how those covariates would function in a statistical model. This approach is especially useful when some of the covariates are not applicable to large numbers of participants, as the propensity score model can include variables intended to capture the pattern of missingness on other variables. By using propensity scores to create groups that are well-balanced on important covariates and have similar patterns of missingness, the treatment effect analysis will be less sensitive to modeling assumptions (Ho, Imai, King, & Stuart, 2007).

Overview of Remaining Chapters

Chapters II, III, and IV of the dissertation take the form of complete manuscripts, structured for future publication as journal articles. In each manuscript, I have endeavored to include the same level of detail as would be found in a traditional dissertation. For that reason, the manuscripts are longer and more detailed than typical journal articles.

Chapter II is a review of empirical research on self-regulated learning interventions for college students. I used a systematic approach, documenting the search procedures and the number of results they produced and using a set of defined inclusion questions to select the articles for review. As previously mentioned, I organized the reviewed articles around Zimmerman's theoretical framework of six psychological dimensions of self-regulated learning (1994, 1998). Because Zimmerman first presented the framework in 1994, I chose 1994 as the start date for the review. The original 1994 framework had four dimensions; two additional dimensions were added in 1998. By systematically reviewing articles over a twenty-year span, 1994–2013, I have attempted to paint a useful picture of the landscape of intervention research, describing what has been done, what has not been done, what we have learned, and what we have yet to learn.

Chapter III presents the mixed methods investigation of the study-journaling intervention for developmental mathematics students at a community college. The research questions addressed by the study have already been presented. The data collection, data analysis, and results sections are each organized into two strands: a confirmatory strand addressing how the intervention affected mathematics success, and an exploratory strand addressing the students' study habits as revealed by the study journals. Zimmerman's dimensional framework serves to organize the results of the exploratory strand. The strands are brought together at the end of the results section.

The third manuscript, Chapter IV, is an extended methodological discussion intended for an audience of community college faculty and administrators. It may also be of use to quantitative researchers interested in options for handling the aforementioned difficulties with community college data analysis. While the entire dissertation is written in the first person, Chapter IV is deliberately written in an informal, conversational tone. (While writing it, I envisioned myself engaged in a leisurely chat, perhaps over coffee, with a community college administrator.) Unlike many methodological discussion articles, this piece is intended to be non-technical. I discuss the need for credible inferences when evaluating programs, the lack of credibility that arises when comparing groups with inherent differences, and the potential of propensity score matching to improve the credibility of community college research. Also in Chapter IV, I describe in detail the matching variables used, how I handled covariates that were not applicable to all the students, and my decision-making process. Using my data as an illustration, I explain how propensity score matching can be used to trim groups that have substantial imbalances on important covariates to slightly smaller groups that have much better balance. The manuscript closes with a section on practical advice for researchers interested in using propensity score matching.

The final chapter of the dissertation, Chapter V, provides an overview and synthesis of the findings and implications presented in the three manuscripts.

CHAPTER II

SELF-REGULATED LEARNING INTERVENTIONS FOR COLLEGE STUDENTS: A REVIEW OF THE LITERATURE

Introduction

In 1984, Rohwer called for a coherent psychology of studying and a deliberate plan for research about academic studying. He laid out a preliminary theoretical framework, in which student, course, and task characteristics influence study activities, which in turn influence academic achievement. Specifically, he suggested researchers use interviews, observational studies, and large-scale quantitative inventories to determine which study activities discriminate between successful and unsuccessful students. Researchers could then test their hypotheses by inducing students to employ certain strategies and then observing their achievement levels. The researchers could then refine the original framework (Rohwer, 1984).

Rohwer's call for a theoretical framework of academic studying was answered, but not in the way he described. Instead of a component-based theory focusing on specific studying activities, a process-oriented theory developed, called *self-regulated learning.* In the self-regulated learning model, learners are viewed as active drivers of the learning process, not as passive recipients of learning activities provided by the teacher. Study activities are not the focus of the model—study activities are merely tools in a toolbox from which the learner can select as needed.

Self-regulation models follow a general framework that includes planning, monitoring, control, and reflection. Researchers have approached self-regulated learning from a variety of theoretical perspectives, including but not limited to information processing, operant theory, social cognitive theory, constructivism, and Vygotsky's developmental theory (Zimmerman, 2001). Self-regulated learning models, though they differ in details, have much in common: they assume learners are active participants who have potential to exert control over the learning process, and they address learners' thoughts and activities before, during, and after the actual learning activity. In a forethought phase, the learner engages in goal setting and planning. The forethought phase is followed by a learning enactment phase, in which the learner selects and carries out the most appropriate learning strategies. Then, in a reflection phase, the learner evaluates the effectiveness of the learning process by comparing outcomes to some standard. Information from the reflection phase, along with internal and external feedback, is incorporated into planning for later tasks. In a recursive feedback loop, the learner continually evaluates the process, makes adjustments, sets new goals, and begins again (Butler & Winne, 1995; Pintrich, 2000; Winne & Hadwin, 1998; Zimmerman, 2001, 2002, 1990; Zimmerman & Schunk, 2001).

Zimmerman's Dimensional Framework

Zimmerman (1994) developed an academic self-regulation framework not only as a theoretical model, but also to facilitate integration of research from different theoretical perspectives. Initially, the framework consisted of four dimensions: motivation, methods (strategies), performance outcomes, and environment. He later added two more: time and social context (Zimmerman, 1998). In each dimension, the learner can choose to enact self-regulatory processes or demonstrate self-regulatory attributes. Self-regulated learning research, whatever the theoretical framework, will address at least one of these dimensions (Zimmerman, 1994, 1998).

Zimmerman's framework is especially helpful because it is easy to remember. A key word is associated with each psychological dimension, allowing us to organize research by the types of questions asked: why, how, when, what, where, and with whom. We can study students' motivation (*why* they learn and study), their strategies (*how* they study), their time (*when* they learn and study), their performance outcomes (*what* they monitor and record to evaluate the effectiveness of their learning), their physical environment (*where* they learn and study) and their social environment (with *whom* they learn and study and from *whom* they seek help).

In each dimension, learners can engage in the forethought-performance-reflection loop. For example, in the Time (When) dimension, students can plan their time, carry out studying activities during the planned times (or at other times), reflect upon their use of time, and then make adjustments to their time management. In the Strategies (How) dimension, students can plan their strategies, carry them out, reflect upon the strategies' effectiveness, and then adjust.

Zimmerman points out that in each dimension, freedom is essential.
Self-regulation cannot occur unless the learner is given choice—if motivation, strategies, time, outcomes, environment or social context are compelled, then any regulation that occurs is other-regulation, not self-regulation. As long as this freedom exists, then for

each dimension, we can study self-regulatory attributes, self-regulatory processes, and the task conditions under which self-regulation occurs (Zimmerman, 1994, 1998).

In academic settings, this element of choice is an integral part of learning. When studying for a class, the learner has control over how, when, and where studying occurs, or whether it occurs at all. Students may have constraints within which they must work, but they are the decision-makers about what tasks to tackle and how, when, and where to approach them. During class time, the teacher has the power to compel students to participate in learning activities, or at least to exert considerable pressure in that direction. As soon as class ends, the decision-making power switches to the student.

As students mature from elementary school to middle school to high school, they carry an increasing amount of responsibility for moving their learning forward through out-of-class studying. For college students, the expectations for out-of-class studying are higher still. In order to be successful, college students must learn to study effectively. This means planning their studying, carrying it out, reflecting on their progress, and making the needed adjustments.

Empirical Research on Self-Regulated Learning

Self-regulated learning research generally falls into two areas: descriptive research, investigating how self-regulated learning works; and intervention-based research, investigating whether self-regulated learning can be taught. Descriptive research examines strategies used by skilled self-regulated learners, and explores relationships between self-regulated learning and other constructs. In intervention-based research, a researcher evaluates the effectiveness of an intervention or program designed to improve learners' self-regulated learning skill.

In academic environments, intervention-based research includes two main categories of self-regulated learning interventions: the stand-alone seminar or program not tied to a particular course, and an embedded approach in which self-regulation training is incorporated into course material. The intervention's effectiveness is typically evaluated by tying it to either a psychological construct or an achievement-related outcome variable, such as exam score, course grade, or grade point average. If the outcome variable is a psychological construct, it can be measured either by observational data or by self-report, often a score on an inventory.

In 1994, Schunk and Zimmerman urged researchers to prioritize intervention research over additional descriptive research. They recommended that the intervention research take place in students' "actual learning settings." In particular, researchers should seek to determine whether students actually engaged in self-regulated learning activities, rather than relying on self-reports or inventories (Schunk & Zimmerman, 1994).

This review attempts to summarize progress made on that research agenda in a postsecondary setting. Specifically, it focuses on interventions designed to improve the self-regulated learning of college students. For college students, self-regulated learning is centered on outside-of-class studying. Therefore, before describing the criteria and procedures used to select articles for review, I will summarize several existing reviews of academic studying research.

Existing Reviews of Intervention Research

Kaldeway and Korthagen (1995) reviewed 20 empirical studies of stand-alone study skills courses. They classified the study strategies taught in the courses as strategic (understanding the problem), operational (creating a plan), executive (enacting the plan), or reflective (looking back at the process). All of the courses focusing on strategic or operational activities produced a positive effect on exam scores or text comprehension. Some courses involving only executive or reflective activities produced a positive effect, while others did not. In nearly all the reviewed studies, the intervention was evaluated by measuring students' comprehension of a text learned specifically for the research. None of the studies addressed whether the students were able to integrate the study strategies they learned into actual courses they were taking for a grade.

Hattie, Biggs, and Purdie (1996) conducted a meta-analysis of 51 study skills interventions for K-12 and college students. Instead of classifying the interventions as stand-alone or embedded in content, the researchers classified them by level of structural complexity and by whether transfer of learning was near or far. The average effect size, based on 270 effect sizes from the 51 studies, was 0.45. When separated by type of outcome measure, average effect sizes were 0.57 for performance, 0.16 for study skills, and 0.48 for affect. Effect size decreased as student age increased, dropping to 0.27 for university students. Largest effect sizes were for near transfer, in which the performance outcome was closely related to the context of the study skills training, and for unistructural interventions, in which a single skill or feature was taught with a narrow aim. This meta-analysis resulted in a recommendation to embed study skills training in the teaching of content rather than in a stand-alone course or seminar.

The need for research tying self-regulated learning interventions to authentic content and achievement was supported by Hadwin and Winne's (1996) review of college study skills interventions, which was based on a theoretical framework of self-regulated learning and cognitive processing. The only studies reviewed were those in which the study strategy instruction or achievement evaluation was related to a course students were taking for a grade. Stand-alone study skills courses were only considered if they involved learning of actual postsecondary material for a grade. Experimental designs in which students completed low-level or artificial activities were omitted, as were designs in which student course grades depended upon participation rather than the learning of content. Only 16 studies met these criteria. Concept mapping, self-questioning, and monitoring of study time were found to have at least moderate positive effects in certain contexts. In this review, Hadwin and Winne found that very little empirical research has been conducted involving the application of study strategies to a local meaningful goal.

Winne and Hadwin (1998) also developed an important model that conceptualized academic studying as self-regulated learning. In the Winne-Hadwin (1998) model, the studying process consists of four stages: task definition, planning, enactment, and adaptation. Each stage follows a five-facet model, conveniently abbreviated COPES, which consists of conditions, operations, products, evaluations, and standards. Conditions are the set of resources and constraints under which cognitive activity occurs. Operations are the tactics and strategies that create products. Some products take the form of externally visible performance; other products are internal constructions, such as information structures, images, or additions to working memory. The student evaluates these products against standards. In the adaptation stage, the learner uses the results of these evaluations to make necessary adjustments to the studying process. The Winne-Hadwin (1998) model of studying, with its emphasis on the learner's evaluations and adjustments, makes an important contribution to the theory of self-regulated learning.

Greene and Azevedo (2007) used the Winne-Hadwin (1998) studying model as a lens for their theoretical review of self-regulated learning research. For each facet of the model (conditions, operations, products, evaluations, and standards) and several subfacets, they examined review articles and representative empirical studies with the goal of demonstrating how the model can help researchers conceptualize and analyze the learning process. Their examination highlighted many promising areas for research. They suggested investigations of the temporal changes in task definitions, goals, plans, and adaptations induced by changes in conditions and standards. The model's separation of task definition and goal setting into separate phases facilitates interventions targeted toward improving the quality of each. Interactions of the model's facets should be examined, including the interaction between task definitions and strategies for different levels of domain knowledge. Self-monitoring is a critical element of this studying model, and the review uncovered a shortage of studies that examine whether students can be taught how to self-monitor.

Purpose and Inclusion Criteria

The purpose of this review is to summarize and critically assess empirical research on self-regulated learning interventions for college students. By using a systematic approach with defined inclusion criteria, I hope to improve the clarity, validity and auditability of the review (i.e., make it easier for the reader to see the overall pattern in the review's results, reduce selection bias, and increase transparency, making it easier to evaluate whether the review's conclusions are grounded in the retrieved data; Booth, Papaioannou, & Sutton, 2012).

The review was guided by the following question: What are the effects of self-regulated learning interventions on college students? The review was not expected to definitively answer this question, but rather to describe the landscape of empirical studies addressing it. I decided in advance to use Zimmerman's dimensional framework (1994, 1998) to organize the results. As previously mentioned, Zimmerman's framework consists of six dimensions of self-regulated learning: motivation, time, strategies, outcomes, environment, and social context. I anticipated that this framework would make it easier to see which dimensions of self-regulated learning have been heavily researched, which dimensions have been relatively ignored, and which dimensions are influenced by the most effective interventions. Because Zimmerman first presented his dimensional framework in 1994, I chose 1994 as the start date for the literature review.

This approach is similar to that taken by Greene and Azevedo (2007), who based their review on a theoretical model of academic studying (Winne & Hadwin, 1998) chosen a priori. However, whereas Greene and Azevedo purposefully chose articles that illustrated particular aspects of the model, I chose articles based on defined criteria and then simply used the model to organize them. This meant that the review could potentially include studies not fitting neatly into the model.

In order to choose the articles, I needed to decide, "What counts as a self-regulated learning intervention?" In this review, a self-regulated learning intervention is any sort of training, project, program or assignment designed to help students improve their self-regulated learning skills and therefore their academic outcomes—a program tacked onto "regular teaching" or different from the "usual way of doing things."

For this review, the intervention needed to target the learning of academic content. Interventions targeting overall behavior (e.g., physical activity) or non-academic skills (e.g., music or swimming) were not included, even if they were explicitly based on a self-regulation framework. Also excluded were interventions targeting preservice teachers in their role as teacher (e.g., interventions intended to help them reflect on their classroom instruction or foster self-regulated learning in their classrooms). If an intervention targeted preservice teachers in their role as students (learning academic content for an education-related class), I included it, as long as it met the other criteria.

The requirement for interventions to target the learning of academic content was relatively straightforward to apply. However, the "self-regulated learning intervention" requirement was not. Self-regulated learning has become a popular topic in educational research, and is considered a desirable outcome for nearly any educational program or curriculum. For that reason, potential improvement in self-regulated learning is often used as a justification for educational interventions, even if those interventions do not directly teach or scaffold it. Self-regulated learning is often considered an outcome of interest, even for interventions whose main aim is something else.

To determine what sorts of interventions qualified as self-regulated learning interventions, it was helpful to rephrase the original guiding question (about the effectiveness of self-regulated interventions for college students) more informally: “Can college students be taught to self-regulate their learning of academic content?” I wanted research studies to be included in the review if and only if they shed light on this question. Presenting self-regulated learning theory in the literature review or using self-regulated learning as an outcome variable was insufficient.

To clarify the definition of “self-regulated learning intervention” for this review, I revisited the three phases common to most self-regulated learning models: forethought, learning enactment, and reflection. Some interventions concentrated on the self-regulated learning model in its entirety, including all its phases and its recursive nature. Other interventions concentrated on one of the three phases.

Because forethought and reflection are trademark practices of self-regulated learners, I included interventions focused on either the forethought phase or the reflection phase. Thus, interventions were included if they taught or scaffolded goal setting or planning, or if they taught or prompted students to reflect upon the learning process. Because self-monitoring is a key ingredient for effective reflection, I included interventions that taught or prompted students to self-monitor their learning. Interventions including one or more of these elements (forethought, reflection, or self-monitoring) were included in the review even if they were not explicitly designed around a self-regulated learning framework.

However, because learning enactment is included in nearly all educational interventions, the learning enactment phase was not helpful as an inclusion criterion. Unlike forethought and reflection, learning enactment is not unique to self-regulated learners. What distinguishes self-regulated learners during the learning enactment phase is their level of control over the process and the way they use the forethought and reflection phases to adjust the learning enactment phase. For this reason, interventions focused on learning enactment were only included if they were specifically situated in a self-regulated learning framework.

Because this review focused on the studying of academic content, study strategies courses and college success courses became candidates for inclusion. Some of these courses are based on a specific theoretical model, while others provide a cornucopia of offerings from a variety of theoretical models. Others are hodge-podges of study techniques and time-management tips, not incorporating any theoretical model at all. Study strategies courses were included in this review if they used or taught self-regulated learning as a theoretical framework. Study strategies courses incorporating either forethought or reflection were also included, even if they were not built on a framework of self-regulated learning theory. Because planning, goal setting, and time management involve forethought, study strategies courses containing these components were included in the review. If the research article did not contain sufficient information to conclude that the course contained instruction on forethought, reflection, or self-regulated learning, the study was excluded.

Interventions were excluded if they did not explicitly teach or scaffold some aspect of self-regulated learning. For example, the following were not considered self-regulated learning interventions: changes in instructional mode (e.g., face-to-face versus online), instructional materials (e.g., videos versus no videos), assessment (e.g., quizzes versus no quizzes), or feedback (e.g., handwritten feedback versus computerized feedback).

I also excluded studies in which the self-regulated learning component was a relatively small piece of a larger intervention with a purpose other than improving self-regulated learning. For example, I excluded an action research project, in which teams of marketing students organized events at a coffeehouse (Young, 2010) and a freshman learning community in which engineering students solved substantial problems and built exhibits (Lipson, Epstein, Bras, & Hodges, 2007), even though both interventions contained a reflective journal component. Because the interventions did so many other things, the studies could not provide useful information about whether self-regulated learning could be taught. For similar reasons, I excluded studies that compared problem-based learning classrooms to traditional classrooms, or that compared classes with and without supplemental instruction.

Before beginning the article screening process, I created a list of inclusion questions. Many of the inclusion decisions described above were not addressed by the initial list of questions, but stemmed from finding an article that provoked a “What on earth do I do with this?" response. When that happened, I made a decision, documented the decision, and attempted to apply that decision consistently throughout the remainder of the screening. When necessary, I tweaked the inclusion questions.

The following is the final list of questions used for screening. In order to be included in the review, the answer to each question had to be "yes." A rationale for the questions follows the list.

1. Is the article in English?
2. Was the article published in 1994 or later?
3. Does the article appear in a scholarly journal in the ERIC database?
4. Does the article describe empirical research targeting college students (without specifically targeting learning-disabled students)?
5. Does the research study evaluate the effectiveness of an intervention designed to improve college students' academic studying or self-regulated learning of academic content?
6. Is the intervention based on a self-regulated learning framework, or does it explicitly teach or scaffold goal setting, planning, self-monitoring, or reflection?
7. Were the participants enrolled in a U.S. college or university?
8. Did the intervention target students' learning of content for a face-to-face class?

Question 1 (written in English) was included for practicality. Question 2 (1994–present) was chosen because Zimmerman's dimensional framework was first published in 1994. Regarding Question 3, I restricted my search to the Education Resources Information Center (ERIC) database, because I anticipated that the majority of relevant articles would be listed in ERIC. The purpose of this review was not to do an exhaustive search, but to provide a useful overview of existing research on self-regulated learning interventions as seen through the lens of Zimmerman’s framework. I anticipated that the ERIC database would suffice for this purpose. Within ERIC, I limited the search to scholarly journals, because these typically undergo a more stringent review process than other ERIC documents.

Question 4 (empirical research on college students) is self-explanatory. After screening began, I added the restriction to eliminate interventions targeting learning-disabled students, who were not my population of interest. Questions 5 (interventions targeting the studying of academic content) and 6 (self-regulated learning framework) were the least clear-cut and have already been discussed. Question 7 (research conducted in the U.S.) was included to keep the review manageable and because higher education structure and culture may differ substantially among countries. If no information was given about the location of the study, and if all the authors were affiliated with institutions outside the U.S., I assumed the study was conducted outside the U.S.

Question 8 (learning content for face-to-face classes) was included for two reasons. First, I decided that face-to-face and online classes provide very different sets of constraints and opportunities for facilitating self-regulated learning. For that reason, self-regulated learning interventions for online classes are best left for a separate review. Second, I discovered that some self-regulated learning interventions were designed to evaluate students’ learning of academic content that was unrelated to their classes. The participants for such interventions are typically recruited from psychology classes, which often contain a research participation requirement. Instead of studying psychology content, the researchers ask the participants to study material on wildcats, the human heart, or some other topic unrelated to their class. Such research may be indeed be valuable, and may provide useful insights into self-regulated learning and how it can best be taught. However, it seems important to distinguish this type of research from research into students' learning of material germane to the subject of the course.

A few limitations should be noted. My decision to restrict the search to the ERIC database could result in bias, if there are systemic differences between ERIC and other databases. A different way of defining "self-regulated learning intervention," especially in regard to interventions addressing the learning enactment phase, might have resulted in different studies being reviewed. Also, because the first step in the search process was to screen a large number of abstracts, the pool of reviewed articles is highly dependent on the information contained in the abstracts.

Search Procedure

To find the articles, I used the ProQuest search interface, restricted to the ERIC database. The overall search strategy was to intersect "self-regulated learning" (or related terms) with "college students" (or related terms), and then remove articles that listed "learning disabilities" (or related terms) as key words. The search terms also required a publication date in 1994–2013 and for the publication to be designated as a "scholarly journal" in ERIC. The exact search terms are listed in Appendix A.

For the “self-regulated learning” portion of the search terms, I used the database’s thesaurus to find the most relevant subject tags, “study skills,” “study habits,” “self management,” and “time management.” I also used the databases ALL operator and truncation character (*) to search all fields except the full text for "study skill*" OR "self regulat*" OR "study habit*." For the “college students” portion, I used the database’s thesaurus to find other subject tags related to “college students” and “higher education.” Through the “explode” function, the search also found articles whose subject tags were subtopics of the subject tags listed in the search terms.

For example, an article would appear in the search if it listed both “study skills” and “higher education” as subject tags, as long as it did not also have the subject tag “learning disabilities” (or a tag listed as a subtopic under “learning disabilities”). An article would also appear in the search if it had “colleges” as a subject tag and “self-regulated learning” in the title or abstract, provided the article did not have “learning disabilities” as a subject tag.

This search strategy resulted in 1,825 articles. To prepare for article screening, I saved the 1,825 references to the RefWorks reference management software. In RefWorks, I set up a profile with fields to record the results of the initial screening (abstracts only) and the secondary screening (full texts). The abstract was visible in the RefWorks profile, which was convenient for screening.

For the initial screening, I read the abstract of each of the 1,825 articles and compared it to the inclusion questions. If it was apparent from the abstract that the article did not meet the inclusion criteria, I recorded this, along with the question number of an unmet criterion. For example, if an article did not describe empirical research on college students, I recorded it as “no 4.” If, based on the abstract, the article appeared to meet the inclusion criteria, I recorded “yes.” If I was not sure, I recorded “maybe.”

When recording the question number of the criterion I used to exclude the article, I chose the question number that made it easiest to ascertain a “no” response from the abstract. Most frequently, this was Question 5, which asked whether the research evaluated the effectiveness of an intervention. If it was clear from the abstract that the research was descriptive and not intervention-based, I recorded “no 5.” Often the absence of an intervention was easier to ascertain than the presence of empirical data on college students. For 947 articles, I recorded “no 5.”

The initial screening of the 1,825 abstracts resulted in 75 “yes” responses, 364 “maybe” responses, and 1,386 “no” responses. For the 439 “yes” and “maybe” articles, I attempted to obtain the full-text articles. I was successful in all but two cases. After skimming the texts of these 437 articles, I classified 356 of them as “no,” again using Refworks to record the question number of the most obvious unmet criterion. This resulted in a pool of 81 tentative “yes” articles to undergo further review.

For each of these 81 articles, I read the full text again. For each, I used a spreadsheet to summarize information about the sample, intervention, analysis method, outcome measure, targeted dimension(s) of self-regulated learning, and results. I also recorded the answers to each of the inclusion questions. During this process, I excluded 39 more articles that I had originally classified as “yes,” because after a more careful reading, I decided they did not meet the inclusion criteria. The final sample for this review contains 42 articles.

Results

As previously mentioned, Zimmerman (1994, 1998) described six dimensions in which students could self-regulate their learning, each with a key word: motivation (why), outcomes (what), strategies (how), time (when), environment (where), and social context (who). Twenty-seven of the articles described an intervention aimed specifically at just one or two of these dimensions. The other 15 articles described an intervention aimed either at the entire self-regulated learning (SRL) process, or at several dimensions. Often, but not always, this intervention was a stand-alone study strategies course not tied to a particular content area. In the text, I will refer to these 15 studies as *overall-SRL* interventions or studies. I will first summarize these overall-SRL studies; then I will summarize the 27 studies that targeted one or two specific dimensions of SRL.

Interventions Targeting Overall SRL

Table 1 summarizes the interventions evaluated by the 15 overall-SRL studies. For consistency, "Study Strategies" has been listed as the subject for all stand-alone courses not associated with a particular content area, regardless of whether the article refers to it as an orientation course, a learning strategies course, a college success course, a study skills course, a self-regulated learning course, or some other name.

Table 1

Interventions Targeting Overall Self-Regulated Learning

Author(s) (Year)	N $(n_1/n_2/\ldots)$[a]	Subject	Intervention
Ahuna, Tinnesz, & VanZile-Tamsen (2011)[b]	9665 (1900/7765)	Study Strategies	Success course based on active and dynamic self-regulation. Lectures plus one-on-one meetings with peer monitor to help students learn to self-assess.
Bail, Zhang, & Tachiyama (2008)	157 (79/78)	Study Strategies	Writing-intensive SRL course, part face-to-face and part computerized. Must develop 3 specific strategies and write major paper about them. First-semester freshmen discouraged from enrolling.
Gerhardt (2007)	223	Management	Four tutorials, each completed partly online and partly in class. Topics include self-assessment, goal setting, self-monitoring, and self-regulation.
Haught, Hill, Walls, & Nardi (1998)	69 (34/35)	Study Strategies	One component of a learning strategies course. One-on-one individualized feedback/counseling on each student's below-50th-percentile LASSI subscales.
Hofer & Yu (2003)	78	Study Strategies	Student success course targeting academically struggling students. SRL plus cognitive psychology. Lectures plus small lab/discussion sections. Students chose a target class for the goals/strategies.
Hopper (2011)	120 (19/101)	Anatomy & Physiology	Optional supplemental class for students in Anatomy and Physiology (A&P). Mandatory for course repeaters. Goal setting, time planning, notebook setup, A&P-specific study skills.
Humphrey (2006)	Not reported	Study Strategies	Voluntary program for probationary students. Small group meetings with facilitators. Students discuss strategies, create weekly written goals, and report on progress toward goals.

Table 1 Continued

Author(s) (Year)	N $(n_1/n_2/\ldots)^a$	Subject	Intervention
Kamphoff, Hutson, Amundsen, & Atwood (2007)	387 (307/80)	Study Strategies	Mandatory program for probationary students. Failing to enroll or missing a session resulted in suspension. Conceptual model was table with 4 legs: personal responsibility, positive affirmations, goal setting/life planning, self-management. On tabletop, interaction/support from group and facilitator.
Lee (2007)	83 (36/47)	Study Strategies	One component of learning strategies course required for conditionally admitted students. Analyzed case studies about academic learners. Students received lesson, analysis template, and rubric; then wrote 5-paragraph essays about the case studies. Treatment students analyzed collaboratively; control students analyzed individually.
Orange (1999)	63 (29/34)	Educational Psychology	Self-regulation videotape portraying student actors at an Academics Anonymous meeting. Students committed to an action plan based on 12 steps of self-regulation.
Reeves & Stich (2011)[b]	243	Study Strategies	Success course based on active and dynamic self-regulation. Lectures plus one-on-one meetings with peer monitor to help students learn to self-assess.
Ryan & Glenn (2003)	1497 (77//66/1354)	Study Strategies	College success seminar based on four strategies: question-asking, goal-setting, task analysis, self-assessment.
Schapiro & Livingston (2000)[b]	342	Study Strategies	Success course based on active and dynamic self-regulation. Lectures plus one-on-one meetings with peer monitor to help students learn to self-assess.
Sweidel (1996)	87	Educational psychology	Study Strategy Portfolio. Written study plan, action commitment, test grade prediction, reflection on study plan and test grade.

Table 1 Continued

Author(s) (Year)	N ($n_1/n_2/\ldots$)[a]	Subject	Intervention
Travers, Sheckley, & Bell (2003)	78 (24/54)	Mathematics (mostly developmental)	Mathematics faculty solicited for training on teaching SRL strategies. Classes taught by trained faculty compared to classes taught by faculty not responding to invitation for training. No details about the strategies.

[a] If applicable, second line of the sample size column describes the breakdown of the total sample N into smaller groups. The group best characterized as *treatment* is listed first; the group best characterized as *control* is listed last. When there are more than two groups, a double slash (//) indicates the division that best separates treatment from control. [b] These three studies evaluated the same course at the same university.

For 10 of the 15 overall-SRL studies, the sample was composed of students in a study strategies course. Because three of these 10 studies evaluated the same course, only eight different study strategies courses are represented. Eight studies evaluated the effectiveness of the study strategies course itself, while two studies evaluated the effect of a single component of the course (individualized counseling on the results of a learning strategies inventory [Haught, Hill, Walls, & Nardi, 1998], and collaborative analysis of case studies [Lee, 2007]). The remaining overall-SRL interventions were associated with courses in a specific content area (two with educational psychology, one each with mathematics, management, and anatomy/physiology).

Table 2 summarizes the results of the fifteen overall-SRL studies, and the outcome measures used to evaluate each intervention. I classified outcome measures as *achievement measures* if they assessed the learning of content other than the content of a study strategies class. Examples of achievement measures are grade point average

(GPA), retention, graduation, and course grade in a course other than a study strategies course.

Table 2

Outcome Measures and Results for Studies Targeting Overall SRL

Author(s) (Year)	Outcome Measure	Results
Ahuna, Tinnesz, & VanZile-Tamsen (2011)[b]	Retention[a] Graduation[a]	Students in Methods of Inquiry (MOI) class more likely to be members of disadvantaged groups and have lower high school GPA and SAT scores. MOI students more likely to be retained or graduate within 2, 3, 4, and 5 years of taking the course. MOI and non-MOI students just as likely to graduate within 4 years, but MOI students had higher 5-year graduation rate.
Bail, Zhang, & Tachiyama (2008)	GPA[a] Graduation[a] Probation/suspension status[a] Number of F grades[a] Acceptance into graduate program[a] Earning of graduate degree[a]	After controlling for gender and prior GPA, students in SRL course had significantly higher cumulative GPA four semesters later. SRL course enrollment was a significant predictor in logistic regression model to predict graduation within 7 years (odds ratio 12.69). Students in SRL course had significantly fewer F grades in subsequent courses. No significant differences on probation/suspension status, acceptance into graduate program, or earning graduate degrees.
Gerhardt (2007)	Helpfulness survey Four-item inventory of self-management skills.	Most students thought the self-management training was very helpful, had a positive impact on their performance, and was useful for the future. Self-reported use of self-management skills increased after training.

Table 2 Continued

Author(s) (Year)	Outcome Measure	Results
Haught, Hill, Walls, & Nardi (1998)	LASSI inventory (Learning and Study Strategies Inventory) Course grade GPA[a]	On the post-course LASSI, individually counseled students scored significantly better then control students on 7 of 10 subscales, and also had significantly fewer below-50 subscales. Number of below-50 subscales decreased significantly (from pretest to posttest) for individually counseled students but not for control students. Individually counseled students had significantly higher intervention semester GPA and subsequent semester cumulative GPA. No significant differences on study strategy course grade or subsequent semester GPA.
Hofer & Yu (2003)	MSLQ inventory (Motivated Strategies for Learning Questionnaire)	Significant improvement for 4 of 6 motivational variables and 6 of 7 cognitive variables between Time 1 and Time 2. Some changes in correlations over time. Final course grade correlated with one pretest motivational variable (extrinsic motivation) and one posttest motivational variable (self-efficacy).
Hopper (2011)	Helpfulness survey Anatomy & Physiology course grade[a]	Higher percentage of students in the Supplement class earned a C or better, compared to students not in Supplement. Lower percentage of Supplement students withdrew. Some survey items showed significant improvement from pre-test to posttest. Comments from students about the Supplement's value were very positive.
Humphrey (2006)	Academic good standing[a] GPA[a]	Percentages of students regaining good academic student were higher for Project Success (PS) students than control students for 3 different semesters (58% vs. 42%, 55% vs. 45%, and 53% vs. 47%). PS students had higher GPAs than control students for some cohorts (ambiguous/inconsistent numbers reported). PS students had higher retention rates than control students during intervention semester and one subsequent semester; then retention rates equalized. More PS students in good standing and fewer suspended than control students.

Table 2 Continued

Author(s) (Year)	Outcome Measure	Results
Kamphoff, Hutson, Amundsen, & Atwood (2007)	Retention[a] GPA[a]	Percentage of Strategies for Academic Success (SAS) students returned to good standing increased from 40% to 58% over a 4-year period, as program was modified and teeth added. SAS students showed significantly higher GPA gains than control students during last 2 years of the 4-year period, but not during first 2 years. For the last cohort of the 4-year period, SAS students gained 0.7309 on the GPA and control students only gained 0.4202.
Lee (2007)	Critical thinking scoring rubric applied to case study analysis essays	Both the collaborative analysis and individual analysis groups showed significant improvements in their ability to analyze case studies about academic learning. No differences between groups.
Orange (1999)	SRI (Self-Regulation Instrument) developed by the researcher	On both pretest and posttest SRI, control students scored significantly higher than students who watched the Academics Anonymous video. Video-watching group showed significantly higher SRI gain than control group.
Reeves & Stich (2011)[b]	DALI-R inventory (Dynamic & Active Learning Inventory-Revised) Helpfulness survey	Students in MOI class showed significant growth in both active and dynamic SRL from pretest to posttest (effect sizes 0.66 and 0.55). No differences in effect among racial subgroups. On helpfulness survey, most responses to open-ended items were positive, a few were negative. Likert item results not reported.
Ryan & Glenn (2003)	Retention[a] Academic good standing[a]	Overall one-year retention rate was significantly better for strategy-based Learning to Learn seminar (74%) than for both theme-based Freshman Seminar (45%) and untreated cohort (56%). No significant differences in probation rates among the three groups. For probationary freshmen, one-year retention rate was significantly better (57%) in Learning to Learn than in Freshman Seminar (21%) or untreated cohort (28%).

Table 2 Continued

Author(s) (Year)	Outcome Measure	Results
Schapiro & Livingston (2000)[b]	DALI-R inventory (Dynamic & Active Learning Inventory-Revised)	Students with high scores for dynamic SRL had significantly higher GPAs. No GPA difference between high- and low-scorers on active SRL. Both high- and low-GPA students showed significant improvements in both active and dynamic SRL.
Sweidel (1996)	Helpfulness survey	Most students found it helpful and thought it improved their study habits.
Travers, Sheckley, & Bell (2003)	ALQ (Approaches to Learning Questionnaire) developed by the authors	Students taught by faculty trained in SRL strategy instruction did not differ significantly from control students on any scale of the post-semester ALQ. Follow-up correlation analysis showed differences in relationships between variables, with stronger correlations in treatment group.

[a]Outcome is classified as an achievement measure. [b]These three studies evaluated the same course at the same university.

Only seven of the fifteen studies used an achievement measure to evaluate the intervention’s effectiveness. All seven reported positive results. Five of them (Ahuna, Tinnesz, & VanZile-Tamsen, 2011; Bail, Zhang, & Tachiyama, 2008; Haught et al., 1998; Kamphoff, Hutson, Amundsen, & Atwood, 2007; Ryan & Glenn, 2003) demonstrated reasonable levels of rigor in both the reporting and the research design, making some effort to control for confounding variables through randomization or statistical analysis. Two are especially notable because they assessed long-term results: 4- and 5-year graduation rates (Ahuna et al., 2011) and 7-year graduation rate (Bail et al., 2008).

The other eight overall-SRL studies did not use an achievement measure to evaluate the intervention. One of these (Lee, 2007) used the student's scores on a learning activity in the learning strategies class. The others used either an inventory or a helpfulness survey. In this context, an inventory is a specific type of questionnaire used to measure students' skill in self-regulated learning or a related construct. It can be created by the researchers themselves or drawn from past research, and its results should be accompanied by reliability and validity statistics. A helpfulness survey is a less rigorous questionnaire intended to elicit students' perceptions of an intervention. Four of the eight studies not using an achievement measure used an inventory as the only means of evaluating the intervention. One study used only a helpfulness survey, and two studies used both an inventory and a helpfulness survey. Most studies showed improvement in inventory scores before and after the intervention. The next section provides additional details about the overall-SRL interventions and the results of the studies.

Studies Using an Achievement Variable to Evaluate Study Strategies Courses

Three of the overall-SRL articles evaluated different aspects of the same elective course, Methods of Inquiry (MOI), at the same university (Ahuna et al., 2011; Reeves & Stich, 2011; Schapiro & Livingston, 2000). Some students took the course on their own initiative, while others were advised to enroll in the course because they had either struggled academically or were in a high-risk group. The class was based on a self-regulated learning framework and included components on goal setting, self-monitoring, and active learning strategies. In addition to lectures, students attended weekly one-on-one meetings with a peer monitor, another student who had earned an A grade in MOI and maintained a 3.0 GPA. The peer monitors were trained to help students self-assess their learning. Students were "required to employ these various study methods continuously in their other academic courses" (Reeves & Stich, 2011, p. 7), but the means of enforcing this requirement were not described.

Ahuna et al. (2011) examined archived data to determine the effects of the MOI course on the graduation and retention of three cohorts of students. The 1,900 students who had successfully completed the course were compared with the 7,665 students who had not. Using chi-square comparisons, they found that MOI students were more likely than non-MOI students to be first-generation college students, student athletes, or members of underrepresented minority groups. MOI students also had had lower high school GPAs and Scholastic Aptitude Test (SAT) scores and were more likely to come from low-income homes and apply for financial aid. Even without controlling for these disadvantages, logistic regression analysis showed that MOI students were more likely to be retained or graduate by the second, third, fourth, or fifth year after the course. The largest retention effect was for the second year, leading the authors to conclude that students should take MOI during the first year. Retention analysis was repeated for various subgroups with similar results, with first-generation college students showing the largest effect size. MOI and non-MOI students were equally likely to graduate within four years, but MOI students had an advantage in the 5-year graduation rate. In the other two MOI studies, inventory scores from the beginning and end of the course showed improvement in students' active and dynamic self-regulated learning, with no notable differences in improvement for racial subgroups or for high- or low-GPA subgroups (Reeves & Stich, 2011; Schapiro & Livingston, 2000).

The study by Ryan and Glenn (2003) compared two versions of the same college success course. The Learning to Learn version focused on academic competencies. Students learned how to apply question-asking, goal-setting, task analysis, and self-assessment strategies to different classroom tasks. The writing-intensive Freshman Seminar version was theme-based rather than strategy-based. Each section of Freshman Seminar explored a different theme, with a focus on collaborative learning and discussion and a faculty member as intellectual mentor. One year after the seminar, significantly more Learning to Learn students (74%) were still enrolled in college than Freshman Seminar students (45%) or students who did not enroll in either course (56%). Although academic probation rates among the three groups did not differ, the Learning to Learn students who went on probation were retained at higher rates (57%) than the probationary students in Freshman Seminar (21%) or not in any college success course (28%).

Bail, Zhang, and Tachiyama (2008) examined the long-term effectiveness of a writing-intensive self-regulated learning course. All students in the course were also in an academic support program for students who showed academic need, and who were either first-generation college students or eligible for financial aid. While many self-regulated learning or college success courses are large lectures, perhaps supplemented with smaller discussion sections, these classes were small, no more than 20 students per class. Also unlike many such courses, first-semester freshmen were discouraged from taking the course. Most students in the course were sophomores. The course emphasized self-awareness during academic tasks, clarifying academic goals, and evaluating and modifying their strategies. The face-to-face classes were supplemented with computer-mediated learning activities and required online meetings with peers. Students were required to write a major paper about three specific strategies they had developed for their other courses.

A control group was created from students who were also in the academic support program but had not enrolled in the SRL course. The control students were chosen by matching them to the SRL-course students on GPA, prior credits, and gender. Preliminary analyses showed the two groups did not differ significantly on prior GPA, prior credit hours, transfer credit hours, or gender. After controlling for gender and prior GPA, the SRL-course students had significantly higher GPAs four semesters after the course than did the control students and were less likely to get an F grade in a subsequent semester. The SRL-course students were also more likely to graduate within seven years (95% vs. 74%). In a logistic regression model, SRL course enrollment and prior cumulative GPA were statistically significant predictors of graduation. The odds ratio estimate for SRL course enrollment was 12.69, meaning that the odds of SRL-course students graduating were nearly 13 times the odds of control students graduating (Bail et al., 2008).

The next two studies evaluated programs aimed at probationary students. In the voluntary Project Success program (Humphrey, 2006), participants met weekly in small groups. Each group had two facilitators (faculty, staff, or graduate student) and one peer facilitator (previous probationary student who had completed Project Success and regained good academic standing). The program was not explicitly based on a self-regulated learning framework, but emphasized goal setting, planning, strategies, and reflection. Each week, students recorded an academic goal, a non-academic goal, and their progress on the previous week's goals. They also completed a weekly report on an academic topic (e.g., class attendance or visiting with professors) and a weekly reflective journal entry on a topic chosen by the group. Using retention, academic good standing, and GPA as outcome variables, cohorts of Project Success students were compared to GPA-matched cohorts of probationary students who had not volunteered for Project Success. This was a secondary summary of an internal research study, and did not describe the statistical analysis or give information on how departing students were handled in the analysis. Possibly for that reason, there were inconsistencies and ambiguities in the numbers provided. Still, there seems to be some evidence that the Project Success students were retained at higher rates and suspended at lower rates. Because the program was voluntary, it is difficult to know whether the improvement was due to the program or to initial attitude or motivation differences between Project Success students and control students.

Kamphoff et al. (2007) evaluated Strategies for Academic Success (SAS), a mandatory program for all students placed on academic probation after their first semester. The conceptual model was a table with four legs: personal responsibility, positive affirmations, goal setting/life planning, and self-management. The legs supported a table top composed of small group interaction and individual interaction with a facilitator. The aim of the goal-setting leg, in particular, was to help students become effective self-regulated learners. During the last two of the four years described in the article, the program had more serious “teeth” than most such programs—if students failed to sign up for the program, or if they missed a SAS class meeting, they were suspended and withdrawn from all their other classes.

For this study, academic-probation students (GPA below 1.50) required to enroll in SAS were compared to a control group composed of academic-warning students (GPA between 1.50 and 1.75) not required to enroll in SAS. The researchers controlled for initial differences through a covariate composed of prior GPA, high school GPA, and SAT scores. Separate statistical analyses were conducted for cohorts from four academic years (2000–2004). Cohorts from the first two years showed no significant differences in next-semester GPA. Cohorts from the last two years showed substantial differences in next-semester GPA, in favor of the SAS students. This improvement coincided with changes in the program, including required meetings with the instructor and severe penalties for nonattendance. A detailed description of the sample and the results was only provided for the Spring 2003 cohort. For this cohort, the SAS students (*n* = 309, academic probation) gained an average of 0.7309 on their subsequent semester’s GPA, while control students (*n* = 80, academic warning), had a GPA gain of only 0.4202. This difference could be partially due to the probationary students’ lower initial GPAs—because they started lower, they had more room to improve than the academic warning students in the control group.

Instead of evaluating the overall effect of a course or program, the next study evaluated a single component: individualized feedback about inventory results (Haught et al., 1998). Students in a study strategies course were randomized into two feedback conditions. All students completed the Learning and Study Strategies Inventory (LASSI) at the beginning of the course. After receiving their results, all students received generalized feedback in a whole-class setting. Using a script, instructors discussed each of the 10 LASSI subscales (e.g., Motivation, Time Management, Self-Testing) and conveyed suggestions on how they could improve in each area. Students were instructed to pay special attention to the subscales on which they scored below the 50th percentile. This whole-class general feedback was the only LASSI feedback the control students received. In addition to the whole-class general feedback, students in the treatment group left class for a short individual counseling session with one of the researchers. For each below-50th-percentile subscale, the researcher used a script to provide detailed suggestions on how the student could improve.

For both groups, student scores on the post-course LASSI were generally higher than on the pre-course LASSI (Haught et al., 1998). However, on the post-course LASSI, the individually counseled students scored significantly higher than control students on seven of the 10 subscales. They also had significantly fewer below-50th-percentile subscales than the control students. Individually counseled students also fared better on GPA. Their intervention semester GPAs and subsequent-semester cumulative GPAs were significantly higher than those of control students. Though the individually counseled students also had higher grades in the study strategies course and higher GPAs in the subsequent semester, these differences did not reach statistical significance. This study seems to indicate that a single counseling session can positively affect student success, if only in the short term. It would have been helpful if the researchers had recalculated the intervention semester GPA and subsequent-semester cumulative GPA with the study strategies course removed from the calculation, so we could see whether the GPA improvement was due solely to the study strategies course.

By including an achievement variable and a reasonably comparable control group, and making some effort to control for confounding variables, the previously discussed studies by Ahuna et al. (2011), Ryan & Glenn (2003), Bail et al. (2008), Kamphoff et al. (2007), and Haught et al. (1998), and to a lesser degree, Humphrey (2006), contribute credible evidence about the value of stand-alone learning strategies or SRL classes. Though the studies vary in both results and rigor, when taken as a whole, they support the hypothesis that coaching or instruction in self-regulated learning strategies can help college students become more effective learners. The remaining nine of the 15 overall-SRL studies are of less value, due to their lack of an achievement variable, lack of a viable control group, or the degree of confounding by other factors. For that reason, they will be discussed in less detail.

Studies Lacking an Achievement Variable, a Viable Control Group, or Control for Confounders

A common approach is to administer an inventory before and after the intervention, and to see whether the scores improved from pretest to posttest. Two such studies (Schapiro & Livingston, 2000; Reeves & Stich, 2011) have already been mentioned. They used learning inventories and helpfulness surveys to evaluate the same Methods of Inquiry course as Ahuna et al. (2011) evaluated for its effect on graduation and retention. In another such study, Hofer and Yu (2003) used the Motivated Strategies for Learning Questionnaire (MSLQ) to evaluate a course covering principles of motivation, research on learning, and strategies for effective learning. In addition to lectures, students attended small discussion sections and kept a reflective journal reporting on their progress in both the learning strategies course and another target course. Paired *t* tests showed statistically significant improvements on four of the six motivational subscales and six of the seven cognitive subscales of the MSLQ, though only one pretest subscale and one posttest subscale were significantly correlated with the final grade in the course (Hofer & Yu, 2003).

Three other studies used pretests and posttests to evaluate an intervention's effectiveness. However, instead of evaluating an entire learning strategies or SRL course, they evaluated a specific learning activity. One such example is Lee (2007). Following a template, students in a learning strategies course analyzed case studies about academic learners. A critical thinking rubric was used to evaluate the quality of their case study essays. Not surprisingly, the quality of the students' case studies improved on the posttest, after they received the template and analyzed three case studies for homework. In another example, some students in education and educational psychology classes watched a videotape lesson called Academics Anonymous (Orange, 1999). Control students did not see the video. In the video, student actors played the roles of students who came to a support group meeting hoping to become higher achieving self-regulated learners. The students committed to an action plan based on twelve steps of self-regulated learning. A researcher-created self-regulated learning inventory was administered before and after the intervention. Though the control group scored higher than the treatment group on both the pretest and the posttest, treatment students exhibited a greater gain in scores, leading the researcher to conclude that the intervention was effective. In another example, students in a management class completed four tutorials, each on a different aspect of self-management: self-assessment, goal setting, self-monitoring, and self-regulation (Gerhardt, 2007). Their scores on a self-management skill inventory improved after the training, and most students rated the training as very helpful.

The last three overall-SRL studies also have limited value as SRL research, for reasons other than the lack of an achievement-based outcome measure. In one, community college mathematics instructors were surveyed about their interest in SRL-enhancing teaching techniques and their willingness to participate in a research study (Travers, Sheckley, & Bell, 2003). The nine teachers who expressed interest were trained by a researcher in several techniques designed to enhance SRL. Students in these SRL-trained teachers' classes formed the treatment group, and students whose instructors did not receive the SRL training became the control group. The groups showed no differences on a self-regulated learning inventory given after the semester. This study had many shortcomings, including the method of treatment assignment and absence of effort to control for student differences, particularly important because all classes were at different campuses.

Hopper (2011) describes a faculty-driven strategies class designed to supplement an anatomy and physiology (A&P) course. The Supplement course was mandatory for A&P course repeaters and optional for first-time A&P enrollees. In it, students learned to set goals, plan their time, and set up their notebooks. They also learned A&P-specific study skills, reinforced with A&P-specific activities, such as using bone boxes and labeling neuromuscular system diagrams. A higher percentage of students in the Supplement class earned a C or better, compared to students not in Supplement; a lower percentage of Supplement students withdrew. On a faculty-created survey about effective behaviors for learning A&P, Supplement students generally improved from pretest to posttest. Some individual items showed significant improvement, others did not. On a helpfulness survey, student comments about the Supplement's value were very positive. This study may very well contribute valuable information to researchers or practitioners interested in interventions to improve A&P performance. However, it is not especially useful as SRL research, because it is impossible to know how much of the Supplement students' improved performance is due to the teaching of SRL strategies and how much is due to the Supplement students' increased exposure to the A&P content. Also, as with all optional courses, the threat of self-selection applies—students signing up for the course may be more motivated than students choosing not to enroll.

Sweidel (1996) describes a study strategy portfolio used in an educational psychology course. For each class test, students completed two short-answer surveys and two journal entries. These involved creating a written study plan, committing to follow it, reflecting upon the study plan, predicting test grades, and reflecting on actual test results. Entries received feedback and were graded for quality. Students also wrote a 3–5 page reflective essay about the whole course. No statistical analyses were mentioned, but student responses to a helpfulness survey were extremely positive. Over 80% of the students felt they had made positive changes to their studying because of the portfolio, and that they would recommend it be used in future classes. These positive responses were especially meaningful considering that 83% said they found the project difficult to complete, only 44% said they enjoyed it, and many expressed criticism of the amount of time the project required. The lack of statistical analysis is acceptable, as this article was primarily intended as a "how-to" article for educators, rather than a serious research article. With its extremely clear descriptions of the surveys, journals, and feedback, it served its purpose well.

Interventions Targeting One or Two Specific Dimensions of SRL

Table 3 summarizes the 27 intervention studies that focused primarily on one or two of Zimmerman's self-regulated learning dimensions. As previously mentioned, Zimmerman's framework has six dimensions, each associated with a key word: Time (When), Strategies (How), Outcomes (What), Motivation (Why), Social Context (Who), and Environment (Where). At least in this application, the boundaries between the dimensions were often fuzzy. The dimension classification was intended to ease discussion and facilitate seeing patterns, not to definitively categorize studies.

Table 3

Interventions Targeting One or Two Zimmerman Dimensions

Author(s) (Year)	N $(n_1/n_2/\ldots)^a$	Subject	Intervention	Dimension(s)^b
Acee & Weinstein (2010)	82 (41/41)	Statistics	Value-reappraisal intervention. Students received computerized messages about the value of learning statistics and completed exercises designed to increase their motivation.	Why
Andrade & Du (2007)	14	Educational psychology	Criteria-referenced self-assessment rubric submitted with assignments. Students recorded whether their work met the criteria and circled the quality gradation they felt best described their work.	What
Bercher (2012)	77	Anatomy & Physiology	Self-assessment sheet at each lab session, with percentages of perceived mastery for each learning objective. Post-exam reflection sheet after discussing exam grade with instructor.	What
Brothen & Wambach (2000)	168	Psychology	Exercise Completion Record to record completion of assignments in a computer-based class.	What
Cao & Nietfeld (2005)	94	Educational Psychology	Weekly self-monitoring sheet. Students rated their understanding, listed difficult concepts, made plans for improving their understanding. Students gave confidence ratings on content questions and predicted their final exam scores.	What
Cho, Cho, & Hacker (2010)	601	Various	Electronic system to scaffold self-monitoring and peer-monitoring of writing. Students compared their own assessments and peer assessments on both their own writing and on others' writing.	What
Cisero (2006)	483 (166/317)	Educational Psychology	Reflective journal over course material. Graded for authentic reflection, clarity, and format.	How

Table 3 Continued

Author(s) (Year)	*N* (n_1/n_2/…)[a]	Subject	Intervention	Dimension(s)[b]
Commander, Valeri-Gold, & Darnell (2004)	127 (20//73/34)	Psychology	Strategic Thinking and Learning (STL) Learning Community geared toward effective study strategies for introductory psychology class. Students analyzed their time management and its connection with results.	When How
Dietz-Uhler & Lanter (2009)	107 (50/57)	Psychology	Web-based interactive learning activity. Computer-guided prompts for the four-question method: (1) analysis, (2) reflection, (3) applying concept to personal life, (4) questioning.	How
Einstein, Mullet, & Harrison (2012)	52 (52/52)	Psychology	Self-testing demonstration. Students studied one passage by simply "studying" and another by "study then self-test." Students analyzed the resulting data and wrote reports.	How
Fitch, Marshall, & McCarthy (2012)	69 (__/__)	Psychology	Self-guided group meetings, with 4-6 students per group. Goals were set, shared, discussed, and tracked using Solution-Focused Goal Setting Worksheet. Students took turns facilitating meetings.	Why Who
Georgianna (2009)	18 (9/9)	Dev. English (Writing)	Students created Implementation Intentions with "if…then…" action plans to complete academic tasks and overcome obstacles.	Why
Goodwin & Califf (2007)	93 (48/45)	Computer Science	Two time-management training sessions by time-management expert (not the instructor). Emphasized goal setting and prioritizing, planning, and recording time. Weekly worksheets to record planned and actual time.	When
Grabe & Flannery (2010)	171	Psychology	Online study questions with confidence ratings about correctness of answers. Points earned or lost based on correctness and confidence levels.	What

Table 3 Continued

Author(s) (Year)	*N* (n_1/n_2/…)[a]	Subject	Intervention	Dimension(s)[b]
Hacker, Bol, & Bahbahani (2008)	137 (2×2)	Educational Psychology	Students predicted (estimated before taking exam) and postdicted (estimated after taking exam) their exam scores with or without extrinsic incentives (extra exam points for accurate predictions/postdictions) and with or without a reflection survey.	What
Hartlep & Forsyth (2000)	52 (_/_//_)	Psychology	Some students instructed to use SQ4R method: Survey, Question, Read, Recite, Review, Reflect. Other students instructed to use only the *Read* and *Reflect* parts of SQ4R: "Reflect about how the reading materials relate to their life experiences." Control students could use any study method.	How
Hilton, Wilcox, Morrison, & Wiley (2010)	162 (21/16/29//96)	Religion & Philosophy	Four different methods for self-grading completion of required class readings.	Why What
Kauffman (2004)	119 (2×2×2)	Educational Psychology	Free-form vs. matrix organizer note-taking conditions, with and without self-monitoring prompts, and with and without feedback designed to boost academic self-efficacy.	How What
Kauffman, Ge, Xie, & Chen (2008)	54 (2×2)	Educational Psychology	Preservice teachers analyzed case studies with authentic classroom problems and wrote solution suggestions, with or without computerized problem-solving prompts and reflection prompts.	How What
Kitsantas & Baylor (2001)	114 (__/__)	Educational Technology	Instructional planning self-assessment rubric for preservice teachers to self-monitor their lesson plans.	What

Table 3 Continued

Author(s) (Year)	N ($n_1/n_2/\ldots$)[a]	Subject	Intervention	Dimension(s)[b]
Kwon, Kumalasari, & Howland (2011)	47 (24/23)	Web Development	Two types of computerized explanation prompts. In both, students rated their confidence in the correctness of their explanations.	How What
MacArthur & Philippakos (2013)	48	Dev. English (Writing)	Writing strategy instruction which included goal setting, monitoring, and reflection. Students wrote journal entries about goals and discussed them in class.	Why How
Schwartz & Gredler (1998)	31 (15/16)	Educational Psychology	Four weekly self-instruction lessons on goal setting, each with an application exercise.	Why
Stanger-Hall, Shockley, & Wilson (2011)	270 (93//90/87)	Biology	Workshop with two demonstration exercises. One showed the value of visualization and the other showed the value of self-testing.	How
Tuckman (2007)	93 (__/__)	Study Strategies	Study skills support groups met online in real time. Students took turns playing role of supporter/sponsor, reviewing another student's weekly checklist, met/unmet goals, and time management. Also real-time meetings and office hours with instructor.	When Who
Williams, Aguilar-Roca, Tsai, Wong, Beaupré , & O'Dowd (2011)	1227 (1227/1227)	Biology	Learn from Exam assignment targeting midterm questions missed by 40% or more of the class. Students summarized when and where the content could be found, and reasons answers were incorrect or correct.	How

Table 3 Continued

Author(s) (Year)	N $(n_1/n_2/\ldots)$[a]	Subject	Intervention	Dimension(s)[b]
Ziegler & Moeller (2012)	168 (85/42//41)	Foreign Language	LinguaFolio system for goal setting, reflection and self-assessment. Students set an achievement goal and a personal goal related to learning the foreign language, and submitted evidence they had met the goals.	Why What

[a] If applicable, second line of the sample size column describes the breakdown of the sample N into smaller groups. The group best characterized as treatment group is listed first; the group best characterized as control group is listed last. If there are more than two groups, a double slash (//) indicates the division that best separates treatment from control. Underscores within the parentheses indicate that the author provided the total sample size but not the sizes of the groups. If the two group sizes are both equal to N, this indicates that each participant served as his or her own control, by applying different conditions to different tasks (e.g., Einstein et al., with $N = 52$ (52/52), had 52 total participants). Notation such as 2×2 or 2×2×2 indicates a crossed design with approximately equal group sizes. [b] Dimensions are represented using keywords from Zimmerman's dimensional framework (1994, 1998): Why = Motivation; When = Time; How = Strategies; What = Outcomes; Who= Social Context; Where = Environment.

About half the dimension-specific studies were associated with psychology-related classes, one study examined a single component of a study strategies course, and all the others were associated with a course in a specific discipline. Fourteen of the 27 studies drew their participants from psychology or educational psychology classes. Preservice teachers formed the sample for one study of educational technology classes and two of the aforementioned studies in educational psychology. Other subject areas for interventions were biology (two studies), anatomy and physiology (one), computer science (two), statistics (one), religion and philosophy (one), and foreign languages (one). Three interventions were intended to improve students' writing skills. Two of these were in developmental (remedial) English classes. The other was a large-scale online intervention aimed at students writing papers for classes in a variety of content areas, including psychology, physics, and history.

Table 4 summarizes the results of the 27 dimension-specific studies and the outcome measures used. As before, an outcome measure was designated as an achievement measure if it was based on the learning of content for a course other than a learning strategies course. As mentioned in the section on search criteria, studies were omitted from this review if they involved students learning content unrelated to the course they were taking (e.g., psychology students learning about wildcats). If the study involved learning content at least marginally relevant to the class the participants were taking, the study was included in this review. However, a content quiz only counted as an achievement measure if it was based on material the students needed to learn for a course grade.

Table 4

Outcome Measures and Results For Dimension-Specific Studies

Author(s) (Year)	Outcome Measure	Results
Acee & Weinstein (2010)	MSLQ inventory (Motivated Strategies for Learning Questionnaire) Endogeneous instrumentality inventory Self-efficacy inventory Exam 3 grade[a] Interest in statistics	Students in value-reappraisal (VR) condition showed higher interest in statistics. Strong condition×time interaction for task value and endogeneous instrumentality, with VR students showing gains from pretest to immediate posttest and from pretest to delayed posttest. No significant group differences in self-efficacy. Significant instructor×condition interaction on postintervention exam performance, with one instructor's VR students outscoring the same instructor's control students.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
Andrade & Yu (2007)	Student perceptions of helpfulness	Qualitative analysis showed (1) attitudes toward self-assessment improved with experience, (2) self-assessment and teacher expectations are inextricable, (3) clear expectations make self-assessment more likely, (4) students self-assess by checking, revising, reflecting, (5) criterion-referenced self-assessment produced better work and higher grades, (6) spotty transfer of self-assessment, (7) tension between teacher assessment and self-assessment, (8) no evidence of gender differences.
Bercher (2012)	Exam grades[a] Helpfulness survey	The 13% of students who said the SSAS was not at all helpful averaged 73% on the exam. The 87% who said it was somewhat helpful, helpful, or very helpful averaged 81%.
Brothen & Wambach (2000)	Course grade[a]	A recording score was calculated for each student: 2 if all quiz scores were recorded correctly, 1 if at least one quiz score was recorded, and 0 if no quiz scores were recorded. The recording score accounted for 0.062 of the variance in final grade.
Cao & Nietfeld (2005)	Judgment of learning Calibration accuracy Bias level of ratings	Judgment of Learning (JOL), Performance, Confidence, and Accuracy improved throughout the course, but Bias did not. Relationships between JOL and Accuracy, and between JOL and Performance strengthened throughout the semester.
Cho, Cho, & Hacker (2010)	Writing quality[a] Average essay score[a] from peer-evaluations Writing improvement[a] Self-monitoring accuracy Self-monitoring improvement	Magnitude of the difference between self-evaluation and others' evaluations showed a significant increase from first draft to second draft, indicating that students' self-monitoring skill worsened. Strong correlation between self-monitoring improvement and writing quality improvement.
Cisero (2006)	Exam average[a]	Exam averages for the semesters in which the reflective journal was implemented were very similar to exam averages from previous semesters. Chi-square analysis showed students in the intervention semester had significantly fewer C and D grades than students in previous semesters.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
Commander, Valeri-Gold, & Darnell (2004)	Course grade[a]	Students in the Strategic Thinking and Learning (STL) Learning Community (LC) had significantly higher psychology course grades than students in the Education LC or not in an LC, and marginally higher psychology grades than students in the Leadership LC. No significant grade differences between the STL LC and the Understand LC.
Dietz-Uhler & Lanter (2009)	Multiple-choice content quiz Helpfulness survey	On both topics, doing the content reflection questions before the quiz resulted in better quiz scores. Students thought the activity was successful in meeting its goals and in improving memory and enjoyment.
Einstein, Mullet, & Harrison (2012)	Content quiz over content not relevant to the course Multiple-choice quiz about the benefits of testing Confidence ratings	Significant main effect for condition, with *Study-Test* scoring better than *Study-Study*. No main effect for order. One week after the exercise, 92% answered the testing-benefit question correctly (compared to 36% before).
Fitch, Marshall, & McCarthy (2012)	MSLQ inventory	Composite MSLQ score was significantly higher for the treatment group (effect size 0.56). When adjusted for pretest scores, treatment group scored significantly higher on self-efficacy and intrinsic motivation subscales but not on test anxiety, cognitive strategy use or self-regulation.
Georgianna (2008)	Mini-essay grades[a] Quality of implementation intentions Final course grade[a] Helpfulness survey	Treatment students scored higher on mini-essays than control students. Groups did not differ in their final grades or success rates.
Goodwin & Califf (2007)	Course grade[a] Final course average[a] Exam grades[a]	No significant difference in overall success rates. When restricted to only students taking Exam 1, or to only students taking Exam 2, success rate was significantly better for treatment students than for control students. Treatment students had significantly higher Exam 1 grades, Exam 2 grades, Final Exam grades, and final course averages. No difference between pre-semester and post-semester Time Management Behavior Questionnaire scores. Perceived Control of Time, Treatment, and GPA were the best predictors in a model to predict the final course average.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
Grabe & Flannery (2010)	Exam performance[a] Exam score prediction accuracy Self-monitoring accuracy for study questions Study question performance[a]	Proportion of low exam performers categorized as nonparticipants (for attempting less than 45 study questions) increased after first exam. High exam performers did better on the study questions; low and middle exam performance groups did not differ from each other. Only Exam 1 showed a difference in self-monitoring accuracy between low, middle and high performers. Exam grade prediction accuracy and study question performance were significant and independent predictors on exam questions covering book content.
Hacker, Bol, & Bahbahani (2008)	Calibration accuracy Attributional Style Questionnaire	Neither the reflection condition nor the extrinsic incentives condition led to improvement in calibration. Students did not improve their calibration accuracy with practice (calibration accuracy did not vary across Exams 2 and 3).
Hartlep & Forsyth (2000)	Multiple-choice test from test bank over the material	No significant differences between groups on the test immediately after the study session. Significant difference between groups on the test taken 2 weeks after the study session. The control group scored significantly lower than both the SQ4R group and the *Read-Reflect* group, but the SQ4R and *Read-Reflect* groups were not significantly different. All groups declined from first to second test, but only the control group showed a significant decline.
Hilton, Wilcox, Morrison, & Wiley (2010)	Survey about motivation and assigned reading completion	Students given a minutes requirement read the most, followed by the students setting their own goals. Students self-grading themselves in some manner completed higher percentages of reading than students not required to self-grade. Significant differences among groups on minutes/day, percentage complete, and motivation due to grades; no difference on days/week or perceived enrichment. When outcomes were combined, students in the various self-grading conditions did not differ from each other, but differed from those not asked to self-grade. Students not asked to self-grade made the fewest positive comments about the readings helping their personal study.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
Kauffman (2004)	Academic self-efficacy inventory MAI inventory (Metacognitive Awareness Inventory) Quizzes over reading materials Amount of information in notes	Students receiving feedback designed to boost self-efficacy did not change their self-efficacy; students not receiving the feedback increased their self-efficacy. MAI showed no effect from self-monitoring prompts. Matrix note-takers recorded significantly more propositions than free-form note-takers. On achievement test, both matrix note-taking and self-monitoring prompts produced significant main effects in the positive direction. Significant self-monitoring×note-taking interaction; self-monitoring prompts made a positive difference for matrix note-takers but not freeform note-takers.
Kauffman, Ge, Xie, & Chen (2008)	Skill in solving classroom problems[a] Writing quality	Both problem-solving skill and writing quality showed a significant main effect for problem-solving prompts; students receiving problem-solving prompts scored higher. Significant interaction between reflection prompts and problem-solving prompts; students receiving reflection prompts scored higher than those not receiving reflection prompts, but only if they also received problem-solving prompts.
Kitsantas & Baylor (2001)	Author-created inventory (on self-efficacy and disposition toward instructional planning, and on its perceived importance) Quality of instructional plan for case study[a]	On posttest achievement quiz (creating instructional plan), treatment group scored significantly higher than control group. Control group improved significantly after the IPSRT rubric was demonstrated. After intervention, treatment students were more positive in their dispositions toward instructional planning. No change between pretest and posttest scores on instructional planning self-efficacy for either group. In treatment group, those who started with low self-efficacy improved, and those with high initial self-efficacy decreased after the intervention; same results in control group after the IPSRT demonstration.
Kwon, Kumalasari, & Howland (2011)	Debugging confidence Correctness of student explanations[a] Midterm exam multiple-choice content questions[a] Midterm exam debugging task[a] Time spent on debugging practice and midterm debugging practice Quiz on HTML concepts[a]	On midterm, Open Self-Explanation (OSE) group solved more errors correctly than the Complete Other-Explanation (COE) group. No difference between groups on multiple choice exam questions, time spent debugging, or percentage of correct explanations during debugging practice. OSE group was significantly more confident in their explanations. During practice phase, OSE group had strong positive correlation between correctness of explanation and debugging performance; COE group did not. Confidence level did not correlate with debugging performance for either group.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
MacArthur & Philippakos (2013)	Quality of persuasive essay[a] Motivation inventory Interviews about goals, strengths, weaknesses, strategies, what they learned	Observations showed that self-regulation strategies were taught with acceptable fidelity in Round 2 but not Round 1. In interviews, 11 of 12 Round 1 students were positive overall about the course and the amount they learned. Round 2 students showed significant gains in writing ability, self-efficacy and mastery motivation, and a significant decrease in performance motivation. Not all students liked writing afterward, but they had a positive experience in the class, thought they learned a lot and had more confidence and better attitude toward writing. Knowledge of writing strategies was better for Round 2 students than Round 1 students.
Schwartz & Gredler (1998)	SESRL inventory (Perceived Self-Efficacy for Self-Regulated Learning) Goal-setting inventory Goal analysis skill	SESRL scores declined for both groups. No significant differences between groups on overall SESRL or on goal-setting habits inventory. Treatment group scored significantly higher on goal analysis skill test. Low-SESRL students were not successful at creating long-term goals with supporting subgoals.
Stanger-Hall, Shockley, & Wilson (2011)	Quality of drawings[a] Final exam questions on workshop topics[a] Helpfulness survey	During self-testing exercise, students' drawings improved from 1st to 2nd attempt. Workshop group scored significantly higher on most life-cycle final exam questions than both the same-semester control group and the previous-year control group. Previous-year control group outperformed workshop group on overall final exam grade. Low- and high-GPA students did not differ in their performance gains from the workshop. Most students thought the visualization exercise was more helpful than the self-testing exercise.
Tuckman (2007)	Overall course average[a] GPA gain[a] Helpfulness survey	No significant effect of treatment condition or procrastination tendency on either course average or GPA. For both course average and GPA, a significant interaction effect between treatment and procrastination tendency (the support groups helped the high procrastinators more). On helpfulness survey, no group differences on numerical ratings. On open-ended questions, one-third of the treatment students characterized the online sessions as not useful.

Table 4 Continued

Author(s) (Year)	Outcome Measure	Results
Williams, Aguilar-Roca, Tsai, Wong, Beaupré , & O'Dowd (2011)	Quality of exam analysis homework[a] Topic-matched final exam questions[a] Helpfulness survey	Each student was in a Topic Analysis group for some questions and in a Control Analysis group for other questions. On 5 of the 11 questions, the proportion of students answering correctly was significantly higher in the Topic Analysis group than the Control Analysis group; the other 6 questions did not show significant differences. Assignment most beneficial when the final exam question had the same emphasis as the midterm question analyzed (as opposed to being identical or having a different emphasis). Students who missed midterm questions and submitted a "strong" exam analysis homework were more likely to get the corresponding final-exam problem correct than students submitting a "weak" exam analysis homework. Most students thought the assignment was helpful and made them look at their midterm more carefully.
Zeigler & Moeller (2012)	MSLQ Inventory PALS Inventory (Patterns of Adaptive Learning Survey) Accuracy of self-assessment Exam grades[a] Course grade[a]	Spanish (treatment) students divided into Extensive LinguaFolio group (ELF; instructors came to at least 1 Q&A meeting with researchers) and Limited LinguaFolio group (LLF; instructors did not come to a Q&A meeting). Significant difference on mastery goal orientation between ELF and LLF with LLF group significantly decreasing their mastery goal orientation over the semester. ELF scored higher on task value than French (control) group. No differences on other subscales. LLF and ELF groups had significant correlations between chapter test scores and self-judgments of their ability. ELF students' self-judgments were more accurate than LLF students.

[a]Outcome is classified as an achievement measure.

Seventeen of the 27 studies incorporated an achievement measure in their evaluation of the intervention's effectiveness. With one exception, the achievement measures were specific to the course addressed by the intervention—exam grades, writing assignment scores, or course letter grades. The one exception was Tuckman (2007), which used semester GPA to evaluate the effect of an online component of a study strategies course. Of the 10 studies not using an achievement measure, two used calibration accuracy as the primary outcome variable, one used goal analysis skill, four used content quizzes over learning materials ancillary to the class, one used a survey about completion of assigned readings, and one used a learning strategies inventory. The tenth used qualitative data to learn about student perceptions of the intervention. Many of the studies, both with and without achievement measures, also used an inventory or helpfulness survey in addition to the primary outcome variable.

The Venn diagram in Figure 1 illustrates how many of the 27 studies addressed each of the six dimensions of self-regulated learning. This Venn diagram visually summarizes the last column of Table 3, which lists the dimension(s) addressed by each study. The layout for the Venn diagram emerged from the data and was not driven by any theoretical model other than Zimmerman's list of dimensions. I chose this particular arrangement for the circles not to show relationships, but to facilitate frequency counts. The Venn diagram should not be used to infer that certain dimensions overlap or do not overlap (e.g., it should not be seen as implying that the What and When dimensions are mutually exclusive). Note that because the Venn diagram is a two-dimensional representation of a six-dimensional framework, it is incapable of capturing the framework perfectly. The six-dimensional framework, in its turn, is only a representation of something more complicated.

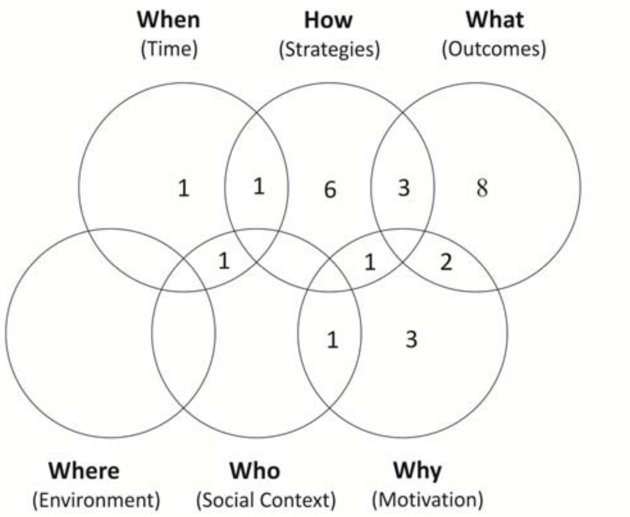


Figure 1. *Venn diagram depicting the 27 studies targeting one or two Zimmerman dimensions.*

The most frequently targeted dimension was the Outcomes (What) dimension. This dimension includes self-monitoring of performance and self-evaluation of learning outcomes (Zimmerman, 1998). Another frequently targeted dimension was the Strategies (How) dimension. Interventions evaluating a specific learning strategy fell into this category. Fewer of the dimension-specific studies addressed the Time (When) and Motivation (Why) dimensions. This does not necessarily indicate neglect of these dimensions by researchers, because nearly all the previously discussed overall-SRL interventions contained both a goal-setting (motivation) component and a time-management component. The same cannot be said of the final two dimensions, Social Context (Who) and Environment (Where). Only two of the dimension-specific studies addressed Social Context, both in conjunction with another dimension, and none addressed self-regulation of the learning environment. In the overall-SRL studies, these two dimensions were also mentioned much less frequently than the other four dimensions.

The following paragraphs provide more detailed information about both the interventions and the results of the studies. The discussion will proceed through the Venn diagram, one circle at a time, starting with the Time (When) dimension. Those interventions that lie within the intersection of two circles will be discussed under whichever circle they fit best. This textual tour through the Venn diagram is an attempt to linearize a pattern that is not linear—at best, this goal can be accomplished only imperfectly.

Time (When) Dimension

The Time circle of the Venn diagram includes three studies. The first, a time-management training intervention for students in a low-success computer programming course (Goodwin & Califf, 2007), concentrated strictly on the Time dimension. The second, peer-accountability meetings for an online study strategies course (Tuckman, 2007), combined two dimensions: Time (When) and Social Context (Who). The third is a learning community that tailored its learning strategies and time-management instruction to an introductory psychology course (Commander, Valeri-Gold, & Darnell, 2004). On the Venn diagram, this study is located in the intersection of the Time (When) circle and the Strategies (How) circle.

Though all produced positive results, the time-management training intervention described by Goodwin and Califf (2007) provides stronger evidence than the other two. In this study, two instructors each taught two sections of a computer programming course. One class in each instructor’s pair of classes was randomly assigned to the treatment condition and the other to control. A trained time-management instructor from the university’s learning center visited each treatment class for two 20-minute training sessions, an initial session the second class week and a refresher during the sixth class week. The trainer emphasized the students’ ability to control their time and the importance of investing 9–12 hours weekly in the class. The trainer explained the need for students to set goals and prioritize tasks, and also to plan, record, and analyze their use of time. Treatment students submitted weekly time-management worksheets, on which they recorded their planned and actual time use. In the treatment classes, the instructors reinforced the training by imposing strict deadlines, penalizing late work, awarding points for the weekly time sheets, and regularly commenting about the importance of time management. To discourage procrastination, they required part of a two-week programming assignment to be submitted after one week. Control students did not receive the training sessions or the time-management worksheets. A chi-square test comparing the success rates of officially enrolled students did not show a statistically significant difference between groups. However, when restricted to students who took Exam 1, the treatment students succeeded (earned an A, B, or C) at significantly higher rates than control students. When restricted to students who took Exam 2, the treatment students were also more successful. On Exam 1 grade, Exam 2 grade, and final course average, *t* tests showed significant differences in favor of the treatment students (Goodwin & Califf, 2007).

One study is included in both the Time (When) and Social Context (Who) dimensions. Tuckman (2007) sought to learn whether social influence could help students improve their regulation of time, particularly students who tended to procrastinate. Students in an online study strategies course were randomized into treatment and control conditions. Those in the treatment condition were required to “meet” weekly in real-time with their instructor and a small support group of their classmates. During the online group meeting, students took turns playing the role of supporter/sponsor and the role of student partner. The supporter's task was to help the student partner manage time by reviewing weekly to-do lists and task accomplishment lists. For accountability, students were required to send these checklists to their partners in advance. A procrastination inventory was used to classify students as high procrastinators or low procrastinators. There was no significant effect of either treatment condition or procrastination tendency on either course average or GPA gain. However, for both outcomes, there was a significant interaction effect between treatment condition and procrastination tendency, with the high procrastinators in the treatment condition outperforming the high procrastinators in the control condition. Low procrastinators’ outcomes did not differ by group. These results indicate that the checklists and social support may have improved the time management of the high procrastinators. Unfortunately, the study did not report sufficient information to ascertain whether the high procrastinators’ GPA gain was due only to the grade in the study strategies course, or whether the students’ other course grades improved also. The study would be more valuable if the researcher had either gathered GPA data for a subsequent semester, or recalculated the intervention semester GPA with the study strategies course omitted.

The learning community examined by Commander, Valeri-Gold, and Darnell (2004) emphasized both the Time (When) and Strategies (How) dimensions. The learning community model, in which students take several courses together as a cohort, has been increasingly used by colleges to promote engagement and a sense of community. In general, this review does not cover research on learning communities. However, I included this study because it evaluated the effect of a particular learning community's strategy instruction, rather than the effect of the learning community model. The Strategic Thinking and Learning (STL) learning community's freshman orientation class was geared specifically toward effective study strategies for an introductory psychology class. As orientation class topics were covered, they were connected to psychology concepts. Students were asked to analyze their time management and connect it to their psychology exam results. Psychology course grades of STL students were compared to psychology grades of students in three other learning communities and those not in a learning community. The STL students had the highest psychology grades of all the groups, significantly outperforming ($p < .05$) the students not in a learning community and the students in one other learning community, and marginally outperforming ($p < .10$) those in another learning community. While these results seem promising, it is difficult to know whether they came from the strategy instruction or from increased exposure to psychology topics.

All the studies in the Time dimension reported positive results. The most solid evidence is found in Goodwin and Califf's (2007) study of time-management training for computer science students. Because the outcome measure was achievement-based and the time-management training was not accompanied by additional content instruction, this study supports the argument that training can help students self-regulate their time and become more successful. Even without a true achievement variable, Tuckman's (2007) study of online support groups also provides reasonably solid evidence that peer support and accountability can help procrastinating students improve their regulation of time. The study by Commander et al. uses a very interesting approach and would probably be useful to a researcher reviewing the characteristics of effective learning communities. It also makes a good argument for connecting the contents of a college orientation class to a specific target class. However, from a "can SRL be taught?" perspective, this study is not especially helpful, because the SRL instruction is so entangled with additional content instruction.

Strategies (How) Dimension

Eleven studies fell into the Strategies circle of the Venn diagram. One was the previously discussed learning community, which incorporated both study strategies and time management (Commander et al., 2004). In addition, the Strategies circle includes three studies of reflection as a learning strategy (Cisero, 2006; Dietz-Uhler & Lanter, 2009; Hartlep & Forsyth, 2000), two studies of learning activities designed to show the value of self-testing (Einstein, Mullet, & Harrison, 2012; Stanger-Hall, Shockley, & Wilson, 2011), and one study of an exam-analysis homework assignment (Williams et al., 2011). The Strategies circle also includes three studies that combined learning strategies with self-monitoring (Kauffman, Ge, Xie, & Chen, 2008; Kauffman, 2004; Kwon, Kumalasari, & Howland, 2011), which will be discussed under the Outcomes dimension, and one study that combined learning strategies with goal-setting instruction (MacArthur & Philippakos, 2013), which will be discussed under the Motivation dimension.

In the three studies evaluating the effectiveness of reflection as a learning strategy, the word *reflection* refers to reflecting on the content of assigned reading, as opposed to the reflection phase of self-regulated learning, in which the learner reflects upon the effectiveness of his or her learning strategies. Psychology students formed the sample for all three of these studies. In two of them, students learned psychology-related material outside of class for research participation credit, rather than learning course material for their course grade (Dietz-Uhler & Lanter, 2009; Hartlep & Forsyth, 2000). The third study involved a reflective journal over actual course material (Cisero, 2006).

In the Hartlep and Forsyth (2000) study, some students were asked to use the SQ4R (*Survey*, *Question*, *Read*, *Recite*, *Review*, *Reflect*) method for studying a psychology text, and other students were asked to use only the *Read* and *Reflect* steps of the SQ4R approach. The *Reflect* step, sometimes called self-referencing, asks students to think about how the reading materials relate to their life experiences. Dietz-Uhler and Lanter (2009) used a similar approach, asking students to answer four questions about the text they read. One of the four questions asked students to state why the concept or theory was important, and another question asked them to apply the reading material to some aspect of their life.

Both studies showed positive results. In Hartlep and Forsyth's (2000) study, both the SQ4R and *Read-Reflect* groups retained information better than control students, outperforming control students on a content test two weeks after the exercise. Immediately after the exercise, the two groups had similar scores on the content test. Control students showed a significant decline in the two weeks between tests, but treatment students did not. In Dietz-Uhler and Lantner's (2009) study, students who answered the four questions before the content quiz had significantly better quiz scores than students answering the questions after the quiz. Students thought the activity was successful in meeting its goals and in improving memory and enjoyment.

In Cisero's (2006) study, students in classes requiring a reflective journal over course material were compared to students from the same instructor's previous semesters, who were not required to submit a reflective journal. The mean exam grades in the two groups were compared and found to be very similar, though no significance values or effect sizes were reported. Chi-square analyses showed the intervention group had significantly fewer C and D exam grades than the control group. However, this finding is of limited value due to the time lag between control and treatment groups (five control semesters followed by three treatment semesters), and the lack of control for confounding variables.

The first two studies of content reflection as a learning strategy (Dietz-Uhler & Lanter, 2009; Hartlep & Forsyth, 2000) utilized a careful research design and randomized treatment assignment. Therefore, they provide credible evidence that content reflection is an effective strategy. However, from a self-regulated learning standpoint, these credible results are not especially valuable, because of the artificial nature of the learning task and the fact that students were compelled to utilize the learning strategy instead of choosing it on their own. Cisero's (2006) study of the reflective journal was conducted in a more authentic learning environment; however, the lack of a comparable control group limits its value.

Two interventions, one in psychology (Einstein et al., 2012) and one in biology (Stanger-Hall et al., 2011), sought to teach students the value of self-testing as a learning strategy. The psychology intervention was a lab activity in which students served as both participants and analysts (Einstein et al., 2012). The students studied two passages that were unrelated to psychology, one in the *Study-Study* condition and one in the *Study-Test* condition. In *Study-Study*, students read the passage for two consecutive 4-minute periods, in which they could highlight, underline, or take notes. In *Study-Test*, they spent the first 4-minute period in *Study*, with highlighting and note-taking allowed. In the second 4-minute period, called *Test*, they flipped the passage over and used the back of the page to write down all the information they could remember. In the next lab session, they took surprise quizzes over both passages. As the researchers expected, the data supported the benefit of self-testing, with students in *Study-Test* outperforming those in *Study-Study*, regardless of order or passage. All the quiz scores were shared with the entire class, and the students used the quiz data to write a report about the experiment's results and limitations. Pretest and posttest quizzes indicated that the activity improved students' understanding of self-testing and increased their likelihood of incorporating self-testing into their studying. This article provided detailed descriptions of the lab experiment and its results, but did not collect or analyze data to evaluate whether it helped the students improve their psychology grades.

The second self-testing intervention was a workshop for biology students, which incorporated two learning activities, one demonstrating the value of visualization and one demonstrating the value of self-testing (Stanger-Hall et al., 2011). Only the self-testing activity is relevant to this review. At the beginning of the activity, students tested their knowledge by diagramming the generalized plant life cycle. After an interactive instructor-led review session, students self-tested a second time, again diagramming the plant life cycle. Student drawings improved from the first to the second attempt. When surveyed, most students found both the visualization activity and the self-testing activity helpful, and planned to implement both techniques in their studying. On most life-cycle final exam questions, the workshop group significantly outperformed both the same-semester control group and the previous-year control group. However, the workshop group did not perform better on the overall final exam grade. The students' inability to transfer the benefits of visualization and self-testing to topics not covered in the workshop suggests that their improved performance on life cycle questions may have come from the increased exposure to life-cycle material during the workshop, rather than from applying the strategies on their own.

In another intervention targeting large biology classes, students analyzed questions missed on the midterm by at least 40% of the class (Williams et al., 2011). For each exam question assigned to them, they wrote a short paragraph summarizing which lecture contained the information for the question, why the correct answer was correct, why the incorrect answers were incorrect, why the student got it wrong (if the student missed it), or why others might have got it wrong (if the student answered it correctly). Students were assigned different sets of questions to analyze, so each student was in the control group for some questions and in the treatment group for other questions. Treatment students significantly outperformed control students on about half the corresponding topic-matched questions, with no difference on the other topic-matched questions. The questions showing a difference were not identical to the corresponding midterm questions, but had the same emphasis. Most students thought the assignment was helpful and caused them to examine their midterms more carefully.

This exam analysis assignment study (Williams et al., 2011) provides useful information because it occurred in an authentic setting, used within-class controls, and connected the characteristics of the exam questions to the results. Because every student was a treatment student for some questions and a control student for other questions, the treatment and control groups were much more similar than in most studies. However, this approach did not allow the researchers to assess whether the students improved their achievement compared to students not receiving the intervention at all. The researchers' detailed coding of the exam question pairs (by topic, Bloom's taxonomy level, and similarity of emphasis) made it possible to see that the assignment helped students' performance on certain types of questions but not others. The other just-described biology intervention, the workshop on visualization and self-testing, was less carefully controlled and thus less useful (Stanger-Hall et al., 2011), but showed the same difficulty in transferring the effects of the exercise to other biology topics. The self-testing activity described by Einstein et al. (2012), in which psychology students compared *Study-Study* to *Study-Test*, was very interesting and may have encouraged students to use the self-testing strategy. Unfortunately, we do not know how effectively it met this objective, because the intervention itself was not seriously evaluated. Though the authors suggest the activity produced long-term changes in students' study habits, the results provided in the article do not substantiate this.

Outcomes (What) Dimension

Learners exercising self-regulation in the Outcomes dimension keep tabs on their behavior and the outcomes of their studying. Through self-monitoring and self-evaluation they become self-aware of their performance (Zimmerman, 1998). The word *monitoring* implies an ongoing continuous process, or at least a system of frequent checks at regular intervals. The previously described studies of self-testing and exam question analysis lacked this element of regularity, or at least possessed it in lesser degree than the interventions in the Outcomes (What) dimension.

Thirteen interventions fell into the Outcomes category. Three of them combined a learning strategy with self-monitoring and thus fall into the intersection of Strategies and Outcomes. These three, along with eight studies emphasizing only self-monitoring, will be discussed here, in the Outcomes section. Two of the 13 Outcomes interventions also included a goal-setting component and will be discussed in the Motivation section.

Three interventions combined computerized self-monitoring prompts with some sort of study-strategy scaffolding, placing them in the intersection of the Strategies and Outcomes circles. In two of them, the participants were recruited from educational psychology classes and completed the activity in a computer lab outside of class (Kauffman et al., 2008; Kauffman, 2004). Though the content they studied was relevant to educational psychology, it does not appear to be course material the students were required to learn for a class. The third such intervention, in a computer science class, was intended to help students learn actual course material for a grade (Kwon et al., 2011).

For one of the outside-of-class educational psychology interventions, students took notes on paper while studying an electronic text (Kauffman, 2004). In a 2×2×2 crossed design, some students were assigned to take free-form notes and others took notes using a matrix organizer. Some students received computerized self-monitoring prompts which asked them to make confidence judgments about the completeness of their notes. Some students received computerized messages designed to bolster academic self-efficacy. Matrix note-taking and self-monitoring prompts seemed to improve scores on the achievement test. The self-monitoring prompts seemed to help the matrix note-takers but not the free-form note-takers. Students receiving the self-monitoring prompts did not improve their metacognitive self-awareness, as measured by a metacognitive inventory.

In a second intervention targeting educational psychology students outside of class, students analyzed classroom case studies (Kauffman et al., 2008). In a 2×2 crossed design, some students received problem-solving prompts that walked them through identifying the problem and describing possible solutions. Some students received self-monitoring prompts that asked them to rate their confidence about how thoroughly they had addressed the problem. Students receiving problem-solving prompts had significantly higher scores for both problem-solving and writing quality. Self-monitoring prompts also helped performance, but only if accompanied by problem-solving prompts.

The third and final intervention in the Strategies/Outcomes intersection combined problem-solving prompts and self-monitoring prompts in an authentic setting (Kwon et al., 2011). Computer programming students used the CatchBugs online tool to practice debugging computer code. Students received problem-solving prompts, such as "what does the error message mean?", designed to walk them through each debugging task. They answered the prompts using one of two randomly-assigned methods, either by generating free-form self-explanations or by using drop-down menus to insert missing words into teacher-generated explanations. All students also received self-monitoring prompts in which they rated their confidence in the correctness of their explanations. The groups performed similarly on the multiple-choice questions on the posttest, but the free-form self-explanation group demonstrated superior debugging performance, correcting significantly more errors. The self-monitoring prompts showed similar levels of confidence for the two groups. Because the main purpose of the study was to evaluate the effect of the two types of explanation prompts, no comparison group was set up to evaluate the effectiveness of the self-monitoring prompts or of the overall program.

Eight interventions focused strictly on self-monitoring, without also addressing any other dimension. One of these involved confidence ratings for online study questions (Grabe & Flannery, 2010), one incentivized self-monitoring accuracy (Hacker, Bol, & Bahbahani, 2008), three used self-monitoring forms (Bercher, 2012; Brothen & Wambach, 2000; Cao & Nietfeld, 2005) and three used asked students for deeper and more involved self-monitoring using rubrics (Andrade & Du, 2007; Cho, Cho, & Hacker, 2010; Kitsantas & Baylor, 2001).

In an intervention designed to promote self-monitoring, metacognition, and improve calibration (accuracy of self-monitoring) , psychology students used an online bank of study questions to earn points in their course (Grabe & Flannery, 2010). For each question, students rated their confidence in the correctness of their answer. Questions with high confidence ratings resulted in a gain or loss of more points. On questions with lower confidence ratings, fewer points were gained or lost. The researchers found that the poorer performing students tended to abandon the system before answering even a minimal number of questions. Perhaps partly for that reason, a consistent relationship between calibration accuracy and exam performance was not found.

In another study focused on calibration (Hacker et al., 2008), psychology students estimated their exam grades before they took the exam (prediction) and estimated their exam grades immediately after they took the exam (postdiction). After the exams were graded, the students received two calibration scores, one for the prediction and one for the postdiction. The calibration scores, expressed as percentages, indicated how closely the student’s estimate matched the actual exam grade (e.g., a 100% calibration score indicated an exact match). In a 2×2 crossed design, four intact classes were randomly assigned to four treatment conditions. Students in the extrinsic incentive condition received extra exam points if both the predicted score and the postdicted score were close to the actual score. Students in the reflection condition completed a questionnaire immediately after receiving their calibration scores; it included questions addressing responsibility for the exam grade, factors influencing accuracy, and strategies to improve accuracy. Neither the extrinsic incentives nor the reflection questionnaire led to improved calibration. In general, students did not improve their calibration accuracy across Exams 2 and 3. However, low-performing students (on Exam 1) improved their accuracy on Exam 2 and 3 if they received extrinsic incentives.

Three interventions used self-monitoring forms. The first of these, Cao and Nietfeld (2005), also focused on calibration. Educational psychology students used weekly self-monitoring sheets to rate their understanding of the day's content (judgment of learning), to list the concepts they found difficult to understand, to describe a plan for improving their understanding, and to answer three multiple-choice review questions. Each review question on the self-monitoring sheet and each exam question was followed by a confidence judgment, in which students rated the accuracy of their answer on a scale from 0% to 100%. During the final week of the semester, students predicted their final exam scores. The researchers examined changes in judgment of learning, confidence, performance, accuracy, and bias throughout the semester. They concluded that students’ accuracy improved with practice, and that through the weekly exercises, students experienced a shift in their judgment of learning and became able to accurately estimate their end-of-semester test performance. Based on my own examination of the constructs described and the data provided, I would not have drawn such strong conclusions. However, the results of this study do support the previously established connection between performance and monitoring accuracy: high-performing students are more capable of accurately predicting their performance (Bol, Hacker, O'Shea, & Allen, 2005).

Two other studies (Bercher, 2012; Brothen & Wambach, 2000) also assessed the value of self-monitoring forms. In one, psychology students used a paper form to record their points from online vocabulary exercises, pretests and quizzes (Brothen & Wambach, 2000). For each chapter, teaching assistants checked the students' recording forms against the online grade book and awarded 0 points if no quiz scores were recorded, 1 point if at least one score was recorded, and 2 points if all scores were recorded correctly. These were combined into a single "recording" score, which was used as a measure of self-regulation. Stepwise regression analysis resulted in three statistically significant predictors of final grade: ACT score, recording score, and practice final exam use.

In Bercher (2012), students in anatomy and physiology labs completed self-assessment sheets on which they rated their percentage of mastery for each learning objective. The students also completed five post-exam reflection sheets, one for each exam, after receiving the exam grade. On these sheets, they rated how much their self-assessment sheet mastery percentages had affected their exam preparation and whether their exam performance was better, worse, or about the same as expected. Though the article did not contain evidence of statistical tests or significance values, it showed that students who thought the self-assessment sheets were helpful performed better on the exam, and were more likely to have earned a grade that was higher than they had expected. Students who felt the self-assessment sheet was not at all helpful or only somewhat helpful scored lower on the exam and were more likely to have scored below expectations.

All three of the aforementioned studies of self-monitoring sheets (Bercher, 2012; Brothen & Wambach, 2000; Cao & Nietfeld, 2005) were among the weakest studies of all the reviewed articles. This criticism is not due simply to limitations of the outcome variables, sampling, or lack of a control group. In addition to these common limitations, which are shared by many studies, these three studies had substantial shortcomings in their analyses and argumentation. To varying degrees, the authors overreached, drawing conclusions that were not warranted by the data presented and analyses performed.

In contrast to the previously mentioned interventions in which students used numerical ratings to rate their mastery of content or their confidence in the correctness of a short answer, three of the eight Outcomes-only interventions used a detailed rubric to scaffold self-monitoring on a larger assignment. Two of these targeted preservice teachers. Kitsantas and Baylor (2001) quantitatively evaluated the effectiveness of a rubric for self-assessing lesson plans. In the only qualitative study in this review, Andrade and Du (2007) used retrospective focus groups to learn about education students' perceptions of criterion-referenced self-assessment. In the third example of intensive self-monitoring, Cho, Cho and Hacker (2010) evaluated an electronic platform designed to facilitate both self-evaluation and peer-evaluation of writing assignments.

In the first study targeting preservice teachers, students in educational technology classes used the Instructional Planning Self-Reflective Tool (IPSRT) to evaluate instructional plans they created (Kitsantas & Baylor, 2001). Seven intact classes were randomly assigned to control and treatment groups. The classes had different instructors, but the same person taught all of the control and treatment classes during the two-week period of the intervention. Using the same three case studies, all participants wrote one instructional plan during class, one as homework, and one as a quiz. The IPSRT was demonstrated for the treatment group prior to the in-class and homework case studies, and was attached to their quizzes. The control students were not exposed to the IPSRT until after the quiz. At that time, they received the same IPSRT demonstration, and were allowed to modify their quiz with a red pen. On the posttest achievement quiz, the treatment group performed significantly better than the control group. The control group improved significantly after the IPSRT demonstration. After the intervention, the treatment students had significantly more positive scores on disposition toward instructional planning, though the groups had been similar initially. The intervention did not seem to improve either group's overall self-efficacy toward instructional planning or the importance they placed in it. However, post-hoc analysis showed that in the treatment group, those who started with low instructional planning self-efficacy improved, and those with high initial self-efficacy decreased after the intervention. The same results occurred in the control group, after the post-quiz IPSRT demonstration.

The preservice teachers in Andrade and Du's (2007) qualitative study also used detailed rubrics to self-evaluate their assignments. This criteria-referenced self-assessment was intended to be formative, not summative. For each assignment, students were required to complete a checklist of criteria and circle the quality gradation that best described their work. The self-assessment form was not graded, but had to be completed for the assigned work to be graded. After the semester ended, some students were contacted and invited to a focus group, in which they would discuss their perceptions of the self-assessment process. The focus groups revealed that as long as expectations were made very clear, the self-assessment resulted in better work and higher grades. If expectations were not clear, the results were frustration and increased tension between self-assessment and teacher-assessment. Although the focus group students generally thought the clear rubric made self-assessment beneficial in the class using the rubric, they had difficulty transferring that self-monitoring skill to classes that did not provide a similar self-monitoring sheet.

The third intervention involving detailed rubrics focused on writing (Cho et al., 2010). Because writing is an essential component for classes in nearly all disciplines, if students can develop skill in self-monitoring their writing, that self-monitoring ability may transfer more readily. Cho, Cho, and Hacker evaluated the electronic SWoRD (Scaffolded Writing and Revision in the Disciplines) system, designed to scaffold self-monitoring and peer-monitoring of writing. Participants came from undergraduate and graduate classes in a variety of disciplines, including physics, history, psychology, writing, and leisure behavior. Students evaluated both their own writing assignments and their peers' assignments on the same three dimensions: prose flow, argumentation, and insight. Students were graded on the quality of their writing, their improvement from first to second draft, and the quality of their reviews of others' papers. In the SWoRD system, students could read others' reviews of the same papers they had reviewed and could visually compare their own ratings to others' ratings using graphs. The SWoRD system provided half of the reviewing grade by assessing consistency. The students' peers provided the other half of the reviewing grade by rating the helpfulness of each review. Student authors could access all the reviews of their papers, the system's assessment of each reviewer's consistency, and their writing and reviewing grades compared to the class mean.

To gauge whether the system helped students improve their self-monitoring, the researchers examined the size of the gap between the student's self-evaluation and the average of others' evaluations of the same writing piece. If the gap was smaller for the second draft than for the first draft, then self-monitoring skill had improved. A paired *t* test showed a significant increase in the mean size of this gap, indicating that the students' self-monitoring skill had worsened from first draft to second draft. The correlation between self-monitoring improvement and writing improvement was 0.66, indicating that gains in self-monitoring skill were associated with gains in writing quality (Cho et al., 2010).

Unlike the three studies of simple self-monitoring forms, the three studies based on detailed self-monitoring rubrics utilized research designs that could adequately support the research questions that were asked. The two studies of self-evaluation rubrics for preservice teachers provided solid evidence that intensive self-monitoring through rubrics can improve performance (Andrade & Du, 2007; Kitsantas & Baylor, 2001); the study of the online platform for self-monitoring of writing did not provide such evidence, but it offered some information about the relationship between self-monitoring development and writing improvement (Cho et al., 2010). Kitsantas and Baylor's (2001) use of a control group and an achievement variable (lesson plan quality) supported the effectiveness of their lesson plan rubric. By providing the rubric to the control group, allowing them to correct their quiz, and then repeating the statistical analysis, the researchers reinforced the conclusions drawn from the original comparison. In addition to benefitting the control students, this additional step defused a potential criticism applicable to most quasi-experimental studies, that of nonequivalence of groups. Andrade and Du's (2007) qualitative study, with its purposive sample, used the students' words to demonstrate the value of the self-monitoring rubric, the importance of clear expectations, and the difficulty of transfer. The study of the SWoRD writing platform (Cho et al., 2010) showed that self-monitoring improvement was positively related to writing quality improvement. However, the students showed an overall decrease in self-monitoring accuracy from first draft to second draft. Without comparing the writing quality for students using the platform to the writing quality for students not using the platform, it is almost impossible to draw meaningful inferences about the platform's effect on writing.

Motivation (Why) Dimension

The remaining seven studies emphasize motivation. Two of these, one evaluating methods of assigning outside reading (Hilton, Wilcox, Morrison, & Wiley, 2010) and one evaluating a portfolio for foreign language learning (Ziegler & Moeller, 2012), also involve self-monitoring and thus are in the intersection of the Outcomes (What) and Motivation (Why) dimensions. The three interventions addressing only the Motivation dimension included computerized motivation-boosting messages for statistics students (Acee & Weinstein, 2010), written goal setting lessons for educational psychology students (Schwartz & Gredler, 1998), and implementation intentions for developmental writing students (Georgianna, 2009). The intersection of the Motivation (Why) and Strategies (How) dimensions contains one intervention, a strategy-based developmental writing curriculum with a goal setting component (MacArthur & Philippakos, 2013). The final intervention, which combines goal setting worksheets with peer group support, lies in the intersection of the Motivation and Social Context dimensions (Fitch, Marshall, & McCarthy, 2012).

Hilton et al. (2010), in one of the two studies combining goal setting with self-monitoring, examined different methods of motivating students to complete outside readings for a religion/philosophy class. Some students were asked to self-monitor their own outside reading as part of their course grade; other students were not. Results were based on self-reports of time spent reading and of the percentage of outside reading completed. One class required students to read thirty minutes per day on weekdays, write reflections in a journal, and self-grade their performance on three occasions. These students read more minutes than any other class. Another class required students to set their own goals for minutes each day, and also to self-grade themselves for meeting the goal and for completing the readings. These students completed the highest percentage of outside reading. A third class did not have a minutes-per-day goal, but gave themselves points for the portion of required readings they completed. The fourth and fifth classes, both taught by the same instructor, had required reading assignments but students did not receive points for completing them. These students read the least and made the fewest positive comments about the readings helping their personal study. Though not a carefully controlled research study, this investigation supports the notion that students may read more if they are asked to set reading goals and self-monitor their reading.

In the second study combining motivation and self-monitoring, Ziegler and Moeller (2012) evaluated the LinguaFolio, a platform designed to encourage goal setting, reflection and self-assessment in foreign language classes. Students were required to set an achievement goal and a personal goal, and submit evidence they had met the goals. Evidence of a completed personal goal might be a signed affidavit from a waiter verifying that the student had ordered dinner in Spanish. Students also used a series of "can do" statements to self-assess their language proficiency. In the research study, three groups were compared. The first group was composed of Spanish classes that used LinguaFolio and had instructors who attended at least one question-and-answer meeting with the researchers. The second group was composed of Spanish classes that used LinguaFolio and had instructors who did not attend a meeting with the researchers.

The third group, which served as the control group, was composed of French classes not using LinguaFolio. Findings were inconsistent and not particularly notable. The lack of equivalent groups limits the value of this research study.

All but one of the interventions classified as Motivation (Why) required students to set goals or provided some sort of goal-setting training. The lone exception was Acee and Weinstein’s (2010) value reappraisal intervention. Instead of asking statistics students to motivate themselves by setting goals, the researchers attempted to increase students’ motivation by presenting them with computerized messages about the value of statistics. The students also completed computer-guided activities designed to increase motivation, such as creating lists of incentives for developing statistics skills, describing how statistics knowledge could be useful in potential careers, and replacing negative statistics-related thoughts with positive thoughts. On an inventory designed to measure the overall importance placed on course-related tasks and the perceived future usefulness of statistics, students receiving the intervention showed gains from pretest to immediate posttest and from pretest to delayed posttest. For one instructor, treatment students also had higher exam grades than control students. Students in the treatment condition were also more likely than control students to access statistics websites their instructor had shared with them. Because accessing these websites did not affect students’ grades, this was used as a measure of continued interest in statistics. This study made reasonable efforts to control for confounding variables, by using stratified random samples, crossing instructors and semesters, and using the first test grade as a covariate. However, unlike most other interventions using course grade as an outcome measure, the study was conducted in a laboratory setting instead of being integrated into the course. The participants completed the exercises to fulfill a research participation requirement, but the exercises did not count toward their course grade.

The remaining interventions asked students to set goals or provided goal-setting instruction. In one, graduate students received four weekly sets of self-instructional materials on goal setting (Schwartz & Gredler, 1998). After reading the written materials, students completed application exercises. No verbal goal-setting instruction was provided. Instead of goal-setting materials, control students received four weekly vignettes about classroom situations, each with application exercises. Though the intervention was an in-class activity, the environment was artificial, with lessons distributed in sealed envelopes and completed in silence. Treatment and control students completed their activities simultaneously in the same room. Exercises were returned with feedback but did not count in the students' grades. On an inventory designed to measure self-efficacy for self-regulated learning, scores declined for both groups; on an inventory of goal-setting habits, the groups showed no differences. However, the students receiving goal-setting instruction scored higher on a goal-analysis quiz. In the treatment group, students with low scores on the self-efficacy for self-regulated learning inventory were not successful at creating long-term goals with supporting subgoals. The study did not compare the groups on any achievement measure.

In sharp contrast to the previously described goal-setting intervention, the goal setting training described by Georgianna (2009) was closely connected to the students' class, developmental writing at a community college. Students learned to create “implementation intentions” for important academic tasks. For each goal set, students had to identify positive results of achieving the goal and obstacles that stood in the way. Students were required to use an “if….then” format to specify when, where, and how they would perform an action (i.e., “if I am to study six hours for my test, then I must…”). They also created “if…then” action plans for overcoming anticipated obstacles. The intervention lessons occurred during five consecutive class meetings. After each session, students created implementation intentions for writing an upcoming mini-essay. The treatment class received the intervention the first half of the semester and wrote mini-essays during the first four of the five intervention weeks. The control class received the intervention the second half of the semester, after their first four mini-essays had already been completed. Control students did not write any mini-essays during the five weeks of the intervention. All students wrote three more mini-essays after the intervention. Statistical analysis showed that treatment students had higher grades on the first four mini-essays than control students ($p = .06$). Control students did not do better on the three post-intervention mini-essays than on the four pre-intervention mini-essays. Treatment and control students did not differ in their final grades or in their success rates. Apparently the training helped, but only if it was immediately applied to the current academic task.

Developmental writing classes at a community college were also the setting for the one intervention classified as both Motivation (Why) and Strategies (How). A curriculum focused primarily on grammar and sentences was replaced with a more writing-intensive curriculum based on the theory of change (MacArthur & Philippakos, 2013). In the first semester, instruction focused on writing strategies. The article mentions that self-regulation strategies were taught but does not include any information about them. In the second semester, instruction on goal setting, progress monitoring and reflection was added. Students wrote journal entries about their goals and strategies, and discussed their journal entries in class. Other writing-specific aspects of the curriculum were also modified. Most students had positive feelings about the amount they had learned. Students in the second semester had better knowledge of writing strategies. Instructors had difficulty integrating goal setting instruction with writing strategy instruction, and students did take not give serious attention to the reflective journal entries about goals. Because there was no control group and because the writing curriculum was revamped at the same time the self-regulated learning instruction was introduced, this study does not contribute useful evidence about whether self-regulation training can help students.

The final goal-setting intervention combined two Zimmerman dimensions: Motivation (Why) and Social Context (Who). Intact sections of several psychology courses were randomized into treatment and control conditions (Fitch et al., 2012). Treatment students completed a goal-setting worksheet. The worksheet had space to list four goals, timelines to reach the goals, action plans to complete them, and numerical ratings of the student's progress toward each goal. In a strength-assessment section, it asked students to list past successes with similar goals, and also list the actions they had taken and the obstacles they had overcome to achieve these past successes. The worksheet included brief guidelines and examples about goal setting and planning, emphasizing that goals should be specific and attainable. Treatment students were divided into groups of 4–6 students. Each group met for at least six self-guided group meetings in which students shared their worksheets and discussed their progress with one another. Each meeting was facilitated by a different student leader, with access to assistance from a teacher or counselor. All students completed five scales of a modified MSLQ inventory. Treatment students scored significantly higher than control students on self-efficacy, intrinsic value, cognitive strategy use, and self-regulation but not on test anxiety.

Taken as a group, the studies in the Motivation dimension do not provide much evidence for the effectiveness of motivation-related interventions upon student achievement. Due to shortcomings in the research design, two of the seven studies in the Motivation dimension contribute almost no useful information about the effectiveness of the interventions described (MacArthur & Philippakos, 2013; Ziegler & Moeller, 2012). The study by Hilton et al. (2010) was similarly confounded but still provided tentative support for the value of setting outside reading goals and self-monitoring the progress toward those goals. Though the three studies of goal setting materials or training were more carefully controlled, two of them used only inventory scores to measure effectiveness. Sealed-envelope goal setting lessons, which were not connected to class content or class discussion, did not seem to improve self-regulated learning (Schwartz & Gredler, 1998). The more intrusive goal setting intervention by Fitch et al. (2012), which required students to regularly share their goal worksheets and progress with a peer group, resulted in higher inventory scores for treatment students on nearly all variables.

Unfortunately, neither of these studies incorporated an achievement variable such as GPA or course grade. The use of an achievement variable and an authentic setting can make even a very small study both valuable and interesting, as shown by Georgianna (2009). In Georgianna's study, the use of essay scores and course success as outcome measures revealed students' difficulty in transferring newly acquired self-regulated learning skills to non-immediate tasks. Implementation intentions were associated with higher essay scores, but only for essays written the same week. In the only other reasonably rigorous study using an achievement variable, computerized motivational messages resulted in improved inventory scores overall, but only one teacher's classes improved their grades (Acee & Weinstein, 2010).

Summary and Discussion

Let us return to our guiding question, "What are the effects of self-regulated learning interventions upon college students?" and the informal version, "Can college students be taught to self-regulate their learning of academic content?" In this section, I will highlight the studies that shed the most meaningful light on the guiding question, summarize what we have learned from them, and suggest directions for future research.

In order to provide credible information, the study must utilize a reasonably solid research design and appropriate data analyses. In order to provide useful information about the guiding question, it should focus on achievement in college content classes. Ideally, the intervention would occur in an authentic context over a protracted period of time. For the context to be authentic, the intervention's activities or instruction should be incorporated into the normal routine of a class. Well-designed studies lacking an authentic context or an achievement variable will answer different questions than studies possessing these features. Because this review is focused on students' learning of academic content in college classes, a study using an authentic context and an achievement variable has more value than a study of similar rigor that occurs in an artificial setting or uses an inventory to measure the intervention's effectiveness.

The 42 studies in this review, particularly the 27 dimension-specific ones, reveal an interesting dilemma: there seems to be a trade-off between rigor and authenticity. It is hard to design a rigorous study that effectively controls for confounding variables and also occurs in an authentic environment. Controlling for confounders is relatively easy if you recruit participants from a convenient population (e.g., psychology students who must earn research participation credits), randomize them into treatment conditions, and ask them to perform prescribed activities outside of class in a carefully controlled environment. When conducting research on students taking courses for grades, it is more difficult to control for everything. In authentic settings, research must take a secondary position behind the realities of class scheduling, student enrollment, teaching loads, academic content instruction, assessments, and course grades.

Keeping this trade-off in mind, I revisited each of the 42 reviewed studies and selected those I felt were sufficiently well-controlled so as to be credible and sufficiently targeted toward the research question so as to be useful. To be selected, the intervention needed to occur in an authentic setting and the outcome measure needed to involve grades for an actual content course. If the intervention was a study strategies course, the outcome measure needed to include performance in content classes, not just the study strategies course. If it was in a content course (e.g., mathematics, psychology, or biology), the intervention needed to involve training that occurred during class time or a learning activity completed outside of class as part of the course grade. The outcome measure could be either the course grade or a component of the course grade (e.g., an exam grade). If the intervention required students to come to a laboratory to complete an activity not incorporated into their class, I did not consider the setting to be authentic and did not count the study among the credible and useful studies.

Overview of the Credible and Useful Studies

Using these criteria, I settled on 13 credible and useful studies in which the intervention occurred in an authentic setting and was evaluated using a grade-based achievement measure. I included the qualitative study in this group, because it used a credible qualitative design to obtain and synthesize student perspectives on how the intervention affected the students' achievement. Seven of these studies evaluated interventions tied to a particular content class, and the other six evaluated the effects of study strategies courses or specific elements of study strategies courses. These 13 studies, which I deemed both credible and useful, will be summarized in this section. Organizing the studies by the nature of the interventions and by the dimensions addressed was helpful for seeing what types of intervention research were being done; however, focusing on the credible and useful studies will make it easier to see what we have learned and what future research is needed.

Four studies of stand-alone study strategies courses showed fairly decisive positive effects on academic achievement variables (Ahuna et al., 2011; Bail et al., 2008;

Kamphoff et al., 2007; Ryan & Glenn, 2003). Ahuna et al. (2011) examined an optional course based on self-regulated learning theory, incorporating goal setting, self-monitoring, active learning strategies, and small groups with peer monitors. Successful completion of the course resulted in improved retention and 4-year graduation rates. An optional writing-intensive self-regulated learning course, targeting sophomores and featuring small class sizes, resulted in improved graduation rates and higher GPA four semesters after the course (Bail et al., 2008).

Research on any optional study strategies course has a critical limitation: students enrolling in the course may be different from students not enrolling in the course. Some studies control for this statistically, by including academic variables as covariates or by matching the students on academic variables. This helps, but the possibility remains that the enrolling students may differ on unobserved variables (e.g., motivation) not accounted for in the statistical control process.

Ryan and Glenn (2003) handled this selection threat by comparing two versions of the same college success course. One version focused on learning strategies, including goal setting, planning, and self-analysis; the other version was centered on a particular theme and emphasized collaboration and intellectual discussion. It seems reasonable that the students enrolling in the two versions of the course may be more similar to each other on unobservable motivational variables than they are to students not enrolling in the success course at all. The study found that students in the strategy-based course had higher one-year retention rates than students in the theme-based course or students not in any success course.

For researchers studying mandatory study strategies courses, the problem of treatment assignment being driven by unobservable variables is replaced by the problem of not having a comparable control group. Kamphoff et al. (2007) handled this difficulty in an intriguing way. The study strategies course was mandatory for students on academic probation but not for students on academic warning (i.e., those who were in danger of being placed on probation). This allowed the researchers to use the academic warning students as a control group. The course was based on a model that included personal responsibility and self-regulation. When this model was combined with strict penalties (withdrawal or suspension) for students who missed a single meeting of the success class, students in the course (probationary students) had greater gains in subsequent semester GPA than matched control students (academic warning students).

While the artificial-setting interventions often were confined to a single learning session, most of the interventions integrated into classes were either ongoing for an entire semester or involved multiple learning sessions over two or more weeks. A notable exception to this was the Haught et al. (1998) study of individual counseling on the LASSI inventory results. During a class session of a study strategies course, each student assigned to the treatment condition left the classroom for an individual counseling session, during which the counselor explained the student's below-50th-percentile subscales, suggested strategies for improvement, and answered questions. This one-time script-based counseling session seemed to result in tangible benefits. Individually counselled students had higher grades in the study strategies course and also higher subsequent semester cumulative GPAs.

All five of the studies just mentioned involved study strategies courses and were included in the overall-SRL category of this review. The remaining eight credible and useful studies were described in the dimension-specific section of this review, because they targeted only one or two Zimmerman dimensions.

The first of these, though also tied to a study strategies course, focused on the Social Context (Who) and Time (When) dimensions. Students met online in real-time with their instructor and a small group of their peers, sharing their goals and progress with a peer partner (Tuckman, 2007). The intervention seemed to help high procrastinators manage their time, at least in the online study strategies course. High procrastinators in the online support groups had higher averages in the course and greater GPA gains than high procrastinators in the control group.

Both the Tuckman (2007) study of online support groups and the Haught et al. (1998) study of individualized LASSI counseling would have been considerably more valuable if they had recalculated the GPA variable with the study strategies course omitted. Because they did not do this, there is no way to know whether the treatment students' GPA improvement was due only to higher grades in the study strategies course, or whether the students also had higher grades in their other courses. Both studies used GPA and course grade as outcome variables, indicating that the researchers probably had access to the data required for this easy calculation.

Goodwin and Califf's (2007) study of time-management training for students in a computer programming course provided evidence that students can be taught to improve their self-regulation within the Time (When) dimension. The researchers used a reasonably solid research design, randomizing treatment assignment within same-instructor pairs of intact classes and checking for group differences on key starting variables. Students receiving the training were more successful in the class, although their perceived control of time did not change. This study illustrates how meaningful self-regulation training can realistically be incorporated into a content class. The training occurred in two 20-minute sessions about a month apart. For the entirety of the semester, the training was reinforced by weekly time log sheets and instructor reminders.

Georgianna's (2009) implementation intentions for developmental writing students, in the Motivation (Why) dimension, also wove self-regulation training into the fabric of a class. Over a 5-week period, students wrote academic goals, identified benefits of achieving the goals, anticipated obstacles, and created action plans. The intervention seemed to help students with the immediate upcoming academic task, but did not transfer to tasks that were temporally further away.

Instead of general training on a dimension of self-regulated learning, the next three studies provided rubrics for students to use in self-evaluating a particular assignment. These studies were placed in the Outcomes (What) dimension, because they focused on the products resulting from the learning process.

Two of these three studies had especially strong designs and showed positive results, indicating that detailed rubrics can help students improve their self-monitoring skills and produce higher quality work. In Andrade and Du's focus group study (2007), preservice teachers indicated that the detailed self-evaluation rubrics had helped them improve the quality of their work, as long as the rubrics were accompanied by clear expectations. The students had difficulty transferring their self-evaluation skills to classes not providing similar rubrics. The benefit of rubrics for preservice teachers was supported by Kitsantas and Baylor (2001), who showed that detailed self-evaluation rubrics can help students write better lesson plans. In addition to positive effects on lesson plan quality, the rubric had an interesting effect on self-efficacy inventory scores. Self-efficacy toward instructional planning increased for students who started low, but decreased for students who started high. Apparently the rubric helped less confident students feel they could complete the lesson planning task. For more confident students, the rubric highlighted shortcomings and demonstrated the difficulty in reaching the standard.

The third study of rubrics, though strong enough to be credible, was less strong than the other two. It used an achievement variable, writing quality, but did not utilize a control group. Students used an electronic rubric, the SWoRD system, to evaluate their own and others' writing assignments for prose flow, argumentation, and insight (Cho et al., 2010). The study showed that students' self-monitoring ability worsened over the course of the semester. It also showed that self-monitoring improvement was associated with writing quality improvement. However, because the students using the SWoRD system were not compared to students not using SWoRD, we do not know how the intervention affected either writing quality or self-monitoring ability. The observed decrease in self-monitoring ability may have been due to a factor other than the intervention.

Like the studies of rubrics, the Williams et al. (2011) study of an exam-analysis assignment was centered on a particular task. Instead of self-evaluating the quality of their work before submitting it, biology students analyzed their exams after they had been graded. Because the exam-analysis assignment was a content learning strategy, I placed it in the Strategies (How) dimension. The assignment, which required students to analyze frequently-missed midterm exam questions, resulted in improved performance on topic-matched final exam questions. However, the improvement did not transfer to final exam questions that differed in topic or emphasis from the midterm problems that were analyzed (Williams et al., 2011).

Of all the credible studies with a grade-based achievement variable and an authentic context, only one used computerized prompts. The CatchBugs online system, addressing the Strategies (How) dimension, walked computer science students through the process of debugging computer code (Kwon et al., 2011). Debugging practice with free-form written explanations resulted in superior debugging performance on the final exam, compared to debugging practice without free-form explanation. Although the debugging software also incorporated self-monitoring prompts (addressing the Outcomes [What] dimension), the researchers did not design their study to evaluate the self-monitoring component.

What the Credible and Useful Studies Tell Us

Positive Results on Achievement

All but one of the credible and useful studies showed positive effects for the interventions they evaluated. Some showed decisive positive results, and others showed very limited positive results. The one study not showing positive results for the intervention was the study of the online writing rubric, which showed that self-monitoring ability worsened during the course of the intervention (Cho et al., 2010). However, the lack of a control group, combined with the finding that self-monitoring improvement was associated with writing quality improvement, limits the weight that should be placed on this negative result.

The four study strategies courses seemed to help students' overall academic success. The two evaluations of single components of study strategies also showed positive effects on achievement. All but one of the seven interventions in content courses had positive effects on achievement, even if only for a particular group of students or on a particular task. This positive consensus is meaningful because, in this group of credible and useful studies, the studies had designs that limited the confounding effects of inherent group differences.

However, these positive results should be tempered with caution. Evaluations of study strategies courses are especially prone to confounding by inherent differences between students taking the courses and students not taking the courses; statistical controls can lessen, but not remove, this threat to validity. The two evaluations of specific components of study strategies courses used randomization to avoid inherent differences, but did not separate out the study strategies course grade from the students' GPA, limiting the value of their positive results. The positive results from content course interventions were confined to a particular course, and in most cases, to a particular assignment.

Still, taken as a group, in spite of their limitations and the negative findings of the writing platform study, the credible and useful studies in this review send a consistent message: self-regulated learning interventions for college students can make a positive difference.

While these positive results are encouraging, the possibility of bias in the review must be considered. Credible studies not producing positive results may not have been accepted, or even submitted, for publication. Choosing the reviewed articles through a systematic approach does not reduce publication bias, but it does reduce the selection bias that is present in many literature reviews. During the screening process, I chose articles based on whether they met the inclusion criteria, not on whether they produced positive results. When culling the 42 reviewed articles down to the 13 credible useful articles, I did not use the presence or absence of positive results as a criterion. Instead, I based my decision on the research design, argumentation and strength of inferences, context of the intervention, and type of outcome measure.

Mixed Results on Transfer

The positive results of the study strategies courses on GPA, retention, and graduation indicate that students were able to transfer the skills they learned in the study strategies course to other classes (Ahuna et al., 2011; Bail et al., 2008; Kamphoff et al., 2007; Ryan & Glenn, 2003). However, the studies of interventions integrated into content courses indicate that students had difficulty transferring the skills addressed by the intervention to less immediate tasks or topics (Andrade & Du, 2007; Georgianna, 2009). The remaining credible and useful studies were not designed to evaluate transfer to other tasks.

Substantial Time Commitment Required

Intensity of time and effort characterized the interventions in the 13 credible and useful studies that utilized an achievement variable and an authentic context. With the exception of Haught et al.'s (1998) study of individualized LASSI counseling, none of the interventions I classified as being both credible and useful were one-time sessions. All the other credible and useful interventions required a substantial time investment from the students. Most of these interventions also required a substantial time investment from the instructors, either in class time, preparation, or grading and support.

High Level of Freedom

Freedom is an essential condition for true self-regulated learning (Zimmerman, 1994, 1998). If an intervention compels students to do something, we should not make inferences about self-regulated learning improvement, unless we remove the compelling force and see evidence of improved self-regulation. In the original pool of 42 studies, the interventions varied widely in the level of freedom they provided and in the level of initiative they required. Some interventions *trained* the students in one or more elements of self-regulated learning, while other interventions *compelled* or *prompted* the students to take some action that skilled self-regulated learners do on their own. Some interventions did both. The 15 overall-SRL interventions generally focused on training. Among the 27 dimension-specific studies, some interventions provided training, but others used worksheets or computerized prompts to scaffold self-regulated learning behaviors.

Computerized self-monitoring prompts, used in several studies addressing the Outcomes (How) dimension, do not allow much freedom in the way the student responds. After each question, a pop-up window asks the student for a numerical rating of his or her understanding, and the student complies. Rubrics and self-monitoring sheets, even when mandatory, provide more freedom and require more initiative from the student. The student can choose whether to put forth only the minimal effort needed to get credit for the worksheet or rubric, or to use the rubric for a detailed self-evaluation. In the Strategies (How) dimension, research that compels one group of students to use a particular strategy also restricts the freedom of those students. When choice is not allowed, the researcher cannot acquire information about whether the students improved their self-regulated learning skills.

A high level of freedom characterized the group of 13 credible and useful studies. For six of these studies, students learned skills in a study strategies course and had complete freedom in how they chose to apply those skills to their content courses. For interventions in content courses, a high level of freedom seemed to go hand-in-hand with a credible research design, an achievement variable, and an authentic setting. Of the seven credible and useful interventions in content courses, only one (Kwon et al., 2011) used computerized prompts, and one (Williams et al., 2011) compelled students to use a particular strategy. Because the Williams et al. study of an exam analysis assignment for biology students also provided training in using the exam analysis strategy, it allowed more freedom than other studies of forced strategies. The students were not mandated to analyze all questions, and had freedom to apply the strategy to non-mandatory questions if they wished. The researchers considered the element of choice, and examined whether the students chose to transfer the strategy.

The absence of freedom does not necessarily devalue intervention-based research, but it limits the inferences about self-regulated learning that can be drawn from the research. As investigations of pedagogical tools with the potential to help students learn academic content, studies of computerized self-monitoring prompts and mandatory strategy use are valuable. As self-regulated learning interventions, they are less valuable, because they do not measure whether students see sufficient value in the strategies to employ them on their own.

Recommendations for Future Research

Four especially pressing research needs emerged from this review. These are not four independent directions for research, but rather they are desirable elements that should be considered when designing an intervention study. First, there is a need for rigorous studies of self-regulated learning interventions integrated into content courses. Quantitative studies should use an achievement-based outcome measure to evaluate whether the intervention improves the students' learning of course content. Second, there is a need for credible research on manageable, practical interventions that could be implemented by an individual teacher in a content class (as opposed to study strategies courses or interventions dependent on a computerized platform). Although we know time-intensive interventions can help, we do not know whether students' self-regulation can be improved by interventions that are less burdensome and do not take too much class time away from content learning. Third, there is a need for qualitative evaluations of self-regulated learning interventions. Qualitative research, by eliciting student perspectives, can provide a deeper and more complete picture of how students implement the self-regulation training or strategies into their studying, and how the intervention affects their achievement. Fourth, we need research in which we examine whether students are able to carry self-regulation skills forward with them after the training or scaffolding ends. If students are prompted to set goals, self-monitor, or use self-regulation strategies, we need to phase out those prompts and measure whether students continue the self-regulation activities on their own.

CHAPTER III

A MIXED METHODS EVALUATION OF A SELF-REGULATED LEARNING INTERVENTION FOR DEVELOPMENTAL MATHEMATICS STUDENTS AT A COMMUNITY COLLEGE

Introduction

There is consensus among researchers that students who can self-regulate their learning are more effective learners (Boekaerts, 1997; Pintrich & De Groot, 1990; Pintrich, Smith, Garcia, & McKeachie, 1993; Zimmerman & Martinez-Pons, 1988). A search of any database of educational research will reveal a plethora of articles devoted to describing, measuring, and improving the self-regulated learning of students. There is little dispute that it is beneficial for students to become proactive and take responsibility for their learning. However, research is needed into how instructors can help students develop self-regulated learning skill. The primary purpose of this article is to present an empirical investigation of a study-journaling intervention for developmental mathematics students at a community college, and to describe how the intervention affected the students' self-regulated learning skill and achievement. The secondary purpose of the article is to illustrate how incorporating a qualitative component into intervention research can provide a multifaceted and fuller picture of the intervention's effect.

Before describing the empirical study, I will provide a short overview of self-regulated learning theory and of past research on self-regulated learning interventions. The large number of empirical studies of self-regulated learning makes it impractical to synthesize the results of all relevant studies. Instead, I will summarize the findings of several review articles and highlight a few individual studies that are particularly relevant to the current investigation. Then I will summarize the most urgent research needs and explain how the current study contributes to them.

Overview of Self-Regulated Learning

Self-regulated learning has been approached from several theoretical perspectives, including but not limited to information processing, operant theory, and social cognition (Zimmerman, 2001). Models based on different theories differ in their nuances but not in their essentials. In a recursive loop, the self-regulated learner plans the learning process, implements the plan, reflects upon the effectiveness of both the plan and the implementation, and adjusts the process (Butler & Winne, 1995; Pintrich, 2000; Winne & Hadwin, 1998; Zimmerman, 2001, 2002, 1990; Zimmerman & Schunk, 2001). Instead of being the focus, specific learning strategies are merely tools in a toolbox, from which the skilled self-regulated learner can select as needed.

Zimmerman (1994, 1998) suggests that learners can exercise self-regulation in each of six psychological dimensions, each associated with a key word: Motivation (Why), Strategies (How), Outcomes (What), Time (When), Environment (Where) and Social Context (Who). In the Motivation dimension, self-regulated learners control their motivation by setting goals, by defining tasks, and possibly by enacting positive or negative consequences. In the Strategies dimension, learners choose appropriate learning strategies for the task at hand. In the Outcomes dimension, learners self-monitor their learning outcomes. By constantly monitoring how well their learning process is working, they can make appropriate adjustments. In the Time dimension, self-regulated learners make decisions about the length, frequency, and spacing of learning sessions. In the Environment dimension, self-regulated learners control the setting for their learning and eliminate distractions. In the Social Context dimension, learners choose whether to learn alone or with others; they also choose from whom to seek help.

In each dimension, the key task condition is *freedom.* For example, if the learner is not free to set his or her own learning goals, then the learner is not self-regulating within the Motivation dimension. If instead of being free to choose the most appropriate learning strategy, the learner is required to use a specific strategy, then the Strategy dimension is being other-regulated, not self-regulated. Freedom is necessary for self-regulation to take place.

Though the principles of self-regulated learning can be applied to the learning of non-academic skills such as swimming or welding, the term *self-regulated learning* is usually applied to the way learners function in school (Dinsmore, Alexander, & Loughlin, 2008). Because academic studying occurs outside of class, students have freedom to choose how, when, and where to study. For that reason, research on academic studying and research on self-regulated learning are inextricably linked. Research has established that self-regulated learning is associated with academic achievement (Boekaerts, 1997; Pintrich & De Groot, 1990; Pintrich et al., 1993; Zimmerman & Martinez-Pons, 1988). If we accept the premise that self-regulated learning ability is beneficial to the student, the next logical question is: Is the level of self-regulatory ability an inherent characteristic of each individual, like height or skin color? Or can self-regulatory skills be taught and learned? The following overview of the literature will address that question in the context of academic studying, and for college students in particular.

Meta-Analyses of Study Skills Interventions

There is evidence that educational interventions can help students learn to study more effectively. A meta-analysis of 51 study skills interventions for K-12 and college students found an overall average effect size of 0.45 (Hattie et al., 1996). Restricting the analysis to interventions for college students resulted in a lower, but still respectable, average effect size of 0.27. The largest effect sizes occurred when the performance outcome was closely related to the study skill training. In a meta-analysis of K-12 self-regulated learning interventions subsequent to those in the Hattie et al. meta-analysis, researchers found an average overall effect size of 0.69 (Dignath & Büttner, 2008). Longer-term interventions resulted in larger effect sizes than shorter-term interventions. Topic-specific average effect sizes varied substantially by educational level, but overall effect size did not. For reading performance, the average effect size was 0.44 for primary school interventions and 0.92 for secondary school interventions. This pattern was reversed for mathematics performance, with an average effect size of 0.96 in primary school but only 0.23 in secondary school. The researchers postulated that this difference may have been due to older students' decreased confidence in their ability to be successful in mathematics.

Shortage of Rigorous Achievement-Based Intervention Research in Authentic Settings

While intervention research supports the notion that self-regulated learning can be improved by training, there is a shortage of studies connecting self-regulation training to a meaningful achievement-based outcome measure. This is particularly true for interventions aimed at college students. Too often, self-regulation training is conducted by researchers in a laboratory setting, instead of by classroom instructors in a classroom. Too often, the intervention's effectiveness is measured either by a questionnaire or by a quiz covering academic content learned just for the research, rather than content learned for a course grade.

As early as the mid-1990s, researchers pointed out the need for postsecondary intervention research to be situated in an authentic context and evaluated with an authentic achievement measure. A systematic review, restricted to the ERIC database and the years 1986–1991, found only 20 empirical investigations of stand-alone study strategies courses (Kaldeway & Korthagen, 1995). Only one of these investigations examined whether the study strategies course affected achievement in students' other courses. The other reviewed studies focused on exam grades in the study strategies courses or quizzes over content learned specifically for research purposes. In another review, covering a similar time frame (1989–1995) and based on a self-regulation framework, the inclusion criteria stressed freedom and authenticity (Hadwin & Winne, 1996). To be reviewed, a research study needed to measure students' learning of legitimate course content for a grade, and to provide empirical evidence about whether the students chose to use the taught strategies on their own. The researchers found only 16 studies meeting these criteria.

In the mid-1990s, self-regulated learning was a relatively new concept. So perhaps it is not surprising that most researchers evaluated study strategy training in a carefully controlled setting, rather than examining whether the students chose to apply the strategies in their content classes. In the nearly twenty years since these reviews, self-regulated learning has mushroomed from a novel idea to an established theory, and is now a popular topic for educational research. However, although researchers are more knowledgeable about self-regulated learning as a process and the characteristics of self-regulated learners, the shortage of authentic intervention research has not been rectified.

As described in Chapter II, I used defined criteria to conduct a systematic review of research reports on self-regulated learning interventions for college students. The search covered the years 1994–2013 and was restricted to scholarly journals in the ERIC database. Zimmerman's dimensional framework (1994, 1998) was used to organize the 42 reviewed articles. The review showed that (1) interventions designed to improve college students' self-regulation can improve achievement, and (2) there is a continued scarcity of credible studies of interventions in content classes.

In general, the reviewed studies showed an inverse relationship between the rigor of the research design and the authenticity of the intervention's setting. Many of the reviewed studies occurring in actual classes did not have solid research designs and analyses, and thus did not provide useful information. Some of these studies measured the intervention's effectiveness by looking at gain scores on an inventory, without using a control group. Other studies had control groups, but used only an inventory score for an outcome measure, instead of course or exam grades. Of the studies that were well-designed and used appropriate analyses to support the researchers' inferences, many were conducted in laboratory settings rather than actual classes. Often the researchers recruited participants from psychology classes that imposed a research participation requirement, and the content the students learned for the research did not count in the student's grade for any course.

Only 13 of the 42 reviewed studies had credible research designs (with appropriate analyses and some level of control for confounders), occurred in an authentic setting (as opposed to a laboratory), and utilized an achievement-based outcome variable (as opposed to an inventory, a grade in a study strategies course, or a quiz over content learned in a laboratory setting). All but one of these 13 studies showed positive effects on achievement, reinforcing the past findings that students' self-regulated learning skill can be developed. Of these 13 studies, only seven occurred in content classes (e.g., biology or computer programming). The other six evaluated stand-alone study strategies courses.

The Chapter II review resulted in a recommendation for rigorous evaluations of interventions in content classes. In particular, we need to examine the effect of manageable interventions that can be implemented by an individual teacher using easily available resources and a reasonable amount of class time.

Interventions Involving Goal Setting, Planning or Reflection Assignments

Before describing the current study, I will review several examples of a specific type of teacher-manageable intervention—worksheets or assignments designed to foster goal setting, time management, planning or reflection. All these interventions were implemented in postsecondary content classes, and all are manageable by an individual teacher. I selected these particular studies because their interventions shared some similarities with the study-journaling intervention used in the current study. Five of these studies (Fitch et al., 2012; Georgianna, 2009; Goodwin & Califf, 2007; Sweidel, 1996; Williams et al., 2011) turned up during the systematic review of Chapter II, and the other two (Fleming, 2002; Zimmerman, Moylan, Hudesman, White, & Flugman, 2011) arose out of less structured database searches. Though this is not an exhaustive list of such interventions, or even necessarily a representative sample, it provides information about the potential value of such interventions.

Two interventions asked students to complete reflective assignments about missed questions on exams or quizzes. The first targeted mathematics students at a technical college, some in developmental algebra and some in the introductory college-level algebra class (Zimmerman et al., 2011). Instructors demonstrated error-detection strategies and asked students to verbally explain their problem-solving and error-finding strategies. Students used self-reflection forms to analyze their quiz errors and work alternative problems. The self-reflection forms also had space for the students to record the number of practice problems they had worked and the time they spent studying for that particular topic. On tests, students were asked to self-monitor their performance by giving numerical confidence ratings to each problem. Treatment students outperformed control students on class exams and on a standardized placement test given after the course. However, these results should be interpreted with caution, due to lack of control for instructor differences, including differences in assessment practices.

In a similar intervention, biology students analyzed midterm questions that were missed by 40% or more of the class (Williams et al., 2011). The exam analysis resulted in improved performance on topic-matched final exam questions. The design of the study, with all students assigned to the control group for some questions and the treatment group for others, did not allow the researchers to evaluate whether the students used the exam analysis strategy on their own. Because the reflective assignments in these two studies (Williams et al., 2011; Zimmerman et al., 2011) involved reworking specific mathematics or biology problems, they provided additional content-specific reinforcement. This content reinforcement distinguishes them from the remaining interventions discussed here and from the intervention in the current study.

Developmental writing students created written action plans, called *implementation intentions*, for upcoming academic tasks (Georgianna, 2009). For each goal, they listed the actions they would take to meet the goal, the benefits of reaching the goal, the obstacles they might face, and the steps needed to overcome each obstacle. Students creating implementation interventions for weekly essays had higher essay scores than students not creating implementation intentions for their essays. However, the training did not result in improved scores on essays written three or more weeks after the training. This indicated the students had difficulty applying the training on their own.

In a study strategy portfolio intervention for psychology students, students submitted reflective journal entries before and after each test (Sweidel, 1996). In the journal entries, they created a study plan, predicted their test grades, reflected on how well they implemented their plan, and described how their study strategies connected to the test grade they received. The portfolio also included a 3–5 page reflective essay over their study strategies throughout the whole course. Both the journal entries and the essay received feedback and were graded for quality. In a helpfulness survey, the students indicated the strategy portfolio had a strong positive impact on their studying. Although the students felt the portfolios were extremely time-consuming, they overwhelmingly recommended the intervention be continued. The intervention's impact on the students' course grades was not evaluated.

In another intervention for psychology students, treatment students completed goal-setting worksheets in which they listed specific and attainable goals, created timelines to meet the goals, and rated their progress toward each goal (Fitch et al., 2012). They also reflected upon past goals, along with the actions they had taken to reach them and the obstacles they had overcome. At regular intervals, they shared their goals and progress in small student-led groups. Treatment students scored higher than control students on several dimensions of a learning strategies inventory. The study would have had considerably more value if it had examined whether the intervention affected the students' course grades or GPA.

In a third intervention targeting a psychology course, students spent the last five minutes of each class setting goals and planning the number of minutes to spend reading, reviewing notes, and studying course material (Fleming, 2002). The students also recorded the actual minutes spent. Control students spent the same amount of time working brainteasers. On three of the four exams, treatment freshmen scored higher than control freshmen and just as well as treatment upperclassmen. Results are not definitive, but indicate that a relatively minor intervention, taking only five class minutes and no teacher grading time, has potential to help students, particularly first-year students, study more effectively.

In a time-management intervention for computer programming students, a time-management expert presented treatment classes with two 20-minute training sessions, spaced four weeks apart (Goodwin & Califf, 2007). Throughout the semester, teachers collected weekly time-management worksheets, on which students planned their time and recorded their actual time use. When analysis was restricted to students taking the first exam, treatment students succeeded (earned grades of A, B, or C) at higher rates than students in same-teacher control classes. Treatment students also outperformed control students on the two midterm exams and on the final course average. This study's consistent positive results and its relatively well-controlled design provide evidence that time-management interventions in content courses may be valuable to students. It would be interesting to learn whether the students' improved achievement was primarily due to the weekly time sheets, or to the time-management training, or if both components were necessary. The weekly time sheets are certainly manageable by an individual instructor; to implement the time-management training, the instructor might need assistance or materials from the college's student success center.

These seven studies provide some evidence that college students can benefit from assignments or worksheets focused on goal setting, planning, or reflection. However, because of the small number of studies and the lack of rigorous design in some of the studies, the evidence is not conclusive.

Research Recommendations From the Literature

The review in Chapter II, along with the additional studies and review articles summarized here, and the recommendations of experts in the field of self-regulated learning, suggest several priorities for intervention research. There is a need for empirical studies that (a) tie study strategies and self-regulated learning to actual content in an authentic setting (Chapter II; Hadwin & Winne, 1996; Schunk & Zimmerman, 1994); (b) examine how teachers can facilitate development of self-regulatory processes (Greene & Azevedo, 2007; Schunk, 2008; Schunk & Zimmerman, 1994); (c) connect self-regulated learning observations and interventions to achievement measures (Chapter II; Schunk, 2008); (d) use qualitative methods to provide a more nuanced picture of the intervention's effect (Chapter II; Rohwer, 1984); (e) use observational data, not self-reports, to assess self-regulation levels (Dinsmore et al., 2008; Schunk, 2008); (f) explore self-regulation growth over time (Schunk, 2008; Schunk & Zimmerman, 1994); and (g) phase out self-regulated learning support and examine whether students continue to apply self-regulation strategies after the supports are removed (Chapter II). The current study contributes to (a), (b), (c), (d), and, to a lesser degree, (e).

Context, Scope, and Purpose of the Current Study

The purpose of the current study was to investigate the value of a study-journaling intervention for developmental mathematics students at a large urban community college in Texas. Specifically, I wished to know how the intervention affected student success in the course and on the final exam, and what it revealed about the students' study habits. The intervention was based on self-regulated learning theory and included goal setting, planning, self-monitoring, and reflection components. The intervention was relatively simple and could be carried out by a teacher without significant preparation or loss of class time.

Nine developmental mathematics classes implemented the study journal project; another nine classes served as the comparison group. For ease of discussion, students in classes receiving the intervention will be referred to as either treatment students or study journal students; their study logs, goal sheets, and reflective writings will be collectively referred to as study journals. Students in control classes did not keep study journals and will be referred to simply as control students.

The following research questions guided this study:

1. Are study journal students more likely to pass the course and the final exam than control students?
2. What are the perceptions of the study journal students regarding the effects of the study-journaling process on their study habits and academic performance?

3. What are the study habits of the study journal students, as shown by their written goals, study logs, and reflective writings?
4. For the study journal students, which of these study habits distinguish successful students from unsuccessful students?

Methods

Preliminaries

Rationale for Mixed Methods Research Design

The nature of the intervention motivated my choice of a mixed methods research design. The need for quantitative analysis of the intervention’s effect on mathematics course success has already been documented. A qualitative component provides additional insights into how the intervention affected the students—information not captured by the quantitative success data. The first two research questions were driven by the need to find out how the intervention affected success.

The third research question arose from a recommendation from the self-regulated learning literature, combined with the nature of the data created through the intervention. When determining the extent to which students apply self-regulated learning strategies, researchers have recommended using observational data rather than relying upon self-reports (Dinsmore et al., 2008; Schunk & Zimmerman, 1994; Schunk, 2008). In a laboratory setting, direct observation of the studying process may be possible. However, studying for college classes rarely occurs in laboratories—it occurs in kitchens, bedrooms, coffee shops, libraries, and campus study nooks. Authentic college studying, by its very nature, is unobservable by researchers.

In this research study, as in others, it was impossible to directly observe the participants' study habits. However, the study journal intervention produced a rich data set: the written goals, reflections, and study records of the study journal students. These study journals could serve as windows, allowing us a glimpse into the students' study habits, their thoughts about studying and about mathematics, their goals and strategies, and the obstacles they face. By qualitatively analyzing the study journals and organizing the qualitative findings around self-regulated learning theory (Zimmerman's dimensional framework), we could form a picture, albeit a fuzzy one, showing how the students were studying and whether they incorporated self-regulated learning strategies. The third research question addresses this picture, and the fourth research question connects this picture to mathematics course success.

For these reasons, I chose a mixed methods design with two strands: (1) a confirmatory strand in which I evaluated the effectiveness of the study-journaling intervention, and (2) an exploratory strand in which I sought information about the study habits and strategies of the study journal students. The first research question anchors the confirmatory strand; the second question supplements it. The third question anchors the exploratory strand. The last research question seeks connections between study habits and success, linking the two strands and thus linking the qualitative and quantitative data, as recommended by Tashakkori and Creswell (2007). Table 5 describes the type of data used to answer each question.

Table 5

Data Types Used to Answer Research Questions

Research Question	Strand	Data Type	How Answered
1. Are study journal students more likely to pass the mathematics course and the final exam than control students?	Confirmatory	QUAN	Statistical comparison of treatment and control students on binary outcome variables representing final exam success and course success
2. What are the perceptions of the study journal students regarding the effects of the study-journaling process on their study habits and academic performance?	Confirmatory	qual + quan	Surveys (beginning and end of semester) Focus groups
3. What are the study habits of the study journal students, as shown by their written goals, study logs, and reflective writings?	Exploratory	QUAL	Qualitative coding and analysis of study journals
4. For the study journal students, which of these study habits distinguish successful students from unsuccessful students?	Connects the two strands	QUAL→quan	Classify themes of study habits as present or absent for each student; use quantitative analysis to determine which study habits are associated with success

Note. Capitalization (QUAL or QUAN) indicates data type with dominant status. Lower case (qual or quan) indicates less dominant status (Creswell & Plano Clark, 2000).

The study has an unusual design feature: the collection of the qualitative data served as the intervention. At first glance, the research design does not fit neatly into the established landscape of mixed methods designs. To classify it, we must consider the two strands separately. The confirmatory strand follows the Embedded Experimental Model, in which qualitative data are used to further explain the results of an intervention. The exploratory strand uses the Data Transformation Model, in which qualitative data are quantified and then analyzed quantitatively (Creswell & Plano Clark, 2007, pp. 62–67). When the strands are combined, the study can be classified as a parallel/simultaneous equivalent status mixed model design (Tashakkori & Teddlie, 1998, p. 43) or a fully mixed concurrent equal status design (Leech & Onwuegbuzie, 2009)[1]. See Figure 2.

[1] The terms *mixed model* and *fully mixed* indicate that the qualitative and quantitative data are mixed in at least one of the three stages prior to interpretation: research inquiry, data collection, and analysis. The terms *parallel*, *simultaneous*, and *concurrent* refer to the fact that both types of data are collected at approximately the same time. The term *equal status* indicates that the qualitative and quantitative data carry approximately equal weight.

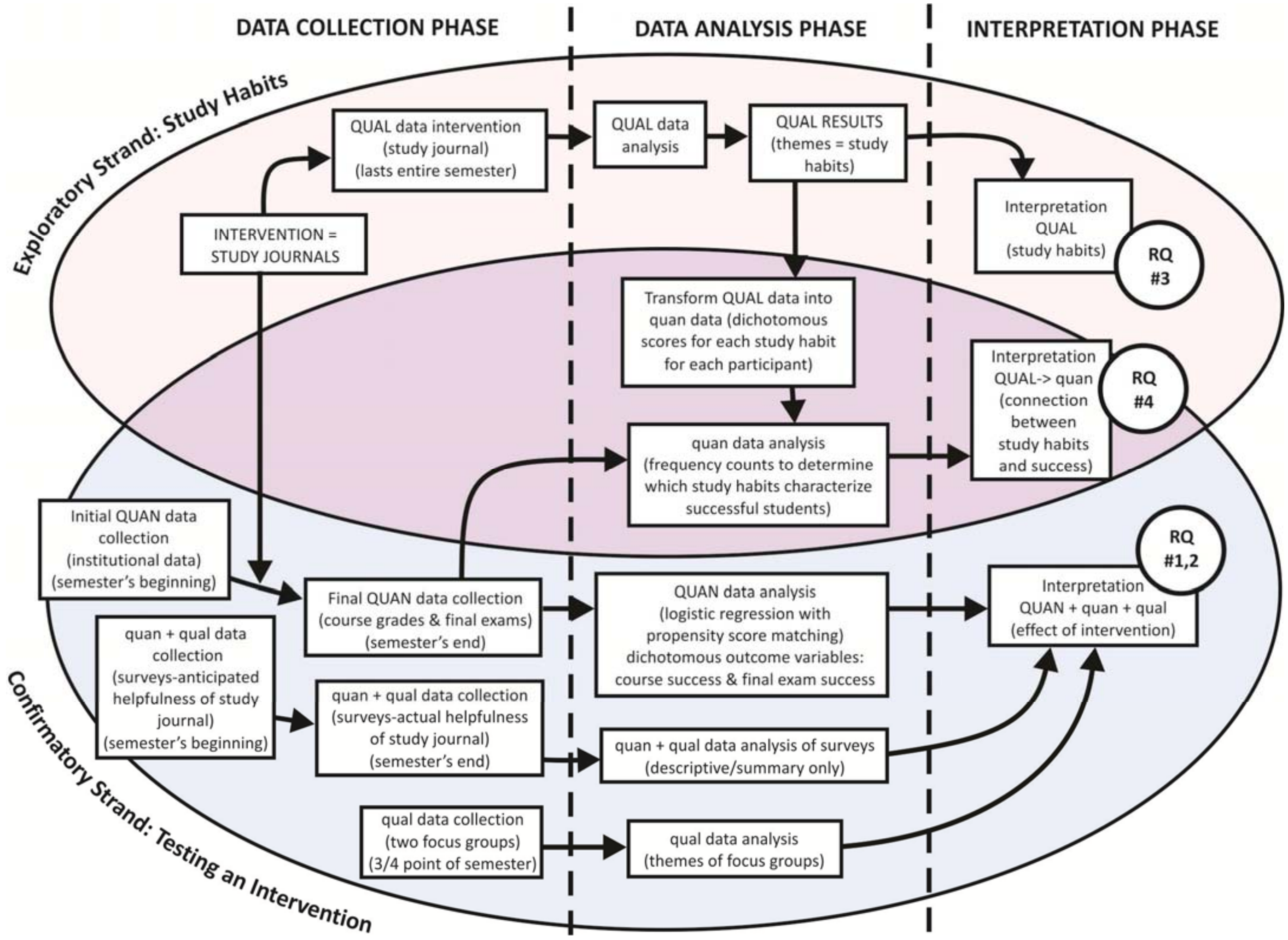

DATA COLLECTION PHASE
DATA ANALYSIS PHASE
INTERPRETATION PHASE
Exploratory Strand: Study Habits
QUAL data intervention (study journal) (lasts entire semester)
QUAL data analysis
QUAL RESULTS (themes = study habits)
INTERVENTION = STUDY JOURNALS
Interpretation QUAL (study habits)
RQ #3
Transform QUAL data into quan data (dichotomous scores for each study habit for each participant)
Interpretation QUAL-> quan (connection between study habits and success)
RQ #4
quan data analysis (frequency counts to determine which study habits characterize successful students)
Initial QUAN data collection (institutional data) (semester's beginning)
Final QUAN data collection (course grades & final exams) (semester's end)
QUAN data analysis (logistic regression with propensity score matching) dichotomous outcome variables: course success & final exam success
RQ #1,2
Interpretation QUAN + quan + qual (effect of intervention)
quan + qual data collection (surveys-anticipated helpfulness of study journal) (semester's beginning)
quan + qual data collection (surveys-actual helpfulness of study journal) (semester's end)
quan + qual data analysis of surveys (descriptive/summary only)
qual data collection (two focus groups) (3/4 point of semester)
qual data analysis (themes of focus groups)
Confirmatory Strand: Testing an Intervention

Figure 2. *Research design. Capitalization (QUAL or QUAN) indicates data type with dominant status. Lower case (qual or quan) indicates less dominant status (Creswell & Plano Clark, 2000).*

Study Journal Intervention

Treatment class instructors were asked to have their students complete two worksheets each week, collectively referred to as *study journals* (see Appendices B and C). The questions on both worksheets were specific to the students' mathematics class—not their English class or their life. The worksheets were based on key elements of self-regulated learning theory: goal setting, planning, self-monitoring, and reflection. On the first worksheet, called the *goal sheet*, Questions 1, 2, and 7 pertained to setting goals and planning; Questions 3, 4, 5 and 6 called for reflection. The second worksheet, called the *study log*, was designed for self-monitoring of the actual study sessions. For each study session, it contained spaces for recording the start time, end time, and location of the session, and any other people who were present; it also contained spaces for recording the goals for the session, the task tackled during the session, and how well the student met the session's goals. The study log was intended to be completed outside of class. Most instructors opted for students to also complete the goal sheet outside of class. Instructors counted the study journals as a small part, three percent or less, of the overall course grade.

Setting, Participants, and Treatment Assignment

The setting for this study was one campus of a large multi-campus community college system. This campus enrolled about 17,000 students during the semester of the study (Texas Higher Education Coordinating Board, 2013, p. 71). The campus is ethnically diverse: about 36% Hispanic, 32% Black, 19% White, and 5% Asian (Lone Star College System Office of Research and Institutional Effectiveness, 2013). About

42% of students are over 24 years old and about 69% attend college part-time (Lone Star College System Office of Research and Institutional Effectiveness, 2013).

As in most community colleges, many students arrive unprepared for college-level classes. In 2011, over 64% of first-time-in-college students did not meet state college readiness standards in mathematics (Texas Higher Education Coordinating Board, 2012). Most of these students, along with about half of the students meeting the college readiness standards, enrolled in developmental mathematics (Texas Higher Education Coordinating Board, 2012). Overall, about 30% of students at this campus are in developmental English or developmental mathematics classes (Lone Star College System Office of Research and Institutional Effectiveness, 2013). During the fall semester of 2012, there were 3,578 students enrolled in developmental mathematics courses, more than twice the 1,548 students in all credit-level mathematics courses combined (Lone Star College System, 2013). In a typical semester, less than 45% of enrolled students complete their developmental mathematics courses with a passing grade (Lone Star College System Office of Research and Institutional Effectiveness, 2011).

This study targeted students enrolled in Introductory Algebra or Intermediate Algebra, the second and third courses in the three-course developmental mathematics sequence offered by Texas community colleges. Intermediate Algebra is the prerequisite for College Algebra, which for most students is the first transferable mathematics course.

After the appropriate institutional review board approvals were obtained, 18 classes were selected. In this project, as in much educational research, randomization of individuals was impossible. The only feasible way to implement the study-journaling intervention was with intact classes. Within this constraint, the goal was the same as for a randomized trial: to create control and treatment groups that did not differ in systematic (nonrandom) ways. For quasi-experimental research to approach this ideal, the researcher must control which group receives the treatment, or at least have no reason to suspect differential recruitment because of the treatment (Campbell & Stanley, 1963). If there are systematic differences between naturally occurring groups, the researcher must decide what treatment assignment mechanism best handles them. Matching and subsequent randomization can distribute systematic differences across different treatment conditions, sometimes resulting in a better design than randomization alone (Campbell & Stanley, 1963)

For this reason, classes were selected in pairs, matched on as many class-level variables as possible, and then randomized into the treatment (study journal) and control conditions. All instructors understood the purpose of the study and what the study-journaling project would require of them. All instructors were willing to participate in either the control or the study journal conditions; all of them understood that the conditions would be randomly assigned. The study-journaling project was not expected to either attract students to particular classes or drive them away, because neither the students nor the advisors would know about the project in advance.

Classes were matched, as far as possible, on the following characteristics: course, number of meetings per week, instructor, and time of day (morning, afternoon, or evening). In an ideal pair, both classes would be the same course, and both classes would be taught by the same instructor, either in back-to-back time slots on the same day or in the same time slot on different days. Not surprisingly, this was not always possible. Priority was given to course and number of meetings per week. The next priority was the instructor, then the time of day. No Introductory Algebra class was paired with an Intermediate Algebra class, and no three-day-per-week class was paired with a two-day-per-week class. After the pairs were created, a dice roll was used to randomly assign one class to the treatment condition and the other to control.

The group of classes included four pairs of Introductory Algebra classes and five pairs of Intermediate Algebra classes. Two of the nine pairs of classes met in the evening. Twelve of the 18 classes were taught by full-time faculty members, and six were taught by adjuncts. There were four same-instructor pairs of classes.

During the first two weeks of the semester, I visited each control and treatment class, provided snacks, and explained that the class had been chosen for a research study. I distributed informed consent forms and explained what information would be used if they agreed to participate. In both groups, consenting students allowed their educational records, course grade, and final exam to be used in the research study. In the study journal classes, consenting students also agreed to share their study journals. In my visits to the study journal classes, I explained the purpose of the research study and described the study-journaling process. I explained that if they opted not to participate in the research, they would still complete the study journals as part of their course grade. In the control class visits, I told the students we were researching study habits and student success, but omitted the details about the study-journaling project. In both groups, I emphasized that participating in the research study was optional, and that if they consented to participate, they could change their minds at any time with no consequence.

There were 117 study journal students and 141 control students who consented to share their data for the research study, for a total sample size of 258. Overall, 48.7% of the 530 enrolled students chose to participate.

Data Collection

Confirmatory Strand: Effect of the Intervention

Final Exams

For each of the three developmental mathematics courses (Prealgebra, Introductory Algebra, and Intermediate Algebra), the mathematics department at this campus has developed a departmental final exam. For each course, the departmental final has three versions. The problems on all versions are the same in their essentials—the numbers differ, but the solution process is identical.

I planned for all classes involved in the research study to use Versions A or B of the final exam. The two versions were distributed as equally as possible across the control and treatment groups by pairing classes and giving each pair the same version(s). For example, if a control class used only Version A, it was paired with a treatment class also using only Version A. If the instructor of a control class wanted to use both Versions A and B, then the class was paired with a treatment class whose instructor also agreed to use both Versions A and B. (Instructors with crowded classrooms often preferred to use two versions, so students sitting next to each other would not have the same test.)

Though the exams were the same in both groups, the instructors graded their own students' exams and made their own partial credit decisions. Thus, inconsistencies in grading could potentially confound the treatment effect estimate. For this reason, the instructors' final exam grades were not used in the statistical analysis. Instead, the final exams of participating control and treatment students were photocopied before the instructors graded them. After the semester was over, the final exams were all graded by me, incorporating interrater reliability checks with another grader on a sample of the exams. In the statistical analysis for the research, my final grades were used. The instructors' final exam grades were used to calculate the students' official course grades.

Surveys

At the beginning of the semester, study journal students completed a short survey, including both Likert-style and open-ended questions, about the study journal's anticipated helpfulness to them. Near the end of the semester, the study journal students (those who remained in the class) completed a similar survey regarding their perceptions of the study journal's actual helpfulness.

College Transcripts

After the semester ended, I obtained unofficial college transcripts for all enrolled students who had signed consent forms agreeing to share their final exams, study journals, and college records. There were a few students who had signed consent forms but who were not on the official roll (they had dropped or switched classes before the official day of record, which occurred approximately two weeks into the semester). I did not use transcripts or other data on these non-enrolled students, and did not include them as participants in the study.

Focus Groups

At about the two-thirds point of the semester, I arranged for two instructors to invite students to stay after class for an informal focus group. There were two focus groups (one for each class), chosen based on classroom availability. (To encourage attendance, I wanted the focus group to meet immediately after class, in the same classroom.) I provided pizza for the students who volunteered. For each focus group, another faculty member observed and took notes. Each focus group student signed a consent form agreeing to be audio-recorded for the research. (This was a different consent form from the previously mentioned consent form, in which students agreed to share their study journals and educational records.) Guided by the questions in Appendix D, I asked the students to share their thoughts on how the study journal project affected them and their studying, what aspects they found most helpful or annoying, how much time it took, and whether they thought their teachers should require study journals the next semester.

Exploratory Strand: Study Habits Revealed by the Study Journals

Study Journals

The goal sheet and study log were typically completed longhand on paper. Whenever an instructor collected the study journal forms, he or she distributed a new set of blank forms. On the mathematics department website, students could access both study journal forms in two ways: (1) as pdf files, so they could print the sheets if they missed class or needed replacements; and (2) as Qualtrics surveys, so they could submit the forms electronically if they felt comfortable doing so. If they wished, students could sign up for text message reminders. These reminders, typically “don’t forget to log your study time” or “turn in your goal sheets Thursday,” continued until about the halfway point of the semester.

Because the study journals would affect students’ grades, the students could possibly feel pressure to fill their study journals with what they thought their teachers wanted to see, rather than their actual goals and study time. I wanted the qualitative data (study journals) to reflect the students’ thoughts, goals, and study sessions as accurately as possible. For this reason, the instructors agreed not to read the study journals. During the recruitment visits, we assured the students that the instructors would not read their journals and stressed the importance of filling out the study journals honestly and accurately, both for the students’ own benefit and the benefit of the research. If they didn’t study their mathematics for an entire week, we directed them to write “I did not study all week” in their study logs, to reflect on the reasons this happened, and to consider whether they should do anything differently the next week.

Because the instructors had promised not to read the study journals, they collected the journals and gave them to me. I recorded which students satisfactorily completed the goal sheet and study log each week, and gave this list of study journal “grades” to the instructors, who then added the grades to their class gradebooks. Study journal grades were counted in the course grade for all students, regardless of whether they decided to participate in the research. If students wrote that they did not study all week, they still received credit for completing the study journal.

After the end of the semester, the study journals of students not participating in the research study were removed and discarded. Then, for each participating student, the goal sheets were put in chronological order and stapled into a packet. The process was repeated for the study logs. As long as a student submitted at least one of each sheet, the student would have two stapled packets (one packet of goal sheets and one packet of study logs.) After this organization process was complete, the participating students' study log and goal sheet packets were photocopied. After labeling the photocopied packets with the same identification numbers that had been assigned for the quantitative analysis, I masked the students' names. I locked the originals in a safe place and used the numbered photocopies for data analysis.

Data Analysis

Confirmatory Strand: Effect of the Intervention

Outcome Variables

To address the first research question, concerning the effect of the intervention, the treatment and control groups were compared on four binary outcome variables: course success and three versions of final exam success. The three versions of final exam success were created using three different cut scores on the departmental final exam. On all four outcome variables, students not finishing the course were counted among the unsuccessful and included in the analysis.

The four outcome variables should be thought of as four different representations of the same construct—mathematics course success. Course success was included as an outcome because it is meaningful to both the student and the institution. A passing course grade is evidence that the student, according to the instructor's professional judgment, has met the minimum standard described in the syllabus; it allows the student to proceed to the next mathematics course in his or her degree plan. However, from a research standpoint, final course grade as an outcome variable has limited value, due to teacher differences in expectations and grading policies. Although the instructors were required to cover the same learning objectives, they wrote their own tests, except for the final exam. Outside of a requirement to count the departmental final exam for at least 20% of the final course grade, instructors had freedom to set their own grading requirements. They wrote their own syllabi and had varied policies on homework, quizzes, and attendance. The inevitable confounding effects of instructor policies on the course success variable motivated the decision to also use final exam success as an outcome. Essentially, departmental final exam success, because it was less confounded by teacher differences, was used as an alternative measure of course success.

The final exam success variables were based on cut scores of 70, 60, and 50 on the 100-point final exam. The first cut score, 70, was chosen because 70% is traditionally regarded as a cutoff to earn a C grade, and also because departmental guidelines require students to have a 70% course average in order to pass with a C or higher. The third cut score, 50, was chosen because departmental guidelines also require students to score at least 50% on the departmental final exam in order to pass with a C or higher. (Students with excellent grades on their midterm exams and other assignments could potentially have a course average of 70% even if they only scored 50% on the final exam.) The cut score of 60 was chosen to provide more nuanced information about the students scoring between 50 and 70. Using three different cut scores provided a fuller picture of the intervention's effect on final exam success.

Why were binary outcomes chosen, instead of course letter grade or percentage grade on the final exam? Dichotomizing a continuous or ordinal variable should not be done lightly, as it discards information by collapsing the variable into only two possible outcomes—one represented by 1 and the other by 0. In this case, the decision to dichotomize was driven by two issues: (1) the large percentage of students who leave the class, either by officially withdrawing or by "disappearing," and (2) the difficulty of distinguishing between various categories of unsuccessful students, due to factors such as financial aid rules and differences in teacher policies.

At some four-year institutions, course withdrawals are strictly limited and therefore rare, perhaps making it reasonable to omit withdrawn students from the analysis. That is not the case for this community college, however, especially for developmental classes. Over 15% of students typically withdraw from these developmental mathematics classes; a significant number of additional students stop attending class without officially withdrawing. Because poor performance may be one reason students leave the class, it is important to include departing students when examining an intervention's effect upon student success. After deciding to include departing students, the next step was to decide how best to do so: was it best to put these students in a category of their own, or lump them together with other categories of students? In this study, students withdrawing or disappearing were combined with other students who did not pass the class. The following descriptions of the outcome variables will also detail the reasons for this choice.

On the course success outcome variable, students were considered successful if they received grades of A, B, or C, allowing them to move to the next class. Students receiving grades of IP (In Progress), F (Failing), or W (Withdrawn) cannot move on and were classified as unsuccessful. Grades of IP, F, and W were considered equivalent because it is nearly impossible to disentangle these grades from one another, or to decide what each grade actually represents in terms of performance. According to the college's official course catalog, students who earn a grade of IP (In Progress), which is only awarded in developmental classes, "have participated fully in the class but have not met all criteria for making progress to the next level of courses." A grade of F means *failing* and a grade of W means *withdrawal* (Lone Star College System, 2012, p. 75). The catalog describes, in reasonably clear fashion, the different situations these grades are intended to represent. In practice, these distinctions are blurred. Enrolled students who do not earn a passing grade will receive either an IP or F, depending on the criteria and philosophy of their instructor. Some instructors have strict attendance and participation requirements for an IP, requiring the student to take all exams, including the final exam, to submit all homework and computer lab assignments, and to accumulate fewer than five absences. Other instructors feel an F in a developmental class is punitive; they give

IP grades to all students who do not pass, even those who stop coming to class or submitting assignments.

Instructor policies also affect W grades, because instructors can award grades of W by dropping students not meeting attendance requirements. Some instructors scrupulously monitor class attendance and drop students the very day they reach the maximum number of absences allowed by the syllabus. Other instructors never drop students, even if they permanently disappear from class after the second week—disappearing students would receive grades of F (or IP, if the instructor doesn't believe in F grades). Students can also drop themselves, by officially withdrawing before the withdrawal deadline. However, many students do not withdraw, even if they are unable or unwilling to continue attending class due to illness, work, or class performance. Instead, because of financial aid rules and requirements for full-time enrollment, many disappearing students prefer to remain officially enrolled in class. Students sometimes contact the dean to ask for a W (dropped by instructor) to be changed to an F, so they can avoid losing financial aid. For all these reasons, grades of W, F, and IP were treated as equivalent.

On the outcome variables representing success on the departmental final exam, students scoring at the cut score or higher were considered successful. Those not taking the final exam or not reaching the cut score were considered unsuccessful, regardless of whether they were still enrolled in the course.

It should be noted that instructor differences in IP requirements can also affect whether enrolled students take the final exam. If an instructor requires students to take the final exam in order to receive an IP, some students will take the final exam even if they have no hope of passing. They will at least come and write their names on the tests, perhaps trying a few problems, or perhaps leaving the whole test blank. If an instructor does not require the final exam for an IP, students who do not expect to pass may stay home. The zero or near-zero grades of these students would have made it difficult to draw meaningful comparisons between the means of the control and treatment groups. Dichotomizing the final exam variable treats all students who gave up on the class as equivalent, whether they officially withdrew, remained enrolled and took the final exam, or remained enrolled and did not take the final exam. Dichotomous outcomes also ensured there would be no attrition on any outcome variable; all students who began in the study were included in the final analysis.

Final Exams

As previously mentioned, the participating students' final exams were photocopied before the instructors graded them. After writing the participants' randomly assigned identification numbers were on all the exam photocopies, I masked the students' names and all references to instructor and class. I created a detailed partial credit rubric and then discussed it with another mathematics instructor in the same community college system, who had agreed to grade a sample of the exams as a reliability check. Random samples of 20% of the Introductory Algebra exams and 20% of the Intermediate Algebra exams were chosen; the exams in the samples were photocopied an additional time so the second instructor could grade them independently.

For consistency, the entire grading process was completed from start to finish for Introductory Algebra before the Intermediate Algebra grading was begun.

The two of us independently graded the Introductory Algebra exams in the 20% sample and recorded the scores for each of the 37 items. Using IBM SPSS Statistics 21, an intraclass correlation coefficient (ICC) was calculated for each item and for the total exam grade. The intraclass correlation coefficient is a measure of interrater reliability for interval data. It captures not only the degree to which one rater's score predicts the other, but also how closely the two raters' scores match (Landers, 2011; Shrout & Fleiss, 1979; Tinsley & Weiss, 2000). The appropriate type of intraclass correlation depends on the situation. For this situation ICC(2,1) with absolute agreement was appropriate, because the same two raters would rate all exams in the sample, because the two raters were a sample from the population of raters who could have made these ratings, because only one rater's ratings would be used in the analysis, and because we were interested not only in consistency but also in the amount of the underlying construct (Landers, 2011; Shrout & Fleiss, 1979; Tinsley & Weiss, 2000). Though total exam score was the variable of interest, individual item scores were also examined for evidence of grading discrepancies. For simplicity, I also used the intraclass correlation coefficient for the items, even though the intraclass correlation coefficient assumes an underlying continuous distribution and does not work well when there is little variance in the scores (Tinsley & Weiss, 2000). This happened on several items, when most students either got the problem completely correct (received the maximum score for that item) or left the problem blank (received the minimum score of 0 for that item). For this reason, I also examined the percent agreement on each item.

The intraclass correlation coefficient and percent agreement for each Introductory Algebra item are recorded in Appendix E, Table E-1. The other rater and I discussed the largest discrepancies, and found and corrected four errors. Even before these corrections, the first grading of the Introductory Algebra sample demonstrated acceptable agreement, with an ICC(2,1) of 0.993 for the total exam grade, and close agreement on item scores. The average difference in total exam score between the two graders was 0.5 points (out of 100) and the maximum difference was 4 points. Satisfied with this level of agreement, I graded the remaining 80% of the Introductory Algebra exams.

After Introductory Algebra grading was complete, the entire process was repeated for the Intermediate Algebra test, which had 35 items. Although the ICC(2,1) for the Intermediate Algebra sample was 0.978, there were unacceptably large differences in total test grade for several students, and also several items that showed high levels of disagreement. Upon discussion, we discovered the primary cause was the absence of advance agreement on how to handle students who showed their work on scratch paper instead of the test, and on how to score items on which the final answer was correct but there was no supporting work. (In this case, we decided to give full credit, assuming that the student's scratch paper was not photocopied when the test was photocopied.) After deciding how to handle missing supporting work and making other minor modifications to the rubric, we both regraded several items on all the tests in the sample, and also regraded several students' tests start-to-finish. I then recalculated the ICC and percent agreement on all items (see Appendix E, Table E-2). After this second grading of the sample, the overall ICC(2,1) was 0.999, the average difference in overall test grade was 0.548 points (out of 100) and the maximum difference was 3.5 points. I then graded the remaining 80% of the Intermediate Algebra exams.

As previously mentioned, I planned for all control and treatment classes to use either Version A of the final exam, Version B of the exam, or both. However, one Intermediate Algebra and one Introductory Algebra class received Version C by mistake. Because all Version C students were in one group, instead of being equally distributed across the control and treatment groups as originally planned, it was important to check whether this affected the results. So, during the final exam grading process, I rechecked the Version C tests, comparing them problem-by-problem with Versions A and B. In particular I examined whether Version C's numbers were bigger or smaller, and whether these differences caused students to receive a different number of points, particularly for students near the cut scores. This did not seem to be the case, and the Version C grades were included in the analysis with no reservations.

Using Propensity Score Matching to Improve the Sample's Credibility

For any evaluation of an intervention's treatment effect to be meaningful, the control and treatment groups should have similar distributions of important starting characteristics, at least on observable variables likely to be associated with the outcome. I made efforts to prevent systematic (nonrandom) differences between the control and treatment groups, by matching the classes, getting as many same-teacher pairs as possible, and randomly assigning one class in each pair to treatment and the other to control. Still, it was possible the two groups were sufficiently different in their starting characteristics to confound the treatment effect estimate. Even when individuals are randomized into treatment conditions, it is wise to check the groups' starting characteristics—dissimilar groups can occur simply because of luck, especially when samples are small. If the groups are unacceptably balanced on important characteristics, the researcher can rerandomize before the experiment begins (Rubin, 2008a).

Matching, when used as a nonparametric preprocessing phase before analyzing outcome data, can reduce bias and model dependence (Ho et al., 2007; Stuart & Rubin, 2007; Stuart, 2010). One approach is to select pairs of participants that are matched on a set of specified individual characteristics (covariates). However, researchers trying to find exact (or close) matches on many covariates, especially when some of them are continuous, will encounter a dimensionality problem. Even with large samples, finding matches on all the covariates may be impossible (Guo & Fraser, 2010, p. 132; Ho et al., 2007; Rosenbaum, 1995, p. 200).

The propensity score, a single scalar that contains information from many covariates, is a useful tool for dealing with this difficulty. Matching on the propensity score—instead of on the covariates themselves—can reduce bias and model dependence without the need for exact matches on many covariates. When used for this purpose, propensity score matching is not intended to produce paired data, but rather to choose a subset of the control group and a subset of the treatment group with similar distributions of the participants' background characteristics (Ho et al., 2007).

The propensity score of an individual is defined to be the conditional probability that the person will be assigned to the treatment group, given that particular person's vector of covariate values. If the groups were created by randomly assigning half the individuals to treatment and half to control, each person's propensity score would be 0.5. If the cases are stratified by covariate values and then a fixed number from each stratum is randomly chosen for treatment, the propensity score may not be 0.5, but could still be calculated. When the individuals are not randomized into treatment conditions, as in most observational studies, the exact value of the propensity score will not be known. Instead, the propensity score must be estimated by fitting a model, which uses the covariate values as predictors and treatment assignment as the outcome (Rosenbaum & Rubin, 1983).

When estimating propensity scores, it is best to select covariates generously, including all background variables that could potentially affect treatment assignment or outcome. Including a variable that turns out to be unrelated to treatment assignment does not cause serious problems, because that variable will have only a small influence on the propensity score. But if the variable is related to treatment assignment, omitting it from the propensity score model will bias the evaluation of treatment effect (D'Agostino, 1998; Luellen, Shadish, & Clark, 2005; Rosenbaum & Rubin, 1983; Stuart & Rubin, 2007; Stuart, 2010).

In the current study, academic history variables were used to estimate the participants' propensity scores. Because I was interested in academic outcomes, I wanted the groups to have similar distributions on the chosen academic history variables. Values for these variables were taken from the participating students' unofficial college transcripts. (During the consent process, participating students agreed to share their college educational records.) Though the transcripts also contained data for the course success outcome variable, these outcome data were not recorded until after the propensity score matching was complete and the resulting groups were satisfactorily balanced in their academic histories, as recommended by Rubin (2007, 2008b). An advantage of using propensity score matching to adjust for group differences is that that it does not involve outcomes. This essentially eliminates the possibility of researcher-created model bias derived from the researcher trying different models and choosing the one that best fits the desired outcome (Hill, Rubin, & Thomas, 2006; Ho et al., 2007; Rubin, 2007; Stuart & Rubin, 2007).

Using a combination of past research and common sense, I chose a set of key academic history variables. A brief description of these variables is provided in Table 6. The rationale for these variables, along with details on how they were calculated, is presented in Chapter IV. Data from the transcripts were entered into a spreadsheet and used to calculate these key variables (see Appendix I).

Table 6

Key Academic History Variables Used for Propensity Score Matching

Variable	Description
HrsAttF2012	Hours attempted during the intervention semester (includes credit and developmental). Distinguishes part-time students from full-time students.
CumHrsAttPreInt	Cumulative hours attempted before the intervention semester (includes all classes—credit-level, developmental, and grade-excluded). Measures the student's amount of college experience.
YrsSinceStartCollege	Difference between the current year and the year of student's first enrollment in this college system.
YrsSinceMath	Difference between the current year and the year of last previous math class, regardless of whether it was successful (either this course or a different course).
DevMathGPA	Cumulative developmental math GPA pre-intervention. W (Withdrawal) and IP (In Progress) grades were counted the same as F grades (0 grade points/3 hours)
GPAPreInt	Cumulative credit-level GPA prior to the intervention, with grade-excluded classes restored (does not include developmental classes, ESOL (English for Speakers of Other Languages) classes, or other non-transferable classes, such as HUMD 0330 (College Success Course: First Year Experience). Also omits classes with a grade of W (Withdrawn).
CredEarnedPreInt	Total credit hours earned (grades of A, B, C, D, or P) before the intervention semester. (P indicates "pass" in a pass/fail course.) Does NOT include hours earned for "grade-excluded" classes, even if those hours were passed with a C or D. Including grade-excluded classes would have meant "double-dipping" for some students (crediting them twice for the same course; for example, if they got a D the first time and then an A.)
CourseCompletionRatio	The proportion of hours attempted that have been passed, prior to the intervention (Hagedorn & Kress, 2008). Includes all classes (credit, developmental, non-transferable, pass/fail, and grade-excluded). This ratio includes grade-excluded classes, because these hours are incorporated into both the numerator and the denominator, eliminating the double-dipping problem.
PrereqStatus	*A*, *B*, or *C* if student passed the prerequisite course to the current course. *Repeat* if student has previously attempted the current course and earned a D, IP, F, or W. *Placement* if this is the student's first math course at this college (Little, 2002).

Table 6 Continued

Variable	Description
AttemptsPerPass	Represents the number of attempts a student takes, on average, to pass a developmental mathematics course. Serves as a measure of how "on-track" the student is in his or her developmental mathematics sequence. Low numbers are better; a student with a score of 1.00 is considered perfectly on-track and has never repeated a course.
ESOL	1 if ESOL/ESL (English for Speakers of Other Languages/English as a Second Language) appears on transcript; 0 if ESOL/ESL does not appear on transcript. (ESOL classes were previously called ESL classes by the college.)
CurrentCourse	A binary variable representing whether the student was in Introductory Algebra or Intermediate Algebra.

Several of the key variables in Table 6 required modifications before they could be used in the propensity score model. So that the upcoming presentation of the study's results will make sense to the reader, I will give an overview of the most important modification in the upcoming paragraph. Additional details on this modification, and descriptions of the other modifications, are provided in Chapter IV.

Some variables were tricky to conceptualize numerically because they combined an ordinal or ratio scale with a categorical value. One example is grade point average (GPA), calculated by dividing the total number of earned grade points by the number of GPA-eligible hours. If the participants had been seniors at a four-year university, this would have presented no problems. But because this study's participants were developmental mathematics students, many of them (about 23.6%) did not yet have any GPA-eligible hours, either because they had not yet taken any credit-level classes, or because they had withdrawn from all their credit-level classes. The college transcripts listed a 0.00 GPA for these zero-denominator students, the same value listed for students who had failed all their GPA-eligible hours. In the propensity score model, it was important to distinguish these two very different groups of students. This difficulty was handled by creating an indicator variable. For students who had a legitimate GPA, the indicator variable had a value of 1; for students who had zero GPA-eligible hours, it had a value of 0. Then, on the original GPA variable, the mean was computed (including only those students who had GPA-eligible hours). This mean was then imputed to the students without GPAs. Both the indicator variable and the mean-imputed GPA variable were included in the propensity score model. A similar approach was used for the DevMathGPA, CourseCompletionRatio, and PrereqStatus variables.

Of the 12 key variables in Table 6, six were inserted into the matching model in their original form. Two others were used in the matching model after a minor modification (truncation of the range). Three other key variables were replaced by mean-imputed versions supplemented with an indicator variable, as described above. The final key variable was converted from a nominal variable to an ordinal variable and then replaced by a mean-imputed version supplemented with an indicator variable (see Chapter IV). This resulted in a set of 16 matching variables, composed of the 12 key variables from Table 6 (or versions of them) along with four indicator variables.

The propensity score matching was conducted using IBM SPSS Statistics 21, along with an R-based propensity score plug-in created for SPSS (Hansen & Bowers, 2008; Hansen, 2004; Ho et al., 2007; Ho, Imai, King, & Stuart, 2011; Thoemmes, 2012). The package uses logistic regression to estimate the propensity score for each participant, then uses nearest-neighbor matching to match treatment cases with control cases. For simplicity, I chose one-to-one matching without replacement, using a caliper of 0.2 to prevent extremely poor matches (measured in standard deviations of the logit of the propensity score; Thoemmes, 2012). Because nearest-neighbor matching without replacement is order-dependent, the order of the participants was randomized first (Caliendo & Kopeinig, 2008; Reynolds & DesJardins, 2009). See Chapter IV for more details about the propensity-score matching process.

Ideally, the final subsample would achieve two goals: (1) satisfactory balance of covariates between treatment and control groups, and (2) not discarding too many treatment students. In matching, there is typically a trade-off between quality of matches and number of unmatched cases. If extremely strict matching requirements are invoked, many participants may have to be discarded, due to the lack of an available match.

The success of a matching algorithm is assessed by examining the balance of the groups. It is acceptable, even encouraged, to try several different matching algorithms and then choose the one that results in the best balance (Ho et al., 2007; Stuart & Rubin, 2007). If nearest-neighbor one-to-one matching without replacement did not produce acceptably similar groups, the next step would have been to try matching with replacement or *k*:1 matching, or use one of the more robust matching algorithms available in R (a programming language used for statistics). If no matching algorithm produced acceptable balance, the data set may not have been sufficient to support the needed analysis.

Balance was assessed through visual diagnostics and the examination of standardized differences (Ho et al., 2007; Stuart, 2010). For each matching variable (covariate), a standardized difference was calculated by subtracting the means of the control and treatment groups and then dividing the result by the standard deviation of the treatment group in the unmatched sample (Stuart, 2010). After several iterations of the propensity score matching process—with tweaks to the matching variables and different random orderings of the participants—I chose the matched sample that, overall, showed the smallest standardized differences on the covariates. Then, for each covariate, I conducted more detailed balance checks to verify that the groups not only had similar means on each covariate, but also similar distributions. These detailed balance checks, shown in Chapter IV, included frequency counts, boxplots, histograms, and quantile-quantile plots (depending on the type of variable).

In each group, some students did not have a match. These students were discarded from the statistical analysis.

Estimating Treatment Effect

To estimate the effect of treatment condition upon binary outcome variables, logistic regression was the appropriate technique. I used the adjusted sample resulting from the matching phase, treating the two groups as independent samples. Applying a parametric model to matched sets with similar covariate distributions, instead of the original (unmatched) sample, reduces dependence on modeling assumptions and makes the results less sensitive to potential model misspecification (Ho et al., 2007; Stuart & Rubin, 2007; Stuart, 2010). However, the final parametric model should still include those predictors expected to be predictive of the outcome (Ho et al., 2007).

So, in addition to the treatment condition predictor, three other predictor variables were included in the logistic regression model: HrsAttF2012, PrereqStatus, and AttemptsPerPass. These three variables were chosen because they were expected to be predictive of the outcome, they were not redundant with one another, and they did not depend on imputed data points. HrsAttF2012 captured students' full-time/part-time status, shown to be an important success predictor (Cartnal, 1999; Serna, 2011). For PrereqStatus, I used the original (nominal) version, in which the values *A*, *B*, *C*, *Repeat*, *Placement* were treated as unordered categories. (The propensity score model used an ordinal version of PrereqStatus, in which the non-*Placement* students' mean was imputed to the *Placement* students, in combination with an indicator variable that captured their *Placement* status.) PrereqStatus, which captured information about the results of the student's previous developmental mathematics attempt, was also expected to be a predictor of success (Little, 2002). AttemptsPerPass was derived from all the student's developmental mathematics attempts and was designed to capture information about the overall pattern of course repetitions (see Chapter IV). AttemptsPerPass was also expected to be related to the outcome variables.

As previously mentioned, four binary outcome variables were defined. One outcome variable represented whether the student officially passed the course; each of the other three outcome variables represented whether the student reached a particular cut score on the departmental final exam. On CourseSuccess, a student earning an official course grade of A, B, or C was assigned a 1; a student earning an official course grade of IP, F, or W was assigned a 0. On ExamSuccess70, a student was assigned a 1 if the student scored at least 70 on my grading of their final exam; a student was assigned a 0 if the student either did not take the final exam or if the student scored below 70 on the final exam. The other two outcome variables, ExamSuccess60 and ExamSuccess50, were created in a similar manner, also based on my grading of the final exam. For each outcome variable, a separate logistic regression analysis was conducted. Each of these logistic regression analyses used the same four predictor variables: Treatment, HrsAttF2012, PrereqStatus, and AttemptsPerPass.

Focus Groups and Surveys

Research Question 2, like the quantitative Research Question 1, basically asks "Was the study journal intervention helpful?" Therefore Research Question 2 pertains to the confirmatory strand. However, it was answered by mostly qualitative data. This question focuses on the students' perceptions of the study journal's usefulness and was addressed in two ways: two informal focus groups about one month before the semester's end, and surveys near the semester's beginning and end.

Two of the nine study journal classes were chosen for the focus groups, based upon the availability of the classroom for students to stay after class. I provided pizza for students who volunteered to stay and share their thoughts on how the study journal project had affected them and their studying. The sessions were audio-recorded, transcribed by a transcribing company, and analyzed for themes.

At the beginning of the semester, study journal students completed a short survey, including both Likert-style and open-ended questions, about the study journal's anticipated helpfulness to them (see Appendix F). At the end of the semester, a similar survey asked students for their perceptions of the project's actual helpfulness (see Appendix G). For the Likert questions, descriptive statistics were recorded. Using scatterplots and correlations, I compared the students' final exam scores to their end-of-semester ratings of the study journal's helpfulness. For the open-ended questions on the end-of-semester survey, I coded the student responses into themes and recorded frequency counts for each theme. Details of the survey analysis procedures will be presented in the section on survey results.

Exploratory Strand: Study Habits Revealed by the Study Journals

Qualitative Analysis of the Study Journals

"Content analysis is the process of identifying, coding, and categorizing the primary patterns in the data" (Patton, 1990, p. 381). In order to answer Research Question 3, about the study journal students' study habits, I conducted qualitative content analysis on the participating students' packets of goal sheets. I first put the packets in numerical order using the identification numbers which had been randomly assigned during the quantitative phase. Then I went through each student's packet in chronological order, starting with the first week. I labeled each chunk of information with a preliminary code, written in the margin. The codes emerged from the data, and were not based on any a priori theory. Often, they reflected the students' own words—"practice" or "review," for example. Other codes were more conceptual, such as "control my time."

After coding 8–10 students' packets, I made a second pass through that batch of packets and summarized the codes in a spreadsheet. In one column of the spreadsheet I listed the codes. In the next column, I listed the identification numbers of the students from whom the codes had come, including duplicates. For example, if Student 10 mentioned “practice” five different times in his packet, and Student 13 mentioned “practice” three different times, I wrote “10, 10, 10, 10, 10, 13, 13, 13” next to “practice” in the spreadsheet. As I coded each subsequent batch of 8–10 packets, I used existing codes if possible, when students used different words to express similar concepts. If a student mentioned a concept that had not yet been coded, I added a new code to the list.

In the spreadsheet, I organized the codes into categories, placing similar codes together (Patton, 1990, pp. 381–382, 402–406). When applicable, I organized the codes according to Zimmerman’s dimensional framework (1994, 1998). As previously mentioned, this framework suggests that learners can exercise self-regulation in six dimensions, each associated with a key word. As I added new codes, I placed them under the appropriate dimension: motivation (*why*), strategies (*how*), time (*when*), outcomes (*what*), environment (*where*), and social context (*who*). If a code did not fit into any of the dimensions, I put it in a separate area. Whenever I added a new code, I placed it next to whichever existing codes were most similar. Using the constant comparison approach, I refined the categorization system as I went along, creating subcategories of codes within each of Zimmerman’s dimensions (Merriam, 1998, pp. 178–185). I also grouped the non-Zimmerman codes into categories and subcategories, giving a descriptive name to each category. After I finished coding the goal sheets, I carefully examined the spreadsheet to see if all the codes were arranged logically. As a result, I made minor adjustments to both the groupings and the category names.

Quality Scoring of the Study Journals

In consultation with a postdoctoral researcher who had agreed to assist with the study journal scoring, I developed a quality rubric for the study journals (see Appendix H). Each student's packet of goal sheets and each student's packet of study logs received a score of 1, 2, or 3 based on level of detail, depth of reflection, and evidence of adaptation.

After discussing the rubric, we each independently scored a random sample of 20 (out of 102) of the goal sheet packets. We compared our scores and examined the packets on which we disagreed. On our initial comparison, we disagreed on 12 of the 20 scores; our scores differed by more than one unit on only one of those disagreements. Most disagreements occurred on the packets that contained only one or two weeks, or on the packets in which some weeks were very detailed and other weeks were not. During this discussion, we reached consensus on all but one of the disagreements; we changed the scores to reflect our consensus. On the remaining disagreement, our scores differed by only one unit. Based on our discussion, we also clarified the scoring criteria by making minor modifications to the rubric's wording.

During the discussion, we also agreed that students should not automatically receive lower scores if they submitted fewer study journals. If a packet contained only one or two goal sheets but they showed reflection, detail, and thoughtfulness, the packet should receive a high score. This was important, because the information from the study journals, including the quality scores, would later be connected to student success data. Because departing students were counted as unsuccessful, and because departing students would also have fewer study journals, the study journal quality score needed to be independent of study journal quantity.

We then selected a second random sample of 20 goal sheet packets, and again scored them independently. On the second sample, we disagreed on 8 packets; our scores did not differ by more than one unit on any of the disagreements. Each of us then scored the remaining goal sheets independently.

For the study logs, we repeated the process, beginning with a random sample of 20 study log packets (out of 103). On the sample, we disagreed on 6 packets; our scores did not differ by more than one unit on any of the disagreements. After discussing our disagreements, we each independently scored the remaining study logs. For both the goal sheets and the study logs, the average of our two ratings was used in the analysis.

Using SPSS, I calculated the ICC(2,2) intraclass correlation coefficient (Landers, 2011; Shrout & Fleiss, 1979; Tinsley & Weiss, 2000). The notation ICC(2,2) indicates that all the study journals were rated by two raters, who were only a sample from the population of all possible raters. In SPSS, I chose the option for *average measures* because I planned to use the average of our two scores in subsequent analysis, rather than using only one rater's scores. I specified *absolute agreement* because we cared not only about consistency but also about the amount of the actual construct (study journal quality). For the goal sheets, the ICC(2,2) was 0.876; for the study logs, it was 0.729.

In the spreadsheet containing the list of study habits, I added two categories: “average quality score—goal sheet” and “average quality score—study log.” Each was divided into three subcategories: (a) 2.5 or 3, (b) 2, and (c) 1 or 1.5. Because the rubric had three scoring levels, and because both raters’ final scores all agreed within one unit, an average score of 2.5 or 3 meant that at least one of the two raters had awarded a 3, the highest possible quality score. An average score of 1 or 1.5 meant that at least one rater had awarded a 1, the lowest possible score. An average score of 2 meant that both raters agreed that the goal sheet or study log packet was of medium quality. For each score-based subcategory, I listed the identification numbers of the students who received those score(s).

Connections Between the Strands

To answer Research Question 4, I need to ascertain which of the study habits—derived from the qualitative analysis of the study journals—were associated with success. For this purpose, I chose to define success using ExamSuccess50. As previously mentioned, in order to pass the course with an A, B, or C, departmental guidelines required students to have both a course average of at least 70% and also a final exam score of at least 50%. Thus, an exam score of 50 indicated the student had reached the absolute minimum final exam grade acceptable by the department and had a chance of passing the course. ExamSuccess50 also resulted in similar sizes for the group of successful students and the group of unsuccessful students, which would aid in analysis.

For this phase, I revisited the previously mentioned spreadsheet containing the codes from the qualitative analysis of the goal sheets. Next to each code were listed the identification numbers of the students who had mentioned that particular concept or study habit, including duplicate listings for multiple mentions. I split these lists of student numbers into two columns of the same width. In one column I placed the identification numbers of the successful students; in the other column I placed the identification numbers of the unsuccessful students (based on ExamSuccess50). I also added leading zeroes to all the one- and two-digit identification numbers. This made all the identification numbers the same physical size (three digits), facilitating visual comparisons between the two groups. In a second version of the spreadsheet, I deleted the duplicate entries for multiple mentions of the same study habit.

Treating the lists of case numbers like a bar chart or stem-and-leaf plot, I visually examined both spreadsheets (with duplicates and without duplicates) for obvious differences between successful and unsuccessful students, flagging the study habits that seemed to distinguish the two groups. I sometimes combined related codes, if there were very few students listed, or if the codes had similar distributions of successful and unsuccessful students. If it appeared that collapsing two codes into a single code would result in the loss of interesting information, I did not combine them.

It became apparent that the spreadsheet including the duplicates was heavily influenced by a small number of students who repeated the same concept many times over many weeks. Because the category of unsuccessful students included many students who left the class partway through the semester, the unsuccessful students, as a group, did not have as many opportunities to mention those concepts. For this reason, I decided to concentrate on the spreadsheet from which the duplicate mentions had been removed.

It should be noted that many students still appeared on the list many times, but each student could only appear once next to a single code. Whenever I combined two codes in the no-duplicate list, I merged the two lists of successful students together, put them in numerical order, and removed any resulting duplicates (duplicates occurred when the same student mentioned both items). I then repeated the process for the two lists of unsuccessful students.

The preliminary visual inspection was very helpful for refining the list of study habit codes. However, for the credibility of the final analysis, a more systematic approach was desirable. Therefore, I created a set of numerical criteria. These criteria allowed me to systematically decide which study habits discriminated between successful and unsuccessful students. The two criteria were (1) at least 10% of the students submitting study journals mentioned that study habit; and (2) the larger group (either the successful group or the unsuccessful group) mentioning that study habit must be at least 50% greater in size than the smaller group. The first criterion was essentially used to screen out "noise." If a particular study habit code appeared in less than 10% of the students' study journals, it was dismissed as "noise," not signifying a meaningful distinction between groups. The second criterion involved the ratio of the successful group to the unsuccessful group, calculated separately for each study habit code. If the ratio was larger than 1.5 or smaller than 2/3, I considered that study habit code as signifying a meaningful distinction between the groups. In the calculation of this ratio, I had planned to use a multiplier to adjust for the size difference between the overall groups (the group of successful students who submitted study journals and the group of unsuccessful students who submitted study journals). However, by coincidence, the two overall groups turned out to be identical in size (53 students each), so this step was not needed.

Results

Implementation

The plan called for all the treatment classes to collect study journals every week from every student. Not surprisingly, this plan was not followed exactly. As can be seen in Table 7, the nine treatment classes differed widely in their implementation of the intervention, and some classes approached this ideal more closely than others. In Classes A, B, F, and J, the instructors managed to integrate the study journals into the classes' normal routine. In these four classes, the instructors collected study journals on the same day every single week, unless that day fell on a college holiday. Even late in the semester, a large proportion of the students in these four classes regularly submitted complete study journals. As expected, due to student attrition, the number of submissions generally decreased as the semester progressed. In Classes C and E, the instructors collected study journals on a regular schedule, but the number of students submitting them dwindled to almost zero. In Classes D and G, the collections occurred on an irregular schedule and often students submitted only the study log or only the goal sheet, instead of both. In Class H, a very small class, the study journal intervention never really got off the ground.

Table 7

Counts of Complete and Partial Study Journal Submissions by Class and Week

Class	Week													
(Official enrollment)	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A (29)	19(2)	15(0)	17(0)	18(2)	16(1)	17(0)	17(0)	19(1)	17(1)	17(2)	12(0)		13(1)	14(2)
B (34)	22(1)	23(1)	21(1)	14(0)	18(0)	14(0)	16(0)	14(0)	12(0)	12(0)	9(0)	10(0)	10(0)	9(1)
C (29)	13(1)	9(0)		5(2)	8(0)	7(1)	1(0)	3(0)	3(0)	2(0)	1(0)		1(0)	1(0)
D (30)		11(4)		3(8)	8(4)	5(5)	3(1)	6(4)	5(2)	3(0)	2(1)	2(0)		
E (29)	10(1)	10(3)		5(3)	8(2)	3(4)	4(2)	8(2)	5(1)	2(0)	3(1)[a] 3(0)[a]			
F (27)	19(1)	17(0)	18(0)	18(0)	15(1)	16(0)	14(2)	15(1)	15(0)	10(0)	8(3)		9(1)	12(2)
G (31)		5(7)			6(16)	0(17)		0(16)	2(8)		4(11)[a] 10(4)[a]			13(1)
H (17)		6(1)	3(1)	1(0)										
J (31)	11(5)	10(4)	18(2)	15(0)	15(6)	19(0)	21(0)	17(1)	18(2)	17(0)	16(0)	12(1)	16(0)	11(0)

Note. Outside the parentheses is the number of complete submissions (goal sheet and study log). Inside the parentheses is the number of partial submissions (goal sheet only or study log only). The college withdrawal deadline occurred during Week 10.
[a] Two sets of study journals were submitted during the same week.

Confirmatory Strand: Effect of the Intervention

Original Sample

The original sample consisted of 140 participants from the nine control classes and 117 participants from the nine treatment classes.

Propensity Score Matching on the Original Sample

After conducting several trials of the propensity score matching algorithm using different random orderings of the participants, I chose the trial producing the best overall balance of the covariates. The matched sample chosen for further analysis had 105 students in each group. A short summary of the balance assessment will be presented in this section. For a more detailed discussion of the balance assessment process, see Chapter IV.

The first step in assessing balance was to compare the means on the covariates in the two groups. For each covariate, the difference in means was standardized by dividing by the standard deviation of the treatment group in the unmatched sample (Stuart, 2010). Because the goal of matching is for the two groups to be as similar as possible on the covariates, the standardized differences should be as close to zero as possible.

The before- and after-matching standardized differences are displayed numerically in Table 8 and graphically in Figure 3. The matching process improved the balance on nearly all the covariates. The standardized difference increased slightly for three covariates: HrsAttF2012, CumHrsAttPreInt, and CurrentCourse. This is not unusual when the standardized differences are very small even before matching (Stuart, 2010). For every covariate, including these three, the magnitude of the after-matching standardized difference was less than 0.05, well below the 0.25 threshold generally considered acceptable for good balance (Stuart & Rubin, 2007).

Table 8

Means and Standardized Differences of Covariates Before and After Matching

	Unmatched Sample			Matched Sample		
Variable	Means Treated	Means Control	Std. Mean Diff.	Means Treated	Means Control	Std. Mean Diff.
Propensity Score	.483	.432	.405	.460	.455	.039
HrsAttF2012	9.650	9.586	.020	9.552	9.438	.036
CumHrsAttPreInt	28.521	28.236	.012	27.924	27.581	.014
CredEarnedPreInt	19.932	18.214	.096	19.257	18.943	.018
YrsSinceStartCollege	1.949	1.843	.047	1.933	1.905	.013
GPAPreint	2.286	2.153	.173	2.259	2.259	-.001
GPAIndicator	.752	.779	-.061	.771	.781	-.022
CourseCompletionRatio	.713	.663	.238	.707	.705	.011
CCRIndicator	.855	.900	-.128	.876	.867	.027
DevMathGPA	1.751	1.621	.128	1.707	1.690	.016
DMathGPAIndicator	.769	.843	-.174	.810	.810	.000
AttemptsPerPass	1.501	1.570	-.113	1.530	1.513	.029
PrereqStatusGradePts	1.529	1.432	.068	1.462	1.529	-.047
PrereqStatusIndicator	.769	.843	-.174	.810	.810	.000
YrsSinceMathTruncated	.359	.314	.046	.371	.343	.029
CurrentCourse	309.111	309.143	-.032	309.143	309.181	-.038
ESOL	.051	.007	.199	.010	.010	.000

Note. Standardized differences were calculated by subtracting the control group mean from the treatment group mean and dividing the result by the standard deviation of the treatment group in the unmatched sample.

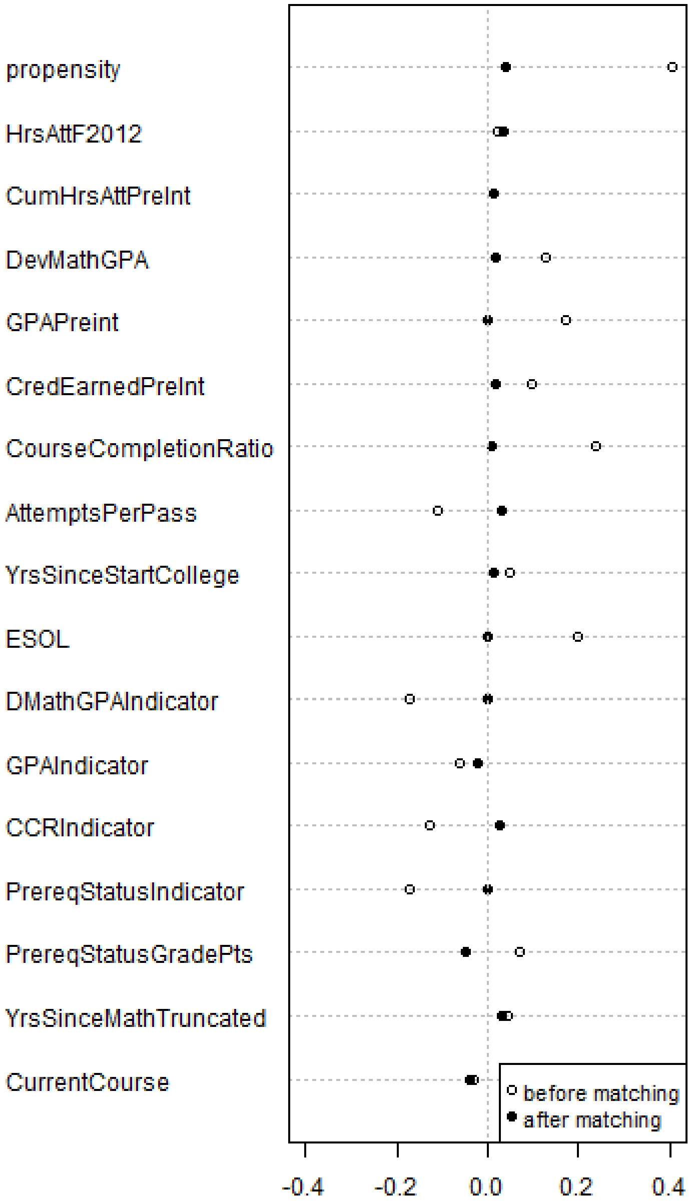


Figure 3. *Dot plot of standardized differences on covariates before and after matching for original sample of 18 classes.*

For the groups to be well-balanced on a covariate, it is not sufficient for the difference in means to be small. For balance, the shapes of the distributions should also be similar. This can be evaluated using histograms, boxplots, or quantile-quantile plots. These detailed balance checks are presented in Chapter IV.

Treatment Effect on Matched Sample Taken From Original Sample

As previously mentioned, I planned to use logistic regression to estimate the effect of the intervention on course success and final exam success, with the outcomes considered as binary variables. I used three different variables to represent final exam success: ExamSuccess50, ExamSuccess60, and ExamSuccess70. Each was derived from a different cut score on the departmental final exam. Exam success was based on my grading of the final exams, incorporating the aforementioned interrater reliability checks. Students not taking the final exam were counted among the unsuccessful, regardless of whether or not they officially withdrew from the class. These outcome variables are summarized in Table 9.

Table 9

Values for Outcome Variables

Variable	Successful	Not successful
CourseSuccess	A, B, or C	IP, F, or W
ExamSuccess70	Scored at least 70 on final exam	Did not take final exam or scored below 70
ExamSuccess60	Scored at least 60 on final exam	Did not take final exam or scored below 60
ExamSuccess50	Scored at least 50 on final exam	Did not take final exam or scored below 50

As a preliminary step in the analysis, I performed a chi-square analysis of the frequencies of success in the two groups. The CourseSuccess outcome variable showed a statistically significant difference in favor of the control group (*p* = .037). ExamSuccess60 and ExamSuccess70 favored the treatment group, though the difference was not statistically significant at the .05 level. On ExamSuccess50, the success rates of the two groups were identical. The results of this frequency analysis are summarized in Table 10.

Table 10

Frequencies for Outcome Variables: Matched Subsample From Original Sample of 18 Classes

Group	CourseSuccess	ExamSuccess70	ExamSuccess60	ExamSuccess50
Treatment Group				
Successful	39 (37.1%)	31 (29.5%)	41 (39.0%)	46 (43.8%)
Not successful	66 (62.9%)	74 (70.5%)	64 (61.0%)	59 (56.2%)
Control Group				
Successful	54 (51.4%)	21 (20.0%)	35 (33.3%)	46 (43.8%)
Not successful	51 (48.6%)	84 (80.0%)	70 (66.7%)	59 (56.2%)
Chi-square (sig.) (1, *N* = 210)	4.342 (.037)	2.556 (.110)	0.742(.389)	0.000 (1.00)

The discrepancy in direction between the course success and exam success variables, evident in Table 10, along with my awareness that the study journal project was not fully implemented in all the treatment classes, caused me to reconsider both the outcome variables and the sample before proceeding. There were two areas of concern, both related to teacher differences: (1) teacher differences in awarding course letter grades potentially confounding the CourseSuccess outcome variable, and (2) teacher differences in implementing the intervention potentially confounding the estimation of the treatment effect. These concerns were sufficiently serious that I decided to forego further statistical analysis on the matched sample taken from the original sample of 18 classes. Instead, I conducted the entire analysis on a subsample composed of eight classes: the four treatment classes (A, B, F, and J) that fully implemented the project, and the four control classes with which they had been paired during the treatment assignment process. Before presenting the statistical results for this subsample, I have included a short discussion of the two concerns that motivated the change, and my rationale for modifying the sample.

The first concern was the relationship between course success and exam success in the control group. As can be seen in Table 10, course success in the treatment group lined up reasonably well with exam success. However, in the control group, the course success rate was higher than the exam success rate for all three cut scores. Also, 54 control students passed the class with an A, B, or C, but only 46 control students received a score of at least 50 on my grading of the departmental final exam. Departmental guidelines called for passing students to have both a course average of at least 70% and final exam score of at least 50%. Because of differences in partial credit rubrics, the instructors' final exam grades and my final exam grades would not be expected to match exactly. Still, it seemed possible, based on these results, that teacher differences in awarding course letter grades may have confounded the course success variable to such an extent that that it did not have value as an outcome variable for research.

To explore this possibility, I considered each class as a separate subsample and conducted a chi-square and correlation analysis to evaluate the strength of relationship between CourseSuccess and each of the exam success variables in turn. For 16 of the 18 classes, there was a strong relationship between CourseSuccess and at least one of the exam success variables. However, there were two control classes in which CourseSuccess was not strongly related to any of the exam success variables. In these two control classes, a relatively large proportion of students received passing course grades even though they scored below 50 on my grading of the final exam.

The second concern was that the treatment effect of the intervention, if any, might have gone undetected due to incomplete implementation of the project in some classes. In the statistical analysis thus far, all students enrolled in the nine treatment classes had been classified as treatment students, regardless of how often the teachers had collected study journals, or how many students submitted them. Because some of the classes had not collected study journals regularly, or not collected very many study journals, this resulted in many students being classified as treatment students even though they did not regularly submit study journals.

Because the objective of this research strand was to determine whether completing weekly study journals affected students' success, I decided to modify the sample, concentrating on the classes in which the treatment was most fully implemented. For the treatment group, I chose to use classes A, B, F, and J, because most students in these classes had submitted study journals each week. Instead of comparing these four classes with the original control group, I compared them with the four classes with which they had been paired during the treatment assignment process. At that time, the classes had been matched on teacher (when possible), days per week, and time of day. In each pair, the two classes had been randomized into treatment and control conditions. By using the four classes from the original pairings as the new control group, I was able to control for as many class-level variables as possible. This choice of control group had the additional advantage of removing from the analysis the two control classes in which the course success outcome was not strongly related to the exam success outcomes.

Modified Sample

The modified sample included 77 participants in the four control classes and 60 participants in the four treatment classes. There were two pairs of Introductory Algebra classes and two pairs of Intermediate Algebra classes. All four pairs of classes met two days each week. Three of the four pairs met during the day, and one pair of classes met in the evening. Two of the four pairs were same-teacher pairs.

Propensity Score Matching on the Modified Sample

As with the original sample, I repeated the propensity score matching several times, using different modifications of the algorithm and different random orderings of the participants. Of the resulting subsamples, I chose the one that had the best balance on the covariates without discarding a large number of treatment participants. The subsample chosen for further analysis was composed of 54 treatment students and 54 control students. It was obtained using nearest neighbor one-to-one matching with no caliper, discarding treatment and control cases outside the region of common support (region of overlap in propensity scores). Although a 0.2 caliper had worked well on the original sample, it did not work well on the modified sample, because it caused too many treatment students to be discarded due to lack of an available match within the caliper.

Before- and after-matching standardized differences for the modified sample are shown in Table 11 and Figure 4. The unmatched groups in the modified sample (8 classes) were less similar on the covariates than the unmatched groups in the original sample (18 classes). Because the groups in the modified sample were less similar initially, the propensity score matching process was unable to produce the excellent balance apparent in the larger matched sample. This can be seen by comparing the balance summaries for the smaller sample (Table 11 and Figure 4) with the balance summaries for the larger sample (Table 8 and Figure 3), noting that the *x*-axes of Figure 3 and Figure 4 use different scales. Still, the matching process improved the balance on nearly all the covariates, and all the after-matching standardized differences (except propensity score, which is not a true covariate) were below 0.25, the threshold recommended by Stuart and Rubin (2007).

Table 11

Means and Standardized Differences of Covariates Before and After Matching for the Modified Sample of Four Treatment Classes and Four Control Classes

	Unmatched Sample			Matched Sample		
Variable	Means Treated	Means Control	Std. Mean Diff.	Means Treated	Means Control	Std. Mean Diff.
Propensity Score	.521	.373	.695	.484	.415	.324
HrsAttF2012	9.933	9.299	.193	9.722	9.796	-.023
CumHrsAttPreInt	30.700	26.338	.185	29.556	29.056	.021
CredEarnedPreInt	21.433	16.143	.311	19.963	18.537	.084
YrsSinceStartCollege	2.100	1.545	.230	1.889	1.704	.077
GPAPreint	2.361	2.128	.301	2.355	2.225	.169
GPAIndicator	.783	.740	.104	.778	.759	.045
CourseCompletionRatio	.712	.649	.288	.697	.661	.167
CCRIndicator	.900	.896	.013	.907	.926	-.061
DevMathGPA	1.914	1.519	.361	1.828	1.600	.208
DMathGPAIndicator	.833	.831	.006	.833	.852	-.049
AttemptsPerPass	1.511	1.615	-.187	1.537	1.599	-.112
PrereqStatusGradePts	1.713	1.431	.183	1.616	1.515	.066
PrereqStatusIndicator	.833	.831	.006	.833	.852	-.049
YrsSinceMathTruncated	.267	.195	.095	.259	.204	.073
CurrentCourse	309.200	309.039	.163	309.148	309.185	-.037
ESOL	.067	.013	.213	.037	.019	.074

Note. Standardized differences were calculated by subtracting the control group mean from the treatment group mean and dividing the result by the standard deviation of the treatment group in the unmatched sample.

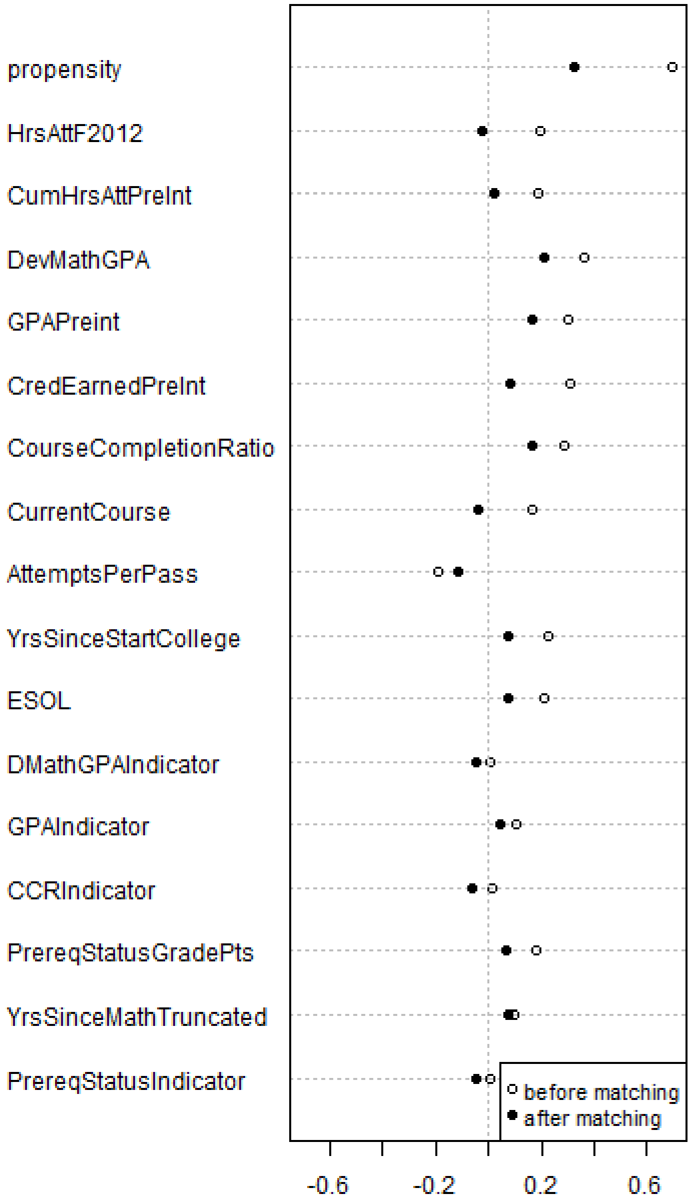


Figure 4. *Dot plot of standardized differences on covariates before and after matching for modified sample of 8 classes*.

Treatment Effect on Matched Sample Taken From Modified Sample

As with the larger sample, I began with a chi-square analysis on the frequencies of the control and treatment students' course success and exam success, using the matched sample taken from the eight classes. The results are shown in Table 12. Control students had higher success rates on all four outcome variables: CourseSuccess, ExamSuccess70, ExamSuccess60, and ExamSuccess50. On the CourseSuccess variable, the difference was statistically significant (*p* = .012). On ExamSuccess50, the difference approached statistical significance (*p* = .054).

Table 12

Frequencies for Outcome Variables: Matched Subsample From Modified Sample of 8 Classes

Group	CourseSuccess	ExamSuccess70	ExamSuccess60	ExamSuccess50
Treatment				
Successful	19 (35.2%)	14 (25.9%)	20 (37.0%)	24 (44.4%)
Not successful	35 (64.8%)	40 (74.1%)	34 (63.0%)	30 (55.6%)
Control				
Successful	32 (59.3%)	18 (33.3%)	27 (50.0%)	34 (63.0%)
Not successful	22 (40.7%)	36 (66.7%)	27 (50.0%)	20 (37.0%)
Chi-square (sig.) (1, *N* = 108)	6.279 (.012)	0.711(.399)	1.846(.174)	3.724(.054)

To estimate the intervention’s effect on course success and exam success, I applied a logistic regression model to the matched sample. As previously mentioned, the model included four predictors: Treatment, HoursAttF2012, PrereqStatus, and AttemptsPerPass. I conducted a separate analysis for each of the four outcome variables (CourseSuccess and the three versions of ExamSuccess using different cut scores). The results are shown in Table 13–Table 16.

Table 13

Logistic Regression With Dependent Variable CourseSuccess

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HrsAttF2012	.063	.071	.796	1	.372	1.065	.927	1.225
AttemptsPerPass	-.342	.519	.434	1	.510	.710	.257	1.965
PrereqStatus[a]			9.483	4	.050			
PrereqStatus (Repeat)	-.415	.786	.278	1	.598	.661	.142	3.083
PrereqStatus (C)	-2.893	1.221	5.612	1	.018	.055	.005	.607
PrereqStatus (B)	-.141	.719	.038	1	.845	.869	.212	3.557
PrereqStatus (A)	1.085	.876	1.537	1	.215	2.961	.532	16.471
Treatment (1)	-1.291	.450	8.242	1	.004	.275	.114	.664
Constant	.758	1.074	.498	1	.481	2.133		

Note. Omnibus fit test: χ^2 (7, 108) = 24.304, *p* =.001. Hosmer-Lemeshow: χ^2 (7, 108) = 3.321, *p* =.913. Cox & Snell R^2 = 0.202, Nagelkerke R^2 = 0.269. Model correctly classified 68.5% of the cases, compared to 52.8% in the model with no predictors.

[a]Reference value for PrereqStatus is Placement.

Table 14

Logistic Regression With Dependent Variable ExamSuccess70

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HrsAttF2012	.066	.075	.788	1	.375	1.069	.923	1.237
AttemptsPerPass	-.303	.571	.282	1	.595	.738	.241	2.262
PrereqStatus[a]			7.191	4	.126			
PrereqStatus (Repeat)	-.882	.809	1.186	1	.276	.414	.085	2.023
PrereqStatus (C)	-2.483	1.201	4.270	1	.039	.084	.008	.880
PrereqStatus (B)	-1.364	.753	3.283	1	.070	.256	.058	1.118

Table 14 Continued

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
PrereqStatus (A)	.053	.776	.005	1	.946	1.054	.230	4.826
Treatment(1)	-.585	.464	1.588	1	.208	.557	.224	1.384
Constant	.043	1.102	.002	1	.969	1.044		

Note. Omnibus fit test: χ^2 (7, 108) = 13.368, *p* =.064. Hosmer-Lemeshow: χ^2 (7, 108) = 3.245, *p* =.918. Cox & Snell R^2 = .116, Nagelkerke R^2 = .166. Model correctly classified 74.1% of the cases, compared to 70.4% in the model with no predictors.

[a]Reference value for PrereqStatus is Placement.

Table 15

Logistic Regression With Dependent Variable ExamSuccess60

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HrsAttF2012	.023	.066	.127	1	.722	1.024	.900	1.165
AttemptsPerPass	.283	.478	.352	1	.553	1.328	.520	3.387
PrereqStatus[a]			5.487	4	.241			
PrereqStatus (Repeat)	-.731	.749	.952	1	.329	.482	.111	2.090
PrereqStatus (C)	-1.966	.971	4.096	1	.043	.140	.021	.940
PrereqStatus (B)	-.471	.685	.473	1	.492	.624	.163	2.390
PrereqStatus (A)	.264	.777	.115	1	.734	1.302	.284	5.968
Treatment(1)	-.619	.410	2.274	1	.132	.539	.241	1.204
Constant	-.080	1.009	.006	1	.937	.923		

Note. Omnibus fit test: χ^2 (7, 108) = 8.532, *p* =.288. Hosmer-Lemeshow: χ^2 (7, 108) = 8.671, *p* =.277. Cox & Snell R^2 = 0.076, Nagelkerke R^2 = 0.102. Model correctly classified 63.9% of the cases, compared to 56.5% in the model with no predictors.

[a]Reference value for PrereqStatus is Placement.

Table 16

Logistic Regression With Dependent Variable ExamSuccess50

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HrsAttF2012	.039	.067	.347	1	.556	1.040	.913	1.185
AttemptsPerPass	.074	.478	.024	1	.878	1.076	.422	2.746
PrereqStatus[a]			6.152	4	.188			
PrereqStatus (Repeat)	-.735	.762	.932	1	.334	.479	.108	2.133
PrereqStatus (C)	-1.871	.901	4.316	1	.038	.154	.026	.900
PrereqStatus (B)	.008	.717	.000	1	.992	1.008	.247	4.109
PrereqStatus (A)	-.163	.795	.042	1	.838	.850	.179	4.039
Treatment(1)	-.807	.413	3.810	1	.051	.446	.198	1.003
Constant	.574	1.028	.311	1	.577	1.775		

Note. Omnibus fit test: χ^2 (7, 108) = 11.430, *p* =.121. Hosmer-Lemeshow: χ^2 (7, 108) = 5.097, *p* =.648. Cox & Snell R^2 = .100, Nagelkerke R^2 = 0.134. Model correctly classified 63.9% of the cases, compared to 53.7% in the model with no predictors.

[a] Reference value for PrereqStatus is Placement

For the CourseSuccess outcome, the logistic regression model fit well. In the omnibus goodness-of-fit test, a significance value below .05 indicates good fit. In the Hosmer-Lemeshow test, a significance value below .05 indicates poor fit (Pallant, 2009). Therefore the chi-square values of χ^2 (7, 108) = 24.304, *p* =.001 (omnibus) and χ^2 (7, 108) = 3.321, *p* =.913 (Hosmer-Lemeshow) both support the model's fit. Two pseudo- R^2 calculations gave similar results (Cox and Snell R^2 = .202, Nagelkerke R^2 = .269), indicating that the model explains somewhere around 20–30% of the variance. The CourseSuccess model correctly classified 68.5% of the cases, compared to 52.8% in the model with no predictors.

The models for ExamSuccess70, ExamSuccess60, and ExamSuccess50 did not fit as well, falling short of statistical significance on the omnibus goodness-of-fit test and producing lower pseudo-R^2 values. Numerical results for the goodness-of-fit tests, pseudo-R^2, and classification percentages are provided at the bottom of each table. Caution must be used when comparing the classification percentages, as they are affected by the overall frequencies. For example, the ExamSuccess70 model correctly predicted 74.1% of the cases. However, this does not mean it is a better model than the others. Because 70.4% of the students failed to score at least 70 on the exam, a prediction correctness level of 70.4% could be reached with no predictors at all, simply by assigning 100% of students to the unsuccessful category. Thus, the 74.1% prediction correctness by the logistic regression model does not represent much of an improvement.

As was expected based on the preliminary chi-square results, treatment condition was a statistically significant predictor of CourseSuccess (p = .004). Treatment condition was not a statistically significant predictor of ExamSuccess70 or ExamSuccess60, but it approached statistical significance for ExamSuccess50 (p = .051). Although overall PrereqStatus was a statistically significant predictor only for CourseSuccess (p = .050), PrereqStatus (C) was a significant predictor for CourseSuccess and for all three versions of ExamSuccess (p = .039 for cut score 70, p = .043 for 60, p = .038 for 50). This indicated that students receiving a C in the prerequisite class were less likely to earn a passing grade on the final or in the course, compared to students who were placed directly into the class.

For CourseSuccess, the odds ratio for Treatment was 0.275. This means the odds for a treatment student earning an A, B, or C were 0.275 times the odds for a control student earning an A, B, or C. Equivalently, the odds for a control student passing the class were 1/0.275 = 3.636 times the odds for a treatment student passing the class. Because odds ratios larger than 1 are generally easier to interpret, I also used reciprocals to invert the confidence interval for Treatment in Table 13, resulting in a confidence interval of [1.51, 8.77]. Thus there was a greater than .95 probability that the true value of the odds ratio was between 1.51 and 8.77. Because 1 was not included in the confidence interval, there was less than a .05 probability that the odds of success were identical for both control and treatment students.

For ExamSuccess50, the odds ratio is 1/0.446 = 2.42, meaning the odds of a control student staying in the class and scoring at least 50 on the final were 2.42 times the odds for a treatment student. The corresponding confidence interval is [0.997, 5.05]. Because this 95% confidence interval included 1, it included the possibility that the groups' odds of success were actually the same. (This was expected, because the significance value was slightly more than .05.)

As we have seen, the chi-square and the logistic regression analyses indicated that treatment students were less likely to pass the course, and were also less likely to score at least 50 on the exam. The group of unsuccessful students included two sets of students: (1) those who remained until the end and attempted the final but performed poorly (either on the final, on the official course letter grade, or both), and (2) students who left the class before the final exam. Therefore, I created a new outcome variable, TookFinal. This variable had a value of 1 for students who took the final and a value of 0 for students who did not. I repeated the chi-square analysis, using the same subsample of 108 students from the eight matched classes. There was a statistically significant ($p = .019$) difference, with more students in the control group taking the final than students in the treatment group. The frequencies are shown in Table 17.

Table 17

Frequency of Students Taking Final Exam

Group	TookFinal
Treatment	
Took final	33 (61.1%)
Did not take final	21 (38.9%)
Control	
Took final	44 (81.5%)
Did not take final	10 (18.5%)
Chi-square (sig.) (1, *N* = 108)	5.475(.019)

I also repeated the logistic regression analysis, using the same predictors as before, with TookFinal as the dependent variable (see Table 18). The model did not fit especially well. While the model with no predictors could correctly classify 71.3% of the cases, the logistic regression model could only classify 69.4% of them correctly. The only significant predictor was treatment condition. Unlike the analyses for CourseSuccess, ExamSuccess70, ExamSuccess60, and ExamSuccess50, PrereqStatus (C) was not a significant predictor for TookFinal. This meant that students receiving a C in the prerequisite class were no more likely to drop out before the final than students placed directly into the class.

Although the logistic regression model did not fit well enough to be useful for prediction, it supported the results from the chi-square analysis by showing that assignment to the treatment (study journal) condition was associated with not taking the final exam. The odds ratio provided by the logistic regression aids interpretation, essentially serving as an effect size. This odds ratio can help us see whether the group differences visible in the chi-square analysis represent an extreme difference or just a slight difference. The odds ratio for treatment condition was 1/0.340 = 2.94, indicating that the odds for treatment students leaving the class before the final exam were nearly three times the odds for control students leaving the class. The corresponding 95% confidence interval was [1.193, 7.246].

Table 18

Logistic Regression With Dependent Variable TookFinal

							95% C.I.for Exp(B)	
Variable	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HrsAttF2012	.065	.072	.812	1	.368	1.067	.926	1.230
AttemptsPerPass	-.246	.514	.230	1	.632	.782	.285	2.141
PrereqStatus[a]			2.867	4	.580			
PrereqStatus (Repeat)	.913	.822	1.233	1	.267	2.491	.498	12.472
PrereqStatus (C)	-.225	.853	.070	1	.792	.799	.150	4.251
PrereqStatus (B)	.367	.746	.242	1	.623	1.444	.334	6.236
PrereqStatus (A)	.752	.872	.745	1	.388	2.122	.384	11.719
Treatment(1)	-1.079	.460	5.493	1	.019	.340	.138	.838
Constant	.786	1.100	.510	1	.475	2.194		

Note. Omnibus fit test: χ^2 (7, 108) = 8.912, *p* =.259. Hosmer-Lemeshow: χ^2 (7, 108) = 4.174, *p* =.841 Cox & Snell R^2 = .079, Nagelkerke R^2 = .113. Model correctly classified 69.4% of the cases, compared to 71.3% in the model with no predictors.

[a] Reference value for PrereqStatus is Placement

Focus Groups

Six students from one class volunteered for the first focus group, and six students from another class volunteered for the second focus group. The classes had different instructors, each a full-time faculty member with over 25 years of teaching experience. Both classes were in the group of four treatment classes used for the final quantitative analysis, whose instructors had fully implemented the project by collecting study journals every week from nearly all the students.

Six main themes emerged from the focus groups:

1. Some students felt there was redundancy, either between the goal sheet and the study log, or between the goal sheet and the student's planner.

2. The students' previous skill in planning affected their view and their use of the study journal.
3. The study journal served as a mirror to show students how much time they were spending on math.
4. The study journal helped illuminate the connection between study time and math grades.
5. The study journal helped some students to study more and to concentrate more.
6. Students found the study journals helpful and would recommend them for future math classes.

The quotations in the following sections are all transcribed from audio recordings. Thus, the words are from the students, but the punctuation decisions were mine. Because this section reflects the perspectives and language of the students, I will refer to mathematics as *math*.

Redundancy

Two types of redundancy arose in the focus group discussions. One was caused by a misunderstanding of the intended study-journaling process and the other was a legitimate critique of the worksheets. Some students had not realized that the time planning grid on the back of the goal sheet (Appendix B) was intended for advance planning, that the study log (Appendix C) was intended for tracking actual time, and that the planned time and the actual time might not match. These students confessed to hurriedly copying their study times from one sheet to the other, five minutes before class. Not surprisingly, students using this approach felt the study log was redundant.

Other students had used the study log as intended, to track their actual time, and did not feel the two sheets were redundant. A different sort of redundancy was experienced by students who already kept a detailed planner, with planned study times for all their classes. For these students, the goal sheet duplicated the information they had already recorded in their planner. Clarifying the purpose of the study journal worksheets would have resolved the first redundancy but not the second.

Differences for Students With Different Levels of Planning Skill

Students who used their planners to schedule their study times found the study log more helpful than the goal sheet (because the goal sheet was redundant). Students not in the habit of setting specific goals or planning their study time found the goal sheet more helpful. Students who used their planners only for writing down due dates and scheduled quizzes, rather than for planning their study time, also saw value in the goal sheet. One such student mentioned that although her planner worked well for reminding her of quizzes, the goal sheet helped her realize "I needed to have a plan for my math."

Mirror to Show Time Spent on Math

Students felt the worksheets helped them to see just how much time they were spending on math. Sometimes this encouraged them, by showing them how hard they were working: "I was impressed to realize how much time I was spending doing the exercise. Because before this, I didn't realize how much time I was spending every other day until I write it down. I was like, 'Really?' " However, seeing a large time investment could also bring discouragement, if the investment did not pay off in success: "So prior to this, I would just—I set a time for this class and I set an amount of time for the other class and I set—but now, with that worksheet in front of you, it's like, 'Man, I have to put in so many hours just for math and I'm still failing.' " Discouragement also occurred when completing their math took far more time than they had anticipated: "Oh yeah because when I look at something, I think, 'Oh I should be able to do that in an hour.' Three hours later and I'm still doing it."

Connection Between Study Time and Grades

In general, the focus group students felt that their grades were connected to their study time, and that the study journal made that connection easier to see: "You actually fill it out and compare it to your grades. It makes sense. I mean, the more you study, the better the grade you're going to get." By making that comparison, students could sometimes diagnose the cause of their difficulties in the class: "It shows me where I lack at in my studying habits because there have been weeks that I've looked at it and I've said, 'Okay. You know what? The reason I did poorly on this paper or this exam or whatever is because obviously I didn't do s—.' "

For some students, the study journal showed the need for spending a large amount of time on math: "I think that for me, getting—it opened my eyes as to how much I really need to study." Sometimes it also revealed that they were not spending sufficient time: "Because it shows how much time, I think. And I see that I don't spend enough time on it, so I'm kind of trying to catch up, so that I need to spend more time."

Improved the Quality and Quantity of Their Studying

Writing down weekly goals helped some students stay focused on what they needed to do: "It's like it keeps you on track like for the week what you want to accomplish or what you want to set your goals. So it kind of keeps you on track because you set the goals, so now you need to know what to do to reach the goals, so it is helpful." For some students, committing to their study hours on paper kept them from getting distracted: "For me, I put—whenever you put the time down, how many hours you studied and from where to where, it's helped me stay concentrated for the two or three hours that I was studying—not stand up and do anything else." Committing to the study hours on paper also helped battle procrastination: "I just want to say the work—this paper, it just makes me keep going—do the homework, just making you concentrate to do that. Not like, 'Oh, I don't want to do that. I will do it later.' 'No, you need to do this. You have the sheet you need to fill out.' So you really do the homework and everything."

One student said the "where?" and "with whom?" questions on the study log prompted her to start utilizing the college's developmental mathematics tutoring center, which she had not used in previous semesters. She realized that if the answer to "where?" was "at home," then the answer to "with whom?" would be "nobody." But if she studied at the tutoring center, people could help her. It also helped that someone from the college distributed information about the tutoring center around the same time the students received the first study journal worksheet. The instructor provided additional impetus, by constantly reminding the students about the tutoring center and other resources.

One student described how, in the past, he devoted his attention to athletics instead of school. After losing a full athletic scholarship due to a knee injury, he realized, “well, if I put enough effort in my academics as much as I did in athletics, I would probably be decent I think.” He understood that if he did the same amount of work he had done in the past, he would get the same results he had gotten in the past. “And this [the study journal], like I said, helped me out to have focus. My intention is to try harder and study harder and more efficiently.”

Keep It for Future Classes

All the focus group students recommended that their instructors should use the study journal worksheets in future classes. At first, some students felt the worksheets were overwhelming and difficult to remember. However, once the study journals became a habit, the time commitment was not too much, and the study journal became more of a help than a burden. Both groups credited their instructors for incorporating the worksheets into the class routine and making them a habit. All the focus group students thought the study journal sufficiently beneficial that it should be used in future semesters: “I don’t think it would be smart to cut it out. I think you should continue.” “Especially when you’re teaching math.”

Though all the focus group students strongly recommended their instructors use the study journals in future classes, a few recommended minor changes. Most students thought the weekly schedule was best, but a few who were already avid planners suggested changing the goal sheet frequency to once a month, because it was redundant with their planners. They suggested making the study journal extra-credit, instead of required, so students did not have to do it if they did not think it would be beneficial.

These particular students said they always took advantage of every extra-credit opportunity and would still have done it.

The students already extremely skilled in planning did not seem to realize that not all developmental mathematics students possessed this skill. Those who lacked experience writing down goals and mapping out study times in advance felt the study journal taught them a valuable skill that would benefit them in the future. Some students said they would continue to use the worksheets next semester, even if they were not collected for a grade. Others said they might not use the worksheet, but they would continue to apply the lessons they learned from it: “Even if we don’t have that worksheet, I think it’s stuck in our minds to set a goal.”

Surveys

Sample

As previously mentioned, two surveys were distributed to the participants, one at the beginning of the semester and one at the end. The initial survey asked students to rate how helpful they anticipated the study journal to be, and the final survey asked them to rate and describe its actual helpfulness (see Appendices F and G).

During the initial recruitment visits to the study journal classes, I gave the initial survey to the students who consented to participate in the research. Instead of using class time for them to complete it, I asked them to return it to their instructor the next day. Only 54 of the 117 study journal students did so. For the final survey, I wrote the participating students’ names on the surveys and asked the instructors to have the students complete the survey in the classroom before or after the final exam. This resulted in a much higher participation rate. Of the 79 study journal students who took the final exam, 73 submitted at least one goal sheet or study log. Of these 73 students, 63 completed the final survey, an 86% participation rate. Six of the 79 study journal students taking the final exam did not submit a single study log or goal sheet the entire semester. Five of these six students completed the final survey about the study journal's helpfulness. These five students' survey responses were discarded and were not included in the analyses. Thus, the final sample for the survey analysis was composed of the 63 students who took the final survey and completed at least one study log or goal sheet.

I chose to focus the survey analysis on those students who had regularly completed the study journals. Because the purpose of the confirmatory strand was to find out whether completing weekly study journals was helpful to students, it made sense to distinguish the students who completed study journals weekly or near-weekly from the students who completed very few. Therefore, I created a subsample composed of those students who submitted at least seven study logs and at least seven goal sheets, as well as the final survey. This requirement meant that the student submitted a complete study journal at least once every two weeks, on average. Thirty-six of the 63 students met these criteria. Not surprisingly, most of these (31 out of 36) were in the four classes used in the final treatment effect analysis—those classes whose instructors had collected study journals from most students every week.

Helpfulness Ratings: Descriptive Statistics

The first three questions on the final survey asked students to rate the helpfulness of writing down weekly goals for their math class, planning their math study time, and tracking their actual math study time (see Appendix G). For these three items, the Likert choices ranged from 1 (*not very helpful*) to 6 (*extremely helpful*). On the fourth question, students used a Likert scale to rate the study journals as less helpful, about as helpful, or more helpful than they expected. The means and standard deviations for these questions, using the subsample of students submitting at least seven study journals, are listed in Table 19.

Table 19

Means and Standard Deviations of Likert Questions on Final Helpfulness Survey (36 Students Submitting at Least 7 Study Journals)

Question	Mean	Std. Dev.
How helpful was it to write down weekly goals for your math class?[a]	3.92	1.574
How helpful was it to plan your math study time each week?[a]	4.06	1.472
How helpful was it to track your actual math study time each week?[a]	3.97	1.521
I found the study journal to be ______[b]	2.14	.723

[a]Response values ranged from 1= *not very helpful* to 6 = *extremely helpful*. [b]Response values were 1 = *less helpful than expected*, 2 = *about as helpful as expected*, 3 = *more helpful than expected*.

For most students, the responses to the first three questions were very similar. Students generally gave a positive response to all three, a neutral response to all three, or a negative response to all three. Therefore, for each student, I averaged the first three questions on the initial survey and the first three questions on the final survey. The correlation between the initial survey average and the final survey average was .689 (p < .001). Because only 36 students took both surveys, this correlation is not especially useful. It indicates that the students who expected to find the study journal helpful found it helpful, and those that did not expect to find it helpful found it not helpful. No further analysis on the initial surveys was conducted.

Relationship Between Final Exam Grade and Helpfulness Rating

For the students who completed at least seven study journals, Figure 5 shows the relationship between the final exam grade and the average of the first three items on the final survey (1 = *not very helpful*, 6 = *extremely helpful*). The distribution of the average helpfulness ratings can also be seen in this scatterplot. Only one of these students rated all aspects of the journal as not very helpful, and that student scored above 70 on the final exam. As expected, final exam score and helpfulness rating were not significantly correlated (r = .218, p = .203). Of the students who did well on the final exam, some thought the study journal was helpful, while others did not. The same applied to students who did not do well on the final. However, for the twenty students scoring at least 70 on the final, there was a significant positive linear relationship between the helpfulness rating and the final exam grade (r = .620, p = .004).

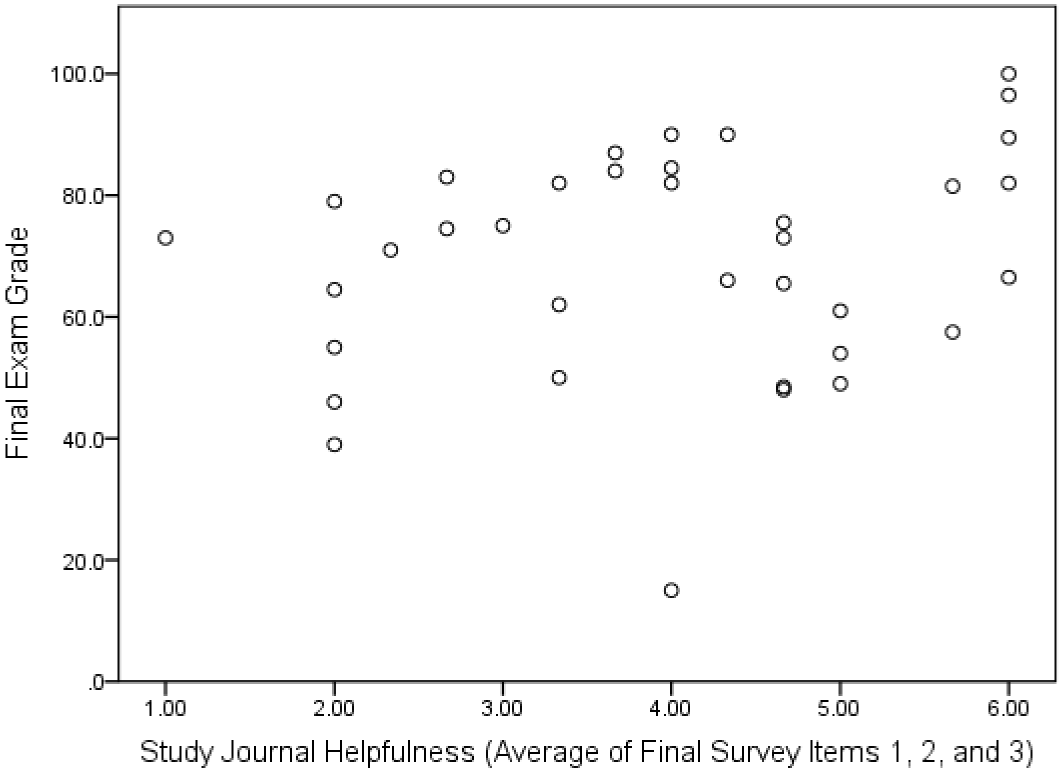


Figure 5. *Scatterplot of study journal helpfulness vs. final exam grade (36 students who submitted at least 7 study journals).*

Figure 6 shows final exam grades and helpfulness ratings for students who completed at least one study journal but fewer than seven. From a visual comparison of Figure 5 and Figure 6, it appears that students submitting at least seven study journals tended to have better final exam grades and higher helpfulness ratings than those submitting fewer than seven study journals. Statistical analysis supports both conclusions: for the 63 students who submitted the final survey and at least one goal sheet or study log, the submission of at least seven complete study journals was positively correlated with both final exam grade (*r* = .319, *p* = .011) and with helpfulness rating (*r* = .288, *p* = .017).

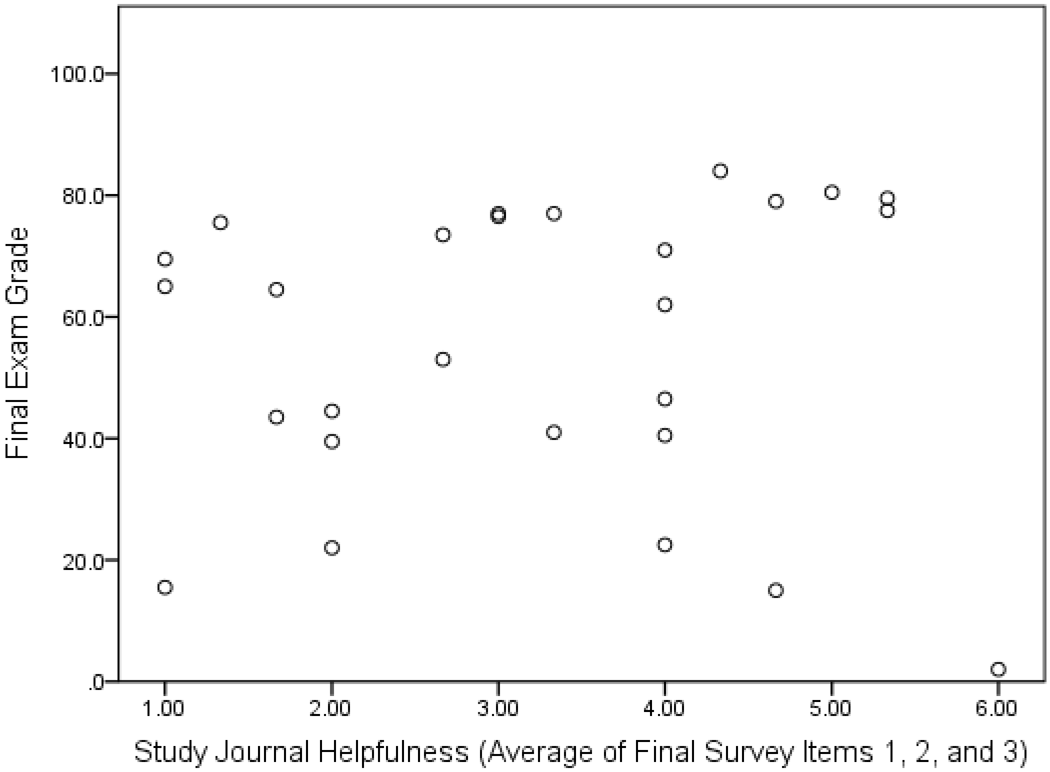


Figure 6. *Scatterplot of study journal helpfulness vs. final exam grade (27 students who submitted at least one study log or goal sheet but fewer than seven).*

Figure 7 combines the final exam grades and helpfulness ratings for both groups of students (those who submitted at least seven journals and those who did not). The most noticeable differences occur near the extremes of the scales. Most students (7 out of 10) awarding an average helpfulness rating above 5 submitted seven or more journals, while most students (6 out of 7) rating it below 2 did not. Only one of the five students scoring below 30 on the final exam submitted at least seven study journals, whereas thirteen of the fifteen students scoring above 80 on the final exam submitted at least seven study journals.

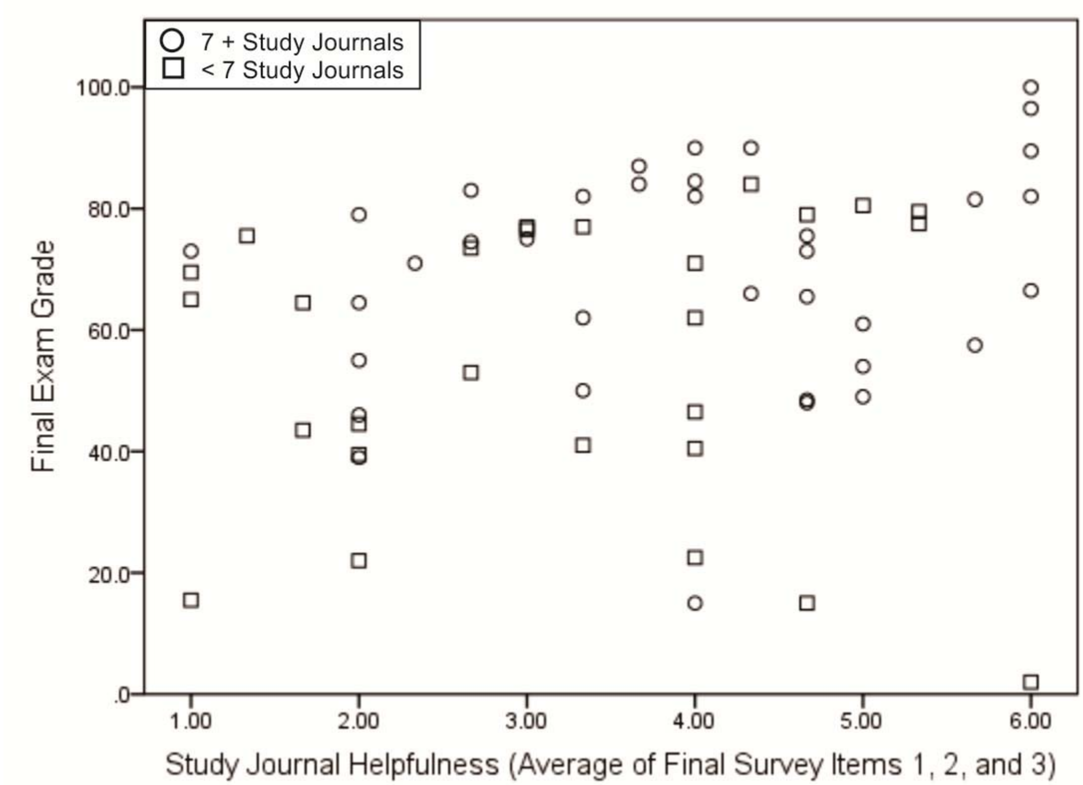


Figure 7. *Scatterplot of study journal helpfulness vs. final exam grade (63 students who submitted at least one study log or goal sheet).*

Responses to Open-Ended Questions

The last two survey questions were open-ended, “How do you think the study journal project affected your success in your math class?” and “Do you expect to do anything differently in future classes because of your experience keeping a study journal? If so, what do you plan to do differently?" Many of the responses to the last question were either redundant with the responses to the previous question, or alluded to future study habit improvements not influenced by the study journal. Therefore, in the analysis of the open-ended responses, I collapsed the last two questions into one, combining redundant responses and ignoring responses that did not refer to the level of helpfulness of the study journal.

For each student's final survey, I listed one theme that characterized the relevant responses to the two open-ended questions. I then examined the resulting list of themes, combined a few similar themes, and made minor wording changes. The theme "helpful-general" includes responses that said the study journal was helpful but did not provide any additional specifics. Although 63 students submitted both the final survey and submitted at least one study journal sheet, one of these did not provide a meaningful response to either of the open-ended questions. Thus, the total sample size was 62. Table 20 lists the frequency counts for each theme (for the students submitting at least one study journal sheet). The information is disaggregated into two subsets: the 36 students submitting at least seven complete study journals, and the 26 students submitting fewer than seven.

Table 20

Theme Frequencies for Open-Ended Questions on Final Helpfulness Survey, Grouped by Journal Submission Frequency

	Frequency		
Theme	< 7 journal submissions	7+ journal submissions	Total
Positive responses			
Helpful-general.	3	3	6
Helped me organize/ prioritize tasks and manage time better.	3	11	14
Helped me remember tasks.	2	2	4
Helped me study more.	2	4	6
Helped me see my weak areas.	1	0	1
Helped me focus.	1	3	4
Helped me see I have too much on my plate.	1	1	2
The free grade was nice.	0	1	1
Total positive responses	13 (50%)	25 (69.4%)	38 (61.3%)
Negative responses			
Trying to schedule my time doesn't help, because I am very busy so I just study when I can.	1	3	4
It was a pain/waste of time/hard to remember.	0	4	4
A planning system should be the student's responsibility.	1	1	2
Total negative responses	2 (7.7%)	8 (22.2%)	10 (16.1%)
Neutral responses			
No effect.	8	3	11
It only helps if you do it.	3	0	3
Total neutral responses	11 (42.3%)	3 (8.3%)	14 (22.6%)
Total responses	26	36	62

Overall, about 61% of the students provided a positive response, saying that the study journal helped them in some way. Many, but not all, of the themes resulted in similar proportions for the students completing at least seven study journals and for those completing fewer. However, 11 of the 14 students commenting that the study journal helped them organize and prioritize tasks and manage their time better completed at least seven journals, indicating that it was difficult to realize this benefit if the study journals were not completed regularly. All four students commenting that the study journals were a waste of time or hard to remember were in the group of students regularly completing them. This makes sense, as students could not have wasted time on study journals unless they did them. It also makes sense that a much higher proportion (42.3%) of students not regularly submitting study journals characterized the journals' effect as neutral, compared to the proportion (8.3%) of regular submitters who chose a neutral response. Again, the only way for the study journals to have an impact, either positive or negative, is for the students to do them.

Students regularly submitting study journals were the source of most comments about specific aspects of the study journals. Unless specified otherwise, the quotes in the following paragraphs came from students turning in at least seven study journals.

Several students thought the study journals not only helped, but helped substantially. A student who earned an 82 on the final exam and an A in the class wrote, "Well, I have always been real good, but I have taken this class twice and wasn't able to pass it. I truly believe that thanks to this project I was able to get it together and understanding [*sic*] math better." Some students thought the study journal motivated them to study more. “I think the study journal helped me alot [*sic*] because knowing that I had to turn something in each week made me study whether I wanted to or not.” From another student, “I believe that it served as a good reminder to put forth effort and also it was helpful to see a written record of how much effort I actually put forth.” A student submitting fewer than seven study journals commented, “It was really helpful in the beginning, but then I stopped.”

As shown in the previously discussed Table 20, a sizable number of students commented that the study journals helped their organization of tasks and time: “It allowed me to stay on track. I sometimes find myself concentrating on all subjects daily but this journal enabled me to arrange things.” “I believe it had a very profound impact. I did very wll [*sic*] in the class and think a lot of it had to do with organizing my study time.” “It certainly improved my studying habits. I feel like I'm more organized now.”

Students who disapproved of the study journals did so for different reasons. Some thought the study journals were a pain or a waste of their time: “I mean it was a great idea, but having a study journal was more of a pain than something helpful. The study journal didn't affect my grade at all, and me doing bad is my own fault for not understanding the work like I should of. At least I'll be one step ahead next semester when I retake this class.” Some students felt their tight schedules made it impossible to plan their study times, and therefore the study journal was not useful: “I did not because I could not plan ahead to study because of my work schedule. I had to study when I could.” One student felt strongly that the study journals contributed to removing responsibility from where it belonged, on the student: “The student must make time to study on their own." "Scrap the journal. Put the emphsis [*sic*] on the student. If they want to learn & pass they will put forth the effort. Holding their hands does not help." Whether for philosophical reasons or not, this student did not submit seven study journals.

Exploratory Strand: Study Habits Revealed by the Study Journals

The confirmatory strand was designed to answer the first two research questions, about the study journal's effect on student success. The exploratory strand was designed to answer Research Question 3, "What are the study habits of the study journal students, as shown by their written goals, study logs, and reflective writings?" It was addressed by the qualitative data in the students' study journals. The exploratory strand and the confirmatory strand were tied together by Research Question 4, "For the study journal students, which of these study habits distinguish successful students from unsuccessful students?" This question was addressed by combining the exploratory strand's qualitative data about study habits and the confirmatory strand's quantitative data about student success. Because Research Questions 3 and 4 pertain to the same set of study habits, their results will be discussed together.

As previously described, I coded the study journal students' goal sheets by themes. I grouped these themes into nine categories. Zimmerman's dimensions of self-regulated learning provided six of the categories: Time (When), Strategies (How), Outcomes (What), Motivation (Why), Environment (Where), and Social Context (Who). The other three categories emerged from the data: Attitudes/Emotions, Obstacles, and Study Journal Characteristics.

For each theme, I recorded the number of students who mentioned that theme at least once. In these frequency counts, I used the ExamSuccess50 variable to classify each student as successful or unsuccessful. Students classified as successful (S) scored at least 50 on my grading of the departmental final exam. Students classified as unsuccessful (U) either scored below 50 on the final exam or did not take the final exam. The successful group and the unsuccessful group each contained 53 students. All of these 106 students submitted at least one study journal. For each theme, I calculated the ratio S/U: the number of successful students mentioning that theme divided by the number of unsuccessful students mentioning that theme. I flagged the themes that met both criteria for distinguishing successful students from unsuccessful students: (1) mentioned by at least 10% of the students submitting a study journal ($S + U \geq 11$), and (2) mentioned by at least 50% more students in one group than students in the other group ($S/U \geq 3/2$ or $S/U \leq 2/3$).

For ease of discussion, I will present the results one category at a time. In each category, the study habits (themes) are summarized in a table. For each theme, the table lists the frequency counts for successful and unsuccessful students along with the S/U ratio. Any themes meeting the criteria for distinguishing the groups are marked with asterisks. Thus, the tables on pp. 199–218 contain the information used to address Research Questions 3 and 4. A short discussion follows each table, highlighting the most notable findings. In the qualitatively derived themes and in this section of the paper, which represent the voices of the students in their study journals, mathematics will be referred to as *math*.

Time (When) Dimension

Table 21

Time (When) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Use of time				
Control my time	25	21	46	1.19
Dedicate study time	14	14	28	1.00
Schedule study times	14	15	29	0.93
Choose optimal times	6	6	12	1.00
Use spare/free time for math	0	3	3	0.00
Time myself	2	0	2	—
Take breaks	2	2	4	1.00
Lack of time	15	10	25	1.50*
Create/make time	11	3	14	3.67*
Work/try/study hard/harder/more	28	24	52	1.17
Study regularly/daily/frequently	15	14	29	1.07
Stay current on math				
Finish on time or early/stay current	16	12	28	1.33
Got behind/ran late	7	2	9	3.50
Start math right away/same day/earlier, don't procrastinate	16	9	25	1.78*
Intentionally delay starting math homework	0	1	1	0.00

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
*S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.)

Time was a frequently mentioned topic for both successful and unsuccessful students (see Table 21). Both groups frequently wrote of the need to take control of their time, to set aside time to work on math, to schedule their study times in advance, and to work on math daily or near-daily, instead of trying to do it all on the weekend.

When coding the study journals, I tried to capture nuances in meaning by using separate codes for different time-related concepts, such as “control my time,” “dedicate (lay aside) time,” and “create time.” I hypothesized that successful students might be more likely to feel that time was within their control, and that unsuccessful students might be more likely to feel they simply did not have enough time. However, the results show that successful students were more likely to cite lack of time as a reason for not meeting their goals. Successful students were also more likely to mention the need for creating more time. Apparently, they viewed “create time” not as the impossible task of adding extra hours to a 24-hour day, but rather as the difficult but possible task of carving extra math time out of a packed schedule.

Staying current on math was another common theme, especially for successful students. Twenty-five students, about a fourth of the total, mentioned the importance of working on their math soon after class or on the same day as their class. This topic met the criteria for distinguishing the groups, being mentioned by more successful students than unsuccessful students. Some students were very aware of their shortcomings in this area. As one student observed, “I need to change my mind set and attitude. I also need to change my study habits, and stop waiting to the last minute to complete my work.” When asked about changes that were needed, another responded, “To try to study more on the days that I actually have class because I noticed that I apply the information better to my homework when I do it the same day as I have class.” Successful students were more likely than unsuccessful students to criticize themselves for running behind or procrastinating.

Strategies (How) Dimension

Table 22

Strategies (How) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Miscellaneous Content Learning Strategies				
Teacher-specific strategies, index cards, teach myself, cram	6	4	10	1.50
Review				
Review (general)	12	3	15	4.00*
Review/study for test	22	9	31	2.44*
Review sheet	12	2	14	6.00*
Review/do handouts/worksheets	5	3	8	1.67
Review homework	6	1	7	6.00
Review/take/use/improve notes	19	9	28	2.11*
Do chapter reviews	1	0	1	—
Take practice test	1	0	1	—
Read/review/use book	10	4	14	2.50*
Review previous course	2	3	5	0.67
Review/study/read before class	2	2	4	1.00
Error-checking	10	10	20	1.00
Practice				
Practice (general)	33	16	49	2.06*
Make practice problems	1	0	1	—
Work extra (unassigned) problems	10	6	16	1.67*
Work more problems	3	2	5	1.50
Rework problems/do them twice	4	3	7	1.33
Practice to perfection/without help	6	3	9	2.00
Focus on my weaknesses	4	6	10	0.67
Attendance				
Don’t miss class/tutoring. Be on time.	13	7	20	1.86*

Table 22 Continued

Study Habit	Frequency			
	S	U	Total	S/U
Organization				
Remember tasks/supplies	15	11	26	1.36
Focus				
Focus/need for focus	10	8	18	1.25
Don’t rush/slow down/be careful	7	6	13	1.17
Pay attention/listen	13	4	17	3.25*
Resource Use				
Videos/websites/etc.	4	11	15	0.36**

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
*S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.). **S + U ≥ 11 and S/U ≤ 0.67 (Met numerical criteria for distinguishing the groups—mentioned by more unsuccessful students than successful students.)

Of all the study habits emerging from the study journals, the most striking difference between groups is seen in the category of review strategies. Throughout this category, successful students were consistently more likely to mention various review strategies (see Table 22). This difference is not just attributable to the same small group of students mentioning all the different review strategies. Taken together, all the subthemes under Review (from Review [General], Review/study for test, all the way down to Review/study/read before class) represent 36 distinct successful students and 22 distinct unsuccessful students, for an S/U ratio of 1.64. This also meets the criteria for distinguishing between the groups.

“Practice” was another frequently mentioned theme. Many students simply listed “practice” as a goal. Others were more specific, describing the need to practice until they could do the problems perfectly by themselves: “Practice doing problems without looking at an example." "I want to be able to finish my homework with as little help possible." "I want to fully understand polynomials in order to be able to work them completely with ease." Some students were hard on themselves for not practicing enough: "I did homework just one time…was not ready for quiz." "Did not work on fractions as much as I should."

The themes of "practice" and "work extra problems" also served to distinguish the groups, both mentioned by more students in the successful group than the unsuccessful group. Interestingly, "attend class," "pay attention," and "remember tasks/supplies" also distinguished the groups. All were more likely to be mentioned by successful students than unsuccessful students. Though one might think that attending class, paying attention, and bringing supplies are such obvious necessities that good students would not need to write them down, that was not the case. The use of videos and websites was the only strategy meeting the criteria for distinguishing the groups that was mentioned more by unsuccessful students than by successful. Without collecting more data, we can only speculate as to the reasons for this. There is no way to know whether the unsuccessful students watched the videos because they did not understand the explanations provided during class, or whether they were using videos as a substitute for coming to class or practicing.

While the aforementioned Time (When) and Strategies (How) categories of study habits were reasonably clear-cut, the Motivation (Why) and Outcomes (What) dimensions are less so, and require some explanation. As described in Chapter II, Zimmerman's Motivation (Why) dimension is concerned with self-motivation and goal setting, and Zimmerman's Outcomes (What) dimension is concerned with the self-monitoring of performance outcomes. If a student response referred to the process of goal setting or planning, I placed it under Motivation. If a student response described the desired outcome of a study session, or a set of study sessions, I placed it under Outcomes.

Motivation (Why) Dimension

Table 23

Motivation (Why) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Described planning/goal setting process	4	8	12	0.50**

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
**S + U ≥ 11 and S/U ≤ 0.67 (Met numerical criteria for distinguishing the groups—mentioned by more unsuccessful students than successful students.)

Only 12 study journals contained entries about the process of goal setting or planning (see Table 23). These entries included comments about adjusting goals or strategies as needed, reevaluating goals, using the goal sheet, meeting written goals, absence of goals, and sticking to the plan. Examples of goal-related comments were "My last week's goals were met but I always add more goals." "The reason that I didn't meet any goals is because I didn't set any goals." and "In the beginning I sticked [*sic*] to my goals but by the end of the week I let it go." One student focused on the enjoyment of the process: “I’m very satisfied and I enjoy setting goals and time limits to complete my work.” Because four students who wrote about the goal-setting process scored at least 50 on the final and eight did not, this theme met the criteria for distinguishing the groups. However, because the numbers were small and the responses varied widely, not much weight should be placed on this finding.

Outcomes (What) Dimension

Table 24

Outcomes (What) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Task				
Complete/do homework	45	32	77	1.41
Complete/do computer labs or online homework	26	22	48	1.18
Miscellaneous: Math vocabulary, math appreciation, better study habits, confidence, speed, etc.	5	5	10	1.00
Topic/chapter goals: e.g., “section 4.3” or “quadratic equations”	28	24	52	1.17
Process				
Specific frequency/ time goals	16	12	28	1.33
Achievement				
Goal of good/better grades on quiz/test/class	27	7	34	3.86*
Expertise				
Understanding/mastery	30	22	52	1.36

Table 24 Continued

	Frequency			
Study Habit	S	U	Total	S/U
Competency				
Retain/memorize content; learn formulas/rules, steps/shortcuts	9	11	20	0.82

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
*S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.)

Most study journal entries listed under Outcomes were responses to the first question on the goal sheet: "What math-related goals do you want to accomplish this week?" Though the question referred to goals, most responses could be thought of as outcomes that could potentially be self-monitored. I divided them into task outcomes, process outcomes, achievement outcomes, expertise outcomes, and competency outcomes (see Table 24). Task outcomes were most frequently mentioned, especially homework and computer labs. Many students listed a book section, such as "Section 3.1," or a mathematical topic, such as "factoring" or "solving equations." If the student's goal was process-related, such as "study at least 30 minutes every day," "work on math 3 times this week," or "study at least 4 hours per week," I listed it as a process outcome. Achievement-related outcomes included comments such as "Make a 100 on the first math test," "maintain an A average in the class," and "pass the test." Expertise and competency outcomes represented different levels of content knowledge. An example of an expertise outcome was "I want to fully understand polynomials in order to be able to work them completely with ease." Even among the entries I counted as expertise outcomes, this entry was unusual for both its specificity and its coherence. Competency outcomes reflected a more superficial approach to learning math, focusing on memorizing rules and formulas, without mention of understanding or mastery.

Achievement-related outcomes showed a decisive difference between the successful and unsuccessful groups. Of the 34 students mentioning some sort of achievement-related goal, 27 of them scored at least 50 on the final. It is important to note that not all of the achievement-related outcomes mentioned by the students represented a high level of accomplishment. Some students simply wanted to "pass the test" or "get a C." Of course, the "successful" standard being used here, ExamSuccess50, also does not represent a high level of accomplishment, as a course average of 70% is needed to earn a passing grade. Still, it is notable that students focused on grades, whether they were striving for an A or a C, were far less likely to either leave the class or to score below 50 on the final exam.

Environment (Where) Dimension

Table 25

Environment (Where) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Choose optimal study setting	8	9	17	0.89
Eliminate/avoid distractions	5	10	15	0.50**

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
** S + U ≥ 11 and S/U ≤ 0.67 (Met numerical criteria for distinguishing the groups—mentioned by more unsuccessful students than successful students.)

Relatively few students mentioned a need to control their study environment (see Table 25). Some students simply stated a plan for working in a particular place: “I plan to work on campus and not do homework at home.” “Stay in my room for more than an hour doing homework.” Other students described specific distractions they needed to avoid: “I will put aside time each week with no tv, radio, computer or any other distractions to focus only on math.” One student mentioned several specific distractions: “Turn off my phone and ipad so I can study.” “Stay off the social network.” “Party later.” More of the unsuccessful students provided journal entries about eliminating or avoiding distractions, meeting the criteria for distinguishing the groups.

Social Context (Who) Dimension

Table 26

Social Context (Who) Dimension: Study Habits Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Seek help – unspecified	25	28	53	1.12
Seek help – official/professional sources: (college tutoring center, instructor, private tutor, etc.)	19	21	40	1.11
Seek help – unofficial sources (family friends, etc.)	3	4	7	1.33
Work alone	3	3	6	1.00
Support network for studying	4	4	8	1.00
Participate/ask questions in class	8	5	13	1.60*

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
*S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.)

Seeking help was mentioned by more students than any other study habit theme, save one (the only study habit mentioned by more students was complete/do homework, with 77 students total). Many students simply mentioned they planned to seek help, without specifying the source. By far the most commonly mentioned sources of help were the two mathematics tutoring centers on campus (one serves developmental mathematics students only; the other serves both credit-level and developmental students). Four students mentioned asking for help from their instructor, three students mentioned a private tutor, and seven mentioned getting help from a friend or family member.

When examining how the social context study habits are distributed between the successful and unsuccessful students, perhaps the most notable feature is the absence of differences (see Table 26). The successful students and the unsuccessful students were nearly equally likely to write about seeking help. Moreover, a sizable number of the successful students seeking help actually did quite well on the final. As previously mentioned, the dividing line between “successful” and “unsuccessful” on ExamSuccess50 does not represent a high level of accomplishment, as 50 is not considered a passing grade. However, the need for help in math was not restricted to those “successful” students with scores near 50. Of the 25 successful (on ExamSuccess50) students who mentioned seeking help (unspecified), 14 scored at least 70 on the final. Of the 19 successful students seeking help from official sources, 11 scored at least 70 on the final. Students doing well in the class were also more likely to mention class participation as a goal or strategy. The theme of participating in class met the criteria for distinguishing the groups, barely, mentioned by more successful students than unsuccessful students. Six of the 8 successful students mentioning class participation scored at least 70 on the final.

Attitudes and Emotions

Table 27

Attitudes and Emotions Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Attitude or Emotion	S	U	Total	S/U
Negative emotions: Laziness, fear, stress, frustration, discouragement, defeat	10	8	18	1.25
Positive emotions: Confidence, positive attitude, love for math	15	6	21	2.50*
Overconfidence	2	1	3	2.00
Discipline, determination, commitment, diligence	9	7	16	1.29

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
*S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.)

Some students expressed strong feelings of discouragement about their math class. A student who eventually dropped the class said, "I was not too successful, b/c I actually got behind in my math homework and wanted to quit. I also felt defeated that math won and I will end up dropping the class. Thank God, I set my mind to finish what I started and I'm back." Another student, who stayed until the end but scored below 50 on the final, was frustrated that she still was not understanding math, even though she was spending a great deal of time on it: "I am focusing so much on math that I'm losing in site [*sic*] of other courses." She expressed dissatisfaction with her progress in the class: "I am really not understanding or getting the concept." "I am not [satisfied] because I have failed every test thus far." She did not plan any changes, "…because I have applied my all."

Feelings of defeat and discouragement were not restricted to unsuccessful students (see Table 27). Successful students were just as likely to mention such feelings, and not just the students who were borderline successful. Of the 10 students scoring above 50 who mentioned negative emotions, 7 of them passed the final with a score of at least 70. However, positive feelings about the math class were much more likely to be expressed by successful students. Not surprisingly, this theme met the criteria for distinguishing between successful and unsuccessful students.

Obstacles

Table 28

Obstacles Mentioned in Study Journals of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Obstacle	S	U	Total	S/U
Trouble with math content: confusion, lack of understanding, weak skills	17	13	30	1.31
Frustration with instruction, institution	6	10	16	0.60**
Unexpected problems or emergencies	12	10	22	1.20
Busy/overcommitted				
Too busy	9	3	12	3.00*
Work	16	12	28	1.33
Other classes	11	10	21	1.10
Other priorities/activities	6	2	8	3.00
Kids	5	2	7	2.50
Family/friends	0	2	2	0.00

Table 28 Continued

Obstacle	Frequency S	 U	 Total	S/U
Tiredness	5	3	8	1.67

Note: S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
* S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.). ** S + U ≥ 11 and S/U ≤ 0.67 (Met numerical criteria for distinguishing the groups—mentioned by more unsuccessful students than successful students.)

Under Obstacles, the most interesting finding is that there are very few differences in the obstacles cited by successful and unsuccessful students (see Table 28). Apparently the successful students were just as likely as unsuccessful students to experience difficulty with mathematical content, emergencies, and physical exhaustion. They were also just as likely to be overwhelmed with other commitments, such as work, family, and other classes. In fact, "too busy" met the criteria for distinguishing the groups, being mentioned by more successful students. Examples of emergencies were illnesses, car accidents, loss of loved ones, and difficulties finding housing. Rather surprisingly, lack of understanding of mathematical content was also more likely to be mentioned by successful students, though it fell short of the criteria for distinguishing the groups. Unsuccessful students were more likely to describe frustration with the college or with the way their class was taught. Though not mentioned by a large number of students, this obstacle met the criteria for distinguishing the groups.

Study Journal Characteristics

Table 29

Study Journal Characteristics of Students Who Were Successful or Unsuccessful on ExamSuccess50

	Frequency			
Study Journal Characteristic	S	U	Total	S/U
Specific fraction of work/homework; numerical self-ratings or completion percentages (e.g. 8.5 out of 10 or 85%)	8	9	17	0.89
Slogans: Just do it, finish strong, keep trying, push myself, don't quit or give up, keep open mind	9	6	15	1.50*
Average quality score (goal sheet)				
2.5 or 3	19	18	37	1.06
2	21	20	41	1.05
1 or 1.5	13	11	24	1.18
Average quality score (study log)				
2.5 or 3	15	15	30	1.00
2	13	9	22	1.44
1 or 1.5	23	28	51	0.82

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
* S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.)

In general, the study journals of successful students were similar to the study journals of unsuccessful students (see Table 29). Seventeen students, split nearly evenly between the groups, used numerical scales to rate their satisfaction and progress toward their goals. Many rated themselves using a percentage, such as "almost 2/3 complete on homework" or "I accomplished 90% of the goals." One student used a scale of 1 to 10 and rated most weeks as 8 or 9, with an occasional 5 or 6 rating. Another used a scale of 100, rating satisfaction with the week's progress as 70/100 or 50/100.

Some students seemed to use the study journals to give themselves a pep talk, with motivating slogans. One wrote, "I'm not completely satisfied, at times I'm overwhelmed, but I'm hanging in there." One listed "Don't give up" and "keep a [*sic*] open mind about math" as strategies. This theme met the criteria for distinguishing the groups, being mentioned by more successful students that unsuccessful students. However, because this theme is a conglomeration of assorted such slogans, this finding should be interpreted not as indicating that successful students favor a particular slogan, but rather that successful students were more likely to use the study journal for positive self-talk.

Unlike all the previously described study habit themes, which were derived from the individual words and phrases used by the students, the quality scores were based on an overall assessment of the study journal's depth. As previously described, another rater and I both scored the quality of the goal sheets and study logs using the rubric in Appendix H. The lowest possible score was 1, indicating the entries looked very similar from week to week and contained minimal or no evidence of planning or reflection. The highest possible score was 3, indicating most entries were tailored to the week at hand and contained evidence of deep reflection or detailed planning. After we discussed inconsistencies, adjusted the rubric, and rescored as necessary, all our ratings agreed within one unit. The scores from the two raters were averaged. For the frequency counts, I grouped ratings of 2.5 and 3 together, because these ratings indicated that at least one rater had awarded the maximum score of 3. The ratings of 1 and 1.5 were also grouped together, because these ratings meant that at least one rater had awarded the minimum score of 1. Because all our scores had agreed within 1 unit, an average score of 2 meant that each rater awarded a rating of 2.

For both the goal sheet and the study log, all three rating categories showed a fairly even split between successful and unsuccessful students. None of the study journal quality characteristics met the criteria for distinguishing the successful and unsuccessful groups.

Note that although the ExamSuccess50 variable resulted in 53 successful students and 53 unsuccessful students, there are fewer than 106 quality scores for both the goal sheet and the study log. This is because 3 students submitted only the goal sheet and 4 submitted only the study log, and only 99 submitted both. Thus there are 99 + 3 = 102 quality scores for the goal sheet and 99 + 4 = 103 quality scores for the study log. Also, there were 4 + 3 + 99 = 106 study journals, which happened to be evenly split on the ExamSuccess50 variable.

Summary of Study Habit Themes Meeting Criteria for Distinguishing the Groups

The nine previous tables contain a total of 21 study habits that met the numerical criteria for distinguishing successful students from unsuccessful students (on ExamSuccess50). Table 30 lists these study habits, separated by category.

Table 30

Study Habits that Met Criteria for Distinguishing Successful and Unsuccessful Students on ExamSuccess50

	Frequency			
Study Habit	S	U	Total	S/U
Time (When) dimension				
Lack of time	15	10	25	1.50*
Create/make time	11	3	14	3.67*
Start math right away/same day/earlier, don't procrastinate	16	9	25	1.78*
Strategies (How) dimension				
Review (general)	12	3	15	4.00*
Review/study for test	22	9	31	2.44*
Review sheet	12	2	14	6.00*
Review/take/use/improve notes	19	9	28	2.11*
Read/review/use book	10	4	14	2.50*
Practice (general)	33	16	49	2.06*
Work extra (unassigned) problems	10	6	16	1.67*
Attendance: Don't miss class/tutoring, be on time.	13	7	20	1.86*
Pay attention/listen	13	4	17	3.25*
Videos/websites/etc.	4	11	15	0.36**
Motivation (Why) dimension				
Described planning/goal setting process	4	8	12	0.50**
Outcomes (What) dimension				
Achievement outcomes: Goal of good/better grades on quiz/test/class	27	7	34	3.86*
Environment (Where) dimension				
Eliminate/avoid distractions	5	10	15	0.50**
Social context (Who) dimension				
Participate/ask questions in class	8	5	13	1.60*
Attitudes and Emotions				
Positive emotions: Confidence, positive attitude, love for math	15	6	21	2.50*
Obstacles				
Frustration with instruction, institution	6	10	16	0.60**

Table 30 Continued

	Frequency			
Study Habit	S	U	Total	S/U
Too busy	9	3	12	3.00*
Study Journal Characteristics				
Slogans: Just do it, finish strong, keep trying, push myself, don't quit or give up, keep open mind	9	6	15	1.50*

Note. S = Successful (scored at least 50 on final); U = Unsuccessful (scored below 50 on final or did not take final).
* S + U ≥ 11 and S/U ≥ 1.50 (Met numerical criteria for distinguishing the groups—mentioned by more successful students than unsuccessful students.) ** S + U ≥ 11 and S/U ≤ 0.67 (Met numerical criteria for distinguishing the groups—mentioned by more unsuccessful students than successful students.)

Seventeen of these 21 significant study habits were mentioned by more students in the successful group. Three of them (lack of time, make/create time, and start math right away) fell under Zimmerman's Time (When) dimension. Nine of the significant study habits mentioned by more successful students fell under the Strategies (How) dimension, including five review strategies, two practice strategies, along with on-time class attendance and listening/paying attention. Under the Outcomes (What) dimension, successful students were more likely to mention achievement outcomes (good, better, or passing grades). Under the Social Context (Who) dimension, successful students were more likely to mention class participation. The other three of the 17 significant study habits mentioned by more successful students fell outside of Zimmerman's dimensions: positive attitude toward math, being too busy, and slogans.

Only four of the 21 significant study habits were mentioned by more students in the unsuccessful group. These were videos/websites (Strategies), describing planning or goal setting (Motivation), avoiding distractions (Environment), and frustration with instruction/institution (Obstacles).

Several of the study habits favored by the successful group featured high total frequency counts with decisive differences. For example, reviewing/studying for a test was mentioned by 31 students (22 successful and 9 unsuccessful). Practice (general) was mentioned by 49 students (33 successful and 16 unsuccessful), and achievement outcomes were mentioned by 34 students (27 successful and 7 unsuccessful). Among the significant study habits favored by the unsuccessful group, the highest total frequency count was 16 (6 successful and 10 unsuccessful), for frustration with instruction or institution.

Discussion

What We Have Learned

Intervention Had an Effect That Could Be Detected Statistically

In this study, I sought to learn how weekly study journals would affect students' success in developmental mathematics courses. While many educational interventions do not result in a statistically significant effect, this one did. And the effect was not minor—it was substantial. The odds ratio for TookFinal was 2.94, indicating that the odds for treatment students leaving the class before the final exam were nearly three times the odds for control students leaving the class.

Of course, the intervention was designed with a different effect in mind. It was designed based on a self-regulated learning framework, in hopes that nudging students toward self-regulated learning behaviors could help them become more successful in mathematics. As educational researchers, we attempt to avoid bias and objectively evaluate the effect of interventions, rather than letting wishful thinking drive our conclusions. However, our desire for objective evaluation does not negate the fact that the interventions are intended to help students' success, not to damage it.

In this study, the hoped-for positive outcome was a passing grade in the course or on the final exam. The four outcome variables, CourseSuccess, ExamSuccess70, ExamSuccess60, and ExamSuccess50, should not be seen as independent constructs or dimensions, but rather as four alternative ways to capture a single construct, "mathematics course success." To facilitate discussion, I have used a single table to combine the previously presented frequency counts for the four success variables and for TookFinal (see Table 31). These frequency counts are based on the matched sample taken from the four treatment classes fully implementing the intervention and the four corresponding control classes.

In this table, students taking the final exam were counted as "successful" on TookFinal; students not taking the final exam were counted as "unsuccessful" on TookFinal. On the CourseSuccess variable and the three exam success variables, the "unsuccessful" category is composed of two groups of students: those who left the class (either officially or unofficially) before the final exam, and students who stayed in the class until the final exam but did not earn an A, B, or C in the course (for CourseSuccess) or reach the cut score on the final exam (for ExamSuccess70, ExamSuccess60, and ExamSuccess50). Thus, the set of students classified as “unsuccessful” on TookFinal is a subset of each of the other sets of “unsuccessful” students.

On all the outcome variables shown in Table 31, the percentage of control students classified as successful was higher than the percentage of treatment students classified as successful. A visual examination of the frequency counts indicates that on the three exam success variables, the treatment students’ lower success rates are primarily attributable to students leaving the class, not to treatment students scoring lower on the final exam. When restricted to students taking the final exam, treatment and control students reached each cutoff score in similar proportions (14/33 = 42.4% and 18/44 = 40.9% for ExamSuccess70, 20/33 = 60.6% and 27/44 = 61.4% for ExamSuccess60, 24/33 = 72.7% and 34/44 = 77.3% for ExamSuccess50).

On the CourseSuccess variable, the treatment students’ lower success rates are due in part to students leaving the class, and in part to differences in course grades among the students who stayed. Of the students taking the final, 19/33 = 57.5% of the treatment students and 32/44 = 72.7% of the control students earned grades of A, B, or C. The fact that the control students’ higher course grades are not reflected in the exam success percentages indicates that they may be attributable to differences in teacher grading policies, rather than differences in mathematical knowledge.

Table 31

Frequencies for Course Success, Exam Success, and Taking the Final Exam: Matched Subsample From Modified Sample of 8 Classes

Group	CourseSuccess	ExamSuccess70	ExamSuccess60	ExamSuccess50	TookFinal
Treatment					
Successful	19 (35.2%)	14 (25.9%)	20 (37.0%)	24 (44.4%)	33 (61.1%)
Not successful	35 (64.8%)	40 (74.1%)	34 (63.0%)	30 (55.6%)	21 (38.9%)
Control					
Successful	32 (59.3%)	18 (33.3%)	27 (50.0%)	34 (63.0%)	44 (81.5%)
Not successful	22 (40.7%)	36 (66.7%)	27 (50.0%)	20 (37.0%)	10 (18.5%)
Chi-square (sig.) (1, *N* = 108)	6.279 (.012)	0.711(.399)	1.846(.174)	3.724(.054)	5.475(.019)

Upon finding a statistically significant result, especially an unexpected one, it is essential to consider the plausibility of that result. Is it plausible that the study journal intervention could have caused some students to leave the class?

In the matched subsample chosen from the restricted sample of four treatment classes and four control classes, the treatment group was noticeably stronger than the control group on most of the academic variables, with the largest differences being in CourseCompletionRatio, DevMathGPA, and AttemptsPerPass. Therefore, it is unlikely that the treatment students left the class because they were having more difficulty with the mathematics than the control students.

If indeed the statistically significant difference on TookFinal was due to the study journals, rather than being a fluke of individual differences or unobserved confounders, then the study journal resulted in increased self-awareness, just as it had been designed to do.

For this argument to stand, we must first rule out alternative explanations. Because the study journal intervention did not require a substantial time commitment, was not graded, and was not seen by the class instructors, it does not seem plausible that students would leave the class because the study journal was burdensome. The students could receive full credit for the goal sheet by jotting a few short notes and turning it in. Some students did just that, leaving some questions blank and writing “yes,” “no,” or “study more” on others. On the study log, the students had been specifically instructed to write “did not study” on their study log if they did not study in a particular week, and they were reassured they would still receive full credit for the study log. The students had also been reassured their teachers would not read either of the study journal worksheets.

However, it seems possible that some study journal students may have left the class due to increased self-awareness, derived from writing down their goals, describing the effectiveness of their study strategies, and logging their study time. On the final survey, one student captured this study's most noteworthy statistical result in a single comment: "Because of external factors, time to study is limited. Being more aware of this limitation is only somewhat helpful."

Unfortunately, as reflected in the above student comment, awareness does not automatically improve success. To produce success, the increased awareness must be followed by action—some change in the students' studying process. The study journal intervention was not combined with training or counseling. If the study journals helped students realize they were not studying effectively, had fallen far behind, or had too much on their plates, they may have had no idea what to do about it. They may have concluded that leaving the class was the only way to resolve the situation.

Positive Effect on Some Students Who Stayed in Class

The focus group students overwhelmingly felt the study journals were helpful. As one student said emphatically, holding his goal sheet up and rattling it, "the first sheet here, the one that has the questions and, then it has the thing on the back [a table for planning the next week's time], was exceptionally helpful." Later, just as earnestly, he repeated his assertion that the goal sheet was “exceptionally helpful.” The students in the focus groups willingly shared their opinions about the redundancy of the two worksheets (at least as used by the students), described their forgetfulness about doing the worksheets, and confessed to sometimes filling out a worksheet five minutes before class. Still, the students wholeheartedly recommended the study journals be continued. The students’ openness about their own imperfect implementation lends an air of authenticity to their appraisal of the study journal’s value.

The written surveys, in both numerical ratings and open-ended responses, also indicate that some students found the study journals helpful. However, the survey responses should be interpreted with caution. Five students completed the final survey but did not submit a single study log or goal sheet the entire semester. Three of these said the study journal was helpful, even offering specific comments “I may actually keep a journal in my next class, depending on the time I have” and “allowed me to see places I needed improvement.” Two of these students awarded high helpfulness ratings to the journal, even though they never submitted it. However, the final surveys contained a sizeable number of thoughtful comments from students who regularly completed the study journals and felt the journals had helped their studying.

Students who regularly submitted study journals tended to have higher final exam grades than students who did not. Nearly all students with extremely high exam scores regularly submitted study journals, while very few of the students with extremely low exam scores regularly submitted the journals. Note that this relationship between exam score and regular submission of study journals is associative, not causal. Students who are conscientious about the study journals are likely to also be conscientious about completing their mathematics assignments and studying for their mathematics exams. Still, this association, combined with the student comments in the focus groups and on the surveys, provides evidence that the study journal intervention helped improve the study habits of some students.

Potential Positive Effects on Departing Students

In the following discussion of success, I will use course letter grades rather than exam grades. Course letter grades have meaning to students and will make it easier to visualize success through their eyes. In this study, exam success variables did not represent alternative philosophies of success; the exam success variables were simply less-confounded proxies for CourseSuccess.

Consider two hypothetical students. During the intervention semester, both students carried a full course load along with substantial work hours and family responsibilities. Both students struggled in the class in the intervention semester, receiving low test grades and falling behind. Both students considered dropping the course. The first student stayed in the class and passed, earning the lowest possible C for his course grade. Next semester, he enrolled in the next higher mathematics course. He was not adequately prepared for the more challenging course material, and still faced the same work and family obstacles. He failed the course. The second student chose to withdraw from her mathematics course during the intervention semester, because she felt too overloaded to succeed. The next semester, she decided to reduce her course load. She repeated the mathematics course, and took just one other course with it. The reduced load, combined with thoughtful improvements in her study habits, helped her to earn a solid A in the mathematics course on her second attempt. The next semester (two semesters subsequent to the intervention), she enrolled in the next higher mathematics course fully prepared, and again did well.

Which of these two outcomes should be considered a “successful” result for the intervention semester? According to the definition of success I used in the quantitative analysis, the first student (who passed with a C) was “successful” and the second student (who withdrew) was “unsuccessful.” This definition of success was the least problematic choice for the quantitative analysis, considering the data available and the difficulty of making meaningful distinctions between students receiving grades of W, F, and IP. However, this definition of success is a researcher’s definition, not a student’s definition.

If we consider a student’s perspective, both long- and short-term, we may see a very different picture. Some students may consider squeaking by with the lowest possible C a success, while others see it as a dismal failure. Time gained by withdrawing from the mathematics class may allow a student to earn higher grades in other classes. Repeating the course may build a firmer foundation for future mathematics courses.

If the study journals caused some students to leave the class by making the students more aware of their constraints and the shortcomings in their study habits, those students may have received a benefit. For some students, leaving the class may have been a strategic move, possibly even a form of self-regulating their learning. For other students, leaving the class may have been reactive rather than proactive—a natural response to a situation they felt was hopeless. These students may not have received any benefit to their study habits, and may find themselves in a similar situation the next semester. Without collecting data about their reasons for leaving and their academic performance the next semester, there is no way to know whether they benefited from the study journals.

Intervention Not as Easy To Implement as It Appeared

This study was designed to ascertain whether students could benefit from a simple study-journaling intervention, which did not take class time away from content instruction or require instructors to invest time in preparation or feedback. All the participating instructors agreed to collect the study journals weekly and to count them in the students' course grades. None of the instructors expressed reluctance or any concern about their ability to implement the project. However, the intervention was only implemented as intended in four of the nine treatment classes. In the others, the study journals were collected infrequently, or were collected at regular intervals but only from a few students.

The four classes fully implementing the project were taught by three instructors, each with at least 25 years of experience teaching college full-time. (Two of the treatment classes had the same instructor.) Two of the three instructors were full-time faculty members at the community college where the study took place. The third instructor was an adjunct faculty member for the community college and also a full-time faculty member at a nearby four-year university. (As the researcher, I was not the instructor for any of the treatment or control classes.)

Other than possibly issuing more frequent reminders during class, the instructors of these four classes did not spend any more time on the study journal project than the instructors of the other five treatment classes. Observation of classes was not incorporated into the research design, so we can only speculate about what these instructors may have done differently. Teaching is a craft—it depends on skill, experience, and a multitude of intangible factors. Instilling a habit in other people is not easy; there is not a formula for it. Making an assignment "required" does not guarantee students will do it.

Whether through skill, experience, or the culture they built in their classroom, these three instructors managed to help most of their students develop a study journaling habit, at least for one semester. Not all their students regularly completed the study journals, but most did. One of the focus group students described his teacher's study journal collection process: "Well, and every Tuesday, he walks in, holds up the big manila envelope. He's like, 'Put your stuff in here. Here's two more [study journal forms—goal sheet and study log].' And he will wait until we've all gone up there and done it."

In one of the five classes not collecting many study journals, several students expressed negativity toward the study journal project during the recruitment visit. I talked with the instructor and we both agreed that she should reassure the students that the study journals were meant to be helpful, not punitive, and that it would not be wise to push the study journals in a heavy-handed manner if the students' reluctance persisted. This may explain the lack of study journal submissions in this particular class, but it does not explain it for the other four classes.

The fact that this was a research project introduced constraints that may have affected the implementation. The students knew this was a research project occurring in multiple classes, rather than an activity initiated by their instructor. This may have affected the students' level of buy-in. Also, the students knew that their instructors were not going to read or assess their study journals. This was done for several reasons: so students would feel free to be honest in their study journals, so the statistical analysis would not be confounded by assessment differences, and so the study journal project would not be burdensome to the instructors. If the instructors had been critiquing the study journals for quality, the students may have taken them more seriously.

Connections With Existing Research

Contributions to the Self-Regulated Learning Research Field

As previously discussed, rigorous studies of self-regulated learning interventions in content courses (as opposed to study strategies courses) are in short supply, especially in content courses not related to psychology. This intervention occurred in the authentic context of a mathematics course. Research constraints necessitated a few compromises with authenticity, but these were relatively minor. By matching the classes on as many class-level variables as possible, randomly assigning treatment to intact classes within each pair, and using propensity score matching to clean up residual differences between the groups, the quantitative findings of this study are more credible than those of many studies occurring in content courses.

As described in Chapter II, previous research has shown that self-regulated learning interventions can improve students' achievement in content courses. This conclusion was partially supported by the current study. The qualitative analysis in the current study showed that the intervention helped the achievement and the study habits of some students, but the quantitative analysis did not show the intervention improved achievement overall. However, the quantitative analysis showed higher departure rates in treatment classes, which may have partially been the result of increased self-awareness about their time use and their study habits. This self-awareness could potentially serve as one step on a path toward improved achievement.

Self-regulated learning interventions vary so widely that it is often difficult to say whether one study supports or refutes the findings of another. Most other studies involving study journals, goal-setting worksheets, or time-management worksheets have combined them with other components, requiring a larger time commitment from the instructors and students. The current study did not support the positive effects on achievement found by Goodwin and Califf (2007), who combined time-use worksheets with time-management training, or Georgianna (2009), who provided in-class training on implementation intentions for academic tasks. However, because these interventions were more involved, the current study also does not provide evidence opposing those studies' findings. More similar to the intervention in the current study is that of Fleming (2002), who found positive effects on psychology class achievement when five minutes of each class were devoted to writing down goals and planning study time. Relevant studies not using achievement as an outcome measure were Sweidel (1996) and Fitch et al. (2012), who found positive results on a helpfulness survey and a self-regulated learning inventory, respectively. Both interventions were much more intrusive and time-consuming than the study-journaling intervention in the current study. Sweidel's study strategy portfolios took the form of reflective essays and were graded for quality; the Fitch et al. goal-setting and planning worksheets were combined with peer-guided discussion groups.

The exploratory strand partially fulfills a recommendation that self-regulated learning researchers rely on observations, rather than self-reports, to determine whether students actually apply self-regulated learning strategies during their mathematics studying (Dinsmore et al., 2008; Schunk, 2008). However, direct observation is almost impossible in authentic settings, and few research studies have attempted it. In the current study, although direct observation of self-regulated learning strategies was not possible, the students' studying process was indirectly observed through their study journals. The study journals served as proxies, providing insight into what the students were thinking as they planned and carried out their studying process.

Contribution to the Methodology Base of Educational Research

This study shows how the use of mixed methods can strengthen intervention research. If the study had relied only on the statistical analysis of success data, we would have concluded that the intervention did not benefit students. If the study had relied only on the qualitative data from the focus groups, we would have concluded the study journal was beneficial, but we would have missed the most interesting result—that the intervention was associated with more students opting out of their mathematics class.

Including a qualitative component is especially important when the quantitative analysis is made difficult by the absence of clear predictive variables and by large numbers of departing students, as was the case in this study. The high attrition rate in these classes made meaningful comparisons of final exam averages almost impossible, as the final exam average would have been heavily influenced by the departure decisions of students having very low grades and no realistic chance of passing. The least problematic option was to dichotomize the outcome variable and count departing students as unsuccessful; however, this meant that achievement improvements could not be detected quantitatively unless those improvements bumped students from below the cutoff score to above it. The qualitative data from the focus groups and surveys provide evidence that the intervention may have helped some students' achievement, even though those improvements were not captured in the statistical analysis.

These qualitative data also serve to explain an unexpected quantitative result—that the study journal students were more likely to leave the class before the final exam.

Some of the students who departed may have experienced the same increased self-awareness noted by the focus group and survey students, who stayed until the end. The qualitative data lend weight to the conclusion that the study journals' impact on departure was real, and should not be dismissed as an aberration.

On the quantitative side, this study exemplifies how propensity score matching can be used to improve the credibility of higher educational research, particularly on community college students. When classrooms contain substantial numbers of students repeating the course, students placing directly into the course through a variety of placement tests, and students receiving a grade in the prerequisite course, traditional achievement predictors such as GPA and SAT score have very limited value. Through propensity score matching, we can trim an initial sample into two smaller groups that have similar proportions of placement students and course repeaters, as well as similar distributions on ordinal variables, making our analysis less sensitive to modeling assumptions (Ho et al., 2007). While the methodological literature contains several frequently-cited examples that show how propensity score matching can create well-balanced subsamples from groups with substantial inherent differences, the current study is one of few examples in which propensity score matching is used to clean up residual differences in two groups that were reasonably similar in the first place.

Where We Go From Here

This study provides evidence that study journals have potential to benefit students and should be further investigated. However, due to the unexpected finding regarding student departures, researchers should proceed with caution. In future studies of similar interventions, we should anticipate the possibility of increased departures and include measures to mitigate any potential negative effects on the students.

From a research standpoint, the effects of departures on studies' inferences should also be considered. In addition to reporting and analyzing the numbers of departures, future studies should be designed to acquire additional data about those departures. Researchers could use interviews or surveys to find out why departing students left their classes and what type of support from the college might help them in their return. Future studies should also incorporate data from the semester after the intervention. For successful students, this would help us find out whether students are able to carry their self-regulated learning skills forward into future classes. For unsuccessful students, whether they departed from the class or remained in class but earned a poor grade, the analysis of subsequent-semester data would provide information on whether students receiving the intervention were able to make the necessary adjustments and succeed in the course on their next attempt.

In addition, we should investigate the effects of combining study journals with training, feedback, counseling, or peer accountability. In the exploratory strand of this study, the study journals revealed that many students saw the importance of self-regulating their use of time, but struggled to make self-regulation of time a reality. Many students, both successful and unsuccessful, felt their work and family commitments left them with insufficient time to study mathematics. Budgeting time for obtaining help outside of class is also important; seeking help was a common theme of the study journals, for both successful and unsuccessful students. Counseling, training, or peer accountability could potentially help students improve their self-regulation of time. In addition to help with the time dimension, the study journals suggest other directions for training. The most noticeable differences between successful and unsuccessful students were in three areas: review strategies, practice strategies, and an emphasis on achievement-based outcomes (grades). All three themes were favored by the successful students. This suggests that that if we can help unsuccessful students integrate review, practice, and achievement into their mindset, we may be able to help them improve.

Action research, in which one or two teacher-researchers analyze the effect of study journals in their own classes, would be especially welcome. If the same study journals were used in a more intrusive manner, with individualized feedback and in-class discussion, the results could be very different. For students who become more aware of shortcomings in their performance and study habits and are considering leaving the class, a face-to-face visit with their instructor may encourage them to stay and may supply them with potential solutions. Placing their study journals into an envelope, to be analyzed later by a researcher, provides neither encouragement nor solutions.

Nearly all college instructors desire that their students not only master and appreciate their course content, but also improve their self-regulation skills. Unlike many educational interventions, study-journaling worksheets can be implemented with relatively small costs—in terms of money, preparation, and instructional time. They also have the benefit of flexibility—they can be tailored to the course content, student needs, and the amount of instructional and assessment time the instructor is willing to allot to them. Therefore, as long as student departure is kept in mind, they deserve further investigation.

CHAPTER IV

USING PROPENSITY SCORE MATCHING TO IMPROVE THE CREDIBILITY OF COMMUNITY COLLEGE RESEARCH

Introduction

Those of us engaged in efforts to improve the success of community college students are continually encountering ideas that could potentially improve student achievement—new support programs, new classroom technology, different pedagogical approaches, or physical changes in the classroom. Whether we are conducting research studies ourselves or reading reports of research studies conducted by others, we must be able to draw credible conclusions about the effectiveness of interventions or programs. In the following discussion, I will use the word "program" to refer to an intervention, policy change, curriculum, pedagogical technique, or computerized platform—in short, anything we want to evaluate statistically for its effectiveness.

The existence of a control group, or comparison group, is often seen as providing credibility to the evaluation of a program. However, too often, the success of a program is evaluated through a simple comparison of outcomes without regard for potential initial differences in student characteristics. Sometimes students in the program are compared to students not in the program. Other times, one naturally occurring group is compared to another naturally occurring group, as when classes using some innovative teaching method are compared to classes taught traditionally. Unfortunately, unless effort is made to control for initial differences between the two groups of students, the comparison of outcomes is of questionable value. We cannot know how much of the difference in outcomes (if any) is attributable to the program and how much is due to inherent differences in the characteristics of the students. The potential costs of failing to control for differences between groups are twofold: we spend valuable resources on ineffective programs, mistakenly believing they are effective; or we discard effective programs, mistakenly believing they are ineffective.

With sufficiently large sample sizes and randomization into groups, values of confounding variables (e.g., individual characteristics) are distributed similarly across groups, removing or reducing their influence upon treatment effects (Guo & Fraser, 2010; Ho et al., 2007). In educational research, this ideal is almost never reached. In the rare case when it is possible to randomly assign students to groups, the sample size is usually small. In most cases, students are not randomized into groups. Instead, they are assigned to groups based on their own decisions to register for a workshop, walk into the tutoring center, or enroll in a class that happens to be using a new curriculum. If a program is optional and students self-select into it, the group of students volunteering for the program will almost certainly be very different from the group of students not volunteering. Even when the groups being compared are composed of sections of the same course, there may still be substantial differences between them.

This chapter describes how I used a statistical technique—propensity score matching—to adjust for group differences when evaluating the effect of a study-journaling intervention for community college developmental mathematics students. Instead of providing a cursory summary of the methodology and then an in-depth description of the results and their implications, I will describe the methodology in detail and use the results only for illustration. Through this methodological focus, the chapter serves two purposes: (1) it provides a practical introduction to propensity score matching techniques and how they can add value to community college research, and (2) it provides ideas for a set of numerical variables researchers can use to represent the academic histories of community college students in a meaningful way.

After providing a rationale and background for the propensity score approach, I will give a detailed explanation of the matching variables and how they were created. The academic histories of community college students, particularly developmental students, present modeling difficulties that are often glossed over by researchers. For example, many research reports mention the use of college grade point average (GPA) as a predictor, without stating what was done for students who have not yet registered for any GPA-eligible classes. By describing in detail how I handled such situations, I hope not only to add transparency and thus credibility to my own study, but also to provide practical guidance for other researchers facing similar data analysis dilemmas. I also present several less traditional success-related measures, including a variable designed to numerically capture students’ course repetition patterns. Such measures will be of interest to anyone studying groups of students who frequently repeat courses.

The rationale, methodology, results, and implications of the research study were described in Chapter III. The study addressed the following research question: Are study journal students more likely to pass the mathematics course and final exam than control students? To answer a question such as this, it is crucial to separate the research process into two distinct phases: (1) design, including assessing the sample's credentials and adjusting the sample if necessary, and (2) evaluation of treatment effect. Only in the second phase is the researcher allowed to examine outcome data (Rubin, 2008a, 2008b).

A nonrandomized sample must be scrutinized very carefully, to decide how much value to place on inferences drawn from it. Researchers should look carefully at the data collection process, note possible sources of bias, and collect data for the specific purpose of assessing the sample's credentials. Ideally, they would also examine the sensitivity of the results to possible hidden bias (Rosenbaum, 1989). If possible, the researcher should complete the first phase—assessing the sample's credibility, making necessary adjustments (e.g., matching or stratification), and checking for satisfactory balance—all before collecting any outcome data. If this is not possible, the careful researcher will take steps to blind himself or herself to all outcome data until the sample-analysis phase is complete (Rubin, 2008a, 2008b). Only after the credentials of the sample have been carefully vetted and documented should the researcher move to the second phase, evaluating the effect of the treatment on the outcome.

This study assigned treatment conditions to clusters of individuals (intact classes). Care was taken in selecting and assigning the clusters (classes). As much as possible, I attempted to avoid systematic (nonrandom) differences between control and treatment groups. I chose the classes in pairs, matching classes based on course, time of day, and teacher (if possible). In all nine of the resulting class pairs, both instructors were willing to participate in the project. In each pair, I used a dice roll to randomly assign one class to the treatment condition and the other to control.

Still, it was possible the groups differed more than would be likely had the individual students been randomly assigned. Even if the treatment assignment process did not directly cause systematic differences (as would be the case, for example, if all night classes were assigned to control and all day classes to treatment), unobserved or unanticipated factors could have caused the groups to attract different sets of students. Possibly one class was particularly desirable to students (because of teacher or class time) and was filled first by the most organized and motivated students. Another class may have been undesirable, for whatever reason, remaining unfilled until right before the semester started—then filled by students who waited until the last possible minute to register, when it was their only option (because other classes were full).

When knowledge about the effect of a treatment is desired but randomization is impossible, a less biased estimate of the treatment effect can be obtained by matching members of the treatment group with members of the control group based on a set of personal characteristics—i.e., a vector of covariates (Ho et al., 2007; Stuart & Rubin, 2007). However, if there are more than a few covariates, especially if some are continuous or nearly continuous (such as GPA), it will be difficult or impossible to find near matches on all the covariates, even when the sample size is large (Guo & Fraser, 2010, p. 132; Ho et al., 2007; Rosenbaum, 1995, p. 200). This is sometimes called the "curse of dimensionality."

An alternative approach is to balance the groups using the propensity score, a single scalar that incorporates information from many covariates. Propensity score techniques are helpful when groups are formed in such a way that they may differ inherently rather than randomly. This can happen when participants choose their own treatment (e.g., enrolling in a certain type of school) or when treatment is forced upon them (e.g., enrolling in college at the time a new mandatory success course is implemented). Inherent differences in groups can also occur when intact groups, such as schools or classrooms, are randomly selected for treatment. Randomizing clusters is not equivalent to randomizing individuals (Campbell & Stanley, 1963; Guo & Fraser, 2010, pp. 16–17). Whether differences between groups are caused by nature, self-selection, or luck, propensity score techniques can often help adjust the sample so that the inferences drawn from it are more credible (D'Agostino, 1998; Ho et al., 2007; Stuart, 2010).

The goal of propensity score matching is to choose subgroups of the control and treatment groups that have similar distributions of important starting characteristics, at least on observable variables likely to be associated with the outcomes. First, I will provide a short summary of propensity score theory. Second, I will describe the matching variables used in this study, the rationale for each variable, and any adjustments that I made in order for the variables to be usable in the propensity score matching model. Third, I will describe the tools and the process I used to perform the propensity score matching. Finally, I will also show how I used primarily graphical means to assess the balance of the resulting groups (and thus the success of the matching algorithm).

Propensity Scores: Background Information

The propensity score of an individual is defined to be the conditional probability that the person will be assigned to the treatment group given that particular person's vector of covariate values. If the groups were created by randomly assigning individuals, each person's propensity score would be 0.5. A propensity score is a special case of a *balancing score*, defined to be any vector-valued or scalar-valued function $b(x)$ of the vector x of observed covariates such that the conditional distribution of x, given a particular value of $b(x)$, is the same across both treatment conditions (Rosenbaum & Rubin, 1983).

For a given value of the propensity score or any balancing score, each observed covariate will be balanced across the two groups. This guarantee of balance applies only to the true propensity score (the actual probability the individual is selected for treatment). In practice, we cannot know the true propensity score; we must estimate it based on the observed covariates (Rosenbaum & Rubin, 1983). Because the estimated scores are based on observed covariate values, matching or stratifying on the estimated propensity score removes imbalances due to bad luck as well as imbalances due to systemic problems, resulting in better balanced groups than would have been created by the true score (if it were available, which it's not). Using the true score would only remove systemic bias (Hill et al., 2006; Joffe & Rosenbaum, 1999). For the same reason, groups created using propensity score techniques are often more similar, at least on the observed covariates, than randomly created groups. However, it is important to note that propensity score matching can only directly balance the two groups across the set of observed covariates, whereas randomization into treatment and control groups theoretically balances both observed and unobserved covariates (D'Agostino, 1998; Joffe & Rosenbaum, 1999; Rubin, 2006; Stuart, 2010). Propensity score matching can improve balance on unobserved covariates if they happen to be correlated with the observed covariates (Stuart, 2010).

Propensity score theory is built on potential outcome theory, in which each individual has a response vector (for two treatment conditions) containing two pieces of information: the individual's response to Treatment Condition A, and that same individual's response to Treatment Condition B. Using 0 for control (or Treatment Condition A) and 1 for treatment (or Treatment Condition B), we could denote the response vector (r_0, r_1). Because the individual will only receive one treatment, we will only have information about one coordinate of the response vector, essentially resulting in a missing data problem. The entire process of estimating treatment effects can then be considered as attempting to resolve this missing data problem (Rosenbaum & Rubin, 1983).

Propensity score methods depend on two assumptions: the stable unit treatment value assumption and the strongly ignorable treatment assumption (Guo & Fraser, 2010; Jo & Stuart, 2009; Stuart & Rubin, 2007). Under the stable unit treatment value assumption, the treatment does not have different versions for different subjects, and the treatment outcome for one person is independent of the treatment outcome for all other people. In other words, there is no interference between units, and the treatment of one subject does not interfere with its surroundings in such a way as to change the treatment for other units. In agriculture, this assumption would be violated if rain carried some of the treated fields' fertilizer into the untreated fields (Guo & Fraser, 2010, pp. 35–36).

The second assumption, strong ignorability of treatment assignment, has two parts. First, to satisfy the assumption, the treatment assignment must be conditionally independent of the response vector, given the set of observed covariates used in the model. Second, for every value of the covariate vector, the probability of being assigned to the treatment group and the probability of being assigned to the control group must both be nonzero (Rosenbaum & Rubin, 1983). The first part of this assumption requires treatment assignment to be *unconfounded*; the second part requires there to be overlap, or *common support*, in the distributions of the covariate values and thus the distribution of the propensity score (Caliendo & Kopeinig, 2008; Stuart & Rubin, 2007).

To be unconfounded, treatment assignment must be "independent of the potential outcomes, given the observed covariates" (Jo & Stuart, 2009, p. 2862). In other words, "conditional on covariates *X*, the assignment of study participants to binary treatment conditions (i.e., treatment vs. nontreatment) is independent of the outcome of nontreatment (Y_0) and the outcome of treatment (Y_1)" (Guo & Fraser, 2010, p. 31). The unconfoundedness assumption would be violated if participants receiving the treatment were more likely to respond to it than similar participants not receiving the treatment. This assumption is similar to the exogeneity assumption in ordinary least squares regression, in which the error term is assumed to be independent from the independent variable (Guo & Fraser, 2010, p. 31).

The second part of the strong ignorability assumption, common support, only holds if participants at every value of the propensity score have some chance of being assigned to treatment and also some chance of being assigned to control. In practice, because we must work with an estimated propensity score rather than the true propensity score, a lack of common support is indicated when there are some areas of the propensity score distribution in which nearly all participants were assigned to the same group. This is generally dealt with by discarding the participants outside of the region of common support (Stuart & Rubin, 2007). This limits the inferences that can be drawn—if a large number of controls are discarded because they are dissimilar from all the treatment participants, we may only able to estimate the average effect on those who are similar to the treatment participants, rather than the overall average treatment effect (Caliendo & Kopeinig, 2008; Stuart, 2010).

Choosing the Matching Variables

Because I was interested in academic outcomes (course letter grade and final exam score), I wanted two groups with similar academic histories. For this purpose, I obtained unofficial college transcripts for all participating students. (During the consent process, the participating students agreed to share their college educational records.) Eleven key academic history variables were chosen for use in matching the students. In the remainder of this section, I will describe my rationale for choosing these particular variables, and provide a detailed description of each. Once the variables were chosen, the necessary data were pulled from the transcripts and entered into a spreadsheet. Though the transcripts also contained data for one outcome variable (course success), these outcomes were not entered until later, after the groups' academic histories were satisfactorily balanced.

How did I decide which academic variables to include? When selecting covariates for propensity score matching, it is best to be generous, including all variables that are expected to be associated with the outcome or treatment assignment. Including a variable that turns out not to be associated with the outcome or treatment assignment does not cause major problems, because such a variable will have only a small influence on the propensity score. However, omitting a variable strongly predictive of the outcome or treatment can result in significant bias (D'Agostino, 1998; Dattalo, 2010; Rosenbaum & Rubin, 1983; Stuart, 2010). So, I began with variables that past research had shown to predict success, and added other variables that I expected to be associated with success.

Past Research on Developmental Mathematics Success Predictors

There has been little solid research on developmental mathematics success predictors. Serna (2011) and Cartnal (1999) found college GPA and enrollment status (part- or full-time) to be the best predictors of developmental mathematics success, with students enrolled full-time succeeding at higher rates than part-time students. Armstrong (2000) found dispositional student variables, including high school GPA, grade in last English or mathematics course, and number of years of English or mathematics taken in high school, were the strongest indicators of developmental mathematics and English grades. Unfortunately, the relative importances of these different dispositional variables were not reported. Placement test scores were not a significant predictor.

In a previous study of developmental mathematics students at the same institution as the current study, Little (2002) found Prerequisite Status and college GPA to be the strongest predictors of success in Introductory Algebra, accounting for 27% of the variance in final course average. These were also the most important predictors in a discriminant analysis predicting membership in the successful or unsuccessful group. The categorical variable Prerequisite Status described whether the student was repeating the course, what letter grade the student earned in the prerequisite course (if applicable), or whether the student placed directly into Introductory Algebra via standardized placement test. Students were classified as *Repeat* if they had previously enrolled in Introductory Algebra but did not earn a passing grade. Students who had not previously enrolled in Introductory Algebra were classified as *A*, *B*, *C*, or *Placement*. A Prerequisite

Status of *A*, *B*, or *C* indicated the student had taken the prerequisite course, Prealgebra, and earned a letter grade of A, B, or C. A Prerequisite Status of *Placement* indicated that the current semester's Introductory Algebra class was the student's first mathematics course at this institution. The statistical analysis showed that students categorized as *A* or *Placement* were more likely to succeed in Introductory Algebra than those categorized as *B*, *C*, or *Repeat*.

Based on these findings from past research, I decided that at a minimum, the matching variables needed to contain information about enrollment status (part- or full-time), GPA, and previous mathematics course grade. These three variables represent, in part, three facets of a college student's academic credentials: time in college, overall achievement, and mathematics-specific achievement. For each facet, additional variables were needed to capture other relevant information.

Before beginning the variable descriptions, a few notes on terminology are in order. In community college discourse, the terms *college-level* and *credit-level* are used interchangeably to indicate that the course is generally transferable to a four-year college. Developmental courses, sometimes known as remedial courses, are not generally transferable. At this particular institution, the transcript uses the terms *credits*, *credit hours*, and *hours* interchangeably, to include both developmental hours and credit-level hours (a bit confusing, admittedly). A course is considered *attempted* if it appears on the student's transcript (even if the student received a W grade, indicating an official withdrawal from the course). In college-level courses, students receive grades of

A, B, C, D, F, or W. Grades of A, B, C, and D are considered passing grades, even though D grades are not transferable to most four-year institutions. In developmental courses, students receive grades of A, B, C, IP, F, or W. In developmental courses, only grades of A, B, and C are considered passing grades. Students receiving grades of IP (In Progress), F (Failing) and W (Withdrawn) did not pass the course and cannot move on to the next course in the sequence. Students receiving IP grades “have participated fully in the class but have not met all criteria for making progress to the next level of courses” (Lone Star College System, 2012, p. 75).

Time in College

I used four different variables to quantify students’ time in college. Instead of using a binary variable to represent students’ part- or full-time status, I used a more detailed variable, HrsAttF2012, equal to the total number of enrolled hours during the intervention semester. CumHrsAttPreInt is the total number of credit hours attempted prior to the intervention semester, including both developmental and college-level classes. CredEarnedPreInt is the number of credit hours (both developmental and college-level) that were successfully completed with a passing grade. YrsSinceStartCollege was obtained by subtracting the year of the student’s first enrollment in this college system from 2012 (the year the intervention occurred).

Though it appeared reasonable that the length of time in college might be related to the probability of success in developmental mathematics, I did not know the nature of that relationship. More experienced students could possibly be stronger, as they had already survived the initial transition to college. On the other hand, spending several years in college without completing developmental mathematics could indicate a problem. If there were a “sweet spot,” an ideal amount of college experience possessed by the most successful students, I did not know what it was. Fortunately, for the propensity score approach, I did not need to know. By including variables representing college experience in the propensity score model, I could force the groups to have similar numbers of students with a lot of college experience, a moderate amount of college experience, and no college experience (assuming the data were sufficient to support such balance).

Overall Achievement

To capture achievement, I began with a traditional success measure, Grade Point Average (GPA); I then considered what important information it omitted and tried to capture that information some other way. This thought process resulted in one additional overall achievement variable, the Course Completion Ratio (described later in this section), and several mathematics-specific achievement variables (described in the next section).

GPA does a reasonably good job summarizing a student’s overall academic success; however, it has some rather serious limitations, especially when applied to students in developmental courses. First, it does not include some courses (e.g., developmental courses and withdrawals). This limitation can be addressed simply by recalculating the GPA to include the omitted courses, if doing so would better address the research purpose. Whether studying credit-level academic history, developmental course history, or both, the researcher can decide whether it is best to count the W grades (and IP grades, for developmental classes) as if they were F grades, or to omit them. Second, because GPA does not consider course volume, students with very different track records can appear equivalent. Third, the GPA does not exist for students who have not yet enrolled in a GPA-eligible course. This limitation applies whether the GPA is calculated for credit-level courses only, or for both developmental and credit-level courses; whether it is calculated for all courses, or just courses in a particular subject area; and whether it is calculated with W grades or without W grades. The course volume and no-GPA limitations might not be serious problems for a researcher studying college seniors, but they certainly should be considered by anyone studying freshmen or community college students. A final caution about GPA applies: some colleges (including the institution in the current study) record a 0.00 GPA for those students who would be more accurately described as "not yet having a GPA." Unless the researcher wants these students to be considered equivalent to students who have failed all their courses, an alternate method of handling the no-GPA students must be devised.

The current study incorporated GPA, but supplemented it with other information to address these limitations. The previously mentioned CumHrsAttPreInt and CredEarnedPreInt serve to capture course volume. The Course Completion Ratio, detailed later in this section, describes overall success level, without omitting withdrawals or developmental courses. The issue of having some students with no GPA was addressed using indicator variables and imputation, as will be explained later.

The pre-intervention college GPA (GPAPreInt) was calculated exactly as the college calculates GPA (excluding withdrawals and developmental courses), with two exceptions. First, I recorded NA (instead of the 0.00 listed on the transcript) for students who had not attempted any GPA-eligible courses. This distinguished the new-to-college or new-to-credit-level student from the student who had failed all attempted GPA-eligible courses. Second, I included those courses that the transcript listed as "exclude from GPA." At this institution, students who repeat a course and earn a higher grade are allowed to exclude the original grade from their GPA, if they submit a grade-exclusion request form to the registrar's office (Lone Star College System, 2012, p. 71). Grade exclusions were rare, listed on only 34 of the 259 transcripts.

The Course Completion Ratio (CCR) measures the proportion of attempted courses that have been successfully passed (Hagedorn & Kress, 2008). To calculate the CCR, I began with the total number of credits earned and divided it by the total number of credits attempted. For students who have never attempted a course, NA was recorded. The highest possible value for CCR is 1. Students with a habit of dropping or failing courses will have a low CCR. Students who have attempted at least one course but have not passed a course will have a CCR of 0. On the transcripts, credits were counted as "earned" if the student received an A, B, C, D, or P (a grade of P indicated a passing grade in a pass/fail course; pass/fail courses were rare).

Mathematics Achievement

Four mathematics-specific variables were used in the matching. The first, YrsSinceMath, was obtained by subtracting the year of the student's most recent mathematics course from 2012 (the year the study occurred). YrsSinceMath was included because a long gap could mean that the student had forgotten previous mathematical knowledge. The second, DevMathGPA, contains information about the letter grades earned in all previous mathematics attempts. The third, PrereqStatus, contains information about the grade earned in the student's most recent mathematics course, if applicable. The fourth, AttemptsPerPass, captures information about mathematics course repetition patterns. To my knowledge, no variable similar to AttemptsPerPass has been used in previous research. The variables DevMathGPA, PrereqStatus, and AttemptsPerPass will be described in the remainder of this section. The YrsSinceMath variable needs no further explanation.

Because the study focused on developmental mathematics success, I calculated a pre-intervention developmental mathematics GPA (DevMathGPA) to capture information about the student's past mathematics achievement. In this variable, I included all attempts at any developmental mathematics course, and considered W, F, and IP grades to be equivalent. Instructor policies vary on the awarding of IP grades (versus F grades) and on the awarding of W grades for non-attendance. Also, due to financial aid reasons, some students choose not to officially withdraw from classes they stop attending. For these reasons, W, F, and IP grades are nearly impossible to disentangle from one another.

In the calculation of DevMathGPA, I counted grades of W (Withdrawn) and IP (In Progress) the same as F grades, awarding them 0 grade points. For the other letter grades, I used the same grade point system this institution uses when calculating the official GPA: 4 points for A grades, 3 points for B grades, and 2 points for C grades. To calculate the DevMathGPA, I divided the total number of grade points by the total number of hours attempted. At this institution, each developmental mathematics course is considered a three-hour course; therefore, the denominator was calculated by multiplying the number of attempts by 3. For example, suppose a student's transcript listed four developmental mathematics course attempts, with grades of W, C, IP, and B. The student's developmental mathematics GPA would be

$$\text{DevMathGPA} = \frac{0 \cdot 3 + 2 \cdot 3 + 0 \cdot 3 + 3 \cdot 3}{4 \cdot 3} = \frac{15}{12} = 1.25\,.$$

For students with no previous developmental mathematics attempts on their transcripts, I recorded NA for the DevMathGPA.

PrereqStatus was based on Little's (2002) research at the same institution, in which it was a strong predictor of success. Little treated Prerequisite Status as a categorical variable with values of *A*, *B*, or *C* (student grade in the prerequisite course), *Placement* (if placed directly into the current course placement test), *Repeat* (if student has previously taken the current course and received a W, F, or IP) or *Other*. *Other* was a very small (12 students out of a total sample of 498) catch-all category that included students who did not fit into any other category. Some *Other* students had passed the class but were repeating it anyway; others enrolled in Introductory Algebra even though their placement test or previous grade should have put them in Prealgebra (S.C. Little, personal communication, March 15, 2013). In the current study, I defined PrereqStatus in the same manner as Little (2002), except without the *Other* category. Because PrereqStatus would be used in the propensity score matching model, I did not want a miscellaneous-style category containing students with dissimilar prerequisite histories. Instead, I individually examined each student who had an unusual prerequisite situation, and chose the category (*A*, *B*, *C*, *Repeat*, or *Placement*) that fit the student best. Typically, these were students who had taken a class out of order, or whose previous unsuccessful attempt occurred several years before the current attempt.

Although the PrereqStatus variable tells us about how the student arrived in the current class, it tells us nothing about what happened earlier in the student's mathematics journey. In the course catalog, the developmental mathematics sequence is straightforward: three sequential courses occurring in a specified order. Someone reading the catalog might expect for students to begin in the first course, proceed neatly through the sequence in three semesters, and then move on to college-level math. In reality, this rarely happens. Students' paths through the sequence vary widely; they do not all begin at the same point in the sequence, and they may remain at one point in the sequence for a long time before moving on to the next.

A new student's first mathematics course is usually chosen by an advisor, based on placement test scores. Some students, those with high placement test scores or strong high school mathematics backgrounds, are placed into credit-level math. However, the majority of community college students are placed into developmental mathematics. Some students begin with Prealgebra, the first course in the three-course developmental mathematics sequence. Other students are placed into the second course, Introductory Algebra, or the third course, Intermediate Algebra. Thus, any given Introductory or Intermediate Algebra class will contain some students who began with the current course and some who began in an earlier course. For this reason, comparisons of students on placement scores are not generally meaningful; the deficits represented by low placement scores should theoretically have been remedied by the previous mathematics course(s). To further complicate comparisons, many students repeat courses because they do not earn a passing grade on the first attempt. The previously described PrereqStatus variable (*A*, *B*, *C*, *Repeat*, or *Placement*) captures information only about the last stop on the student's route to the current class. Before that last stop, the student may have had repeats of the prerequisite courses, or multiple repeats of the current course.

AttemptsPerPass is a numerical variable that captures overall course repetition history. It allows meaningful comparisons between students, even if they entered the three-course developmental mathematics sequence at different levels. It is not course-specific—students can be compared on AttemptsPerPass even if they are enrolled in different developmental mathematics courses. This was important in the current study, which included both Introductory Algebra and Intermediate Algebra students. For ease of discussion, the following explanation of AttemptsPerPass refers to the courses by number instead of name. Math 0306 is Prealgebra, Math 0308 is Introductory Algebra, and Math 0310 is Intermediate Algebra.

AttemptsPerPass is the average number of attempts a student takes to pass each developmental math course (based on past track record). A student receives a 1.00 in one of two ways: (1) being placed directly into this course in the current semester, or (2) being placed lower in the developmental mathematics sequence and passing every course on the first attempt (and never previously attempting the current course). Essentially, AttemptsPerPass represents how close to “on-track” the student is in his or her developmental mathematics sequence. High scores are “bad”; low scores are “good”; the “best” score is 1.00 (perfectly on-track).

As in the calculation for DevMathGPA and CCR, I considered each separate listing of a course on the on the student’s transcript as an attempt. For the course to be listed on the transcript, the student had to be registered for it on the official day of record (approximately two weeks after the start of the semester). An attempt was considered “successful” if the student received a passing grade of A, B, or C. An attempt was considered “unsuccessful” if the student received a grade of F (Failing), IP (In Progress), or W (Withdrawn). (As previously mentioned, this institution does not award D grades in developmental courses.) For example, suppose a student had the following grades:

F in 0306, B in 0306, W in 0308, C in 0308, and is now taking 0310 for the first time. This student has 4 past attempts; 2 of those 4 attempts were successful.

If the student has attempted (and passed) previous math courses, but is attempting the current course for the first time, then the AttemptsPerPass ratio is based on previous courses, using the following formula:

$$\text{AttemptsPerPass} = \frac{\text{Number of Previous Dev. Math Attempts}}{\text{Number of Successful Previous Dev. Math Attempts}}.$$

For our example student, we want the AttemptsPerPass score to be 2.00, because the student has taken two tries to pass each course (so far). Following the formula, we get

$$\begin{aligned}\text{AttemptsPerPass} &= \frac{\text{Number of Previous Dev. Math Attempts}}{\text{Number of Successful Previous Dev. Math Attempts}}\\ &= \frac{4}{2} = 2.00.\end{aligned}$$

This formula must now be adjusted to cover two sets of students: those who are repeating the current course, and those who have never passed a course. (These sets intersect—the intersection contains students who began at the current level but did not pass in their first attempt.) For repeaters, we want to include repetitions of both the current course and of previous courses. For students with no successful attempts, the previously described AttemptsPerPass formula will result in a zero denominator (as alert readers undoubtedly noticed). Fortunately, the same adjustment can cover both situations.

For repeaters of the current course and for students with no past successful attempts, we modify the numerator and denominator to include the current course, by adding 1. (i.e., we assume, for the sake of the formula, that the student will pass the class this time.)

$$\text{AttemptsPerPass} = \frac{\text{Number of Previous Dev. Math Attempts +1}}{\text{Number of Successful Previous Dev. Math Attempts + 1}}.$$

To illustrate that this formula captures what we want, we will consider two more hypothetical students. For ease of comparison, these new example students both have an AttemptsPerPass score of 2.00 (the same score as in the previously mentioned example). First, consider a student with grades of F in 0306, B in 0306, IP in 0308, C in 0308, W in 0310, who is currently enrolled in 0310 again. This student is consistently (so far) taking two tries to pass each course, and we want AttemptsPerPass to reflect that number. The modified formula gives us

$$\begin{aligned}\text{AttemptsPerPass} &= \frac{\text{Number of Previous Dev. Math Attempts +1}}{\text{Number of Successful Previous Dev. Math Attempts + 1}} \\ &= \frac{5+1}{2+1} = \frac{6}{3} = 2.00.\end{aligned}$$

Next, consider a student who arrived at college last semester and was placed (via standardized placement test) into 0310. Suppose the student received a W in 0310 last semester, and is now taking 0310 for the second time. Applying the formula,

$$\text{AttemptsPerPass} = \frac{\text{Number of Previous Dev. Math Attempts +1}}{\text{Number of Successful Previous Dev. Math Attempts + 1}}$$

$$= \frac{1+1}{0+1} = \frac{2}{1} = 2.00.$$

Of course, we have no idea how many tries this student will actually take to pass 0310. The student could pass with flying colors this semester, or could flounder in 0310 for two more years. But an AttemptsPerPass of 2.00 (requiring two attempts in order to pass) is a fair reflection of the student's record so far. With a score of 2.00, all three students mentioned so far are considered equivalent on the AttemptsPerPass variable: although they started at different points in the developmental math sequence, their course-repetition patterns are similar.

So far, we have satisfactorily computed AttemptsPerPass for students whose transcripts include previous attempts at developmental mathematics—either attempts at previous courses, or attempts at the current course, or both. For students who have never before attempted developmental mathematics, we must revisit our original vision for the AttemptsPerPass variable. We wanted it to measure how close to "on-track" the students are, regardless of where they started the sequence or what course they are currently enrolled in. A score of 1.00 means perfectly on-track with no unsuccessful course attempts. Thus, the new-to-mathematics student should also get a 1.00. We could simply assign this 1.00 to all students with no previous developmental math attempts; or, equivalently, we could use the same formula we used for the current-course repeaters (with the +1 in numerator and denominator). For a student with no previous attempts at any course, this gives us

$$\text{AttemptsPerPass} = \frac{\text{Number of Previous Dev. Math Attempts} + 1}{\text{Number of Successful Previous Dev. Math Attempts + 1}}$$

$$= \frac{0+1}{0+1} = 1.00.$$

In summary, the AttemptsPerPass variable is defined as follows. In the spreadsheet, I calculated it using an “if” formula, with the previously described PrereqStatus as the “if” indicator.

$$\text{AttemptsPerPass} =$$

$$\begin{cases} \dfrac{\text{No. of Prev. Dev. Math Attempts}}{\text{No. Successful Prev. Dev. Math Attempts}} & \text{if PrereqStatus} = A, B, \text{ or } C \\ \dfrac{\text{No. of Prev. Dev. Math Attempts} + 1}{\text{No. Successful Prev. Dev. Math Attempts + 1}} & \text{if PrereqStatus} = \textit{Repeat} \text{ or } \textit{Placement}. \end{cases}$$

For a few students with unusual academic histories, a decision was made to redefine the value of a variable to better represent the student’s track record. Typically, these were students who had taken a class either out of sequence or very long ago. For example, one student had received an IP in 0306, an F in 308, a B in 306, and a B in 308, in that order. In the AttemptsPerPass ratio, I ignored the out-of-order 308 F, because the student had not passed 0306 and should not have been allowed to register for 0308. Another student received a W in 0308 in 2006, then started over in 2011, earning a B in 0306. I ignored the old 0308 attempt and classified the student’s Prerequisite Status as *B* instead of *Repeat*. Such decisions were carefully considered and rare, affecting only 10 of the 258 students.

ESOL Anomaly

While combing through the transcripts, I noticed that ESOL (English for Speakers of Other Languages) students seemed to have a very different academic history than non-ESOL students. Often they did not take a mathematics course until their second or third year of college, after all their ESOL classes were complete. Unlike non-ESOL students, the long gap without a mathematics course did not seem to damage their success. Therefore, I included ESOL as a variable, recording 1 for students who had at least one ESOL class on their transcripts and 0 for those students who did not. When I eventually ran the propensity score matching package, removing the ESOL variable from the model caused the balance on the other variables to worsen noticeably. For this reason, I kept the ESOL variable, even though it didn't fit into the existing variable categories (time in college, overall achievement, and mathematics achievement).

Adjustments to Variables for the Propensity Score Model

As previously mentioned, I recorded NA whenever the variables DevMathGPA, GPAPreInt, and CCR resulted in a zero denominator. Essentially, these variables combined an interval or ratio scale (for some participants) with a single categorical value (for other participants). In order to proceed with the propensity score matching, I needed to transform these variables into a form that could be interpreted by a statistical computer package.

Using the “recode into different variables” function in IBM SPSS Statistics 21, I created indicator variables for DevMathGPA, GPAPreInt, and CourseCompletionRatio. In each indicator variable, a 1 indicated that the student had a number for the ratio (nonzero denominator). A 0 indicated that the student had NA on the original variable (no applicable hours attempted, so a zero denominator). Once the indicator variable was populated with the correct values, the original NAs were recoded as “system-missing.” This allowed SPSS to calculate a mean for the original variable (incorporating only the students who actually had a GPA, a DevMathGPA, or a CCR). Then, this mean value was imputed to the “system-missing” cases of the original variable. Essentially, for those students who did not yet have a track record, I assumed they were average. I used the indicator variables to preserve the information that they did not yet have a track record.

PrereqStatus was handled in a similar manner. However, because it was a categorical variable, I first had to convert it to numerical values. Starting with the categorical PrereqStatus variable (*A*, *B*, *C*, *Repeat*, or *Placement*) developed by Little (2002), I recoded it into a new ratio variable, PrereqStatusGradePts, using the previously mentioned system of grade points (4.00 for *A*, 3.00 for *B*, 2.00 for *C*, and 0.00 for *Repeat*). Students with a PrereqStatus value of *Repeat* had taken the current class before and received a W, IP, or F. Because I had decided to consider all these grades as equivalent on both the outcome variable and on the DevMathGPA variable, I did so here as well, recoding them as 0.00. Because the grade point scale was not applicable to the

Placement students, I recorded *Placement* as NA. I then computed the mean for the non-NA students, and imputed this mean value to the NA (*Placement*) students. By using this process, I invoked the assumption that *Placement* students are average. However, based on past research, *Placement* students are probably better than average, because they are more successful than other students in developmental mathematics (Donovan & Wheland, 2008; Little, 2002). For this reason, I also tried an alternative imputation scheme for the *Placement* students, in which I imputed a value one grade point higher than the mean. This did not seem to change the results. In fact, the propensity score matching algorithm gave good results even when I imputed 99 to the *Placement* students. Although 99 was meaningless on the grade point scale (0 to 4 points), the matching algorithm could use the 99 to recognize the *Placement* students and balance them across the two groups.

Table 32 lists the variables upon which the matching process was based. Some of these variables were used as matching variables with no modifications. Others required modifications before they could be used (e.g., applying a numerical scale, imputing values, creating indicator variables, or truncating the range). Table 32 describes the variables along with the modifications.

Table 32

Variables Used to Evaluate (and Adjust) the Sample

Variable	Description	Adjustments made before entering into propensity score model (if any)
HrsAttF2012	Hours attempted during the intervention semester (includes credit and developmental). Distinguishes part-time students from full-time students.	
CumHrsAttPreInt	Cumulative hours attempted before the intervention semester (includes all classes—credit-level, developmental, and grade-excluded). Captures volume of college experience.	
DevMathGPA	Cumulative developmental math GPA pre-intervention. W (Withdrawal) and IP (In Progress) grades were counted the same as F grades (0 grade points/3 hours). Students with no previous developmental mathematics classes were listed as NA.	Replaced by two new variables, an indicator variable and a mean-imputed version of DevMathGPA. The indicator variable DMathGPAIndicator had value 1 if transcript showed a previous attempt at a developmental mathematics class, value 0 if not. Computed the mean DevMathGPA for the non-NA students, then imputed this value to the students with NA.
GPAPreInt	Cumulative credit-level GPA prior to the intervention, with grade-excluded classes restored (does not include classes with W grades, pass/fail classes, developmental classes, ESOL classes, or other non-transferable classes.) Students with no GPA-eligible classes were listed as NA.	Replaced by two new variables, an indicator variable and a mean-imputed version of GPAPreInt. The indicator variable GPAIndicator had value 1 if transcript contained a GPA-eligible class, value 0 if not. Computed the mean GPAPreInt for the non-NA students, then imputed this value to the students with NA.

Table 32 Continued

Variable	Description	Adjustments made before entering into propensity score model (if any)
CredEarnedPreInt	Total credit hours earned (grades of A, B, C, D, or P) before the intervention semester. Includes all types of classes—developmental, credit-level, ESOL, and classes taken Pass/Fail. (P indicates “pass” in a pass/fail course.) Does NOT include hours earned for grade-excluded classes, even if those hours were passed with a C or D. Including grade-excluded classes would have meant "double-dipping" for some students (crediting them twice for the same course; for example, if they got a D the first time and then an A.)	
CourseCompletionRatio	The proportion of hours attempted that were passed, prior to the intervention. Includes all classes (credit, developmental, non-transferable, pass/fail, and grade-excluded). For letter-graded classes, a grade of D or higher was considered passing. Students with no prior classes were listed as NA. (The grade-excluded classes were included in both the numerator and denominator, eliminating the double-dipping problem.)	Replaced by two new variables, an indicator variable and a mean-imputed version of CCR. The indicator variable CCRIndicator had value 1 if transcript contained at least one class prior to the intervention semester, value 0 if not. Computed the mean CCR for the non-NA students, then imputed this value to the students with NA.

Table 32 Continued

Variable	Description	Adjustments made before entering into propensity score model (if any)
AttemptsPerPass	Represents the number of attempts a student takes, on average, to pass a developmental mathematics course. Serves as a measure of how "on-track" the student is in his or her developmental mathematics sequence. Low numbers are "better"; a student with a score of 1.00 is considered "perfectly on-track" and has never repeated a course.	
YrsSinceStartCollege	Years elapsed since student's first enrollment in this college system. (0 for students whose first enrollment was sometime in 2012 (fall, spring, or summer), 1 for students whose first enrollment was in 2011, etc.)	Students with YrsSinceStartCollege of 8 or greater were all recoded as 8 (i.e., all students starting college 8 or more years ago were considered equivalent).
ESOL	1 if ESOL/ESL appears on transcript; 0 if ESOL/ESL does not appear on transcript. (ESOL classes were previous called ESL classes by the college.)	

Table 32 Continued

Variable	Description	Adjustments made before entering into propensity score model (if any)
PrereqStatus	Had value *A*, *B*, or *C* if student passed the prerequisite course to the current course, *Repeat* if student had previously attempted the current course and earned an IP, F, or W, and *Placement* if the current course was the student's first math course at this college.	Replaced by two new variables, PrereqStatusGradePts and PrereqStatusIndicator. PreqStatusIndicator had value 0 if original PrereqStatus was *Placement*, 1 if original PrereqStatus was *A*, *B*, *C*, or *Repeat*. PrereqStatusGradePts had value of 4 for PreqStatus value *A*, 3 for *B*, 2 for *C*, 0 for *Repeat*. The mean PrereqStatusGradePts was calculated for the *A*, *B*, *C*, *Repeat* students. This mean was then imputed to the *Placement* students (as the new PrereqStatusGradePts value).
YrsSinceMath	Years since last previous math class, regardless of whether it was successful (either this course or a different course). Value was 0 for students whose first math class was in 2012, either in the intervention semester (fall 2012) or earlier in 2012. Value was 1 for students whose last math class in 2011, value was 2 if last math class was in 2012, etc.	Created a new variable, YrsSinceMathTruncated, in which I recoded all students with a YrsSinceMath of greater than 5 as 5. (i.e., all students with more than a 5-year math gap were considered equivalent.) Note: The PrereqStatusIndicator variable captured the new-to-math status of students for whom the intervention semester was their first math class.
CurrentCourse	308 for Introductory Algebra, 310 for Intermediate Algebra. Was included as a variable so the two groups would have similar proportions of students in the two courses.	

The Sample

In my empirical study (see Chapter III), the original sample had 117 students in the treatment group and 140 students in the control group. The propensity score matching process trimmed the original sample to a matched sample with 105 students in each group. Due to shortcomings in how the intervention was implemented, I chose to replace the original sample by a modified sample of 60 treatment students and 77 control students; propensity score matching trimmed this modified sample to a matched sample with 54 students in each group. The conclusions of the empirical study (see Chapter III) were based on the treatment effect analysis for this matched sample of 54 treatment students and 54 control students.

However, the forthcoming discussion of propensity score matching will be based on the original sample of 117 treatment students and 140 control students (trimmed to a matched sample of 105 treatment students and 105 control students). The larger sample was more suitable for propensity matching and was therefore a better illustration of the matching process. Because the change in sample was driven by problems with the intervention rather than problems with the propensity score matching, sticking with the original sample is consistent with the purpose of this chapter, which is to show how propensity score matching can add value to community college research.

Adjusting the Sample Using Propensity Scores

The propensity score matching was conducted in SPSS 21, using the R-based propensity score plug-in, which uses two R packages: Essentials of R for SPSS and

Match-It (Hansen & Bowers, 2008; Hansen, 2004; Ho et al., 2007, 2011; Thoemmes, 2012). As of this writing, the only matching algorithm available in SPSS uses logistic regression to estimate the propensity scores, followed by nearest-neighbor matching. To prevent poor matches, the user has the option of choosing a caliper (to prevent units from being matched if their propensity scores are too far apart) or discarding units outside the area of common support (where the two groups have insufficient overlap in their propensity scores). The user can also choose whether to match one-to-one or one-to-*k* (with specified *k*), and whether to match with or without replacement. Choosing the matching one-to-many or matching with replacement options will require the researcher to incorporate weights when estimating the treatment effect (Dehejia & Wahba, 2002; Reynolds & DesJardins, 2009; Stuart, 2010). Thoemmes (2012) provides an expert overview of the SPSS propensity score plug-in, its options, and its output. More robust matching algorithms, such as full matching and optimal matching, are available in R (Ho et al., 2011).

If the groups are reasonably similar to begin with, as in the current study, one-to-one nearest neighbor matching without replacement will probably suffice, and offers the advantage of simplicity. I chose one-to-one matching without replacement, using a caliper of 0.2 to prevent extremely poor matches (measured in standard deviations of the logit—logarithm of the odds ratio—of the propensity score; Thoemmes, 2012). Nearest-neighbor matching without replacement is order-dependent; units at the top of the list are matched first. As soon as a unit is matched, that unit and its mate both become unavailable. Therefore, it is essential to randomize the order of the participants first (Caliendo & Kopeinig, 2008; Reynolds & DesJardins, 2009).

The success of a matching algorithm is based on how well it balances the covariates. Trying several different matching procedures is not only allowed, but encouraged (Ho et al., 2007; Stuart & Rubin, 2007). The process is analogous to that used by experimenters who randomize their cases several times, rejecting randomizations resulting in groups with unacceptable disparity on a critical variable (Bruhn & McKenzie, 2009; Ho et al., 2007; Rubin, 2008a). I ran the propensity score matching algorithm several times while I fine-tuned the adjustments to the variables. Once the set of matching variables was finalized, I repeated the propensity score matching for several different random orderings of the participants. For the final adjusted sample, I chose the matched set with the best overall balance on the covariates.

Evaluating the Balance of the Groups

The original sample had 140 students in the control group and 117 students in the treatment group. In the matched sample, each group had 105 students.

The propensity score is a one-dimensional summary of information from the other covariates. Therefore, examining the propensity score distributions before and after matching can give an idea of how the matching process, and the strategic discarding of units, affected overall balance. However, it is important to keep in mind that balancing the propensity score is a means to an end, not an end in itself. Because the units without a good match on the propensity score were discarded, the propensity score is almost certain to be much better balanced after matching than before. It is still possible for some of the covariates to be badly out of balance, or to have worse balance after matching than before. It is essential to also examine the balance of each individual covariate.

Because the propensity score represents the conditional probability a participant is assigned to treatment based on that person's covariate values, the propensity scores of treatment participants will generally be higher than the propensity scores of control participants. This expectation is reflected in Figure 8. Each dot represents the propensity score of one student. At the upper end of the treatment group, several outliers with high propensity scores were unmatched due to the absence of control students with similar propensity scores. At the low end of the distribution, there were more control students than treatment students. For this reason, some controls were unmatched even though their propensity scores were not extreme.

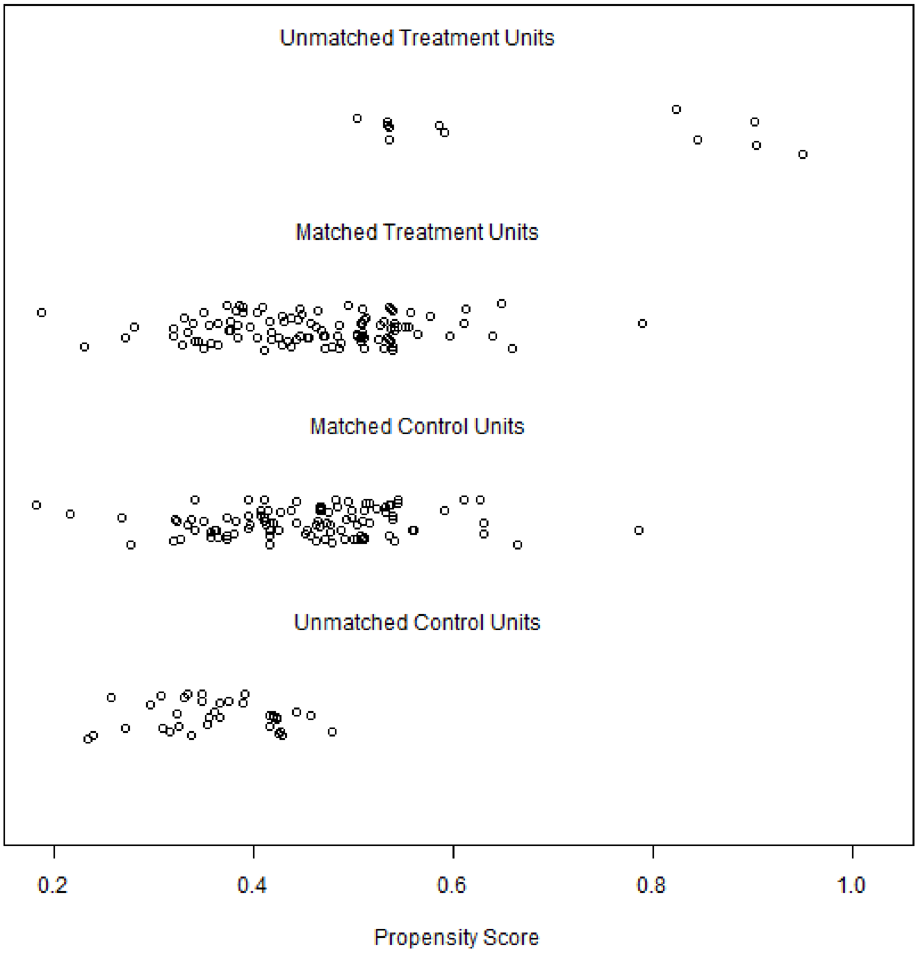


Figure 8. *Jitter plot of propensity score distribution for matched and unmatched treatment and control units.*

Figure 9 shows how the matching algorithm trimmed the control group in such a way that its propensity score histogram more closely matches the histogram of the treatment group. The blocky histogram of a small sample can make it difficult to compare shapes, especially if the histograms being compared have different interval widths. For visual comparisons, it is more helpful to examine the shape of the kernel density overlay—a nonparametric estimate of the probability density function, which uses a smooth curve to represent the information contained in the histogram (Hazelton, 2005). In Figure 9, we can see that the kernel density curves of the matched treatment and control groups have similar shapes. In the unmatched sample, they do not; the control group more closely resembles a normal distribution than does the treatment group.

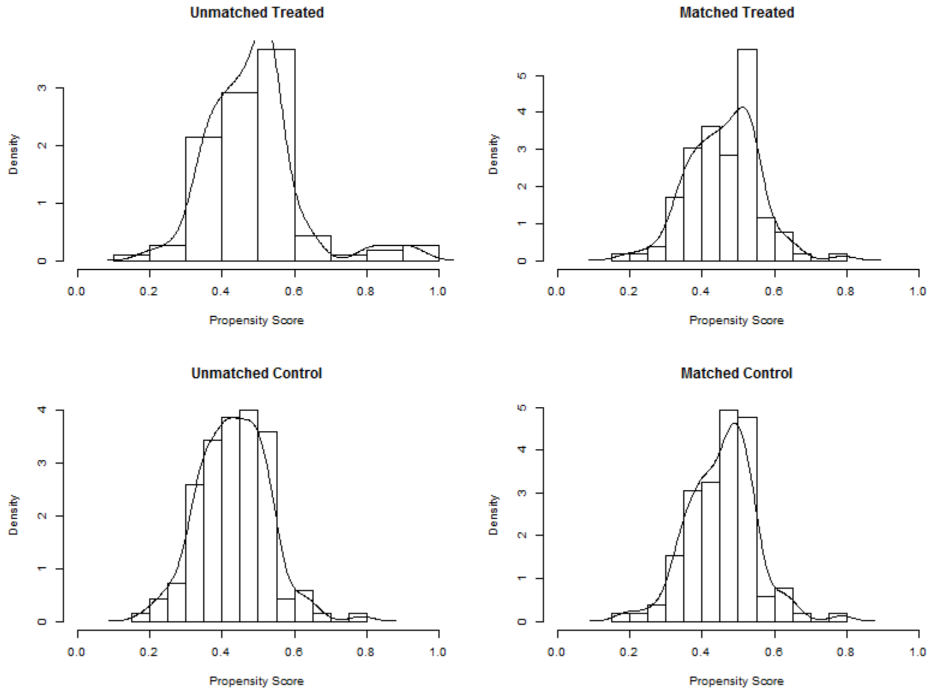


Figure 9. *Propensity score histograms for control and treatment groups before and after matching.*

The purpose of matching, in this context, is to create two groups with similar distributions of each covariate. Thus, the first step in evaluating the success of a matching algorithm is to compare the means of the control and treatment groups on the covariates. For each covariate, the difference in means can be standardized by dividing it by the standard deviation for that variable. To facilitate comparison of the before- and after-matching standardized differences, both calculations should use the standard deviation from the unmatched sample; by keeping the denominator the same, we can see whether the matching process improved the balance (Stuart, 2008).

The SPSS propensity scores package automatically generates a table of standardized differences, along with several graphical summaries of those standardized differences. For both the matched and unmatched samples, the difference between the means is divided by the standard deviation of the treatment group in the unmatched sample[2] (Stuart, 2010).

For the sample to be acceptably balanced, the standardized differences should be below 0.25 (Stuart & Rubin, 2007), but smaller standardized differences are better. If the covariate happens to be one that has a large effect on the outcome, then even a standardized difference much smaller than 0.25 could influence the results (Ho et al.,

[2] As of May 2013, the SPSS propensity scores package uses the standard deviation of the treatment group in the unmatched sample when calculating standardized differences. However, due to a coding glitch, the standard deviations of the control group are listed in the output. This discrepancy should be fixed in future versions (F. Thoemmes, personal communication, June 7, 2013). Note also that Thoemmes (2012), which serves as documentation for the program, mentions the use of the control group standard deviation in the calculation (though the program actually uses the treatment group standard deviation).

2007). The goal of matching is for the two groups to be as similar as possible on the covariates, and thus the standardized differences should be as small as possible.

Table 33 lists the means and standardized differences for the covariates before and after matching, using the standard deviation of the treatment group in the unmatched sample as the denominator. As previously described, the sample for this study was obtained by pairing similar classes and then randomly assigning one class to treatment and one to control. As a result, the groups were fairly similar initially. Except for the propensity score, which is a function of the covariates rather than a true covariate, all the standardized differences were below 0.25 even in the unmatched sample. After matching, the balance improved considerably, and all standardized differences were below 0.05.

Table 33

Means and Standardized Differences of Covariates Before and After Matching

	Unmatched Sample			Matched Sample		
Variable	Means Treated	Means Control	Std. Mean Diff.	Means Treated	Means Control	Std. Mean Diff.
Propensity Score	.483	.432	.405	.460	.455	.039
HrsAttF2012	9.650	9.586	.020	9.552	9.438	.036
CumHrsAttPreInt	28.521	28.236	.012	27.924	27.581	.014
CredEarnedPreInt	19.932	18.214	.096	19.257	18.943	.018
YrsSinceStartCollege	1.949	1.843	.047	1.933	1.905	.013
GPAPreint	2.286	2.153	.173	2.259	2.259	-.001
GPAIndicator	.752	.779	-.061	.771	.781	-.022
CourseCompletionRatio	.713	.663	.238	.707	.705	.011

Table 33 Continued

	Unmatched Sample			Matched Sample		
Variable	Means Treated	Means Control	Std. Mean Diff.	Means Treated	Means Control	Std. Mean Diff.
CCRIndicator	.855	.900	-.128	.876	.867	.027
DevMathGPA	1.751	1.621	.128	1.707	1.690	.016
DMathGPAIndicator	.769	.843	-.174	.810	.810	.000
AttemptsPerPass	1.501	1.570	-.113	1.530	1.513	.029
PrereqStatusGradePts	1.529	1.432	.068	1.462	1.529	-.047
PrereqStatusIndicator	.769	.843	-.174	.810	.810	.000
YrsSinceMathTruncated	.359	.314	.046	.371	.343	.029
CurrentCourse	309.111	309.143	-.032	309.143	309.181	-.038
ESOL	.051	.007	.199	.010	.010	.000

Some researchers (Austin, 2008) suggest dividing by the pooled standard deviation (or an estimate of it), instead of the standard deviation of the treatment group. For that reason, once the final matched sample was chosen, I manually calculated the standardized differences using the pooled standard deviation of the unmatched sample, and compared them to the standardized differences calculated using the standard deviation of the treatment group in the unmatched sample. This made very little difference—all the standardized differences remained very small and all the covariates were still acceptably balanced. On a few covariates, the standardized difference decreased slightly (indicating the balance improved slightly). On all the others, the standardized difference increased by less than 0.005.

Before matching, it appears that the treatment group was slightly stronger academically than the control group. The treatment group had a higher mean on DevMathGPA, GPAPreInt, CourseCompletionRatio, and PrereqStatusGradePts, all variables for which higher numbers are "better." On AttemptsPerPass, smaller numbers are "better," and the treatment group had a lower mean. The treatment group also had a higher percentage of students with a PrereqStatusIndicator of 0, indicating the students were placed directly into the current class without a previous enrollment in developmental mathematics. Because past research has shown that placement students are more successful (Donovan & Wheland, 2008; Little, 2002), this is another indicator that the treatment group was stronger than the control group before matching. After matching, these differences were not only much smaller, but also occurred in both directions, rather than all being in favor of one group.

Though tables are helpful for examining means and differences of individual variables, large arrays of numbers can be overwhelming and difficult to compare with one another. Fortunately, several graphical displays are available to visually depict how the matching process changes the overall balance of the covariates. These graphical summaries were the primary means for determining which of several trials (using different random orderings of the participants) resulted in the best overall balance. The matched sample with the best balance is depicted here.

The dot plot of standardized differences, Figure 10, makes it easy to see which variables improved after matching, which variables got worse, and the direction of the change. The white and black dots represent the standardized differences in the unmatched and matched samples, respectively. Ideally, all the black dots would be very near the center line, or at least they would be closer than the white dots. In this case, the matching improved the balance for nearly all the variables. On the variables that did not improve, the standardized differences were already very small.

In Figure 11, each dot on the left represents the before-matching standardized difference for one variable. A line connects each before-matching dot on the left to the corresponding after-matching dot on the right. If the standardized difference is smaller after matching, the slope of the line will be negative. These negative-slope lines are shown in gray. Positive-slope lines are highlighted in black, and indicate the standardized difference was greater after matching than before. Ideally, all lines would be gray and slope downward from left to right. If standardized differences are very small before matching, it is not unusual for a few of these standardized differences to show a small increase (Stuart, 2010). If the black positive-slope line is nearly flat, then that particular standardized difference increased only slightly. In our example, there was a slight increase in standardized difference for three covariates: HrsAttF2012, CumHrsAttPreInt, and CurrentCourse. The after-matching standardized differences were still very small. Compared to other trials, this trial had fewer positive-slope lines, and the slopes of those lines were smaller.

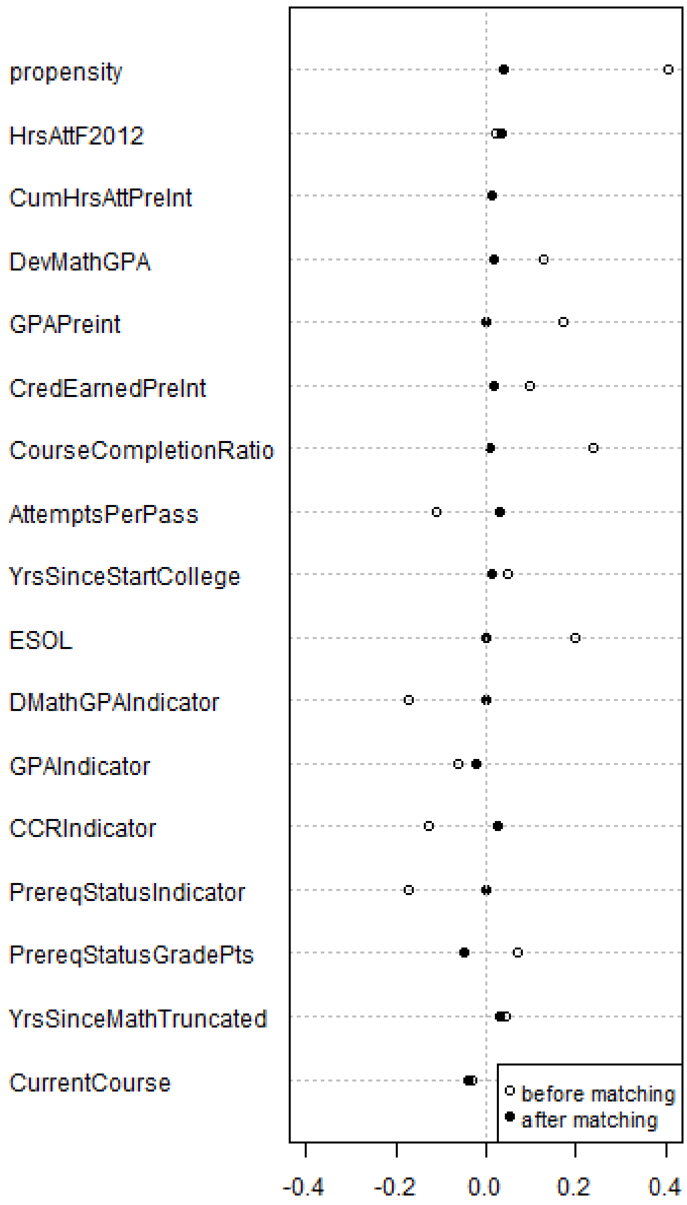


Figure 10. *Dot plot of standardized differences on covariates before and after matching.*

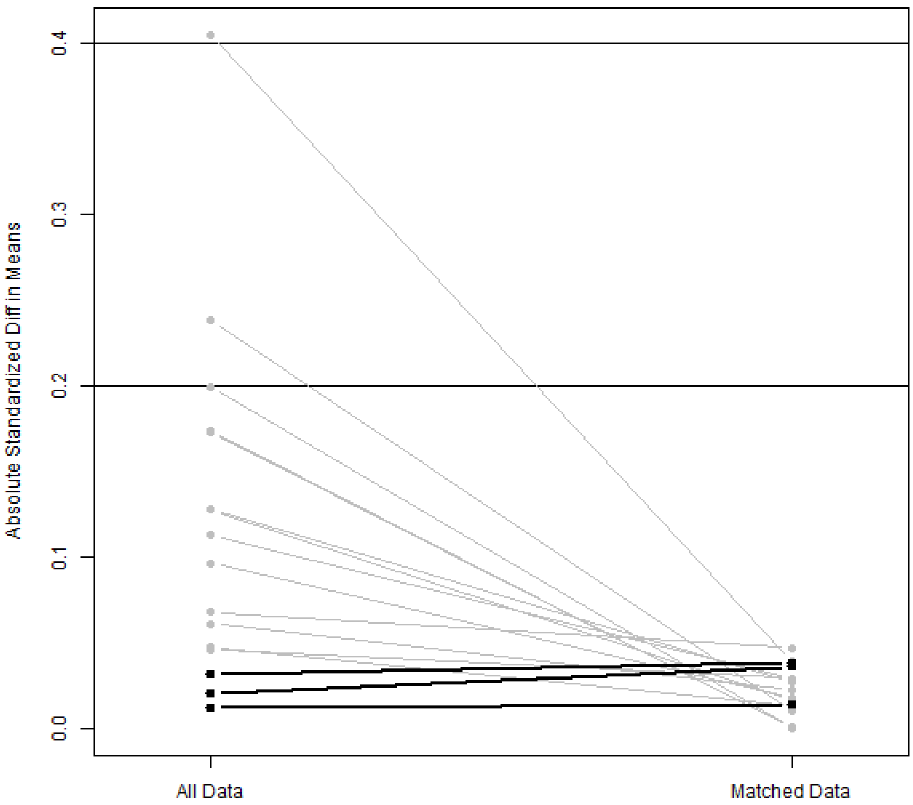


Figure 11. *Line plot of standardized differences before and after matching (without interactions).*

Figure 12 is the same type of figure as Figure 11, a line plot of standardized differences. However, Figure 12 also includes standardized differences for two-way interactions. Again, there were a few positive-slope lines, but none of these resulted in large after-matching standardized differences.

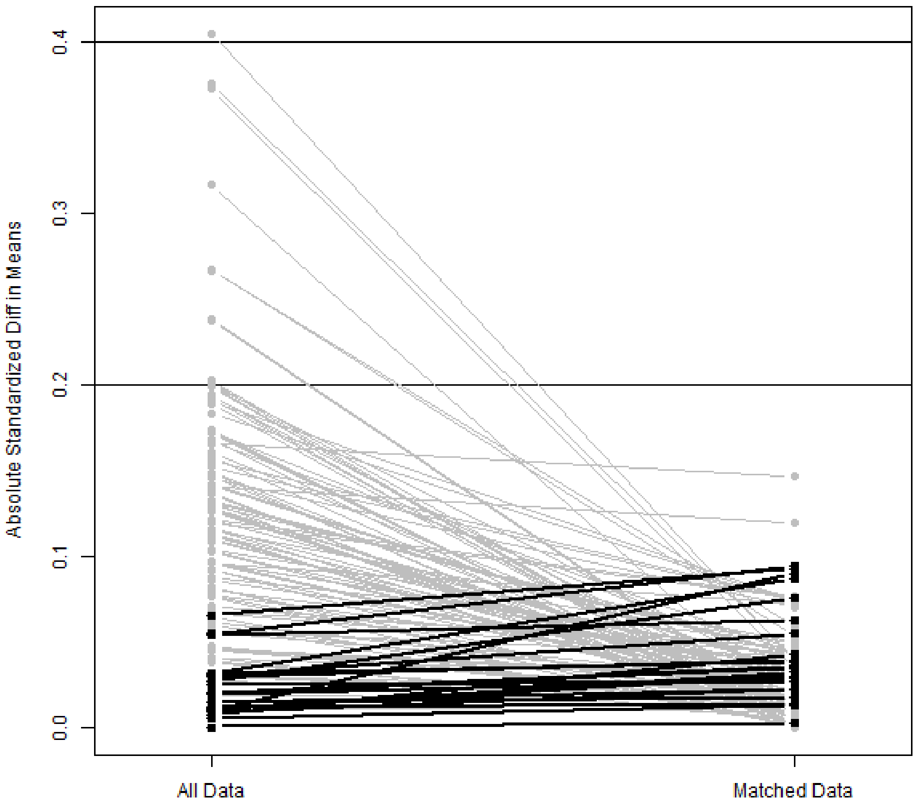


Figure 12. *Line plot of standardized differences before and after matching (with interactions).*

The collection of standardized differences can also be summarized in a histogram, as shown in Figure 13. The density on the *y*-axis is simply a rescaling of the frequencies, so that the area under the histogram is always 1. For a matching that results in good balance, the after-matching histogram will be squeezed in very tightly around 0.

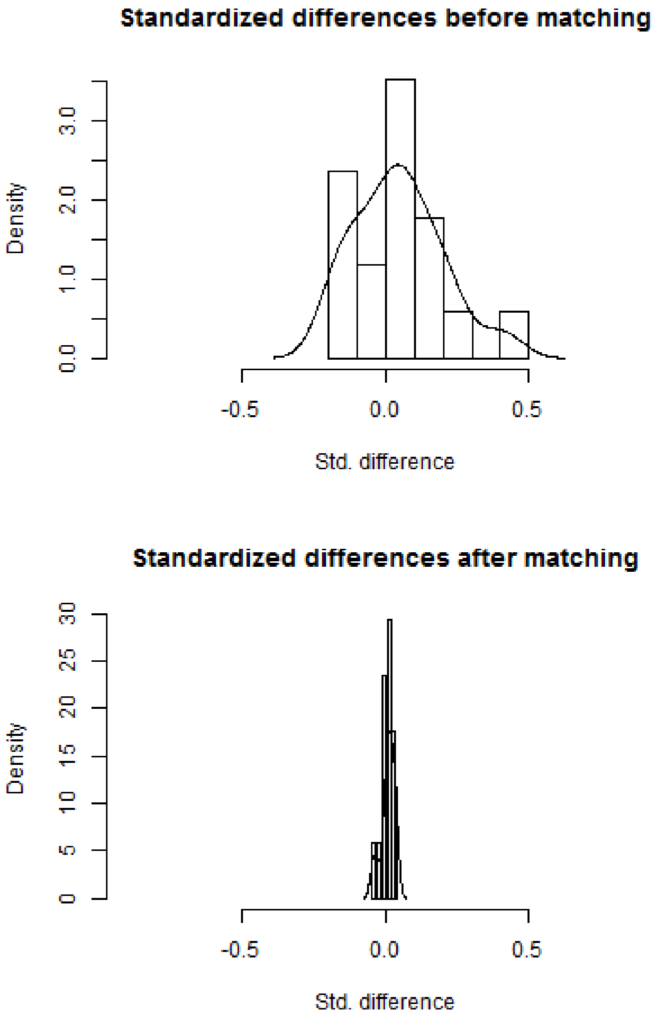


Figure 13. *Histograms of standardized differences before and after matching (without interactions).*

If the graphical summaries and standardized differences indicate the matching process has produced well-balanced groups, the next step is to individually examine the balance of each covariate. Even if overall balance is good, it is possible for one or more covariates to be severely out of balance, or to have gotten worse after matching. If an unbalanced covariate happens to be strongly associated with the outcome, the estimation of the treatment effect can still be very biased, even after matching (Ho et al., 2007).

If balance on a covariate is good, the two group means on that covariate will be nearly equal and the shapes of the control and treatment groups' distributions will be similar. There is not a definitive test for deciding whether the two means are sufficiently close. Though widely used, the *t* test is not appropriate for assessing balance on covariates (Ho et al., 2007; Stuart, 2010). Covariate balance is a characteristic of the sample, not the population; the purpose of the *t* test is to make inferences about two populations. The objective of matching is not to bring the difference in means below some threshold, or to show that the groups are no more different than would have been expected if treatment had been randomly assigned. Instead, the objective of matching is to produce two groups that are as similar as possible on the important covariates.

The quantile-quantile (QQ) plot is a recommended tool for comparing two univariate distributions (Ho et al., 2007). For a given quantile, the *x*-coordinate represents the value of that quantile in the control group, and the *y*-coordinate represents that quantile in the treatment group. If the two distributions are identical, all the points will lie on the line $y = x$, which runs diagonally through the square plot. For example, suppose the top 10% of the values in the control group lie above 8, and the top 10% of the values in the treatment group lie above 9. Then the ordered pair (8, 9) would be plotted on the QQ plot. This point would lie above the line $y = x$. If all the quantile-quantile points lie above the line, the treatment group has higher scores on that variable throughout the distribution.

Figure 14 shows the QQ plots of the propensity score before and after matching. After matching, the points are very close to the line, indicating excellent balance on this variable. QQ plots for the other variables were not expected to show as dramatic an improvement as the propensity score, because the cases were not directly matched on those variables. Still, the other variables generally moved closer to the line after matching, or at least did not move away. QQ plots for other covariates are shown in Figure 15–Figure 21.

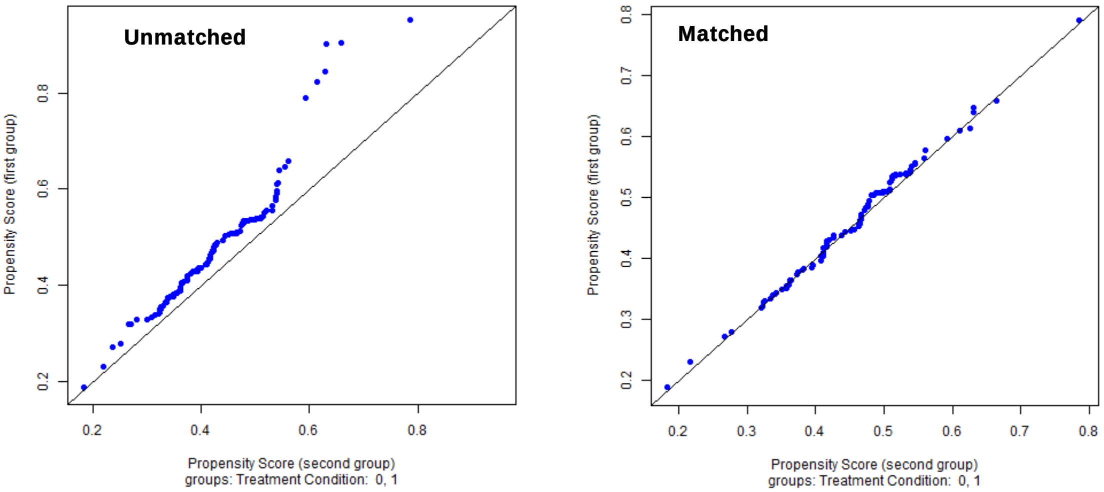


Figure 14. *QQ plots of propensity score distributions for matched and unmatched samples.*

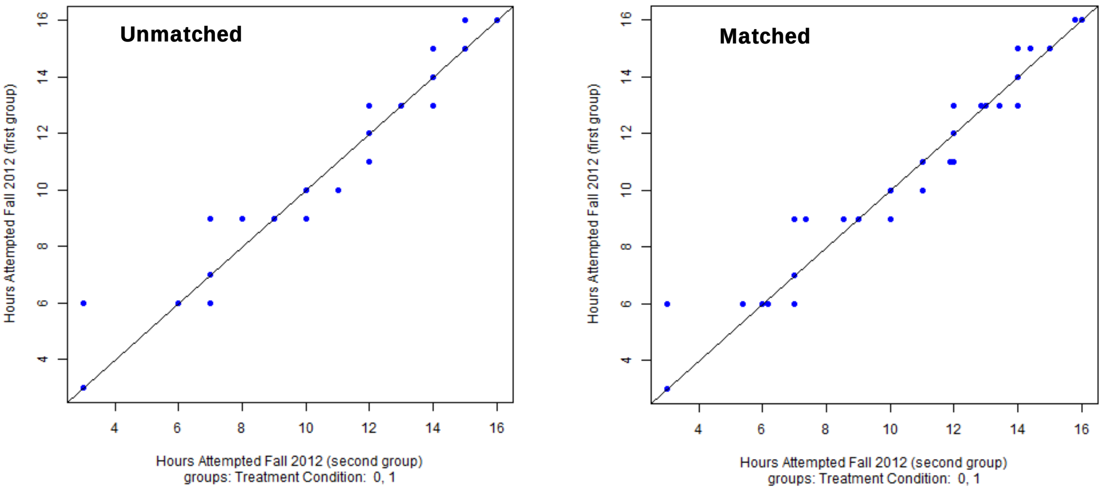


Figure 15. *QQ plots of HrsAttF2012 for matched and unmatched samples.*

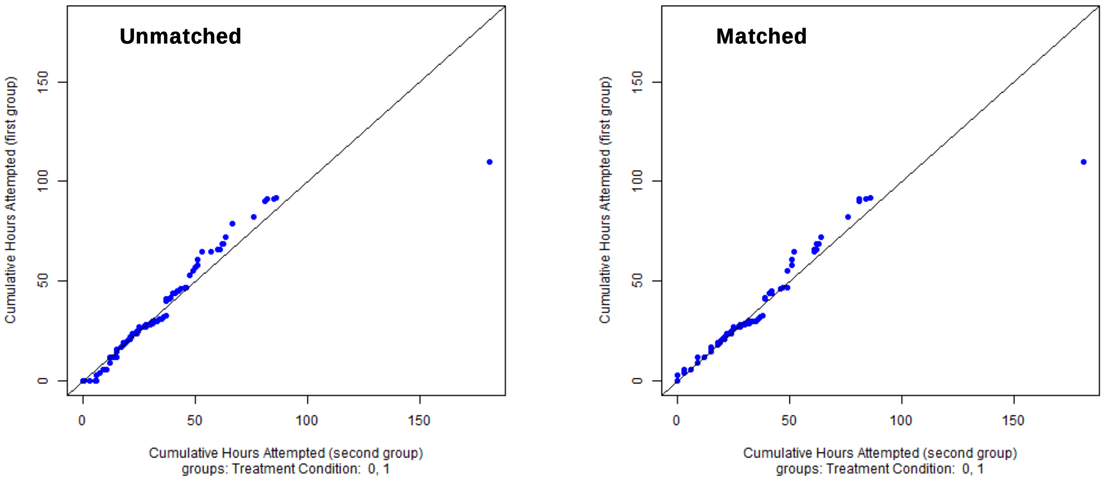


Figure 16. *QQ plots of CumHrsAttPreInt matched and unmatched samples.*

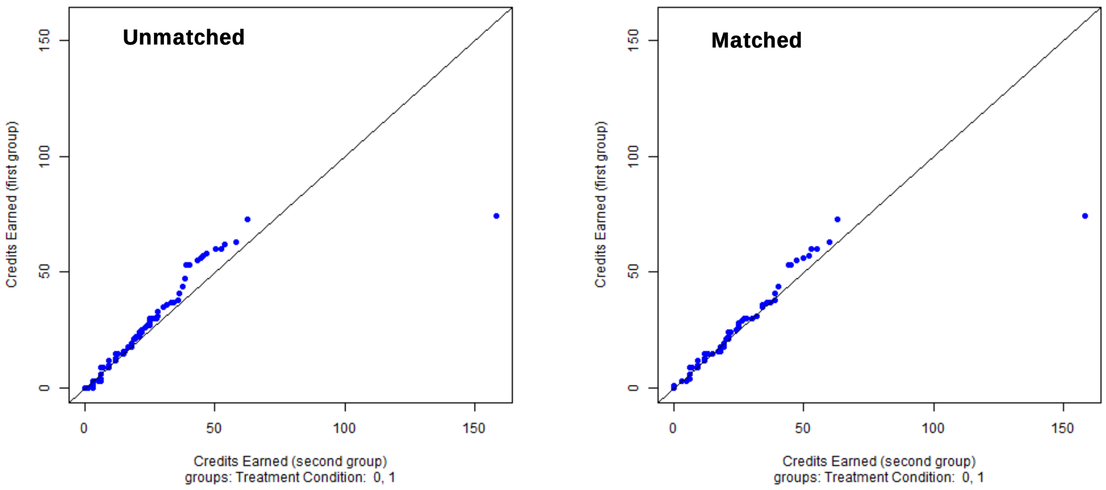


Figure 17. *QQ plots of CredEarnedPreInt for matched and unmatched samples.*

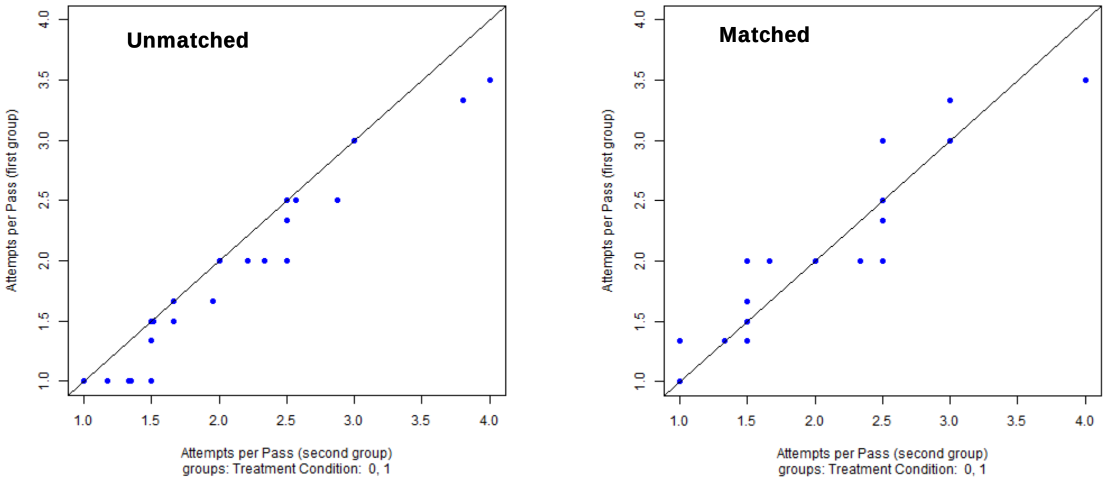


Figure 18. *QQ plots of AttemptsPerPass for matched and unmatched samples.*

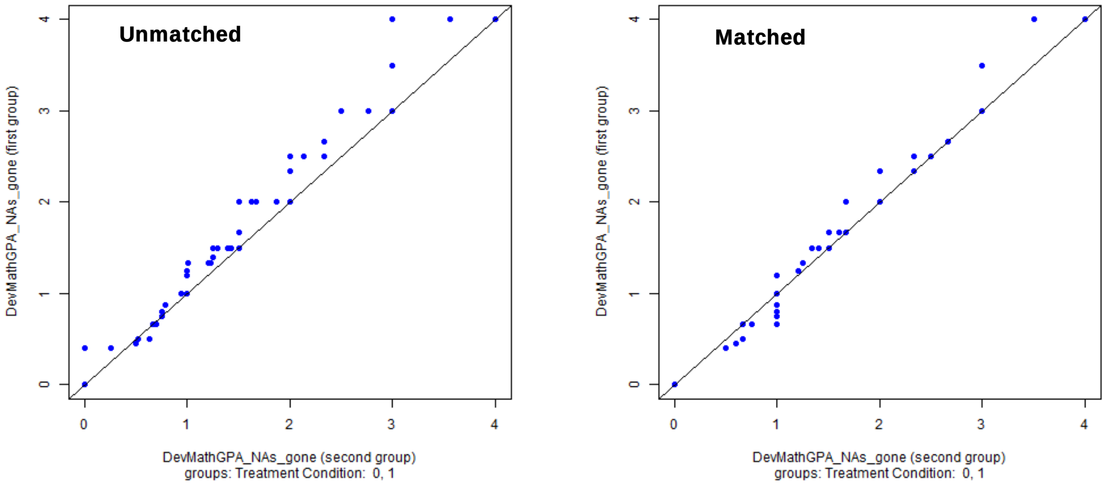


Figure 19. *QQ plots of pre-intervention DevMathGPA for matched and unmatched samples (students with no prior enrollments in developmental mathematics were omitted).*

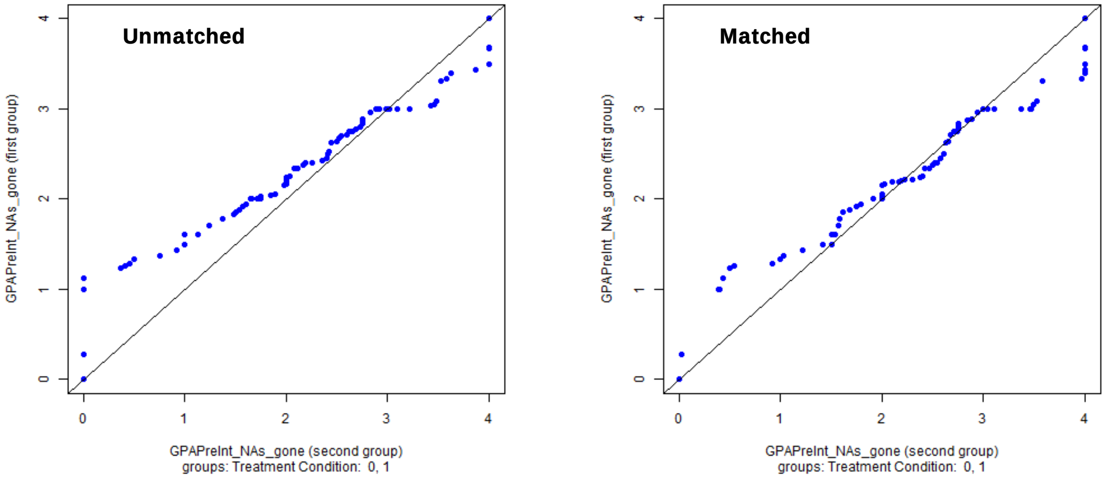


Figure 20. *QQ plots of pre-intervention GPA for matched and unmatched samples (students with no prior GPA-eligible hours were omitted).*

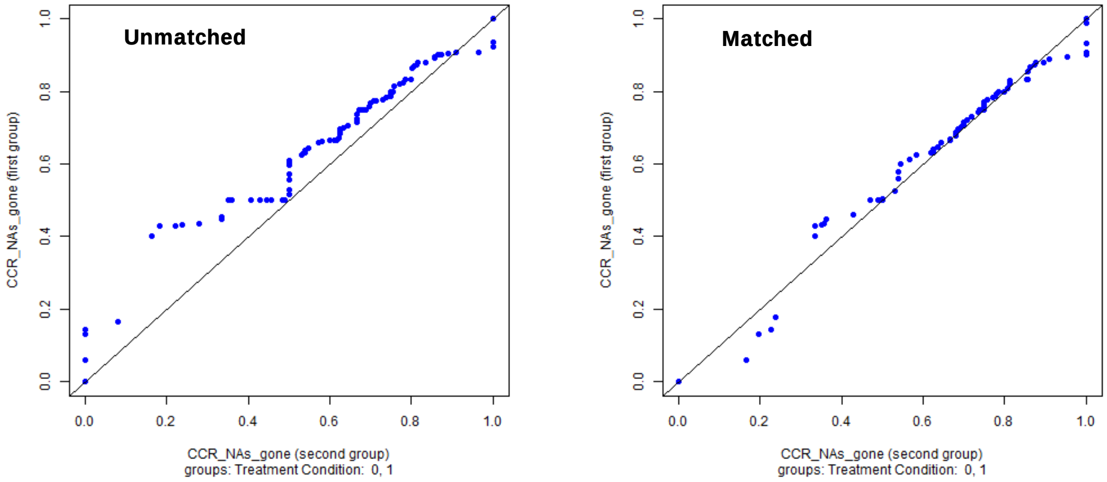


Figure 21. *QQ plots of CourseCompletionRatio for matched and unmatched samples (students with no prior attempted hours were omitted).*

Boxplots or histograms can also be used to visually assess balance. Figure 22–Figure 28 show the boxplots for key covariates before and after matching. In these boxplots, the box length represents the interquartile range (difference between the first and third quartiles). The “whiskers” represent the distance from the first (or third) quartile to the most extreme data point that lies within 1.5 times the interquartile range of the first (or third) quartile. Data points beyond the whiskers are considered outliers and are plotted individually. If the distance (from the point to the first [or third] quartile) is more than 3 times the interquartile range, the point is marked with a star.

In this study, the boxplots in Figure 22–Figure 28 mostly serve to show that the groups were not drastically different even before matching. The boxplots for DevMathGPA and CourseCompletionRatio look slightly more similar after matching;

the boxplots for Hours Attempted Fall 2012 look slightly less similar after matching. Someone starting with less similar groups should expect to see more improvement.

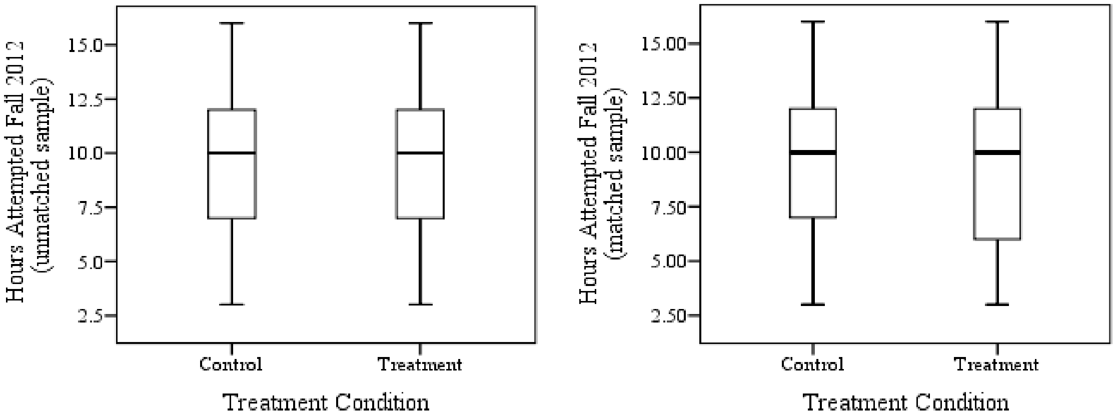


Figure 22. *Boxplots of HrsAttF2012 for matched and unmatched samples.*

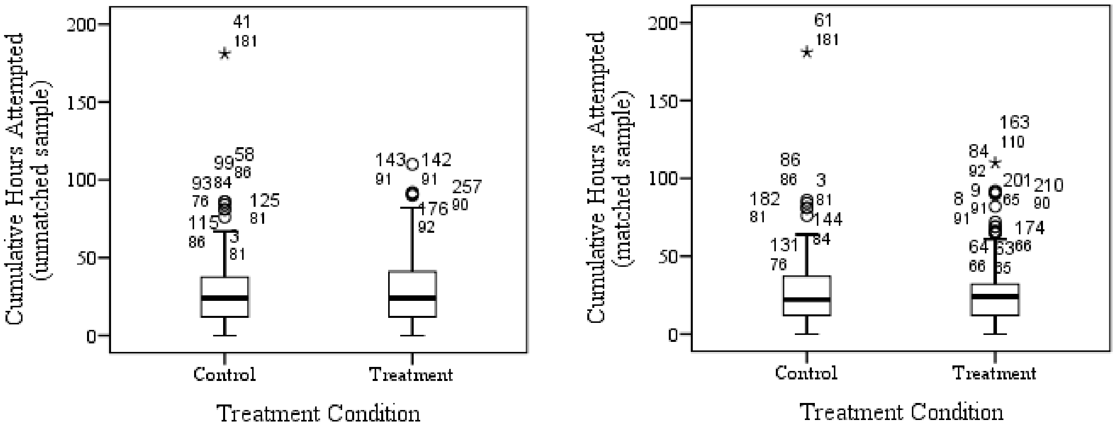


Figure 23. *Boxplots of CumHrsAttPreInt for matched and unmatched samples.*

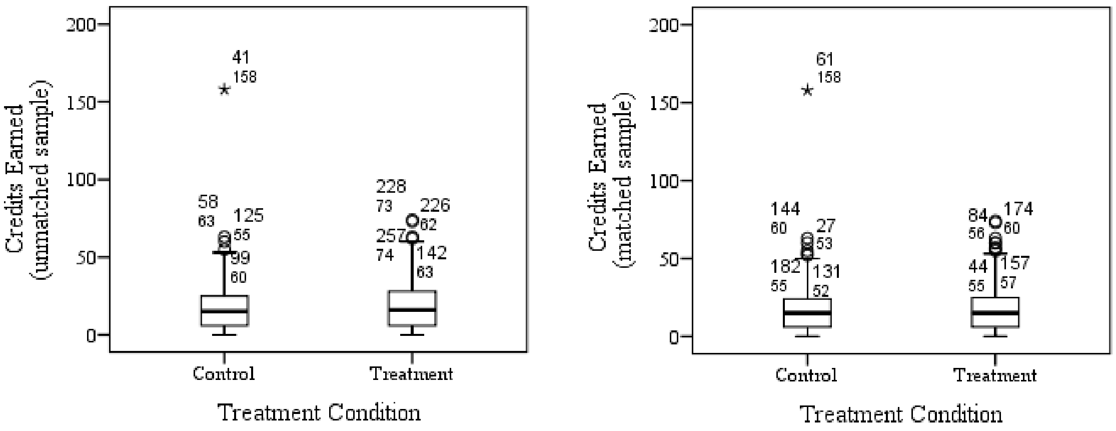


Figure 24. *Boxplots of CredEarnedPreInt for matched and unmatched samples*.

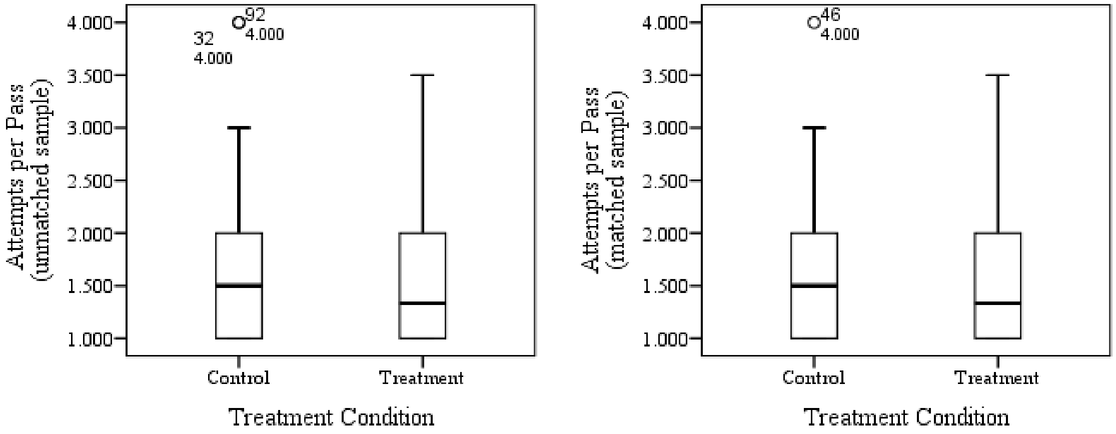


Figure 25. *Boxplots of AttemptsPerPass for matched and unmatched samples.*

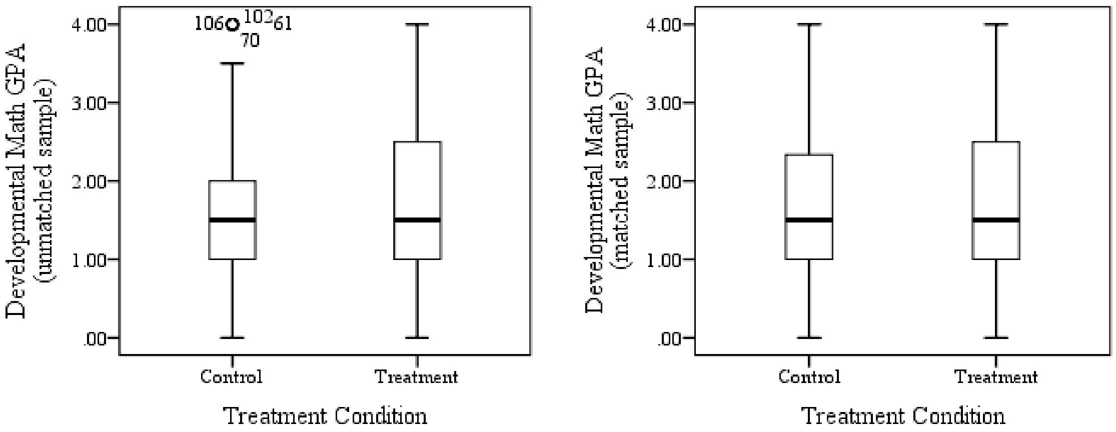


Figure 26. *Boxplots of pre-intervention DevMathGPA for matched and unmatched samples (students with no prior enrollments in developmental mathematics were omitted).*

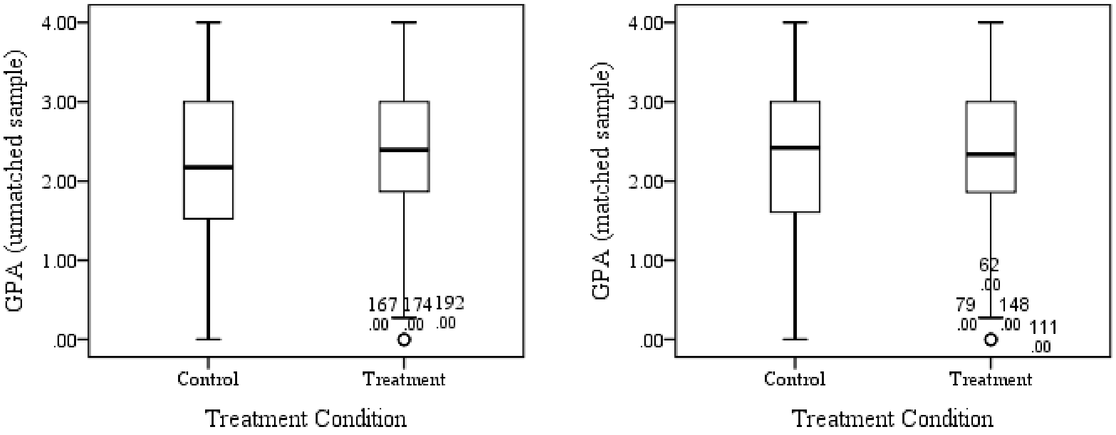


Figure 27. *Boxplots of pre-intervention GPA for matched and unmatched samples (students with no prior GPA-eligible hours were omitted).*

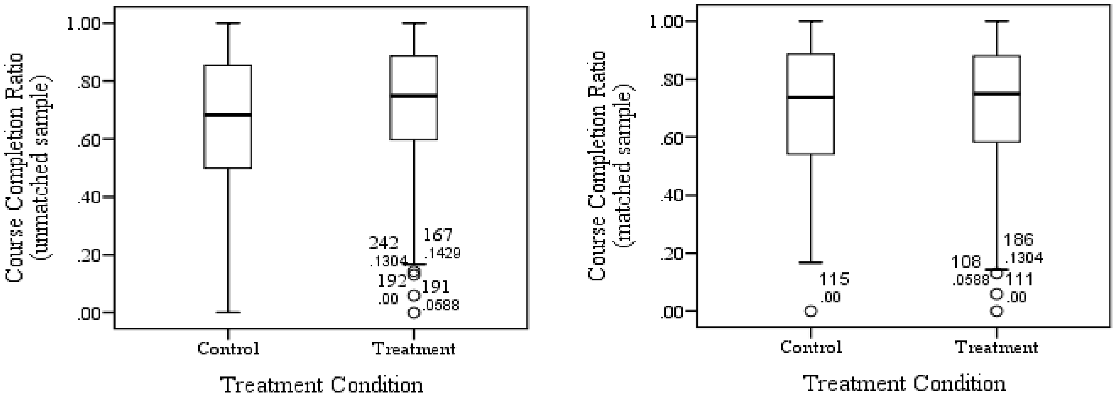


Figure 28. *Boxplots of CourseCompletionRatio for matched and unmatched samples (students with no prior attempted hours were omitted).*

The QQ plots and boxplots are helpful for assessing balance on continuous and near-continuous variables, but do not help on categorical variables. In this study, some of the near-continuous variables (DevMathGPA, GPAPreInt, and CourseCompletionRatio) were also combined with a categorical NA value, used to represent students with no eligible hours. The QQ plots and boxplots above did not include these students. For the distributions to match, the groups also needed to have similar percentages of these NA students. So, for the categorical variables, including the indicator variables for DevMathGPA, GPAPreInt, and CourseCompletionRatio, I created frequency tables and compared the percentages for each categorical value in the control and treatment groups. The almost-ordinal variable PrereqStatusGradePts was handled in the same way.

Table 34 lists frequencies and percentages for the binary variables before and after matching. Before matching, the groups showed obvious imbalance in the percentages of students with NA on DevMathGPA and CourseCompletionRatio, and in the percentage of ESOL students. The matching process fixed these imbalances. On all these binary variables, the matched groups were either identical, or were only off by one or two students. Interestingly, all but one of the six ESOL students in the treatment group were discarded in the matching process.

Table 34

Frequencies of Binary Covariate Values by Treatment Condition for Unmatched and Matched Samples

	Unmatched		Matched	
Variable	Control (*n*=140)	Treatment (*n*=117)	Control (*n*=105)	Treatment (*n*=105)
ESOL				
Not ESOL	139 (99.3%)	111 (94.9%)	104 (99.0%)	104 (99.0%)
ESOL	1 (0.7%)	6 (5.1%)	1 (1.0%)	1 (1.0%)
Developmental Mathematics GPA				
Not NA	118 (84.3%)	90 (76.9%)	85 (81.0%)	85 (81.0%)
NA	22 (15.7%)	27 (23.1%)	20 (19.0%)	20 (19.0%)
GPA				
Not NA	109 (77.9%)	88 (75.2%)	82 (78.1%)	81 (77.1%)
NA	31 (22.1%)	29 (24.8%)	23 (21.9%)	24 (22.9%)
Course Completion Ratio				
Not NA	126 (90.0%)	100 (85.5%)	91 (86.7%)	92 (87.6%)
NA	14 (10.0%)	17 (14.5%)	14 (13.3%)	13 (12.4%)
Current Course				
Introductory	60 (42.9%)	52 (44.4%)	43 (41.0%)	45 (42.9%)
Intermediate	80 (57.1%)	65 (55.6%)	62 (59.0%)	60 (57.1%)

As previously explained, the ordinal variable PrereqStatusGradePts was created by applying the traditional grade point system to the categorical variable PrereqStatus (*A* = 4, *B* = 3, *C* = 2, *Repeat* = 0). Values of *Placement* were captured using an indicator variable. The standardized difference for PrereqStatusGradePts was –0.047. Though well below the benchmark of 0.25, considered the maximum allowed for acceptable balance (Stuart & Rubin, 2007), it was the largest in magnitude of all the after-matching standardized differences (see Table 33 and Figure 10). For this reason, and because PrereqStatus was expected to be associated with the outcome (Little, 2002), it was important to look at this variable carefully. Table 35 shows the frequencies and percentages for each value in the matched and unmatched samples. After matching, the control and treatment groups had identical percentages of *Placement* students. The treatment group had more *A* students than the control group, but it also had more *Repeat* students.

Table 35

Frequencies of Prerequisite Status Values by Treatment Condition for Unmatched and Matched Samples

	Unmatched		Matched	
Value	Control (*n* = 140)	Treatment (*n* = 117)	Control (*n* = 105)	Treatment (*n* = 105)
A	9 (6.4%)	15 (12.8%)	7 (6.7%)	12 (11.4%)
B	30 (21.4%)	17 (14.5%)	23 (21.9%)	16 (15.2%)
C	21 (15.0%)	14 (12.0%)	17 (16.2%)	14 (13.3%)
Repeat	58 (41.4%)	44 (37.6%)	38 (36.2%)	43 (41.0%)
Placement	22 (15.7%)	27 (23.1%)	20 (19.0%)	20 (19.0%)

When converting this categorical variable to ordinal, I chose the coding scheme based on tradition and consistency, choosing the same value for each letter grade that the college uses when computing GPA. I coded values of *Repeat* as 0, because the college awards 0 grade points for F grades. Thus the categorical values *A*, *B*, *C*, *Repeat* were converted to the ordinal values 4, 3, 2, 0. The last two in the sequence are two units apart, but the first three are only one unit apart. So, I recalculated the means and standard deviations using a simpler coding system, one that preserved the order and put a consistent gap between all the values (*A* = 3, *B* = 2, *C* = 1, *Repeat* = 0). As before, I calculated the mean of the non-NA students and imputed it to the NA students. In the matched sample, this resulted in a mean of 0.965 in the treatment group and 0.984 in the control group; the standard deviation of the treatment group in the unmatched sample was 1.021. This resulted in a standardized difference of –0.019, smaller in magnitude than the –0.047 produced by the 4, 3, 2, 0 coding. Because the original choice of 4, 3, 2, 0 was somewhat arbitrary, and because neither group was obviously better on this variable after matching, I decided the groups were sufficiently well matched on PrereqStatusGradePts.

Estimating Treatment Effect

Once the groups were satisfactorily balanced on the covariates, the next step was to use the appropriate statistical technique to estimate the treatment effect (the effect of the intervention on the outcome variable). Whether using linear regression, logistic regression, structural equation modeling, or some other parametric technique, it is important that the model include those predictors that are expected to be associated with the outcome. The fact that we are starting with similar groups does not mean it is sufficient to just look at a simple difference in means. However, starting with well-balanced groups makes the parametric analysis less sensitive to modeling assumptions (Ho et al., 2007).

In this study, the outcome variables were dichotomous, so logistic regression was the appropriate technique. In addition to treatment assignment, three other predictors were included in the model: AttemptsPerPass, HrsAttF2012, and PrereqStatus (the categorical version, with values *A*, *B*, *C*, *Repeat*, *Placement*). These were chosen because they were expected to be associated with the outcomes (course success and final exam success), they were not overly redundant with one another, and they did not involve data imputation. The results of this analysis were described in Chapter III.

Discussion

In this section, I will first provide practical advice on how to record data from college transcripts. Transcripts contain a wealth of information—so much information that my first reaction was "where on earth do I start?" Next, I will discuss some of the factors I considered when choosing matching variables. I will mention other matching variables you may wish to consider if you are studying a different population. This section includes an extensive discussion of the ESOL variable and how it affected the balance of my groups. Following the section on matching variables is a short discussion of the trade-offs necessary when choosing a matching algorithm. Next, I provide suggestions for handling missing data and Incomplete grades in propensity score matching. I then give a short overview of issues involved in estimating the treatment effect for samples created from propensity score matching. In the concluding section, I describe how the matching illustration described in this article makes a unique contribution to the propensity score literature, and review how propensity score matching can improve the credibility of statistical inferences.

Practical Tips on Transcripts

If you are matching the students based on responses to survey data, perhaps on a demographic survey, you will presumably have included all the important matching variables in your survey questions. If you plan to match the students based on academic data from their transcripts, as I did, you will need to decide what variables to use so that you can pull the correct information from the transcripts. Be prepared—recording data from the transcripts is tedious and time-consuming. It may take longer than you expect. (According to my time-use spreadsheet, I spent about 46 hours creating Excel formulas and inputting data from the 258 transcripts.)

If you are not already familiar with the institution's transcripts, plan to spend some time with a calculator, adding and dividing, to learn what types of courses are included in the various totals and calculations. For example, are developmental courses included in the credits earned, credits attempted, and GPA? It is a good idea to select a sample of transcripts and compute key quantities listed on them, such as credits earned, credits attempted, and GPA, and verify that you can duplicate the numbers calculated by the institution. Be sure to look at a few students who have never enrolled in a class, or who have never enrolled in a credit-level class, and see what the transcript lists for GPA.

Once you have chosen your matching variables and become comfortable with the information contained in the transcripts, you'll need to record the information required to calculate each of the matching variables. For each of the 258 participants, my Excel spreadsheet contained values for 38 variables (see Appendix I for descriptions of variables recorded from student transcripts). Values for 27 of these were recorded directly from the transcripts; values for the other 11 were calculated using Excel formulas. If you are not skilled at spreadsheet formulas and if/then operators, you will want to learn, or else to collaborate with someone who has this skill. Only 12 of the 38 variables were used in the propensity score matching; the others were intermediate values used to generate the values of the matching variables. Later, after transferring the spreadsheet data to SPSS, I tweaked the existing matching variables (as previously described) and added four indicator variables, for a final set of 16 matching variables.

Most educators are familiar with the concept sometimes known as “horizontal grading,” in which the teacher grades Problem 1 for all students, then grades Problem 2 for all students, then Problem 3, and so on. This increases efficiency, prevents errors, and improves consistency. I used the same approach to pulling numbers off the transcripts. Instead of recording all the necessary variables off the first student’s transcript and then moving to the second student, I went through the stack of transcripts many times, recording only one or two things at a time. For example, in the first pass, I recorded all students’ values for HrsAttF2012 and GradePointsF2012. Then I went through the stack again and recorded all students’ values for CumCredPts2013. In some passes through the stack, I recorded no data at all, but instead used a colored highlighter to mark certain types of classes (e.g., orange for grade-excluded classes, yellow for developmental mathematics, pink for withdrawals).

It is a good idea to build in redundancy, to double-check your calculations by arriving at the same number in two different ways. For example, instead of simply copying the institution’s GPA, I recorded the total hours and total grade points, removed classes that were omitted from the GPA, and calculated the current GPA myself (using an Excel formula). If the GPA I calculated did not match the GPA listed on the transcript, I searched until I found the error, usually a W or developmental class I had failed to subtract. Because my calculation of current GPA incorporated many of the quantities that would be used in the matching variables, this served to verify that I had recorded those quantities correctly.

If the student records were generated after the intervention, as mine were, it is essential to back out the intervention semester's hours and grades to obtain pre-intervention values for GPA and other matching variables. If at all possible, matching variables should be measured before the treatment (Stuart & Rubin, 2007) so that they cannot be affected by it. In my case, after all my values for the current (post-intervention) GPA matched the college's values for current GPA, I subtracted the intervention semester's hours and grades and calculated the pre-intervention GPA.

After calculating all the variables, ask someone else to spot-check your values. For this, I selected several transcripts, some with long academic histories and some with short. My helper used a calculator to compute DevMathGPA, GPAPreInt, AttemptsPerPass, and several other variables directly from the transcripts, confirming that the formulas in my spreadsheet were correct.

Factors to Consider When Choosing Matching Variables

Before you start recording and calculating variables from the transcripts, surveys, or other data source, you should have a good idea of the variables you plan to use for matching. Though the choice of matching variables should be driven by previous literature, it will also depend on the data that is available to you, and will be filtered by your knowledge of the institution and your professional judgment as an educator and researcher. Keep in mind that propensity score matching can only balance observed variables. As you make choices about which observed variables to incorporate, it is important to consider what variables are unobserved, and whether those unobserved variables are likely to confound the results.

My matching variables incorporated postsecondary academic data only, not demographic data or high school academic data. It may very well be that unobserved demographic or high school variables are associated with the outcome of interest, developmental mathematics success. In that case, those demographic and high school variables should also be associated with some of the postsecondary academic variables I used. Of course, this is only the case for students with some college experience. For some students, the intervention semester was their first semester in college and thus their pre-intervention college transcripts were "blank slates." Because my study involved only the second and third courses in the three-course developmental mathematics sequence, and because the majority of developmental mathematics students at this institution are placed into the first course, most (225 out of 257) participants had at least one semester of college academic history. If your study involves the first course of the sequence, a larger proportion of students will be in their first college semester, making it more important to incorporate other variables in your matching procedure. If your participants are freshmen at a four-year university, it will be important to incorporate demographic or high school variables. If your participants are juniors and seniors, you may want to consider transfer credits. In my study, 43 of the 257 participants had other colleges listed on their transcript. Because the transcripts did not list the classes taken at those other colleges, I did not incorporate this information. My matching variables only included information from classes taken in the six-campus community college system where the study took place.

If you are matching on college academic variables as I did, you will need to decide how to handle withdrawals. If the institution awards In Progress (IP) grades, or the equivalent, you'll also need to decide how to handle those. For reasons described earlier, I decided to treat IP, F, and W grades as equivalent in the DevMathGPA, counting them in the denominator and awarding them 0 grade points. This was consistent with my choice to count IP, F, and W grades as unsuccessful on the outcome variable. I chose this route because I felt that instructor and financial aid policies had so confounded the IP, F, and W grades that separating them would not provide meaningful information. If I were conducting research at a different institution with different policies, or if I were researching credit-level classes, I might make a different choice. Another option would be to create two separate variables to capture the information, one in which IP = F = W = 0, and one which excludes W grades (or excludes both W and IP grades). If the data set was sufficiently large to support both these variables, this could be a good solution. As long as the groups have similar distributions of both variables, it doesn't really matter which variable is "right."

In my AttemptsPerPass variable, designed to capture course repetition information, I also treated W grades as unsuccessful attempts. Though AttemptsPerPass was useful as a matching variable, it was not a significant predictor of course success or final exam success (see Chapter III). This indicates that non-repeaters, as a group, are not necessarily better students than repeaters. This observation is supported by the results found for PrereqStatus in the treatment effect estimation model (see Chapter III). As previously described, the PrereqStatus variable took on values of *A*, *B*, or *C* if the student took the prerequisite class immediately preceding the current class, *Placement* if the student was placed directly in current class without enrolling in the prerequisite class, and *Repeat* if the student was repeating the current class after receiving a W, IP, or F. A PrereqStatus of *C* was a statistically significant predictor of final exam success, with *C* students being significantly less likely to pass the class than *Placement* students. This did not hold true for *Repeat* students, who were no less likely to pass the class than *Placement* students. It would be interesting to create a second version of AttemptsPerPass, one which excludes W grades, and see if it was a stronger predictor of success.

Impact of ESOL Variable

If you notice something unexpected on the transcripts, go ahead and record it—it may prove to be important. I included ESOL on a hunch. I did not plan in advance to include it, and it was the only academic variable based on a specific content area other than mathematics. Going through the transcripts, I noticed that ESOL students seemed to have different mathematical histories than other students. They did not enroll in developmental mathematics until after they completed their ESOL classes, often several semesters after they began college. However, unlike non-ESOL students who delayed their mathematics enrollment, these ESOL students generally succeeded in their first developmental mathematics class.

I ran the matching algorithm with and without the ESOL variable. When I removed the ESOL covariate, balance on the other covariates worsened noticeably. Why would this be? First, I examined the original unmatched data set. Six of the seven ESOL students were in the treatment group, and only one was in the control group. Next, I looked at the propensity scores generated during the process of creating the matched set described in this chapter. This trial included the ESOL variable and had excellent overall balance on the covariates. The matched set contained only two ESOL students, one in the treatment group and one in the control group. The other five ESOL students in the treatment group were discarded. The seven ESOL students had the seven highest propensity scores in the entire data set (i.e., they had the highest probabilities of being assigned to the treatment group given their values on the covariates). The propensity scores of the two ESOL students in the matched set were .790 and .785, and the scores of the five discards ranged from .823 to .951. The next highest propensity score was .665, belonging to a matched non-ESOL student. As can be seen in Figure 8, the high-propensity score ESOL students, all in the treatment group, could not find a match with a similar propensity score. Because the two groups were extremely lopsided on the

ESOL variable, ESOL was a very strong predictor of treatment assignment for these 7 students (though not a good predictor at all for the other 250 students).

Next, I examined the matched set generated when the ESOL variable was omitted. I began with the identical data set and identical randomization of cases that was used to generate the well-balanced matched set of 105 pairs described in this chapter. When I ran the matching algorithm without ESOL, the resulting matched data set still had 105 matched pairs but had very poor balance on the other covariates. This time, six of the seven ESOL students were matched and only one was discarded. Apparently these ESOL students were so atypical in their academic histories that they threw off the balance when they were included.

Finally, I compared the propensity scores generated during the creation of the two matched sets (ESOL included and ESOL omitted). The seven ESOL students showed wide swings in their propensity scores. In the most extreme case, a student's propensity score dropped from 0.790 (when ESOL was included as a covariate) to 0.383 (when ESOL was omitted from the set of covariates). The non-ESOL students showed only moderate changes in their propensity scores.

The effect of the ESOL variable upon my data set illustrates the importance of performing extensive balance checks and of using balance to determine which matched set is best. It also shows that while the propensity score can work very well as a tool for creating similar groups, using it as a predictor can be hazardous. The propensity score is highly sample-dependent. Other predictors, such as GPA or credits earned, are characteristics of the individual and will remain unchanged if that individual is placed in another sample. However, a participant's propensity score could change drastically if the sample or the set of covariates is modified. If the ESOL students had been more evenly distributed across the two groups, the ESOL variable would probably not have had nearly as strong an effect either on the propensity score or on the resulting matched sets.

The purpose of propensity score matching is to create two groups that have similar distributions of important starting characteristics. If a variable is expected to be associated with the outcome, it should be included in the matching procedure (Stuart, 2010). If it is unknown whether a variable is associated with the outcome, it is best to include it, at least until it becomes evident that your set of matching variables is so extensive that good matches are impossible. If a variable turns out not to be associated with the outcome, it will cause no harm for the groups to have similar distributions on that variable; but if it turns out to be associated with the outcome, omitting it could seriously confound your estimation of the treatment effect (D'Agostino, 1998; Dattalo, 2010; Rosenbaum & Rubin, 1983; Stuart, 2010).

The more variables you have, the larger your data set needs to be to support good matches. Also, the more different your groups are, the harder it will be to get good balance on a large number of variables. By forcing the matching algorithm to match on a large number of variables that are unrelated to the outcome, your groups may be matched less well on variables that are associated with the outcome. Omitting a less important variable may help the balance on a more important variable. As we have seen, the opposite can also happen—adding the ESOL variable actually improved the balance on the other variables, by causing outliers to be discarded. It is essential to carefully examine the groups' balance on the covariates both numerically and visually. As long as the groups are acceptably balanced on all the matching variables, it is fine if some of the variables are unassociated with the outcome.

Trade-off Between Two Goals

Propensity score matching involves a trade-off between two goals: achieving the best possible balance on the covariates and keeping as many treatment cases as possible. Imposing a tight caliper improves the quality of the matches, but often results in a smaller sample size. You will need to experiment with variations of the matching algorithm, such as different calipers, and then examine the sample size and covariate balance for the resulting matched samples. If you are using the R-based propensity score program for SPSS, then you are limited to nearest-neighbor (greedy) matching, with a choice of caliper. If you cannot achieve acceptable balance using nearest-neighbor matching, you may wish to try some of the more robust matching algorithms available in R. If it proves impossible to achieve reasonable balance without drastically trimming your sample size, the two groups may be too dissimilar to support credible inferences.

If you expect the groups to be extremely different from one another, having a large control group will help. If a large number of control cases are available, relative to the number of treatment cases, it may still be possible to get two well-balanced groups while discarding very few treatment students. A large pool of controls may occur if the treatment is some small program into which the students self-select, allowing the rest of the school or cohort to serve as controls.

Potentially Sticky Issues to Consider in Advance

If you are using high school data or demographic data from surveys, some values will inevitably be missing. If you have survey data, consider in advance how you will handle missing data. Because the propensity score is a function of all the covariates, it cannot be calculated if some covariate values are missing. In many situations, an acceptable option is to use indicator variables to capture the pattern of missingness, and then fill in the missing values through some sort of imputation technique (D'Agostino & Rubin, 2000; Stuart & Rubin, 2007; Stuart, 2010). This is similar to the approach I used for GPA, DevMathGPA, and CourseCompletionRatio.

Because I used college transcripts as the only data source for matching variables, I did not face a missing data problem. The information on transcripts is clean and presumed accurate. The absence of the GPA for some students was not due to the GPA being "missing"—it was due to the fact that those students had not yet taken any GPA-eligible courses. If GPA data were obtained through self-reports on a survey, the GPA would be legitimately "missing" for those students not answering the question.

The SPSS propensity score plug-in cannot handle missing values, even if the variables with the missing values are not included as covariates. (Future versions of the program may change this.) For example, my HrsAttSpr2013 variable was not a matching variable, but was only used to back-calculate the pre-intervention GPA from the current

GPA. When I recorded values for the HrsAttSpr2013 variable, I left the cell blank (instead of entering 0) for the students not attempting any hours in Spring 2013. This worked fine for the GPA calculation, but caused the propensity score matching program to balk. In order to run the program, I had to copy the data set and delete the variables with missing values.

If the institution allows Incomplete grades, plan in advance how to handle them. For my study, I strictly avoided recording outcome data until after the matching phase was completed, as recommended by the literature. An important advantage of propensity score matching is that it does not involve outcome data and so cannot be biased by the outcome data (Ho et al., 2007; Stuart & Rubin, 2007; Stuart, 2010). After running many trials with different random orderings of the participants, I chose the matched sample that resulted in the best balance of the covariates without discarding too many treatment participants. Only then did I return to the transcripts to record the course success data. One control student had received a grade of Incomplete. My first hope was that the student was among the 35 discarded controls. After this hope was dashed, I contacted the instructor to obtain information about the student's exam grades and the reason for the Incomplete. If the student had clearly been failing at the beginning of the semester, prior to whatever calamity prompted the Incomplete, perhaps I could legitimately count the student as unsuccessful. I had trouble obtaining the information and ultimately decided it would be cleaner to remove the student and rerun the propensity score phase of the analysis. I repeated the process, performing visual diagnostics on various random orderings of the participants, and settling on a matched sample just as well-balanced as the first matched sample (the matched sample presented in this chapter is the one in which the Incomplete student was removed). Not surprisingly, the process went much more quickly the second time. However, I could have saved myself a substantial amount of time by planning for this possibility in advance. It would have been easy to flip through the transcripts and remove any Incompletes before beginning the matching process.

Estimating the Treatment Effect

Propensity score matching, extensive balance checks, and the selection of the best-balanced sample constitute only the first phase of analysis. The second phase will be to estimate the effect of the treatment on the outcome. By using propensity score matching as a nonparametric preprocessing phase, the treatment effect estimation will be less dependent on modeling assumptions (Ho et al., 2007; Stuart & Rubin, 2007). If you used one-to-many matching or matching with replacement, your treatment effect estimate will need to incorporate weights for the cases (Dehejia & Wahba, 2002; Reynolds & DesJardins, 2009; Stuart, 2010). If you used one-to-one matching without replacement, as I did, all the cases will have the same weight.

When calculating the treatment effect, some researchers argue for the use of matched pairs analysis techniques, such as dependent sample *t* tests or conditional logistic regression (Austin, 2008, 2011). Others maintain that the trimmed groups can be treated as independent samples (Hill, 2008; Stuart, 2008, 2010). I chose the latter approach. Because my groups were created using the propensity score, they are more accurately characterized as “matched samples” than as “matched pairs.” Though the distributions of the covariates in the treatment group are similar to the distributions of the covariates in the control group, the groups do not necessarily contain any pairs of individuals who have similar values on the covariates. Propensity score theory establishes that the propensity score balances distributions, not that it creates matched pairs.

Contribution of This Study to the Propensity Score Literature

Most illustrations of propensity score matching use large data sets in which the unmatched treatment and control groups have extremely dissimilar distributions of the covariates. In one frequently cited example, the researchers (Dehejia & Wahba, 2002) began with the same sample used in a previously published randomized trial of the National Supported Work (NSW) program. They compared the original treatment group (n = 185) with subsamples selected from two large databases (n = 2,490 and n = 15,992) using several different propensity score matching procedures. Though the starting characteristics of the original large data sets were substantially different from those of the NSW sample, the propensity score methods were able to select subsamples that were comparable to the treatment group. Using these subsamples as control groups, the estimates of the treatment effect upon earnings were comparable to the treatment effect estimate in the original randomized trial. As long as the original (unmatched) control group contained a sufficient number of units that were similar to the treatment units, nearest-neighbor matching without replacement performed well. When the original control group had very few units similar to the treatment units, matching with replacement worked better. Another example used data from the National Educational Longitudinal Survey, which tracked approximately 12,000 eighth graders for 10 years; the educational outcomes of the 3,770 students who began in two-year colleges were compared to those of the 4,890 students who began in four-year colleges (Reynolds & DesJardins, 2009). Propensity score matching was used to control for the extreme differences in academic, demographic, and family characteristics of the two groups of students.

When applied to large data sets with substantial systematic (nonrandom) differences between the groups, propensity score matching results in dramatic improvements in balance on the covariates. It also causes dramatic differences in sample size. For example, from Dehejia and Wahba's (2002) original comparison group of 15,992, the nearest-neighbor propensity score matching algorithms selected subsamples of 119 and 105; the caliper-matching algorithms selected subsamples of 325 participants, 1,043 participants, and 1,731 participants. The purpose of the Dehejia and Wahba study was to use a huge control group as a starting point, and then use matching to select a fairly small subsample that was comparable to the treatment group.

The study presented in this dissertation, on the other hand, is unusual in that it illustrates how propensity score matching can be used to clean up residual differences between groups that are reasonably similar in the first place. To minimize inherent differences between the groups, I matched classes as far as possible on course, teacher, and time, and then randomized the treatment assignment within each pair. Still, there were differences. Before matching, the treatment group was slightly stronger on most of the academic variables and contained six of the seven ESOL students. Whether these imbalances were due to systemic factors or random variation, propensity score matching was able to remove the imbalances, or at least substantially reduce them. The matched sample (105 control, 105 treatment) was only slightly smaller than the original sample (140 control, 117 treatment).

I am aware of no other detailed discussion of matching variables for developmental mathematics students at a community college. This is a challenging population to study. While students at a four-year university may proceed through their college years in somewhat neat cohorts, community college students do not. A community college does not have a well-defined freshman class or sophomore class. Developmental mathematics students enter the three-course sequence at different levels, and only around half of them, or less, successfully pass their mathematics course. The different entry levels and the high number of course repetitions make modeling difficult. The usual college success predictors, such as GPA, have limited value because so many students have not yet taken any GPA-eligible courses. By using propensity score matching to create groups that are as similar as possible on as many variables as possible, we can reduce dependence on modeling assumptions (Ho et al., 2007). If the groups are sufficiently similar, we may gain valuable information even from a very simple analysis, such as a chi-square comparison of frequencies or a regression with very few predictors.

Closing Thoughts

The purpose of propensity score matching, as described here, is to improve the credibility of statistical inferences. As educational researchers, we want to avoid two types of errors: concluding a program is beneficial when in fact it is not, and concluding a program is not beneficial when in fact it is. If the groups used in our statistical analysis have similar distributions of important variables, these errors are less probable. However, the importance of balancing the groups reaches beyond the mere avoidance of errors. Statistical analysis should do more than simply choose the "correct" answer out of three possibilities (it helps, it hurts, it makes no difference). If the intervention has an effect, statistical analysis needs to estimate the size of that effect. Unbalanced groups can bias that estimate, causing us to overstate or understate the effect size.

Statistical inferences about a program's effectiveness, no matter how credible, do not provide a definitive answer as to whether a program should be continued. We may choose to discontinue a program that produces strong positive effects—perhaps the benefits are outweighed by the costs or negative side effects. Or, if it is cost-effective to do so, we may choose to continue a program even if it does not result in a statistically significant improvement of outcomes. There is nothing inherently wrong with continuing a program based on anecdotal evidence that it helps a few students. However, if we make such a decision, it is crucial that we recognize we are doing so. We do not want imbalances in the groups to convince us that the decision is justified by statistical inference, if that is not the case. For example, in my study, the original (unmatched) treatment group was slightly stronger academically than the original control group. If I had used the raw data and found a statistically significant effect in favor of treatment, it could very well be attributable to the group's starting characteristics rather than the intervention.

As consumers of research reports, we critically assess each study's design and analyses so we can decide how much weight to place upon its conclusions. As writers of research reports, we are obligated to provide sufficient information for our readers to assess the strengths and limitations of our studies. By conducting extensive balance checks and providing numerical and visual summaries of those balance checks, we shine a spotlight on our sample, illuminating any flaws it may have. Hopefully, the propensity score matching will successfully remove most of the imbalances, making our statistical evidence more convincing. If some imbalances remain, highlighting them will help the reader decide what caveats apply to the study's conclusions.

CHAPTER V

SUMMARY AND CONCLUSIONS

The purpose of this study was to investigate the effects of a study-journaling intervention for developmental mathematics students at a community college. The intervention was built on self-regulated learning theory and motivated by my interest in helping students become more effective at self-regulating their learning, with the goal of improving their mathematics achievement.

In conjunction with the empirical study of the study-journaling intervention, I conducted an extensive review of research on self-regulated learning interventions for college students. In this chapter, I will first summarize the methodology, results, and limitations of this literature review, which were presented in Chapter II of this dissertation. Next, I will summarize the methodology and results of the empirical study, which were presented in Chapter III. This will be followed by a short description of Chapter IV, which is an extended discussion of the methodology used in the empirical study of Chapter III. Next, I will describe how the results of the empirical study and the results of the literature review fit together. The chapter will close with a discussion of the limitations of the empirical study and recommendations for future research.

The Literature Review

Before designing the research study, I had conducted a preliminary literature review, examining empirical studies of interventions designed to improve students' self-regulated learning skill. This review indicated there was a need for more research on self-regulated learning interventions, particularly interventions embedded in content courses. This preliminary review gave me confidence that the current study would have value and partially fill a gap in the literature. However, the preliminary review was neither exhaustive nor systematic, and was not sufficient to show clearly where the current study fit or to connect its results to prior research.

For that reason, I conducted a more thorough literature review, based on the research question, “What are the effects of self-regulated learning interventions on college students?” I used a systematic approach, documenting the search terms and inclusion criteria. This literature review, structured as a stand-alone manuscript, is presented in Chapter II.

The inclusion criteria, in the form of a list of questions, defined the scope and direction of the literature review. The start date for the review, 1994, was chosen based on the initial publication date of a dimensional framework of self-regulated learning (Zimmerman, 1994, 1998), which I planned to use in organizing the results of the review. In order to keep the literature review manageable, I restricted the review to studies conducted in the U.S. and listed in the ERIC database. I chose to review only interventions that targeted college students’ learning of academic content in face-to-face courses. This meant that stand-alone study strategies courses were generally included, as long as they contained key elements of self-regulated learning and were intended to improve the students’ study habits in their other courses.

In a three-stage screening process, I culled the initial pool of candidates to the final set of articles for review. In each stage, I applied the inclusion criteria, eliminated some articles, and documented the reasons for elimination. In the first stage, I read the 1,825 abstracts generated by my search in the ERIC database. In the second stage, I obtained and skimmed the full texts of 437 articles. In the third stage, I carefully read the full texts of 81 articles. For each, I used a spreadsheet to record the answers to the inclusion questions and summarize information about the sample, intervention, analysis method, outcome measures, and results. More articles were excluded during this stage, resulting in a final pool of 42 articles to be reviewed.

Results of the Literature Review

A dimensional framework from existing self-regulated learning theory (Zimmerman, 1994, 1998) was used to organize the reviewed studies. Twenty-seven of the studies addressed one or two of Zimmerman’s six dimensions of self-regulated learning: time, strategies, outcomes, motivation, social context, and environment. The other 15 studies addressed the overall process of self-regulated learning.

Of the 42 studies reviewed in Chapter II, only 13 studies had credible research designs and analyses, occurred in the authentic context of a college class, and used an outcome measure involving grades in a content course (e.g., grades in mathematics or biology, as opposed to grades in a study strategies course). These studies, considered as a group, showed that self-regulated learning interventions can have a positive effect on students’ achievement, at least in the short term.

The review revealed a need for credible achievement-based research on interventions integrated into the normal activities of content courses. In particular, we need solid studies of small-scale interventions manageable by an individual teacher—interventions not consuming too much precious instructional time and not requiring specialized software. Additionally, there is a need for studies incorporating qualitative data to obtain a fuller picture of the intervention's effect, and for studies examining whether students carry their self-regulated learning skills forward into future semesters.

Limitations of the Literature Review

There are two main limitations to this literature review. First, it was delimited by my decision to restrict the review to studies conducted in the U.S. and listed in the ERIC database. Second, within those boundaries, the set of reviewed articles was highly dependent on my decisions about how to apply the inclusion criteria. Two researchers, beginning with the same inclusion criteria and the same initial pool of candidates, might generate different final lists of articles—due to philosophical differences, judgment differences, or screening errors.

The Empirical Study

The primary research study, a mixed methods empirical investigation of a study-journaling intervention for developmental mathematics students, was described in detail in Chapter III. Here, I will provide a short overview of the intervention, the research process, and the major findings.

The Intervention

Drawing on self-regulated learning theory, I designed two study-journaling worksheets, which were to be submitted weekly by the participating students. The first worksheet, called the *goal sheet*, focused on goal-setting, planning, and reflection (see Appendix B). It asked students to describe their mathematics-related goals for the week and their strategies to reach the goals. It also asked students to reflect on their progress in the class, their success in meeting the prior week's goals, and the reasons behind any unmet goals. It contained a table for them to use in planning their mathematics study time for the next week. The second worksheet, called the *study log*, took the form of a large table and was intended for use in tracking the students' actual study time (see Appendix C). For each study session, it had spaces for them to note the starting and ending times, the location, other people who were present, and their goals. It also asked for a short reflection on their level of satisfaction with the study session.

The research project took place at a large urban community college in Texas, where I am a mathematics faculty member. I selected nine pairs of developmental mathematics classes. Some of the classes were Introductory Algebra and some were Intermediate Algebra (the second and third courses in the three-course developmental mathematics sequence). After creating the class pairs by matching on as many class-level variables as possible, I randomly assigned one intact class in each pair to the treatment (study journal) condition and the other to control.

The Research Process

I chose a mixed methods design with two strands: (1) a confirmatory strand in which I evaluated the effectiveness of the study-journaling intervention and (2) an exploratory strand in which I sought information about the study habits and strategies of the study journal students.

In the confirmatory strand, I quantitatively evaluated the intervention by examining its effect on mathematics success. The quantitative analysis was divided into two phases. In the first phase, I controlled for initial differences between the treatment and control groups by matching students using the propensity score—a one-dimensional summary derived from key academic variables from the students' college educational records. In the second phase, I used logistic regression to determine the intervention's treatment effect on four binary outcome variables representing mathematics course success. I supplemented the quantitative evaluation with qualitative data from focus groups and surveys.

The outcome of interest—mathematics course success—was represented by four binary outcome variables. In addition to course success, I used three binary exam success variables, each using a different cut score on the departmental final exam. For the research study, I graded all the final exams myself, incorporating interrater reliability checks on a sample of exams graded by another researcher. On all four variables, students not finishing the course were included among the unsuccessful. I also applied the logistic regression analysis to an additional binary outcome variable representing whether the students took the final exam.

In the exploratory strand, I qualitatively analyzed the students' writings contained in the study journals. I grouped the student responses into themes of study habits, then organized the themes around Zimmerman's (1994, 1998) dimensional framework of self-regulated learning theory, the same framework used to organize the studies in the Chapter II literature review. Each student's study journal also received a numerical rating for overall quality and depth of reflection. The quality scores were the averages of scores awarded by me and another researcher; we utilized a rubric and interrater reliability checks. The qualitative analysis of the study journals provided additional explanation to the confirmatory strand by showing how the students were implementing self-regulated learning strategies.

I connected the strands by converting the qualitative data from the study journals to quantitative data. Using defined criteria, I analyzed the frequency counts to determine which study habits were associated with success.

Results of the Empirical Study

The study journal project was carried out in the nine treatment classes. The sample included 117 treatment students and 140 control students who agreed to participate in the research project. However, there were only four classes in which the intervention was fully implemented as planned. In these four classes, the instructors collected study journals from nearly all the students every week. In the other five classes, either the study journals were only collected sporadically, or they were collected regularly but from only a few students.

In the first round of quantitative analysis, I carried out both phases of the analysis on the original sample of 117 treatment students and 140 control students. In the first phase, propensity score matching resulted in two well-balanced groups of 105 students each. In the second phase, logistic regression showed no effect of the intervention on exam success. A comparison of frequencies on the outcome variables revealed that none of the exam success variables had a meaningful relationship with the course success variable (in which students were classified as successful if they earned an official grade of A, B, or C, and unsuccessful if they did not). This discord between course success and exam success indicated that the course success variable was severely confounded by differences in teacher expectations, and therefore had almost no value for the research.

In the second round of quantitative analysis, I focused on the four classes in which most students regularly completed study journals. I repeated both phases of the quantitative analysis, restricting the sample to those four classes and the corresponding four control classes with which they had been matched during the treatment assignment process.

In the first phase, the propensity score matching process produced two groups of 54 students each. The two groups' balance on the matching variables was within recommended guidelines (Stuart & Rubin, 2007); however, due to the smaller sample size, the balance was not as good as in the larger matched sample.

In the second phase of analysis (using the two groups of 54 students each), the logistic regression analysis did not show a positive effect on achievement, as measured by exam success and course success. However, the quantitative analysis showed an unexpected effect of the study journal intervention: students in treatment classes were more likely to leave the class before the final exam. In the logistic regression analysis, the odds ratio for treatment assignment was 2.94 ($p = .019$), indicating that the odds of treatment students leaving the class were nearly three times the odds of control students leaving the class.

The quantitative analysis on the effect of the intervention was supplemented by qualitative data from two focus groups and an end-of-semester survey. In the focus groups, volunteer students shared their perspectives on how the study journals affected their study habits and achievement. Both the focus groups and the surveys provided evidence that the intervention had a positive effect on some students' study habits and achievement.

In the Chapter III manuscript, I argued that the qualitative and quantitative results, taken together, showed that the study-journaling intervention did indeed have an effect on the students: it increased their self-awareness. The qualitative data from the focus groups, surveys, and study journals support this assertion. However, self-awareness does not automatically result in improved achievement. When self-awareness does not produce achievement, it is plausible that it could incite students to leave the class, either because of discouragement or because of an objective evaluation of their time constraints. As one student noted, “Because of external factors, time to study is limited. Being more aware of this limitation is only somewhat helpful.”

A Side Note About Chapter IV

In the empirical study, the propensity score analysis turned out to be far more involved than I expected. Traditional predictors, such as GPA, standardized test score, or previous mathematics grade, were not well-suited for the academic backgrounds of developmental mathematics students at a community college. To adequately capture the students’ academic histories, I had to combine a little creativity with a lot of trial-and-error. The final set of matching variables included several indicator variables and several variables incorporating data imputation. It also included a difficult-to-explain variable designed to capture course repetitions.

The details of the propensity score analysis and the matching variables required a far lengthier explanation than would be appropriate for the methodology section of the mixed methods empirical study. Instead, I placed the explanation into a third manuscript, targeted toward other community college researchers facing the same task as I faced—creating a sample of two comparable groups in order to draw credible inferences about the effects of an intervention.

This manuscript (Chapter IV) was written as a friendly how-to article, and includes an open discussion of the dilemmas I faced and how I reached my decisions.

My data from the first round of propensity score analysis (the original sample of nine treatment classes and nine control classes) serves as an illustration. Also included in this manuscript are the graphs used in the visual balance checks, as recommended by Ho et al. (2007).

Connection Between the Literature Review and the Empirical Study

Self-regulated learning interventions vary so widely that it is often difficult to say whether one study supports or refutes the findings of another. While the credible studies reviewed in Chapter II generally showed positive effects on achievement, the current study did not. However, because those interventions involved other components besides study journal worksheets, and because they targeted different populations, the current study also does not directly refute those studies' findings.

Unlike the intervention in the current study, the interventions evaluated in the credible studies reviewed in Chapter II generally required a substantial time investment by the students either inside or outside of class. Most of these interventions also required a substantial time investment from the instructors. The results of the current study indicate that simple study-journaling worksheets alone are not enough to improve achievement—at least not for developmental mathematics students at this community college. Perhaps achievement would improve if the study journals were combined with some of the elements used in other studies, such as time-management training (Goodwin & Califf, 2007), goal-setting training (Georgianna, 2009), peer group support (Tuckman, 2007), or individualized counseling (Haught et al., 1998).

The literature review and the empirical study also showed the value of Zimmerman's dimensional framework for organizing research results. Although not every research effort or study habit could be slotted neatly into the framework, many could, and the framework made it easier to see the connections between them. Much of the framework's value came from its simplicity—the framework can be comprehended and used by someone who is not an expert in cognition or psychology. As an additional plus, the framework's key words (when, how, what, why, where, who) make it easy to remember the dimensions (time, strategies, outcomes, motivation, environment, social context).

Limitations of the Empirical Study

Internal Validity

It is possible that the treatment classes' high departure rate was attributable to some unobserved factor other than the study journals. The first possibility to consider is that individual differences between students were not well-distributed between the two groups. Because shortcomings in implementation forced me to restrict the analysis to four treatment classes and four control classes, the sample size was smaller than anticipated. Even if the balance had been perfect, it is possible that a different set of matching variables could capture student characteristics that my set of matching variables did not.

Teacher differences are the second possibility for an alternate cause of the difference in student departure rates. However, this seems less likely. The restricted sample of eight classes included two same-teacher pairs of classes and two different-teacher pairs of classes. The differences in departure rates within the different-teacher pairs were relatively small; the most striking difference in departure rates occurred within one of the same-teacher pairs.

External Validity

For the remainder of the discussion of limitations, I will assume that the difference in departure rates was indeed attributable to the study journals. In other words, the remaining limitations are not alternate explanations for the departure rates—they are caveats that apply even if the study journals' effect on departure was real.

By asking instructors to collect the study journals without reading or assessing them, I removed one possible source of confounding. However, this compromise with authenticity also limits the conclusions that can be drawn. If an instructor were to implement study journals on his or her own initiative, without being part of a research study, the instructor would probably read them. The knowledge that the instructor was reading the study journals would probably affect what the students wrote in them. Also, if an instructor reading the study journals sensed a student was becoming discouraged and considering leaving the class, the instructor might initiate a conversation with the student. That conversation could provide the student with alternative strategies and options that might avert the student's departure. In other words, implementing study journals in a more authentic situation might produce different results.

If the difference in departure rate was indeed due to the study journals, it is possible that the effect was specific to this particular population. Perhaps developmental mathematics students, because of their non-school commitments and their relatively weak academic backgrounds, were more likely than other students to opt out of the class due to increased self-awareness. If a similar intervention were implemented at a four-year university, or in credit-level classes at a community college, the results could be different.

Among the students who remained in the class, it is possible that gains in achievement may have gone uncaptured due to the dichotomization of the outcome variables into two categories. This choice was driven by two factors: the expectation that a large proportion of students would not pass the class, and instructor differences in whether students must take the final exam to earn an IP (In Progress) grade. Because the final exam average is heavily influenced by the departure decisions of students who are struggling in the course and have no realistic chance of earning a passing grade, I chose to dichotomize the exam success variables and count departing students as unsuccessful. Though this resulted in a loss of information, it was the least problematic option.

Recommendations for Future Research

In this study, the qualitative data indicated the study journals had a positive effect on the study habits and achievement of some students; the quantitative data indicated the study journals may have increased self-awareness, which could potentially serve as one step on a path toward improved achievement. These results indicate that further research on study journals is warranted. However, this study also shows the need for researchers to proceed cautiously. When implementing interventions that have potential to increase students' self-awareness of their performance, we should consider the possibility that the intervention could cause students to become discouraged and leave the class. Future studies should include measures to analyze and mitigate any potential detrimental effects on the students.

As long as the possibility of student departures is kept in mind, the results and limitations of this research study suggest several directions for future research. One option would be to combine study journals with other supports, such as training, counseling, or peer support. Whether study journals are used alone or combined with other components, empirical studies should attempt to ascertain the reasons for student departures, perhaps with follow-up surveys or phone calls. Longer-term intervention studies are also needed in which researchers collect data on the students' achievement in one or more semesters subsequent to the intervention. If possible, subsequent-semester data should be collected on both the students who remained in the class during the intervention semester and students who departed. While more research on developmental mathematics students would be welcome, study journals' effects should also be investigated for other populations, such as developmental English students, credit-level students, and students at four-year universities.

In research on other populations, quantitative comparisons of exam averages may be less problematic, and may give valuable information about whether the study journals affected the achievement of the students. For research on developmental mathematics students or other groups with high proportions of extremely low non-passing grades, researchers need to think creatively. There may be better options than dichotomizing the final exam grade. For example, analyzing exam grades from earlier in the semester could dampen the effects of student attrition on the outcome measures. When quantitative evaluation of achievement is difficult, including a qualitative component can be especially valuable.

Future research on study journals should feature a higher level of teacher involvement. Action research, in which an individual teacher-researcher implements and evaluates the intervention in his or her own classes, is a possibility. Active involvement by the instructor would have two potential benefits. First, by reading the study journals and offering individual feedback to struggling students, instructors could provide students with both encouragement and with practical advice on how to improve their study habits. If a student is considering leaving the class, this personal connection with the instructor could make a crucial difference. Second, increasing the level of teacher involvement would provide a more authentic picture of how a study-journaling intervention functions in an actual class.

REFERENCES

Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *The Journal of Experimental Education. 78*, 487–512. doi:10.1080/00220970903352753

Ahuna, K. H., Tinnesz, C. G., & VanZile-Tamsen, C. (2011). "Methods of Inquiry": Using critical thinking to retain students. *Innovative Higher Education*, *36*, 249–259. doi:10.1007/s10755-010-9173-5

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education. 32*, 159–181. doi:10.1080/02602930600801928

Armstrong, W. B. (2000). The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community College Journal of Research and Practice*, *24*, 681–695. doi:10.1080/10668920050140837

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*, 2037-2049. doi:10.1002/sim.3150

Austin, P. C. (2011). Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in Medicine*, *30*, 1292-1301. doi:10.1002/sim.4200

Bail, F. T., Zhang, S., & Tachiyama, G. T. (2008). Effects of a self-regulated learning course on the academic performance and graduation rate of college students in an academic support program. *Journal of College Reading and Learning*, *39*(1), 54–73.

Baum, S., Little, K., & Payea, K. (2011). Trends in community college education: Enrollment, prices, student aid, and debt levels. The College Board. Retrieved from The College Board website. http://trends.collegeboard.org/sites/default/files/trends-2011-community-colleges-ed-enrollment-debt-brief.pdf

Bercher, D. A. (2012). Self-monitoring tools and student academic success: When perception matches reality. *Journal of College Science Teaching*, *41*(5), 26–32.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*, 161–186. doi:10.1016/S0959-4752(96)00015-1

Bol, L., Hacker, D. J., O’Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73*, 269–290. doi:10.3200/JEXE.73.4.269-290

Booth, A., Papaioannou, D., & Sutton, A. (2012). *Systematic approaches to a successful literature review*. London, England: Sage.

Brothen, T., & Wambach, C. (2000). A beneficial self-monitoring activity for developmental students. *Research and Teaching in Developmental Education*, *17*(1), 31–37.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4): 200-232. doi**:**10.1257/app.1.4.200

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245–281. doi:10.3102/00346543065003245

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cao, L., & Nietfeld, J. L. (2005). Judgment of learning, monitoring accuracy, and student performance in the classroom context. *Current Issues in Education*, *8*(4). Retrieved from http:// http://cie.asu.edu/articles/index.html

Cartnal, R. (1999, August). Preliminary success and retention rates in selected math courses. San Luis Obispo, CA: Cuesta College Matriculation and Research Services.

Cho, K., Cho, M., & Hacker, D. J. (2010). Self-monitoring support for learning to write. *Interactive Learning Environments*, *18*, 101–113. doi:10.1080/10494820802292386

Cisero, C. A. (2006). Does reflective journal writing improve course performance? *College Teaching*, *54*, 231–236. doi:10.3200/ctch.54.2.231-236

Commander, N. E., Valeri-Gold, M., & Darnell, K. (2004). The Strategic Thinking and Learning community: An innovative model for providing academic assistance. *Journal of the First-Year Experience and Students in Transition*, *16*(1), 61–76.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

D’Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281. doi:10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

D’Agostino, R. B., Jr., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, *95*, 749–759. doi:10.1080/01621459.2000.10474263

Dattalo, P. (2010). *Strategies to approximate random sampling and assignment.* New York, NY: Oxford University Press.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84*, 151–161. doi:10.1162/003465302317331982

Dietz-Uhler, B., & Lanter, J. R. (2009). Using the four-questions technique to enhance learning. *Teaching of Psychology*, *36*, 38–41. doi:10.1080/00986280802529327

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

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

Donovan, W. J., & Wheland, E. R. (2008). Placement tools for developmental mathematics and intermediate algebra. *Journal of Developmental Education*, *32*(2), 2–11.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*, 190–193. doi:10.1177/0098628312450432

Fitch, T., Marshall, J., & McCarthy, W. (2012). The effect of solution-focused groups on self-regulated learning. *Journal of College Student Development*, *53*, 586–595. doi:10.1353/csd.2012.0049

Fleming, V. M. (2002). Improving students' exam performance by introducing study strategies and goal setting. *Teaching of Psychology*, *29*, 115–119. doi:10.1207/S15328023top2902_07

Georgianna, S. (2009). Fostering student success: An exploratory study in English writing classes. *Journal of Applied Research in the Community College*, *17*(1), 20–29.

Gerhardt, M. (2007). Teaching self-management: The design and implementation of self-management tutorials. *Journal of Education for Business*, *83*, 11–18. doi:10.3200/joeb.83.1.11-18

Goodwin, M. M., & Califf, M. E. (2007). An assessment of the impact of time management training on student success in a time-intensive course. *Journal on Excellence in College Teaching*, *18*(1), 19–41.

Grabe, M., & Flannery, K. (2010). A preliminary exploration of on-line study question performance and response certitude as predictors of future examination performance. *Journal of Educational Technology Systems*, *38*, 457–472. doi.10.2190/et.38.4.f

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*, 334–372. doi:10.3102/003465430303953

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage Publications.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*, 101–121. doi:10.1007/s11409-008-9021-5

Hadwin, A. F., & Winne, P. H. (1996). Study strategies have meager support: A review with recommendations for implementation. *The Journal of Higher Education*, *67*, 692–715.

Hagedorn, L. S., & Kress, A. M. (2008). Using transcripts in analyses: Directions and opportunities. *New Directions for Community Colleges*, *2008*(143), 7–17. doi:10.1002/cc.331

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609–618. doi:10.1198/016214504000000647

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*, 219–236. doi:10.1214/08-sts254

Hartlep, K. L., & Forsyth, G. A. (2000). The effect of self-reference on learning and retention. *Teaching of Psychology*, *27*, 269–271. doi:10.1207/S15328023TOP2704_05

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, *66*, 99–136. doi:10.3102/00346543066002099

Haught, P. A., Hill, L. A., Walls, R. T., & Nardi, A. H. (1998). Improved Learning and Study Strategies Inventory (LASSI) and academic performance: The impact of feedback on freshmen. *Journal of the First-Year Experience & Students in Transition*, *10*(2), 25–40.

Hazelton, M. L. (2005). Kernel Smoothing. In *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd. doi:10.1002/0470013192.bsa329

Hill, J. L. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*, 2055–2061. doi:10.1002/sim.3245

Hill, J. L., Rubin, D. B., & Thomas, N. (2006). The design of the New York School Choice Scholarships Program evaluation. In D. B. Rubin (Ed.), *Matched sampling for causal effects* (pp. 328–346). New York, NY: Cambridge University Press.

Hilton, J. L., III., Wilcox, B., Morrison, T. G., & Wiley, D. A. (2010). Effects of various methods of assigning and evaluating required reading in one general education course. *Journal of College Reading and Learning*, *41*(1), 7–28.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236. doi:10.1093/pan/mpl013

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Hofer, B. K., & Yu, S. L. (2003). Teaching self-regulated learning through a "Learning to Learn" course. *Teaching of Psychology*. *30*, 30–33. doi:10.1207/s15328023top3001_05

Hopper, M. (2011). Student enrollment in a Supplement course for Anatomy and Physiology results in improved retention and success. *Journal of College Science Teaching*, *40*(3), 70–79.

Horn, L., Nevill, S., & Griffith, J. (2006). *Profile of undergraduates in U.S. postsecondary education institutions: 2003-04, with a special analysis of community college students.* (Report No. NCES 2006184). Retrieved from National Center for Education Statistics website: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006184

Humphrey, E. (2006). Project Success: Helping probationary students achieve academic success. *Journal of College Student Retention: Research, Theory & Practice*, *7*, 147–163. doi:10.2190/amq4-13ve-rbh7-6p1r

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*, 2857–2875. doi:10.1002/sim.3669

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *150*, 327–333.

Kaldeway, J., & Korthagen, F. A. J. (1995). Training in studying in higher education: Objectives and effects. *Higher Education*, *30*, 81–98. doi:10.1007/bf01384054

Kamphoff, C. S., Hutson, B. L., Amundsen, S. A., & Atwood, J. A. (2007). A motivational/empowerment model applied to students on academic probation. *Journal of College Student Retention: Research, Theory & Practice*, *8*, 397–412. doi:10.2190/9652-8543-3428-1j06

Kauffman, D. F. (2004). Self-regulated learning in Web-based environments: Instructional tools designed to facilitate cognitive strategy use, metacognitive processing, and motivational beliefs. *Journal of Educational Computing Research*, *30*, 139–161. doi:10.2190/ax2d-y9vm-v7px-0tad

Kauffman, D. F., Ge, X., Xie, K., & Chen, C. (2008). Prompting in Web-based environments: Supporting self-monitoring and problem solving skills in college students. *Journal of Educational Computing Research*, *38*, 115–137. doi:10.2190/ec.38.2.a

Kitsantas, A., & Baylor, A. (2001). The impact of the Instructional Planning Self-Reflective Tool on preservice teacher performance, disposition, and self-efficacy beliefs regarding systematic instructional planning. *Educational Technology Research and Development*, *49*(4), 97–106. doi:10.1007/bf02504949

Kwon, K., Kumalasari, C. D., & Howland, J. L. (2011). Self-explanation prompts on problem-solving performance in an interactive learning environment. *Journal of Interactive Online Learning*, *10*, 96–112.

Landers, R. N. (2011, November 16). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS [Web log post]. Retrieved from http://neoacademic.com/2011/11/16/computing-intraclass-correlations-icc-as-estimates-of-interrater-reliability-in-spss/

Lee, K. (2007). Online collaborative case study learning. *Journal of College Reading and Learning*, *37*(2), 82–100.

Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity: International Journal of Methodology*, *43*, 265–275. doi:10.1007/s11135-007-9105-3

Lipson, A., Epstein, A. W., Bras, R., & Hodges, K. (2007). Students' perceptions of Terrascope, a project-based freshman learning community. *Journal of Science Education and Technology*, *16*, 349–364. doi:10.1007/s10956-007-9046-6

Little, S. C. (2002). *Factors influencing the success of students in introductory algebra at a community college* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Order No. 3056473, University of Houston).

Lone Star College System. (2012). *Lone Star College System 2012–2013 catalog*. Retrieved from http://www.lonestar.edu/departments/curriculuminstruction/2012-2013_Catalog-Web.pdf

Lone Star College System. (2013). *Faculty class information* [internal electronic archive]. Lone Star College, The Woodlands, TX.

Lone Star College System Office of Research and Institutional Effectiveness. (2013). *LSC-North Harris fast facts 2012–2013*. Retrieved from http://www.lonestar.edu/images/LSC-North_Harris_Fast_Facts_Fall_2012-Fall_2013(2).pdf

Lone Star College System Office of Research and Institutional Effectiveness. (2011). *Math and Natural Sciences Division grade distribution* [internal memo]. Lone Star College, The Woodlands, TX.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, *29*, 530–558. doi:10.1177/0193841x05275596

MacArthur, C. A., & Philippakos, Z. A. (2013). Self-regulated strategy instruction in developmental writing: A design research project. *Community College Review*, *41*, 176–195. doi:10.1177/0091552113484580

Merriam, S. B. (1998). *Qualitative research and case study applications in education* (2nd ed.). San Francisco, CA: Jossey-Bass.

National Student Clearinghouse Research Center. (2012). *Term enrollment estimates: Fall 2012*. Retrieved from http://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2012.pdf

Orange, C. (1999). Using peer modeling to teach self-regulation. *Journal of Experimental Education*, *68*, 21–39. doi:10.1080/00220979909598492

Pallant, J. (2009). *SPSS survival manual* (3rd ed.). Crows Nest NSW, Australia: Allen & Unwin.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekarts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33–40. doi:10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*, 801–813. doi:10.1177/0013164493053003024

Provasnik, S., & Planty, M. (2008). *Community colleges: Special supplement to the condition of education 2008*. Retrieved from National Center for Education Statistics website http://nces.ed.gov/pubs2008/2008033.pdf

Reeves, T. D., & Stich, A. E. (2011). Tackling suboptimal bachelor's degree completion rates through training in self-regulated learning (SRL). *Innovative Higher Education*, *36*, 3–17. doi:10.1007/s10755-010-9152-x

Reynolds, C. L., & DesJardins, S. L. (2009). The use of matching methods in higher education research: Answering whether attendance at a 2-year institution results in differences in educational attainment. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 47–97). Netherlands: Springer. doi:10.1007/978-1-4020-9628-0_2

Rohwer, W. D., Jr. (1984). An invitation to an educational psychology of studying. *Educational Psychologist*, *19*(1), 1–14. doi:10.1080/00461528409529277

Rosenbaum, P. R. (1989). Safety in caution. *Journal of Educational Statistics. 14*, 169–173. doi:10.3102/10769986014002169

Rosenbaum, P. R. (1995). *Observational studies*. New York, NY: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi:10.1093/biomet/70.1.41

Rubin, D. B. (2006). *Matched sampling for causal effects*. New York, NY: Cambridge University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*, 20–36. doi:10.1002/sim.2739

Rubin, D. B. (2008a). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*, 1350–1353. doi:10.1198/016214508000001011

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*, 808–840. doi:10.1214/08-aoas187

Ryan, M. P., & Glenn, P. A. (2003). Increasing one-year retention rates by focusing on academic competence: An empirical odyssey. *Journal of College Student Retention: Research, Theory and Practice*, *4*, 297–324. doi:10.2190/kunn-a2ww-rfqt-py3h

Schapiro, S. R., & Livingston, J. A. (2000). Dynamic self-regulation: The driving force behind academic achievement. *Innovative Higher Education*, *25*, 23–35. doi:10.1023/a:1007532302043

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review*, *20*, 463–467. doi:10.1007/s10648-008-9086-3

Schunk, D. H., & Zimmerman, B. J. (1994). Self-regulation in education: Retrospect and prospect. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 305–314). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwartz, L. S., & Gredler, M. E. (1998). The effects of self-instructional materials on goal setting and self-efficacy. *Journal of Research and Development in Education*, *31*(2), 83–89.

Serna, A. D. (2011). *Remediation to college algebra: Factors affecting persistence and success in underprepared students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Order No. 3453831, University of South Dakota).

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:10.1037/0033-2909.86.2.420

Stanger-Hall, K. F., Shockley, F. W., & Wilson, R. E. (2011). Teaching students how to study: A workshop on information processing and self-testing helps students learn. *CBE Life Sciences Education*, *10*, 187–198. doi:10.1187/cbe.10-11-0142

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*, 2062–2065. doi:10.1002/sim.3207

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21. doi:10.1214/09-sts313

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. W. Osborne (Ed.), *Best practices in quantitative social science* (pp. 155–176). Thousand Oaks, CA: Sage.

Sweidel, G. B. (1996). Study strategy portfolio: A project to enhance study skills and time management. *Teaching of Psychology*, *23*, 246–248. doi:10.1207/s15328023top2304_14

Tashakkori, A., & Creswell, J. W. (2007). Exploring the nature of research questions in mixed methods research. *Journal of Mixed Methods Research*, *1*, 207–211. doi:10.1177/1558689807302814

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Texas Higher Education Coordinating Board. (2012). Developmental education accountability measures data: Lone Star College-North Harris (Fall 2011 Cohort). Retrieved from http://www.txhighereddata.org/reports/performance/deved/inst.cfm?inst=000722&report_type=2&report_yr=2012

Texas Higher Education Coordinating Board. (2013). Texas public higher education almanac: A profile of state and institutional performance and characteristics. Retrieved from http://www.thecb.state.tx.us/index.cfm?objectid=26B0039A-944A-C4D9-C6092B25A2C7BA27

Thoemmes, F. (2012). Propensity score matching in SPSS. Retrieved from http://arxiv.org/abs/1201.6385

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego, CA: Academic Press. doi:10.1016/B978-012691360-6/50005-7

Travers, N. L., Sheckley, B. G., & Bell, A. A. (2003). Enhancing self-regulated learning: A comparison of instructional techniques. *The Journal of Continuing Higher Education*, *51*(3), 2–17. doi:10.1080/07377366.2003.10400260

Tuckman, B. W. (2007). The effect of motivational scaffolding on procrastinators' distance learning outcomes. *Computers and Education*, *49*, 414–422. doi:10.1016/j.compedu.2005.10.002

Williams, A. E., Aguilar-Roca, N. M., Tsai, M., Wong, M., Beaupré, M. M., & O'Dowd, D. K. (2011). Assessment of learning gains associated with independent exam analysis in introductory biology. *CBE Life Sciences Education*, *10*, 346–356. doi:10.1187/cbe.11-03-0025

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ: Lawrence Erlbaum Associates.

Young, M. R. (2010). Transforming the initial marketing education experience: An action learning approach. *Journal of Marketing Education*, *32*, 13–24. doi:10.1177/0273475309335353

Ziegler, N. A., & Moeller, A. J. (2012). Increasing self-regulated learning through the LinguaFolio. *Foreign Language Annals*, *45*, 330–348. doi:10.1111/j.1944-9720.2012.01205.x

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*, 3–17. doi:10.1207/s15326985ep2501_2

Zimmerman, B. J. (1994). Dimensions of academic self-regulation: A conceptual framework for education. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 3–21). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zimmerman, B. J. (1998). Academic studying and the development of personal skill: A self-regulatory perspective. *Educational Psychologist*, *33*, 73–86. doi:10.1080/00461520.1998.9653292

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 1–37). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*, 64–70. doi:10.1207/s15430421tip4102_2

Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, *80*, 284–290. doi:10.1037/0022-0663.80.3.284

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, *53*(1), 141–160.

Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 289–307). Mahwah, NJ: Lawrence Erlbaum Associates.

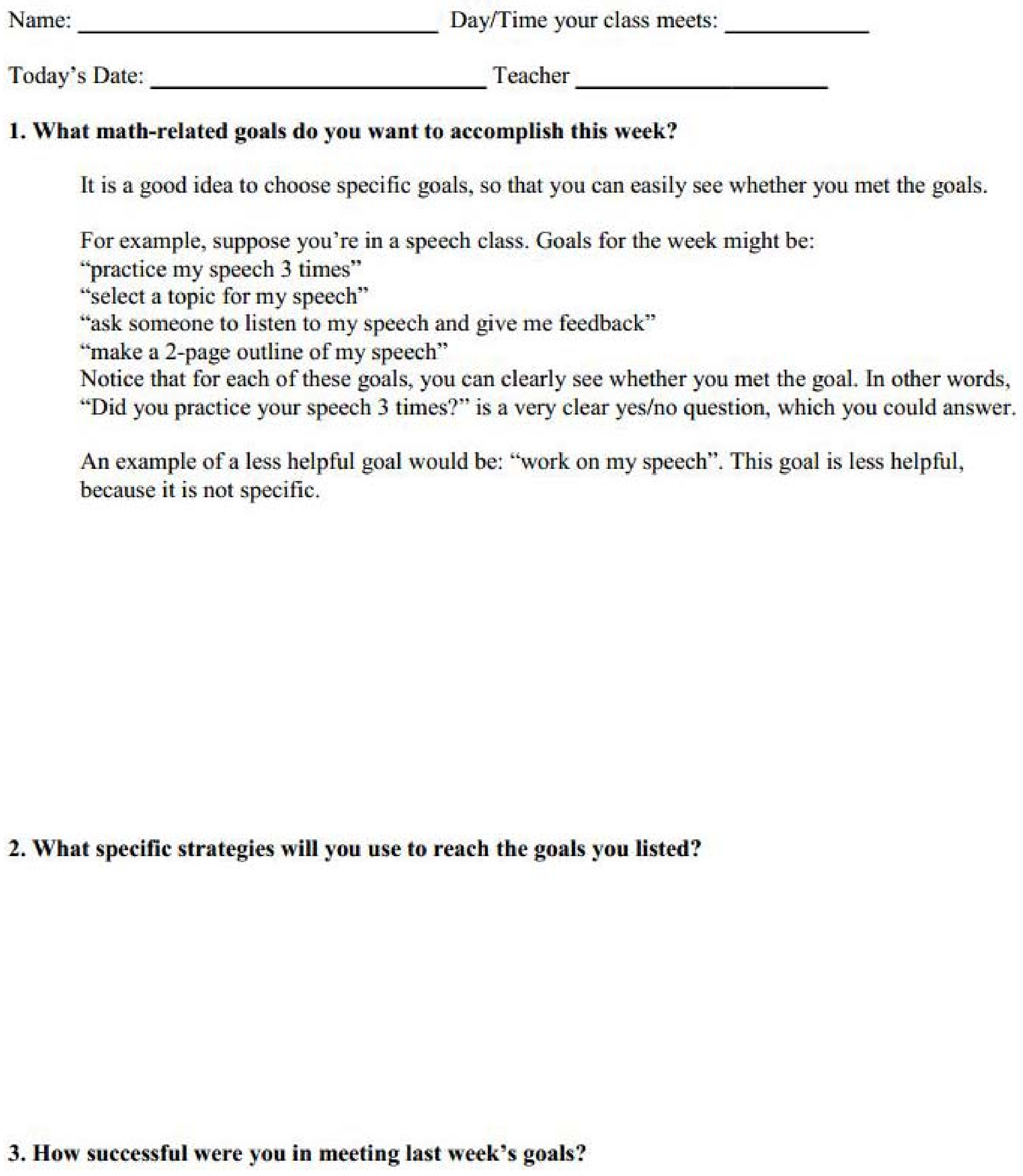
APPENDIX A

SEARCH TERMS FOR LITERATURE REVIEW

((SU.EXACT.EXPLODE("Study Skills" OR "Self Management" OR "Time Management" OR "Study Habits") OR all("study skill*" OR "self regulat*" OR "study habit*")) AND (SU.EXACT.EXPLODE("College Freshmen" OR "College Students" OR "Graduate Students" OR "Law Students" OR "Medical Students" OR "Premedical Students" OR "Preservice Teachers" OR "Two Year College Students" OR "Undergraduate Students" OR "Graduate Study" OR "Higher Education" OR "Undergraduate Study" OR "Graduate Medical Education" OR "Postsecondary Education" OR "Undergraduate Study" OR "College Instruction" OR "Colleges" OR "Community Colleges" OR "Dental Schools" OR "Law Schools" OR "Medical Schools" OR "Technical Institutes" OR "Two Year Colleges" OR "Universities")) NOT SU.EXACT.EXPLODE("Learning Disabilities")) AND (stype.exact("Scholarly Journals") AND pd(19940101-20121231))

APPENDIX B

GOAL SHEET



Name: ______________________________ Day/Time your class meets: ____________

Today's Date: ____________________________ Teacher _____________________

1. What math-related goals do you want to accomplish this week?

It is a good idea to choose specific goals, so that you can easily see whether you met the goals.

For example, suppose you're in a speech class. Goals for the week might be:
"practice my speech 3 times"
"select a topic for my speech"
"ask someone to listen to my speech and give me feedback"
"make a 2-page outline of my speech"
Notice that for each of these goals, you can clearly see whether you met the goal. In other words, "Did you practice your speech 3 times?" is a very clear yes/no question, which you could answer.

An example of a less helpful goal would be: "work on my speech". This goal is less helpful, because it is not specific.

2. What specific strategies will you use to reach the goals you listed?

3. How successful were you in meeting last week's goals?

4. If you did not meet some of last week's goals, what were the reasons?

5. How satisfied are you with your progress in this class so far?

6. What changes, if any, do you plan to make to your study strategies for this math class?

7. List the times you plan to study your math this week:

	Start Time	End Time	Other Notes
Monday			
Tuesday			
Wednesday			
Thursday			
Friday			
Saturday			
Sunday			

If you need assistance handling academic difficulties, you may contact the Math and Natural Sciences Division Counselor, Rhonda Cannon, at Rhonda.K.Cannon@lonestar.edu or 281-618-5480.

APPENDIX C

STUDY LOG

Name: ______________________________ Week beginning: __________________ Instructor/Class Time: _________________

Date & Day of Week	Start time	End time	Where?	With whom?	Goal(s) for Study Session	What did you do or work on?	How successful were you in reaching your goals?

APPENDIX D

FOCUS GROUP DISCUSSION GUIDE

On how the study journal project affected student success
Do you think writing down your goals has affected your success? How?
Do you think planning your study time has affected your success? How?
Do you think recording your study time has affected your success? How?

On study habits
Before this project, did you ever write down goals for your math class? Did you plan your study sessions in advance?
How well has your actual study time matched your planned study time?
What obstacles have hindered your studying?
How effective has your studying been so far in this class?
What has been your most helpful study strategy?
Has your approach toward studying for math changed as the semester progressed? How?
What can you do to improve the effectiveness of your studying?
What would you change about the way you have studied for this math class?
What advice would you give future Math 0308 students?
Do you think you will create written goals for other classes?

On the mechanics of the study journal process
How honest have you been in filling out the logs of your actual study time?
How often have you filled out the study log retroactively, rather than right at the time you studied?
How much of a burden/headache was it to log your study time?
How much of a burden/headache was it to complete the goal/reflection form each week?
How many of you chose to complete the study journal forms electronically?
If you used it, how difficult is the electronic process? Do you have any suggestions for improvement?
For those of you who chose to complete the study journals on paper: Why did you prefer paper?

Recommendations for future classes
Has filling out the goals/planning form been a good use of class time?
Would you recommend this study journal project be done in future math classes?
If this study journal project is done in future math classes, what are your suggestions for improving it?

APPENDIX E

INTERRATER RELIABILITY SCORES FOR FINAL EXAM ITEMS

Table E-1

Introductory Algebra (Math 0308) final exams: Intraclass correlation coefficients and agreement percentages for individual items and total exam grade.

Item Number	ICC(2,1)	Percent Agreement
1	0.923	93.75
2	0.505	87.50
3	0.888	87.50
4	0.839	81.25
5	0.889	75.00
6	1.000	100.00
7	1.000	100.00
8	0.903	81.25
9	0.838	62.50
10	1.000	100.00
11	0.974	93.75
12	0.894	68.75
13	1.000	100.00
14	0.980	87.50
15	0.762	68.75
16	0.776	93.75
17	1.000	100.00
18	1.000	100.00
19	0.842	81.25
20	0.982	93.75
21	0.868	93.75
22	0.929	75.00
23	1.000	100.00
24	1.000	100.00
25	0.859	81.25
26	0.618	56.25
27	0.933	75.00
28	0.922	81.25
29	0.953	81.25
30	0.957	93.75
31	0.981	75.00
32	0.964	87.50
33	0.750	68.75
34	1.000	100.00
35	1.000	100.00
36	1.000	100.00
37	1.000	100.00
Total Exam Grade	0.993	

Note. Intraclass correlations were calculated in SPSS specifying Two-Way Random, Absolute Agreement. and Single Measures

Table E-2

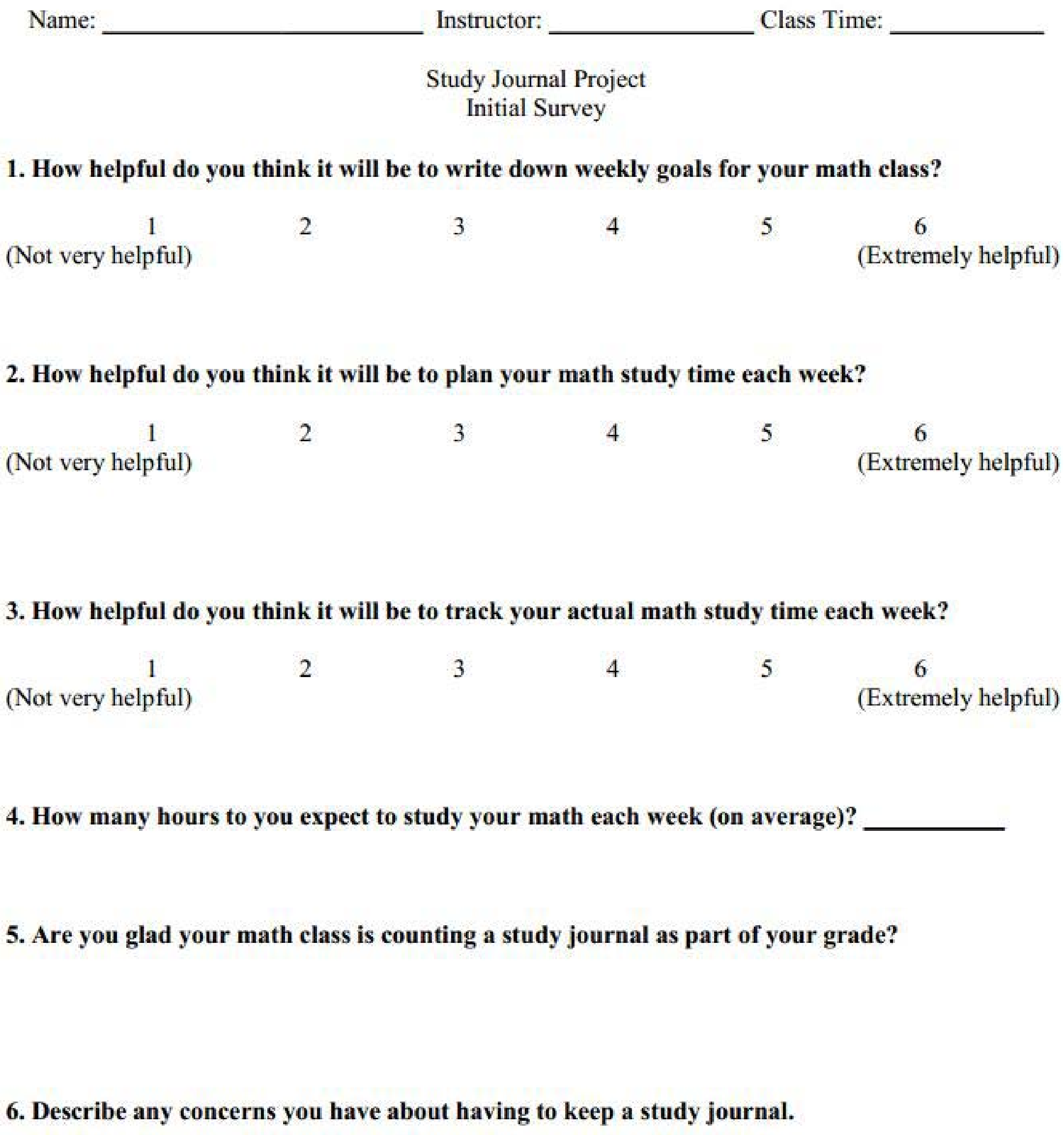
Intermediate Algebra (Math 0310) final exams: Intraclass correlation coefficients and agreement percentages for individual items and total exam grade

Item Number	ICC(2,1)	Percent Agreement
1	0.974	85.71
2	0.996	95.24
3	0.982	90.48
4	0.950	95.24
5	0.981	85.71
6	0.995	95.24
7	0.984	85.71
8	1.000	100.00
9	0.965	95.24
10	1.000	100.00
11	0.971	95.24
12	1.000	100.00
13	0.978	90.48
14	0.967	95.24
15	0.947	90.48
16	1.000	100.00
17	0.993	95.24
18	0.983	95.24
19	0.996	95.24
20	0.981	90.48
21	0.989	90.48
22	0.987	95.24
23	1.000	100.00
24	1.000	100.00
25	0.967	85.71
26	0.981	95.24
27	0.972	90.48
28	0.995	95.24
29	0.931	80.95
30	0.993	95.24
31	0.960	80.95
32	0.990	85.71
33	0.998	95.24
34	0.979	95.24
35	0.995	95.24
Exam Grade	0.999	

Note. Intraclass correlations were calculated in SPSS specifying Two-Way Random, Absolute Agreement, and Single Measures

APPENDIX F

BEGINNING-OF-SEMESTER SURVEY



Name: ________________________ Instructor: _______________ Class Time: ___________

Study Journal Project
Initial Survey

1. How helpful do you think it will be to write down weekly goals for your math class?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

2. How helpful do you think it will be to plan your math study time each week?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

3. How helpful do you think it will be to track your actual math study time each week?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

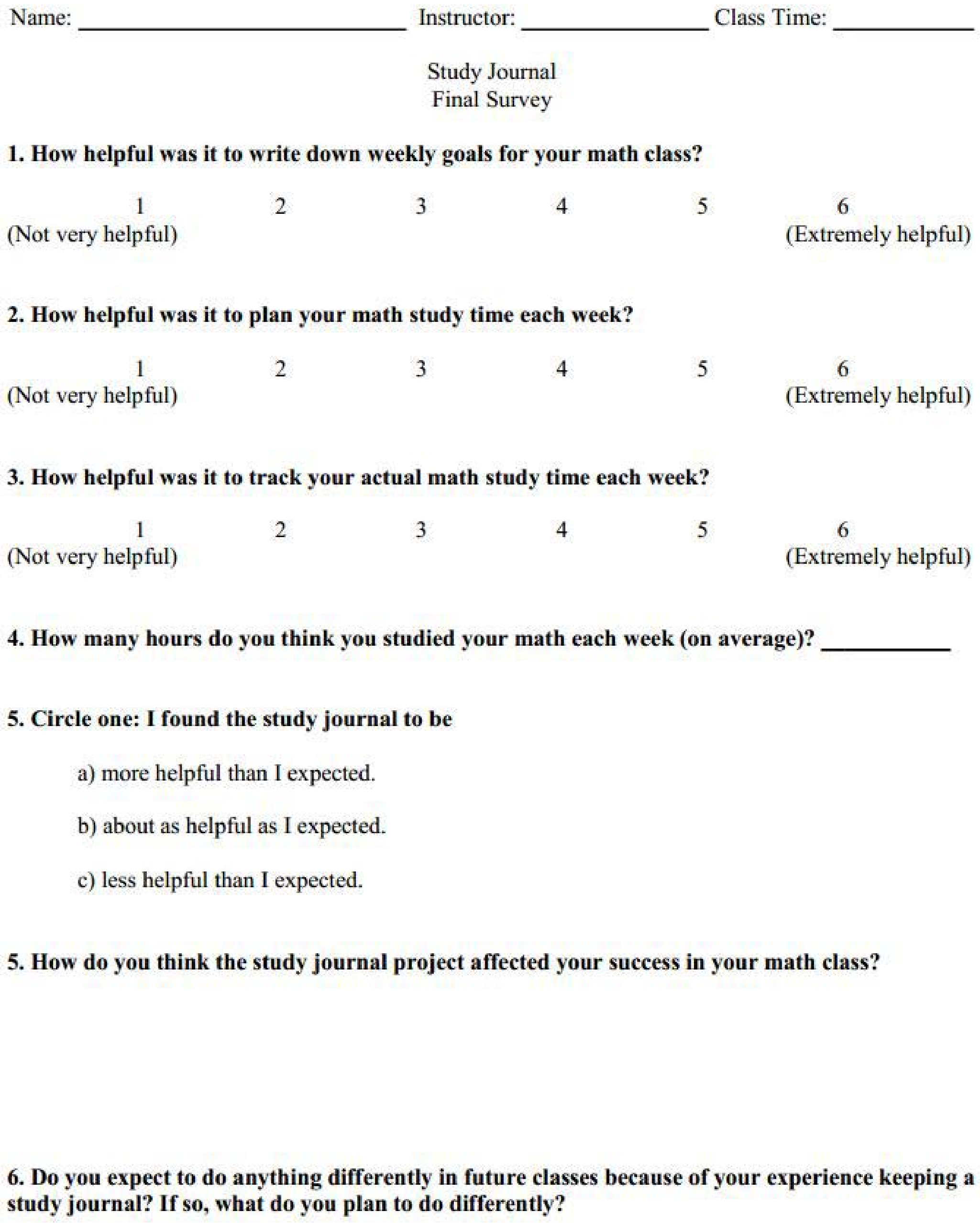
4. How many hours to you expect to study your math each week (on average)? __________

5. Are you glad your math class is counting a study journal as part of your grade?

6. Describe any concerns you have about having to keep a study journal.

APPENDIX G

END-OF-SEMESTER SURVEY



Name: ____________________ Instructor: ______________ Class Time: __________

Study Journal
Final Survey

1. How helpful was it to write down weekly goals for your math class?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

2. How helpful was it to plan your math study time each week?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

3. How helpful was it to track your actual math study time each week?

1 2 3 4 5 6
(Not very helpful) (Extremely helpful)

4. How many hours do you think you studied your math each week (on average)? __________

5. Circle one: I found the study journal to be

a) more helpful than I expected.

b) about as helpful as I expected.

c) less helpful than I expected.

5. How do you think the study journal project affected your success in your math class?

6. Do you expect to do anything differently in future classes because of your experience keeping a study journal? If so, what do you plan to do differently?

APPENDIX H

QUALITY RUBRIC FOR STUDY JOURNALS

	1	**2**	**3**	**Score**
Goal Sheet	Most weeks show evidence of minimal or no reflection, planning, or adaptation. Most entries look nearly the same from week to week. (Example: students write “study more” as their strategy every week).	Most weeks show some evidence of reflection, planning, or adaptation. Most entries look different from week to week indicating thoughtfulness about the student’s current goals and strategies.	Most weeks show evidence of deep reflection, detailed planning, or specific adaptations. Entries clearly indicate students are tailoring entries to the week at hand.	
Study Log (Entry = row that contains some writing, representing a study session.)	Most entries contain minimal information (few words, repetitive ideas, blanks).	Most entries show information specific to that study session (entries are not identical). Most entries contain some evidence of planning or reflection.	Most entries contain detailed information specific to that study session, including detailed planning or deep reflection.	

APPENDIX I

VARIABLES RECORDED FROM STUDENT TRANSCRIPTS

Table I-1

Variables Recorded From Student Transcripts

Variable Name	Description
HrsAtt2013	Hours Attempted as of February 2013. Includes grade-excluded hours (Cum Total Att at bottom of transcript).
HrsAttF2012[a]	Hours Attempted Fall 2012 (includes all subjects).
HrsAttSpr2013	Hours Attempted Spring 2013 (includes all subjects).
CumHrsAttPreInt [a]	Cumulative Hours attempted pre-intervention (all subjects) (HrsAtt2013-HrsAttF2012-HrsAttSpr2013).
CumCredPts2013	Cumulative credit-level points earned as of February 2013 (Cum Total Points from bottom of transcript)
CredPtsF2012Spr2013	Credit-level points earned during Fall 2012 and Spring 2013, not excluding any earned grades (includes A, B, C, D, F in credit-level courses) (could have Spring points due to minimester).
CredPtsPreInt	Credit-level points earned prior to intervention (CumCredPts2013-CredPtsF2012Spr2013)
HoursOmitted2013	Hours omitted from 2013 official college GPA except for pre-intervention developmental math hours (includes W's, IP's, all developmental English, Fall 2012 developmental math, all unfinished Spring 2013 classes, HUMD 0330 and other pass/fail classes, grade-excluded classes).
GradeExclusionHrs	Credit-level grade exclusion hours pre-intervention.
GradeExclusionPts	Credit-level grade points earned pre-intervention.
DevMathHrsPreInt	Developmental math hours attempted (A, B, C, F, IP, W) pre-intervention (does not include Fall 2012).
DevMathPtsPreInt	Developmental math grade points earned pre-intervention (does not include Fall 2012).
DevMathGPA [b]	Cumulative Developmental Math GPA Pre-intervention (NA if DevMathHrsPreInt=0, otherwise DevMathPtsPreInt/DevMathHrsPreInt). W's and IP's are counted the same as F's, as 0 grade points.
CredMathHrsPreInt	Credit-level math hours attempted (A, B, C, D, F, W) pre-intervention (does not include Fall 2012).
CredMathPtsPreInt	Credit-level math grade points earned pre-intervention (does not include Fall 2012).

Table I-1 Continued

Variable Name	Description
GPADenom2013	Credit-level hours Attempted as of February 2013 (HrsAtt2013-HoursOmitted2013-DevMathHrsPreInt).
GPAHrsF2012Spr2013	Credit-Level Hours used in GPA during Fall 2012 and Spring 2013 (A, B, C, D, F in credit-level classes F2012/Spr2013). Do not include hours of W, I, or grade exclusions, as these have already been subtracted from GPA denominator. Includes Spr2013 hours that already have a grade (minimester).
GPAPreInt [b]	Cumulative Credit GPA Pre-intervention, no exclusions (NA if GPADenom-GPAHrsF2012Spr2013+GradeExclusionHrs=0, otherwise (CumCredPts2013-CredPtsF2012Spr2013+GradeExclusionPts)/(GPADenom2013-GPAHrsF2012Spr2013+GradeExclusionHrs)).
CumGPA2013	Cumulative GPA as of February (NA if GPADenom=0, otherwise CumCredPts2013/GPADenom2013). Should equal GPA on transcript except for NA students. (NA students have 0.00 on transcript.)
CredEarned2013	Total Earned Hours (A,B,C,D,P) as of February 2013 (Total Earned at bottom of transcript. Includes credit, developmental, HUMD, and ESOL/ESL classes). Does NOT include hours earned for "grade excluded" classes, even if those hours were passed with a D.
CredEarnedF2012Spr2013	Total Hours Earned Fall 2012 and Spring 2013 (Total Earned bottom of Fall 2012 semester on transcript, plus as any Spring 2013 classes that already have a passing grade. Includes credit, developmental, HUMD, ESOL/ESL classes). Does NOT include hours earned for grade-excluded classes, even if those hours were passed with a D.
CredEarnedGradeExcluded	Credits earned in grade-excluded classes (classes excluded from GPA). (Typically grades of D and C).
CredEarnedPreInt [a]	Total hours earned prior to intervention. Does NOT include hours earned for grade-excluded classes, even if those hours were passed with a D. (CredEarned2013-CredEarnedF2012Spr2013). I did not include the grade-excluded classes, because that would have meant "double-dipping" (getting credit for the same class twice, if they got a D the first time and an A the second, for example.)
CourseCompletionRatio [b]	The proportion of hours attempted that have been passed pre-intervention. Includes Credit, Developental, HUMD, Pass/Fail, and Grade Excluded classes. (NA if CumHrsAttPreInt=0, otherwise (CredEarnedPreInt+CredEarnedGradeExcluded)/CumHrsAttPreInt). I included grade-excluded hours in this, because these hours are incorporated into both the numerator and the denominator, eliminating the "double-dipping" issue.

Table I-1 Continued

Variable Name	Description
Current Course [a]	Fall 2012 Math Course (0308 for Introductory Algebra, 0310 for Intermediate Algebra).
PrereqStatus[c]	*A*, *B,* or *C* if student passed the prerequisite course to the current course. *Repeat* if student has previously attempted current course and earned a D, IP, F, or W. *Placement* if this is student's first math course at this institution.
Attempts0306	Total attempts at Math 0306.
Attempts0308	Total attempts at Math 0308 (not including current attempt).
Attempts0310	Total attempts at Math 0310 (not including current attempt).
NumPassed0306	Number of times 0306 was passed prior to Fall 2012 (A, B, or C).
NumPassed0308	Number of times 0308 was passed prior to Fall 2012 (A, B, or C).
NumPassed0310	Number of times 0310 was passed prior to Fall 2012 (A, B, or C).
AttemptsPerPass [a]	Number of attempts per passed developmental math course pre-intervention. If PrereqStatus="Repeat" or "Placement"), then AttemptsPerPass=(Attempts0306+Attempts0308+Attempts0310+1)/(NumPassed0306+NumPassed0308+NumPassed0310+1),If PrereqStatus= A, B, or C, then AttemptsPerPass=(Attempts0306+Attempts0308+Attempts0310)/(NumPassed0306+NumPassed0308+NumPassed0310)
StartYear	Starting year at this institution.
YrsSinceStartCollege[d]	2012-StartYear.
LastMathYear	Year of last previous math class, regardless of whether it was successful (either this course or a different course).
YrsSinceMath[d]	2012-LastMathYear
ESOL[a]	1 if ESOL/ESL appears on transcript; 0 if ESOL/ESL does not appear on transcript .

[a]This was used as a matching variable. [b]This was transformed into a matching variable by creating an indicator variable (to capture NA values) and then imputing the mean to the NA students. [c]This was transformed into a matching variable by creating an indicator variable (to capture NA values), converting the original nominal variable to a numerical variable, and then imputing the mean to the NA students. [d]This was used as a matching variable after truncating the range.