2015

An Analysis of Student Learning Strategies in High Enrollment Computer-Based College Courses

gordon c. anderson
Umass-Amherst

AN ANALYSIS OF STUDENT LEARNING STRATEGIES IN HIGH ENROLLMENT COMPUTER-BASED COLLEGE COURSES

A Dissertation Presented

By

GORDON C. ANDERSON

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

Ph.D. in Computer Science

May 2015

School of Computer Science

AN ANALYSIS OF STUDENT LEARNING STRATEGIES IN HIGH ENROLLMENT COMPUTER-BASED COLLEGE COURSES

A Dissertation Presented

By

GORDON C. ANDERSON

Approved as to style and content by:

_____
Robert Moll, Chair

_____
David Jensen, Member

_____
Beverly Woolf, Member

_____
Stephen Sireci, Member

_____
Lori A. Clarke, Department Head
Computer Science

DEDICATION

This dissertation is dedicated to my patient and loving wife Katie, my daughter Lydia and son Gordy.

ACKNOWLEDGMENTS

I would like to thank my advisor, Robert Moll, for his many years of guidance and support. His contribution to my professional development in computer science, academics and research have been invaluable and will forever be appreciated. I would also like to extend my gratitude to the members of my committee, David Jensen, Beverly Woolf, and Stephen Sireci, for their infinite patience and helpful comments and suggestions. Additionally, I have benefitted immensely from Professor Woolf's course in Building Intelligent Tutoring systems, as well as from Professor Jensen's seminar in research methods and causality.

I also wish to thank the members of the Center for Educational Software Development (CESD) at UMass, especially Steve Battisti and Dave Hart, for providing support with the OWL system and database access, and in so many other ways for so many years.

I wish to give thanks to Professor Bee Botch for her time and advice on the chemistry course and her knowledge about chemistry education.

ABSTRACT

AN ANALYSIS OF STUDENT LEARNING STRATEGIES IN HIGH ENROLLMENT COMPUTER-BASED COLLEGE COURSES.

MAY 2015

GORDON C. ANDERSON, B.A., UNIVERSITY OF MASSACHUSETTS AMHERST

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Robert B. Moll

The subject of this dissertation is an observational investigation of the effect on outcomes of three behavior patterns students follow during the course of a fourteen week semester in two large, STEM (Science, Technology, Engineering, and Math) college courses: an introductory Computer Science course and an introductory Chemistry course. Both courses employed a computer-based system that recorded student textbook and homework activity. The behavior patterns we investigated are: 1) “book-first”- reading and interacting with the textbook material before working homework problems, 2) “infrequent-session”- long stretches of time between short working sessions with book or chapter homework material, and 3) “working-late”- submitting homework close to the due date. In order to assess the effect of these patterns on outcomes, i.e. final exam scores, we created features to measure the amount of textbook interaction, the relative length of time between work sessions, and the amount of work submitted close to due dates. Our analysis showed a statistically significant, positive effect of following a book-first strategy for students in all courses, and a greater positive effect for novice students. For the second study, we found evidence that the pattern of short working sessions with long intervals was related to lower exam scores. In the third study, we found a

negative effect of late work on final exam scores. We found novice students were more negatively affected than experienced students.

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

BACKGROUND

In this chapter, we provide a description of the context in which our research took place so the reader may better understand what we studied, our methods, and our findings. We start by describing the general problem domain we are interested in addressing and the benefits of doing research in this domain. We then describe that the organization of this dissertation is based on three studies of the effects of student behavioral patterns on outcomes. Next, we explain the structure of the courses we studied and the type of computer-based system that presented and evaluated the course material and logged the data. We then describe the demographics of the student populations and how they relate to outcomes. This is followed by a section on relevant research in our field and a section on methodology.

1.1 Problem Statement

An increasing amount of research has been directed at discovering how students learn in large college courses. By large, we mean courses with enrollments of 200 to 1500 or more students. The goal of much of this research is to evaluate the effectiveness of student learning behavior by relating patterns of student activity to outcomes. Most large-scale college courses employ a computer software application called a Learning Management System, or LMS, to help instructors manage the presentation, assignment and grading of course material. The LMS also records student activity in a database. The use of Learning Management Systems in higher education are particularly important for coping with the phenomenon of mass education [Johnson et al., 2012]. In the past, it was difficult to know how students used course materials; what activities they engaged in and when they did so. The fact that large courses typically

employ computer-based learning management systems (LMS) means researchers have an unprecedented amount of observational data to work with.

There are many benefits to knowing the effect certain behavioral patterns have on outcomes. By “knowing”, we mean that if the effects of these patterns are scientifically established rather than anecdotally reported, there is a reasonable basis for advising students about them and what has been found about their effects. Patterns that are positively related to outcomes could be encouraged and supported by modifications to an instructor’s pedagogical methods as well as to the design of course content. Software modules could be designed to automatically detect these patterns and provide feedback that encourages students to follow beneficial patterns if they are not already doing so. Patterns with a negative effect on outcomes could be discouraged. Furthermore, the patterns can be used to evaluate the effectiveness of new teaching techniques and content on changing student learning behavior for the better. Given the prevalence of large-scale, semi-automated courses in higher education, an understanding of student usage patterns is crucial for realizing the goal of improving the learning experience of many thousands of students.

This dissertation investigates the effect on exam scores of three student behavior patterns observed during the course of several fourteen-week semesters in two large, STEM (Science, Technology, Engineering, and Math) college courses: “CS121 Introduction to Problem Solving with Computers”, and “Chem111 Introductory Chemistry”. Both courses have large enrollments, approximately 450 and 550 respectively. Both courses use the same LMS and use an electronic text as well as automatically graded homework problem sets, both provided by the LMS. The material in both courses is presented to students via a series of assignments, where each assignment consists of sections of the text to read followed by a set of homework problems. The electronic text contains automatically evaluated problems, which are

"embedded" in specific locations to foster engagement with the material. These problems provide a means of observing student interaction with the text. This level of instrumentation means that we can track student activity in the textbook as well as on homework problems, allowing us to create features that represent the ways in which students interact with available online learning resources.

There are many questions one might attempt to answer given a large amount of LMS data. We chose three questions: 1) is it important for students to read their textbooks before starting homework problems? 2) is it detrimental for students to "lose touch" with the material by working in short sessions with long times between sessions? 3) should students avoid working close to a due date? In order to study these questions, we defined three patterns of student activity we can observe from the available LMS data. We refer to these patterns as: the "book-first" pattern, the "infrequent-session" pattern, and the "working-late" pattern. Each of these patterns serves as the basis for a study within this dissertation.

It is important to note that an LMS records only the work students do within the system. There are other, unobserved factors that influence outcomes. Students may do work outside the system, such as visiting an instructor or tutor, working with peers in study groups, referencing other sources of help, etc. Another unobservable factor is a student's "in class" experience. Both courses held on-campus lectures and discussion sections. The quality of these in-class experiences as well as the amount of attendance may have an impact on outcomes as well as the activity the LMS records. Another unobserved factor is the student's ability to learn the material, or "aptitude". We have no pretest or other data such as standardized test scores that would serve as a proxy for aptitude at this time.

Another set of factors that could affect outcomes are "demographics", such as class level, first year, second year, etc., gender, major, any previous experience with the subject

matter, and other socio-economic factors. We have survey data for class level, gender, major, and, in the Computer Science course, a measure of previous experience with programming. We did not have access to any other demographic data.

We next describe the three patterns we studied and our hypotheses about their effect on outcomes.

1.1.1 The Book-First Pattern

The material in the courses we studied is presented in units that contain textbook and corresponding homework assignments. Students who follow the book-first pattern read and interact with their textbook before they attempt the associated homework problems. In both courses, the textbook includes embedded problems designed to provide an immediate recall of the topic presented in the preceding text. The goal of this recall is to facilitate the acquisition of knowledge. In contrast, the homework problem sets are designed to exercise the acquired knowledge by presenting more thought-provoking problems to be solved. Thus, the content was designed to support two learning phases: an acquisition phase and an application phase, with the acquisition phase preceding the application phase.

We hypothesize that students who follow the book-first pattern will exhibit higher outcomes as evidenced in exam scores than students who do not. We further hypothesize that novice students will show a higher increase in exam scores than students with previous experience.

These hypotheses are predicated on the belief that students learn better if they first acquire the skills and knowledge from teachable content, such as the textbook in our courses, which is designed to facilitate acquisition, before they are asked to apply this acquired knowledge in homework activities. In contrast, an alternative pattern we refer to as book-last is

characterized by working on homework problems first, using the textbook as a reference in service of answering the homework problems. We believe the book-last strategy is inefficient for novice students because they spend a lot more time looking up information needed to solve homework problems than they would if they had studied the book material first. The textbook material is presented as a narrative, and we surmise that students learn best when new material is acquired in a narrative form. Working in book-last mode means the student does not receive the information in narrative form because they are accessing it in discrete pieces as needed to solve a problem with little context to the rest of the material. We do not investigate the book-last pattern in this dissertation, but mention it as an aid to understanding our rationale behind our hypothesis about the book-first pattern.

1.1.2 The Infrequent-Sessions Pattern

The second question we posed is based on our opinion that most students will find learning more difficult if they do not "stay in touch" with the course material. Results from research into retention [Roediger & Karpicke, 2006] has shown that learners retain more when they frequently have to recall new facts and skills. In our courses, students log in and work in blocks of time called sessions. Long intervals between relatively short sessions could mean that newly acquired learning will not be reinforced in a timely manner by another session and will not be retained. It could also be that this pattern is a proxy for a general lack of engagement with the course. We refer to this pattern of relatively short working sessions with long intervals between them as infrequent-sessions. We hypothesize that students who consistently follow the infrequent-sessions pattern will show significantly lower exam scores. We describe how we determine what constitutes "short" sessions and "long" intervals in section 3.3.2.

1.1.3 The Working-Late Pattern

The third study investigates the effect of working "at the last minute" on outcomes. Do students who do the majority of their work right before due dates suffer because they are not allowing themselves the time to deal with problems that crop up. After all, most homework style problems are designed to require more time as they involve the application of knowledge and skills, which requires time and consideration. Anxiety about finishing the work on time may also contribute to a lower quality learning experience. Students who so the majority of their homework within hours of the due date are following what we refer to as the *working-late* pattern. In chapter 4 we define more precisely the amount of time before a due date that we consider to be "late". We investigate whether working-late does affect learning outcomes and if so, how late is too late and who is most affected by working-late.

1.1.4 Organization of this Dissertation

The organization of this dissertation is as follows: First, we present an overview of the structure of the computer science and chemistry courses, CS121 and CHEM111 respectively, as well as a description of the automated course content managed by the OWL learning management system. We then provide an overview of the demographics of the student populations in the data we examined in this dissertation, followed by a section that describes previous relevant academic work. In the second chapter we present our study of the book-first strategy. We create a feature that represents the level at which students interact with the online text before doing their homework problems and analyze its effect on exam scores. In chapter 3 we present work on clustering student work sessions in an effort to understand the variance in outcomes for those students who follow distinct patters of working, such as working in infrequent short sessions. In chapter 4 we present a study of the effect of working-late- close

to due dates. We assess how many students work late, how late they work, and what effect, if any, late working has on exam scores. Chapter 5 contains conclusions and discussion of the results of this dissertation with plans for future work.

1.2 Overview of the Course Structure and Learning Management System

In this section we explain the way educational material is organized and presented in the courses we studied and describe the learning management system used for both courses. This information will give the reader a context for better understanding the studies that follow.

1.2.1 Course structure

We used data from two college courses given at the University of Massachusetts, Amherst. The first is Computer Science 121, or CS121, an introduction to problem solving in the Java programming language, and Chemistry 111, or CHEM111, an introductory course to general chemistry. Both courses have been taught using the OWL LMS (OWL for Online Web-based Learning), a system that has been developed at UMass-Amherst. Both courses consist of these major components: lectures, discussion sections, an electronic textbook, homework problem sets, and exams. The textbook and homework problems are delivered by a web-based LMS. We now describe the way the LMS content is organized and presented.

The course content is organized in a hierarchical fashion, as are most textbooks, with chapters, sections, and subsections. Chapters cover main topics, with sections and subsections covering smaller topics. Exams are given after a number of chapters have been covered. The Computer Science course has a shallower organization, with chapters as the main organizational unit, while the Chemistry course uses sections as its organizational unit. The figure below

provides a high-level overview of the amount and grouping of content between each exam in CHEM111 and CS121.

CHEM111
4 Chapters, 13 Sections, Exam1 4 Chapters, 12 Sections, Exam2 4 Chapters, 12 Sections, Exam3 3 Sections: Final Exam
CS121
Chapters 1 to 6: Exam1 Chapter 7 to 13: Final Exam

Figure 1: Course content organization: the sequence of chapters and exams.

The course content is presented to students as a sequence of assignments with due dates and a grade weight, or number of points the assignment is worth. For example, a chapter in CS121 (or a section in CHEM111) would consist of two assignments, a textbook assignment and a homework assignment.

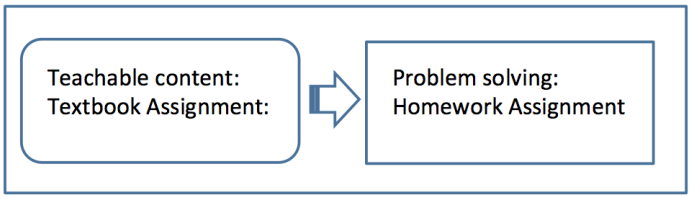


Figure 2: Assignment sequence for a CS121 chapter or a CHEM111 section.

The assignments are intended to structure the way students work through the material during the semester. The textbook and other teachable content is sequenced ahead of homework problems. Figure 3, below, provides a look at a typical assignment sequence for a chapter in CS121. Note that the textbook assignment consists of four sections.

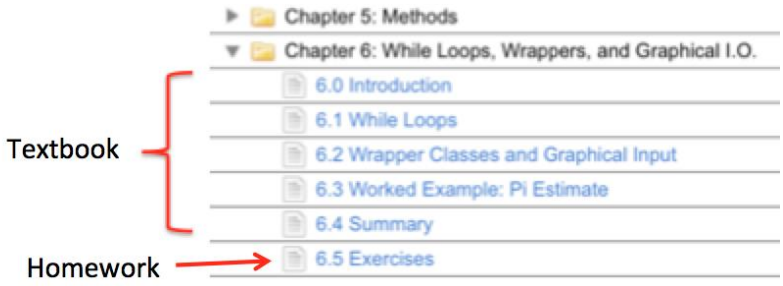


Figure 3: CS121 Assignments for chapter 6.

Next we describe the learning management system used to deliver and evaluate course content, the OWL system.

1.2.2 Learning Management Systems

A learning management system (LMS) is a software system that provides the functionality for administering and delivering educational technology [Ellis, 2009]. There are several designs for delivering course material online. One design is the Massive Open Online Course, or MOOC, which is a relatively new type of online educational experience [Mackness, 2010; Vaughan, 2010]. The enrollment is open and theoretically unlimited, hence the "massive" term in its definition. The curriculum is not as structured as in a traditional course. Students may choose to customize their "path" through a set of content, which is often accessed by links to distributed learning resources. These courses may have tens or hundreds of thousands of participants. Although some MOOC courses offer certificates of completion, the issues of grading and individual student authentication create difficulties for this model to be used for high stakes situations, such as earning college credit or professional licensing [Yuan, 2013; Garrison, 2004].

Another model of online learning follows the traditional college course paradigm, where enrollment is controlled by an institution such as a college or university, and a credential such as

credit and a grade is given for successful completion of the material. The model is high stakes because students must achieve a passing grade to receive the credential. In this model, the content is determined by an instructor and is the same for all participants. An LMS provides the functionality for course management and content delivery. This is the type of system that many academic publishers offer [Garrison, 2008].

This dissertation will study the latter model, where a LMS presents and evaluates course content for college credit. Individual students must authenticate to use the system and usage activity is tracked by the LMS. The course content is posted and controlled by the instructor; outside resources may be included as ancillary material. Students may access the system at any time under any circumstances. Collaboration is usually encouraged, however; students are cautioned to turn in their own work for grading. Final exams are proctored. From this point forward, all discussion pertains to this type of high stakes model.

1.2.3 The OWL System

The OWL (Online Web-based Learning) system is the LMS used in this research. OWL is used at the University of Massachusetts, Amherst in approximately 20% of first and second year students, mostly in large enrollment STEM courses. OWL provides a set of comprehensive services including an instructor tool set, content authoring capabilities, as well as a student interface. OWL records student actions in one database, making data mining easier than collecting data from many sources, which is the case in many computer-based systems.

In the OWL system, students see their assignment list, which lists their textbook and homework assignments. The figure below depicts a collapsible assignment menu in the OWL system. Each folder is a chapter. The chapter contains approximately one week's worth of work.

As you can see, within each chapter folder are links to two assignments: a textbook (OWLBook) assignment and a homework assignment (Exercises in Figure 3).

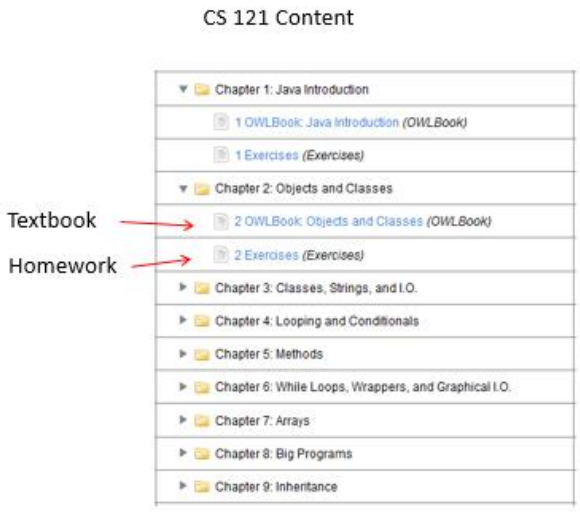


Figure 4: Partial view of course structure of CS121.

A note on terminology: we may use the term “module” interchangeably with “assignment”. A module is a unit of assigned work. A module, or assignment, in OWL may be linked to various types of activities such as tutorials, textbook pages, projects, quizzes, exams, and homework problems. A module has setting for due date, points worth, maximum attempts allowed, and time limits. All of the assignments in our courses have unlimited time and attempts until the due date. In CS121 courses, assignments are called Chapters, where in CHEM111 they are called Sections.

OWL is used by large courses in part because it allows content authors and developers to write custom evaluation software for course content. This means that most of the questions in OWL are automatically graded, with immediate feedback provided to students—a boon for instructors of highly enrolled courses. A developer and content author (they could be the same individuals) may create a custom input and evaluator for use in the OWL system. For example,

we developed an evaluator for the CS121 course that runs a Java process. That means that the OWLBook and homework questions are running the Java language. Students type actual code as responses to the questions. Their code is compiled, run, and evaluated for correctness by our custom evaluator. In chemistry, various custom evaluators have been developed. For example, a custom evaluator can evaluate the correctness of graphical representations of chemical formulas.

1.2.4 OWLBook electronic text

In a previous section we described types of activities one might find in the acquisition phase of a unit of content. A central activity in this phase is reading the course textbook. If a paper version of a textbook is used, we have no way of knowing how and when students access their book. Anecdotal evidence suggests that many students do not read their text, rather, they use it as a lookup reference in service of doing the homework problems. This is understandable, as the text pages may be "assigned" but are not counted for points. Another aspect is time management: it takes more time to read the text before doing homework problems, which are worth points. It's more efficient, in the minds of many students, to just do the homework. It would be interesting to study the efficiency of working homework with and without having read the text before.

At this time, many college courses use an electronic version of the textbook; either as an alternative to the paper version, or as the sole version. An electronic text can take many different forms. The simplest form is a collection of documents or pages in a document, such as pdf or html files. This form of electronic text does not allow for detailed information about student's use of the book. If the text is downloaded, the download event is the only information we can get. For web-based pages, we can record each page access, or "hit"; however, the time

the student has spent reading the page is difficult to assess. After a page hit, a student may go to get a coffee, or chat with a friend, or do other tasks.

In the OWL system, we have developed a version of an electronic text called an "OWLBook" that is fully integrated into the LMS. Furthermore, the OWLBook pages contain "embedded" questions that are automatically evaluated in real time. Textbook authors may embed questions in strategic locations to help reinforce key concepts. Typically, the questions are asking the student to echo a skill or concept that was presented in the text above. Almost without exception the embedded problems are designed to be relatively easy – they are primarily a device to encourage students to read the text. Embedded questions come with a time stamp for completion. Each embedded question submission triggers communication with the OWL server; the book responds in real-time with a correctness judgment on the submitted solution, along with some diagnostics when incorrect solutions are submitted. Student interactions with the text are logged, and statistics on aspects of student performance are provided to instructors. An example of an OWLBook page is given below.

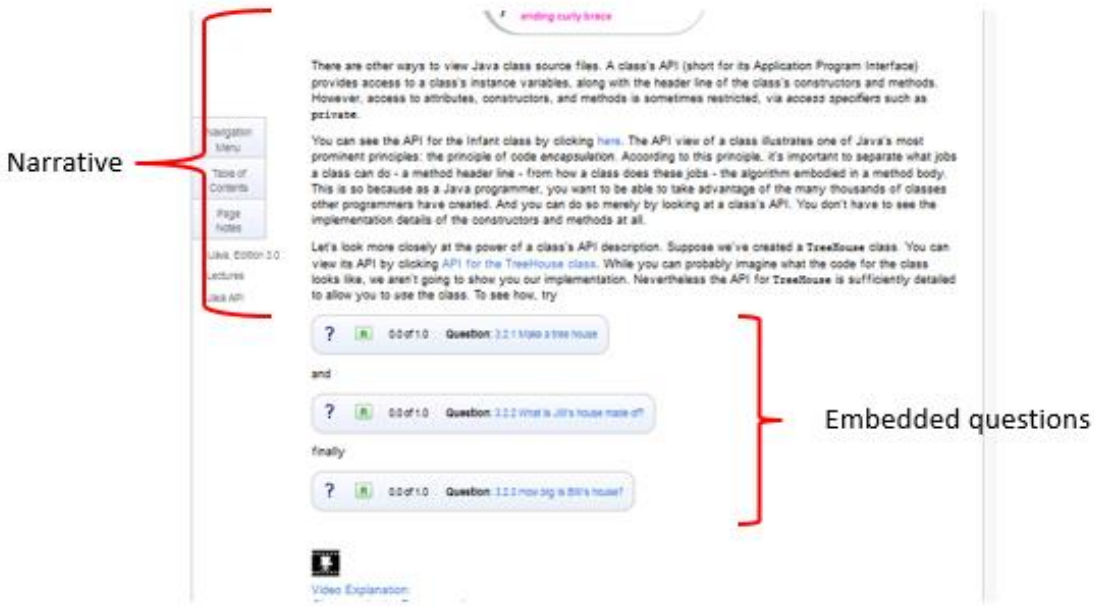


Figure 5: A section of an OWLBook page showing the narrative and embedded questions.

The idea is that students will interact with their text rather than passively encounter the material. The principle here is consistent with the concept of "test-enhanced learning" [Roediger et al., 2006], which claims that learning styles are not as important for learning as frequent recall of acquired material. The fact that the pages and embedded questions are assignable means that the textbook is now in a different status compared to the homework problems, the typical focus of attention.

For our study, the key aspect of the OWLBook is that all student submissions to the embedded questions are recorded in the system database. This allows us to gather data on student book usage in a way that was not possible before. Although we still do not know if the student is fully engaged with the text, the fact that the student is answering embedded questions is a good indication that they are engaged with the text on some level. This time stamped, event data is only the beginning of our work in attempting to give meaning to the way students use the system, but it is crucial.

In addition to embedded questions, each course has more than 200, somewhat harder "chapter problems"—these end-of-chapter exercises, we refer to as "homework problems" are also automatically graded. With both embedded problems and end of chapter exercises students are permitted to make as many attempts as they like until a correct answer is entered.

1.3 Overview of Course Data

In this section we describe some of the general statistics about the courses we use in our study to familiarize the reader with the "landscape" of the courses. We gather data from two high-enrollment courses taught on the University of Massachusetts campus: Chemistry 111, or CHEM111, and Computer Science 121, or CS121. We have data for four semesters of CS121 and one semester of CHEM111. The chemistry data is a useful check on the generality of our findings as it differs in many ways from the CS course, mainly in the type of materials and student population. All data in this dissertation are for students who took the final exam and responded to the post course survey. We have a better than 90% survey response rate for each course. A brief summary of the five courses and their enrollment numbers shows the growing enrollment in the CS c121 course.

1.3.1 Course Enrollment and Assignment Characteristics

Table 1: CS121and CHEM111 enrollments for students who took the final exam and responded to course surveys.

Semester	Enrollment
CS121 Fall 2012	345
CS121 Spring 2013	387
CS121 Fall 2013	465
CS121 Spring 2014	454
CHEM111 Fall 2012	556

The course content includes homework and textbook material, assigned over the fourteen week span of the semester. The general view of chapters (modules) and exams was given in a previous section. Below we provide the details on the number of assignments and questions in both courses. It is important to have an idea of how much work is assigned and to know how much work, in time and attempts on questions, that students are putting in to the course. We calculate the percent of assigned work attempted before the due date as well as the number and duration of the working sessions for each student as a way to judge their level of engagement, a key factor in their success in the course. In this dissertation we are concerned with analyzing the effects of certain student strategies. A student's level and quality of engagement is an important factor to control for in analyzing these effects.

Table 2: The number of assignments and individual questions in each course.

Course	Textbook Assignments	Textbook Questions	Homework Assignments	Homework Questions
CS121	13	160	13	197
CHEM111	40	335	40	373

From the above table, we can see that there is much evaluated content in both courses. In both courses, the OWL system automatically evaluates student question submissions.

Given the amount of assigned material, how much time and effort do students put in to these courses? The table below shows the sheer number of attempts and time in hours recorded for a CS121 and CHEM111 course (one section). Students access the course via a web application. Their work is defined in periods called sessions. To give a sense of the number of these sessions for both courses, we report the number of work sessions per student as well. The table below summarizes the attempts and time for students in the CS121 and CHEM111 courses. The four CS121 course values are averaged. Only students who took the final exam are included in this table.

Table 3: Summary of total number of attempts and time on questions per student. The number of work sessions is also included. The CS121 values are averaged over the four courses.

Course	Textbook question attempts/per student	Homework attempts/per student	Time (hours) /per student	Number of work sessions/per student
CS121	504	621	103	351
CHEM111	373	505	119	214

From the above table, it seems CHEM111 students put in more time but fewer attempts for OWLBook questions, though they have almost twice the number of questions to answer. There are differences in the type and style of questions and activities between CHEM111 and CS121. The chemistry material has many more, smaller assignments called “Sections”. The CS121 courses have fewer assignments called “Chapters”. The vast majority of CS121 questions are based on a live Java environment. That means that student submissions must have perfect syntax to pass by the Java compiler, and then must execute correctly. Therefore, we expect that CS121 students would make more attempts at their questions, many on account of syntax errors. This means we should not draw conclusions about the numbers in the table above alone, without considering the impact that the type of activities has on attempts and time. The figure below shows a typical distribution of time for one semester of CS121.

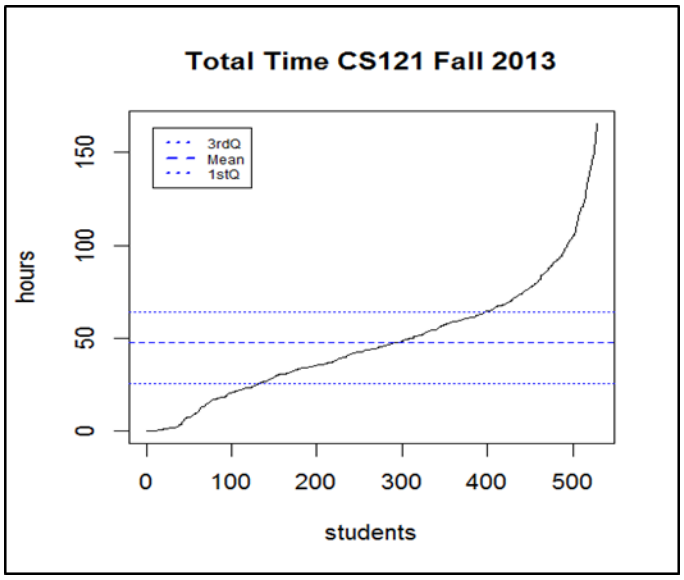


Figure 6: Total hours spent in Owl CS121 Fall 2013.

This plot shows that in extreme cases, a student would have been putting in 10 hours or more per week on OWL work alone. It is often interesting to look at subpopulations with extreme values as they may reveal an effect of the extreme behavior. For example, taking a relatively large amount of time to do a problem could mean that the student is struggling with the material, or may be suffering from not having learned prerequisite material. The above distribution is typical of many features we analyze. The figure below shows an example of the percent of assigned homework done before the due dates for CS121 Spring 2014.

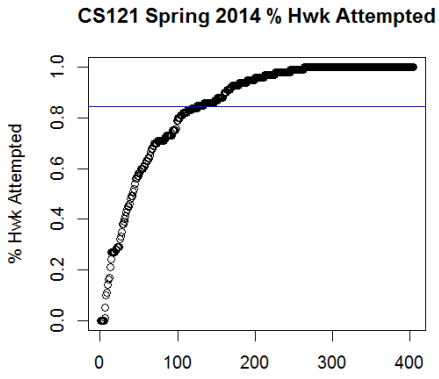


Figure 7: Percent of assigned homework done before the due dates for CS121 Spring 2014

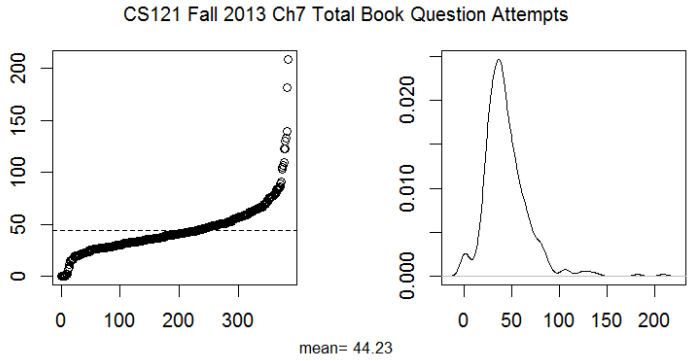


Figure 8: Example of a distribution of OWLBook attempts for a chapter in CS121.

1.3.2: A measure of student engagement: PC_HWK

The amount of assigned work done by a student is likely correlated with patterns of behavior and outcomes and needs to be adjusted for in our studies. We calculated the proportion of assigned homework attempted for each student as a proxy for their level of engagement, of participation. We call this variable PC_HWK. The homework is worth more points than the book questions, and therefore is given a higher priority than the book questions by many students. Therefore, PC_HWK should be a reasonable proxy for a student's participation level.

a = distinct homework questions attempted before due date
b = total number of homework questions assigned

PC_HWK = a/b

The PC_HWK measure is calculated for each unit (chapter or section). These scores are averaged over all assignments in the courses. The aggregate version is called PCAvg. The following plots depict the distributions of these averages and indicate a level of student participation.

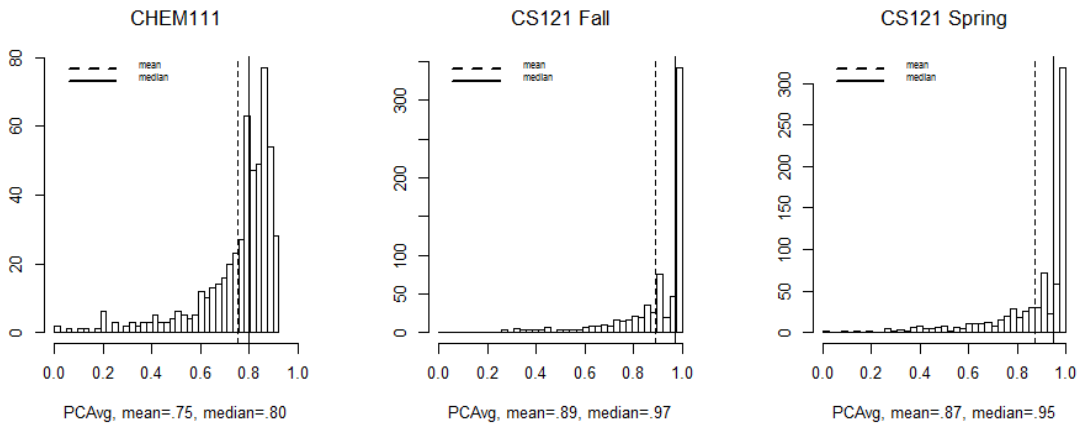


Figure 9: Distribution of percent of assigned homework attempted (before the due dates). The values are course averages.

It is clear that most student are participating at a high level. There are many more, smaller, assignments in CHEM111 which may explain the more disperse distribution. The relationship between PCAvg and final exam scores is plotted in the figure below.

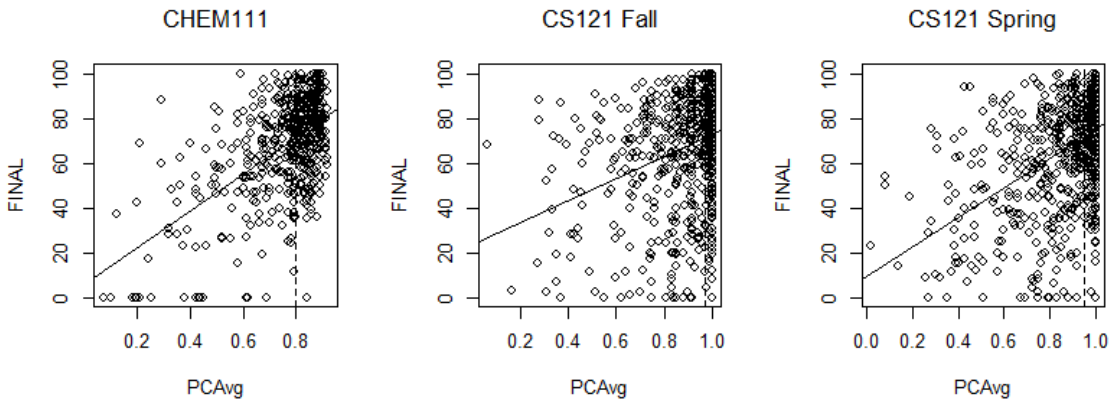


Figure 10: Plots of PCAvg vs. final exam scores. Regression and median lines are included.

It is evident that there is a positive relationship between the amount of the assigned homework attempted and final exam scores, though fairly noisy and not especially linear. Correlation calculations are presented in the following table.

Table 4: Correlation between PCAvg and final exam scores.

	Pearson's r, 95% conf.
CHEM111	.57, [0.51 0.62]
CS121Fall	.30, [0.23 0.36]
CS121Spring	.48, [0.42 0.53]

1.3.3 Subpopulations

Next we present some data on the composition of subpopulations of interest in the student data for CS121 and CHEM111. The subpopulations are defined by variables that represent student attributes we believe to have an effect on learning behavior as well as outcomes. These variables are also likely to have some effect on how students engage with the material. The values of these variables were obtained by survey responses. The survey was given for the CS121 courses at the end of each semester with more than a 90% response rate. The CHEM111 course gave a similar survey with a similarly high response rate. The variables we are

interested in for our study include class level, gender, major, and, for CS121, previous programming experience. The survey questions and response levels are presented below.

Table 5: Survey questions for CS121 and CHEM111.

Survey Question	Response Range
What is your class level?	1 1st-year undergraduate 2 2nd-year 3 3rd-year 4 4th-year
What is your gender?	1 Female 2 Male
What is your major?	1 Chemistry or Biochemistry 2 Biology 3 Food, health, exercise, or other life science 4 Other physical science, such as physics, geosciences 5 Environmental science 6 Engineering 7 Computer science 8 Mathematics 9 Humanities 10 Social sciences 11 Business 12 Undeclared undergraduate 13 post-grad or graduate student
Which of these statements most fits your situation?	1 Before I took this class, I had never programmed before in any language. 2 Before I took this class, I had programmed before, but not in Java. 3 Before I took this class, I had programmed before, and I have had some exposure to Java.

Survey data is problematic in that it is self-reported, and thus open to bias. Students may not answer truthfully, or may not interpret a question in the way it was intended. Furthermore, by using only students who took the survey (it was optional), we are potentially introducing more bias into our population as the students who participate in the survey may also be more likely to follow a pattern of usage we are studying, and to do better on exams. We looked at the population of students who did not take the survey versus the students who did take it with

regard to their final exam scores and the percentage of homework attempted as a gauge of the level of participation in the assigned material.

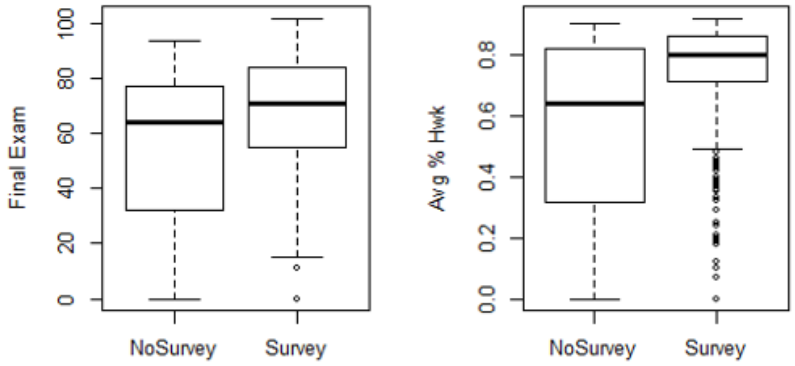


Figure 11: Comparison of final exam and percent homework attempted for the survey taking population and the non-takers for CHEM111.

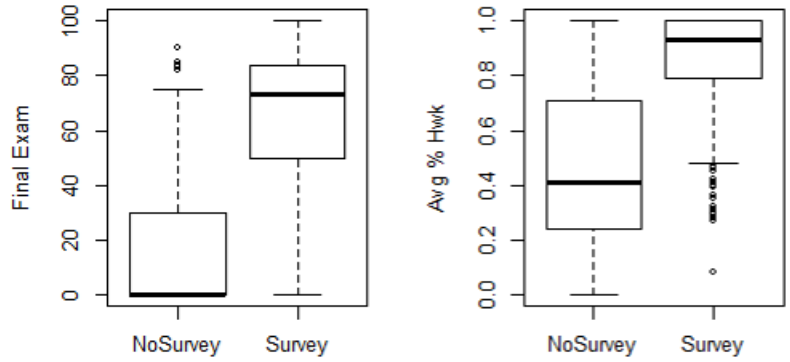


Figure 12: Comparison of final exam and percent homework attempted for the survey taking population and the non-takers for CS121- all courses combined.

Fall 2012 CHEM111 92% took the survey, CS121 95% took the survey. It is clear that the survey is "selecting" the students who are participating at a high level, which is the majority of the population. This is especially true for the CS121 courses. It is also clear that the non-survey takers suffered in the final exam, most likely due to low participation in the assigned work. The information presented above suggests that we are removing the lowest participating students

from our study by considering only survey respondents. This means that we are studying mostly engaged students, which is the population we are interested in because they provide us with more and fuller examples of patterns of learning behavior.

We next describe each survey variable in terms of the proportion of its levels in the data.

1.3.3.1 Class level

First year students represent the largest of the class levels (years 1-4) in all data sets. We believe they are mostly novices in that they have less experience with the study habits and amount of material presented at the college level, while third or fourth year students have had time to learn how to be students, and we suspect that first year students who struggle with how to study are more likely to drop out.

The pie charts that follow give a graphical view of the proportions of subpopulations in CS121 and CHEM111 courses. All data presented in these charts include only students who took the final exams and responded to the survey.

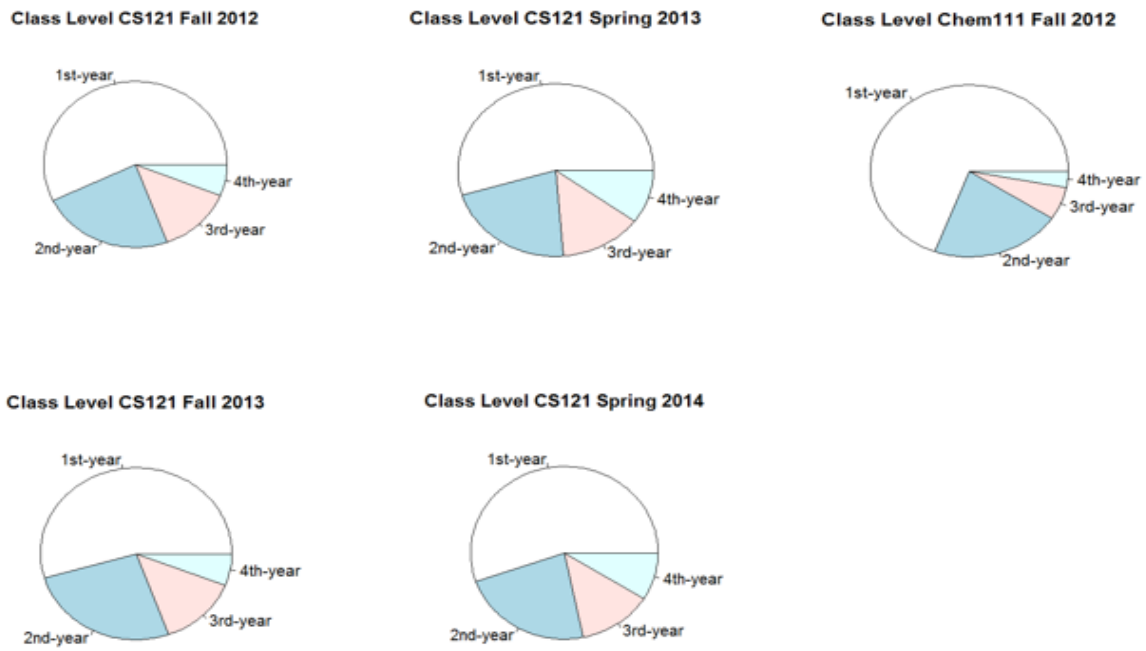


Figure 13: Composition of CS121 and CHEM111 courses by class level.

It is interesting to note that first year students make up a larger proportion of the CHEM111 course. The class level proportions are very consistent in the CS121 courses.

Table 6: Percentages of class levels.

Course	Class Level
CS121 Average	1- 56% 2- 24% 3- 13% 4- 5%
CHEM111	1- 70% 2- 21% 3- 6% 4- 2%

1.3.3.2 Gender

With regard to gender, we have observed differences in the strategies that females take versus males. Female students tend to attend lecture at a greater proportion, and tend to do

slightly more of the homework than males, although this is not consistent over the four courses we include in this study.

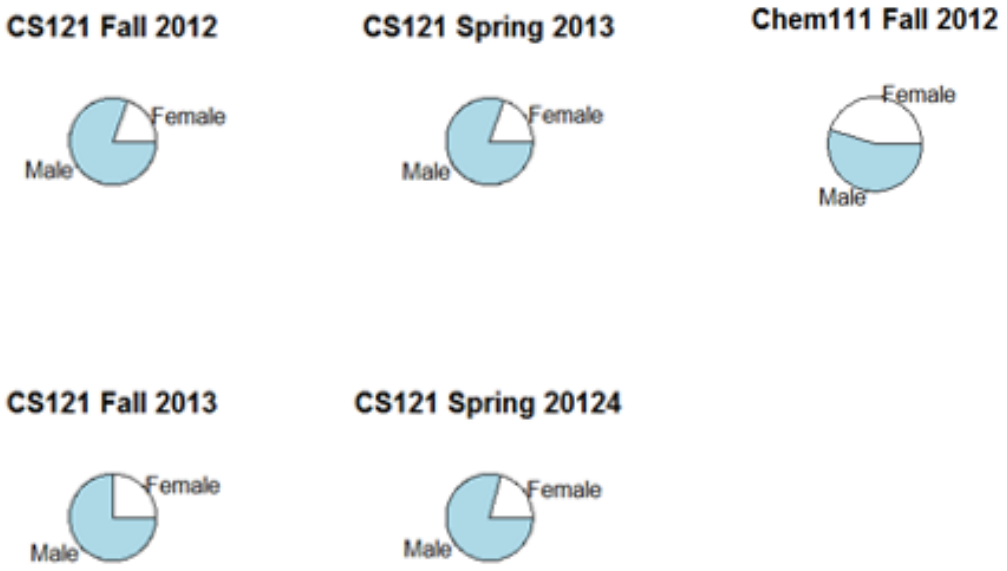


Figure 14: Composition of CS121 and CHEM111 courses by gender.

Gender in CS121 courses is consistently about 20% female, which contrasts with 44% females in CHEM111.

Table 7: Percentages of females and males.

Course	Gender
CS121 Average	20% female
CHEM111	44% female

1.3.3.3 Major

We next look at the composition of majors in our data. From the figures below it is apparent that the majority of CS majors take the fall semester of CS121, while engineering majors take CS121 in the spring semester.

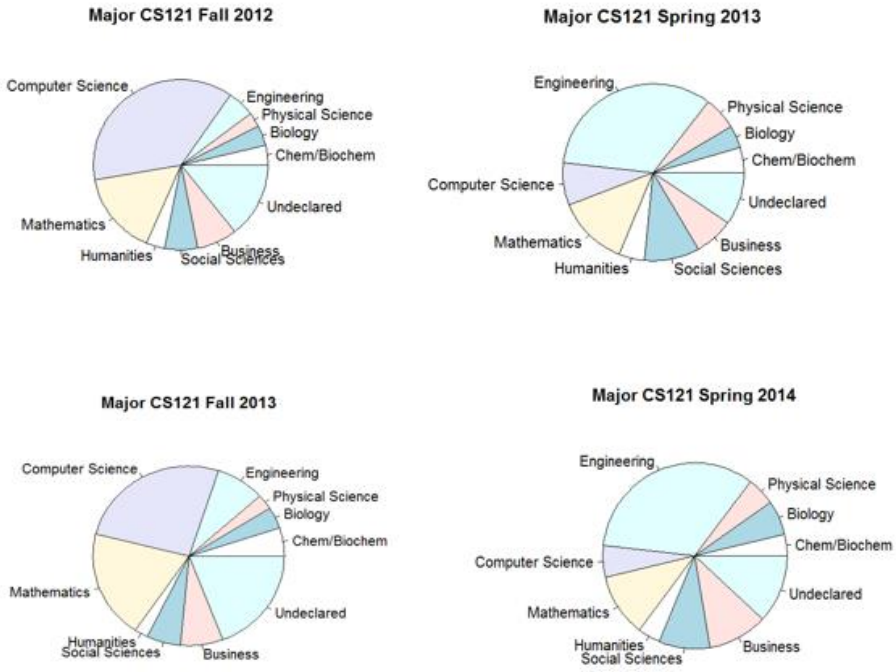


Figure 15: Composition of the CS121 courses by major.

There is a large mathematics major contingent in both semesters. This is because CS121 is a required course for that major. We tend to see second or third year students under this major. The undeclared category is also large in both semesters. These students are probably applicants to the CS major, and they usually do well as a group on outcomes. In our view, it's likely that computer science and engineering majors will have more motivation and perhaps aptitude to learn the material in CS121 and probably tend to have experience with some kind of computer programming.

Table 8: Percentages of majors.

Course	Top Four Majors
CS121 Fall Average	CS- 37% Math- 13% Undeclared- 13% Engineering- 5%
CS121 Spring Average	Engineering- 33% Math- 12% Social Science- 10% CS- 7%
CHEM111	Biology- 70% Engineering- 12% Undeclared- 11% Chem/Biochem- 6%

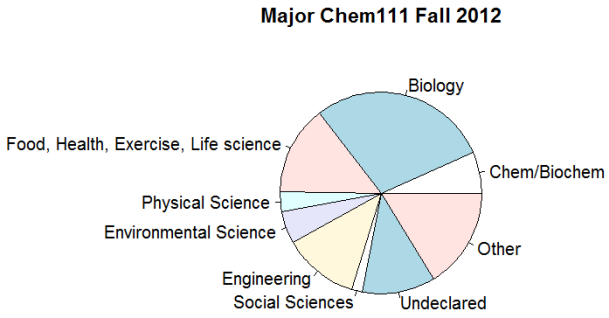


Figure 16: Composition of the CHEM111 course by major.

In the CHEM111 data, we see that Biology and life sciences are the largest component. It is interesting to note that chemistry and biochemistry students make up only 6% of the CHEM111 population in the fall semester. This is because chemistry majors typically take a different course at the beginning level. We therefore do not think that chemistry majors are a category we have to be concerned about in terms of adjusting for any advantage in outcomes.

1.3.3.4 Categorization of Major

The MAJOR variable has 13 levels, many of which represent quite small subpopulations with similar exam score distributions. We combined these subpopulations into two or three groups to make analysis easier. Our method was to group majors together by their median and mean final exam scores. The following graphs show the distributions of scores for each major, along with a plot of their median vs mean scores. The right hand plots includes lines we used to form groups. The groups are represented by numerical levels for a variable named MAJ_GROUP.

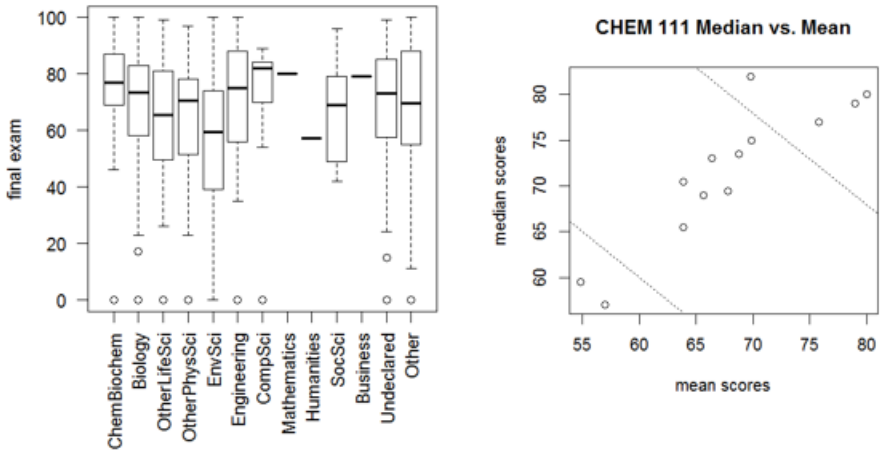


Figure 17: Final exam distributions and median vs mean plot of MAJOR for CHEM111.

The following table summarizes the majors in their new groupings.

Table 9: Summary of major groupings for CHEM111.

Group (MAJ_GROUP)	Major
1 N=44	Mathematics, Business, ChemBiochem, CompSci
2 N=442	Engineering, Biology, Other, Undeclared, SocSci, OtherLifeSci, OtherPhysSci
3 N=27	Humanities, EnvSci

The following graphs and tables depict the new groupings for the two CS121 courses. Note that we are pooling both fall semesters and spring semesters to make two data sets: CS121 Fall and CS121 Spring.

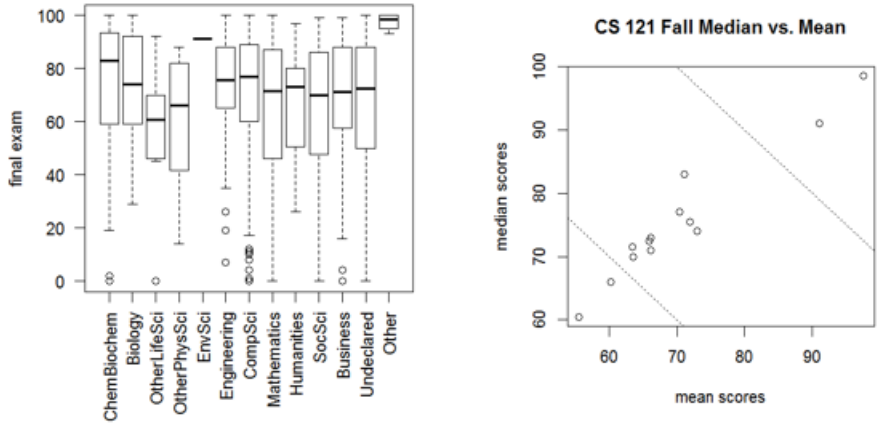


Figure 18: Final exam distributions and median vs mean plot of MAJOR for CS121 Fall.

Note: The majors: OtherPhysSci and OtherLifeSci, with 4 and 3 students respectively, were combined with group 2. We choose to merge these data points into the second category due to their low numbers.

Table 10: Summary of major groupings for CS121 Fall.

Group (MAJ_GROUP)	Major
1 N=18	EnvSci, Other
2 N=690	Mathematics, Business, ChemBiochem, CompSci, Engineering, Biology, Undeclared, SocSci, Humanities,
3 N=7 (merged with group 2).	OtherLifeSci, OtherPhysSci

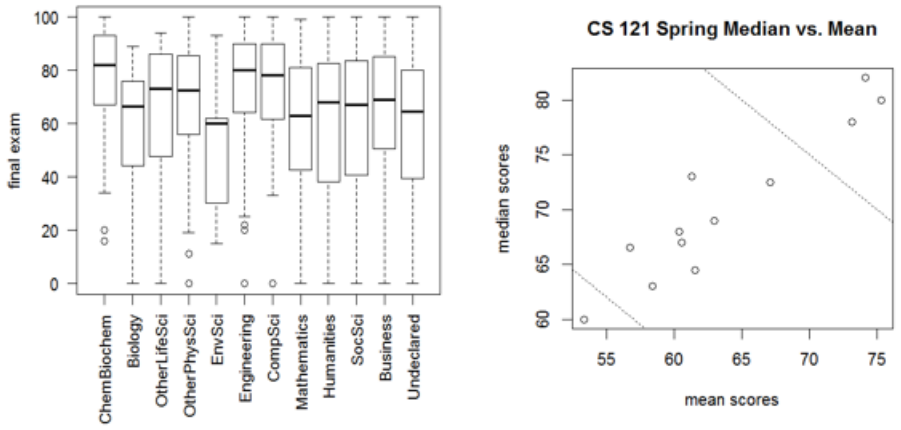


Figure 19: Final exam distributions and median vs mean plot of MAJOR for CS121 Spring.

Note: We chose to add group 3 to group 2 due to the low membership of this major.

Table 11: Summary of major groupings for CS121 Spring.

Group	Major
1 N=324	Engineering, ChemBiochem, CompSci
2 N=423	Mathematics, Business, Biology, Undeclared, SocSci, Humanities, OtherLifeSci, OtherPhysSci
3 N=5 (merged with group 2)	EnvSci

To summarize: we created groupings of the MAJOR variable based on final exam scores. We refer to this new variable as MAJ_GROUP. We combined the majors into three groups for CHEM111, and two groups for CS121, as shown in the table below.

Table 12: Summary of groupings of MAJOR for each data set.

Data set	MAJ_GROUP levels
CHEM111	1 N=44, 2 N=442, 3 N=27
CS121 Fall	1 N=18, 2 N=697
CS121 Spring	1 N=324, 2 N=428

1.3.3.5 Previous experience

As with CS majors, we believe that students with some level of programming experience, especially Java experience will have an easier time of the first portion of the CS121 course, which teaches Java. We do not have a measure about chemistry student's previous experience. The survey for the CHEM111 course did ask about previous chemistry courses taken, however, 95% of the respondents had answered that they had taken a chemistry course in high school. Therefore, we do not include any measure of previous chemistry experience in this dissertation.

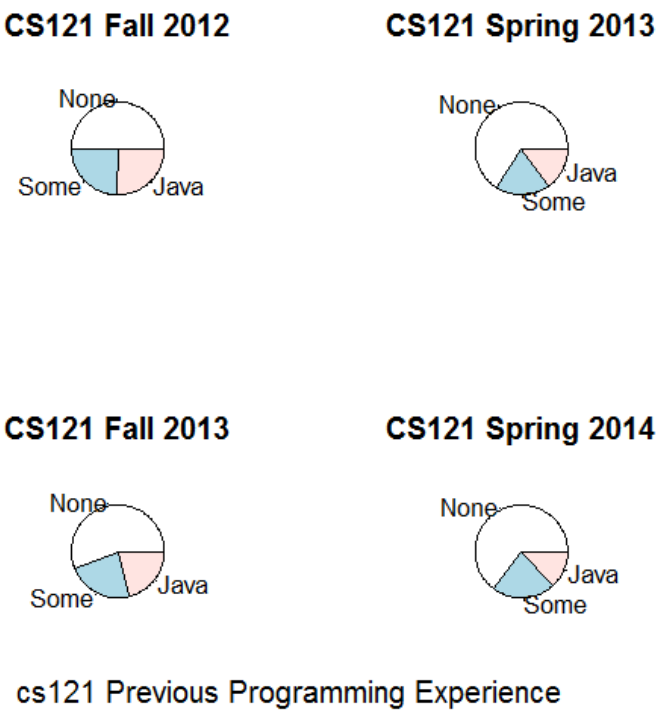


Figure 20: Previous programming experience in CS121.

In the figure above shows that the spring semesters contain a larger proportion of students who have no previous programming experience than the fall semesters. This is probably because the computer science majors are more likely to have previous experience, and relatively few of

them take CS121 in the spring semester. Also of note is that the number of students who claim "some" programming experience are roughly the same as those who claim Java experience. The survey question states: "Before I took this class, I had programmed before, but not in Java". It may not be clear what programming means to students. There are many computer activities which might count as programming to a student, but would not fit our idea of what the question is getting at. Since it is not clear how students may interpret this question, we will leave this group out of our analysis.

Table 13: Percentages of students with previous programming experience.

Course	Programming Experience
CS121 Fall Average	54% none 25% some 25% java
CS121 Spring Average	66% none 20% some 14% java

1.3.3.6 Combinations of interest.

We present graphs of two combinations of the survey variables of interest: gender vs class level, and gender vs. previous experience (in the CS121 courses). We are curious about the distribution of female and male students as there is much research about differences learning styles based on gender [Lau & Yuen, 2009; Arroyo et al., 2006. 2000; Murphy, 2006].

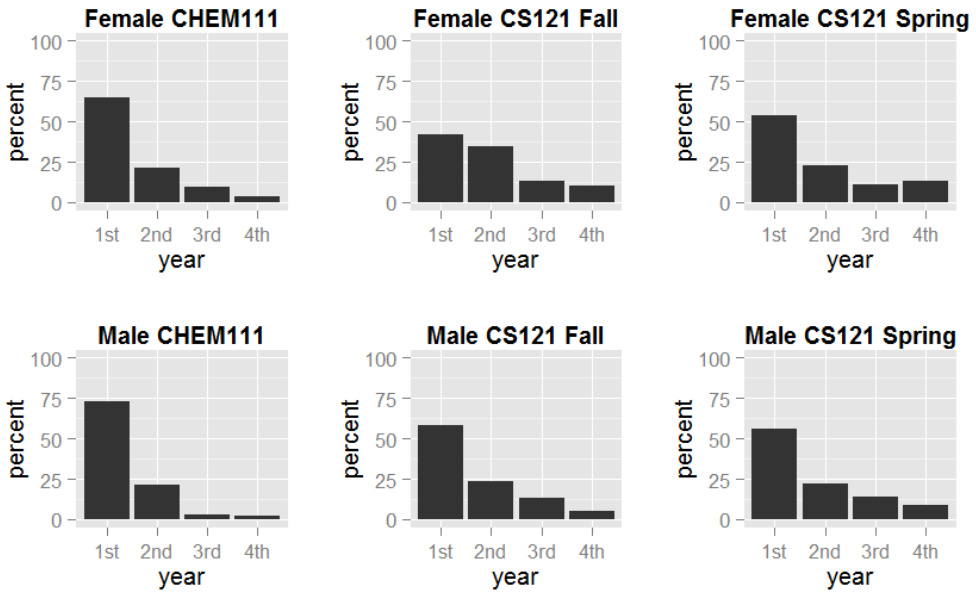


Figure 21: Proportion of the percentages of female and male students in each class level.

The figure above shows that there are proportionally slightly more 1st year males than females in CHEM111. The proportion of 2nd year females is larger than males in CS121 fall semester courses. The proportions are quite similar for the spring semesters of CS121.

The figure below shows that a much higher proportion of females reported no previous programming experience than males for both semesters of CS121. This means that the percentage of female students who are taking CS121 have somewhat less experience programming than their male counterparts. In the following section we see that females do as well as males on average on the final exam.

Figure 22: Proportion of the percentages of female and male students in each category of previous experience for the CS121 courses.

We next look at the distributions of outcomes, i.e. final exam scores, for each subpopulation described above.

1.3.3.7 Subpopulations and Outcomes

The following box plots show how each subpopulation fared on the final exam. In the first plot, below, we see that there are slight differences in gender across all courses. This is heartening for CS121 courses as we found above that a greater proportion of females reported less programming experience.

The result for class level is also displayed in the first plot. We see that first year student also do as well or better than the rest of the students. Most students take these courses as first year students as they are required for several majors. Students from the latter years may be taking the courses as electives or may have postponed taking what they perceive to be more challenging courses.

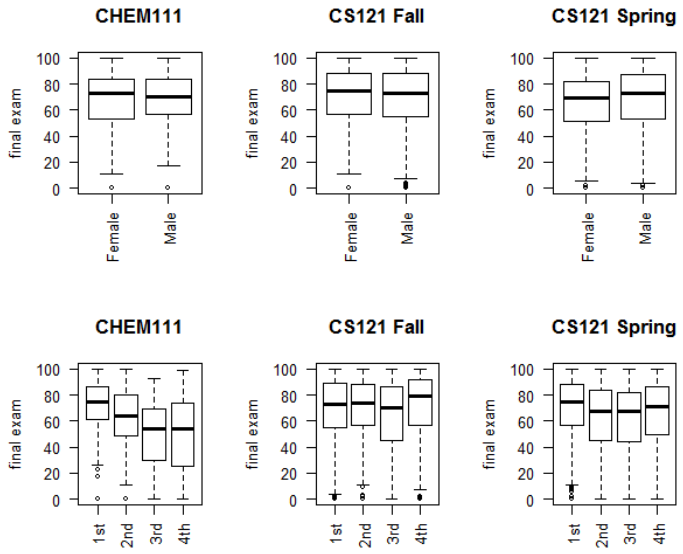


Figure 23: Box plots for gender and class levels vs. final exams.

The plot below shows the final exam distribution for our grouping of majors, which we created to have these distributions in section 1.3.2.4.

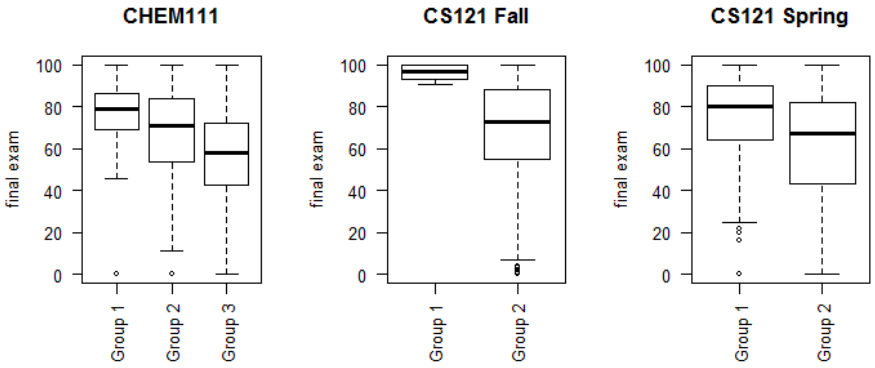


Figure 24: Box plots for major groups vs. final exams.

The next plot of previous programming experience in the CS121 courses shows a higher median for Java-experienced students for the spring but not the fall semesters. The reason for this difference may be the fact that the fall semester has a large contingent of CS majors, who, as

the plot in figure 24 shows, do better on average that the rest of the students according to major.

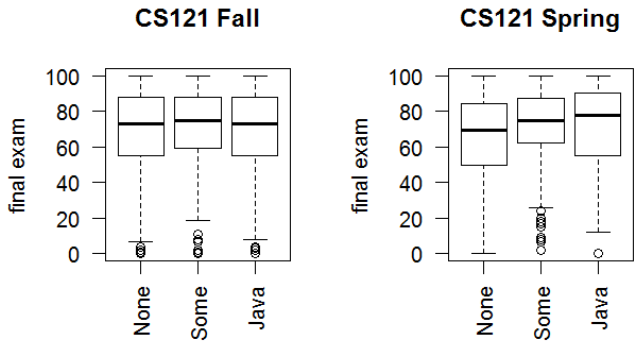


Figure 25: Box plots for previous programming experience vs. final exam for CS121 courses.

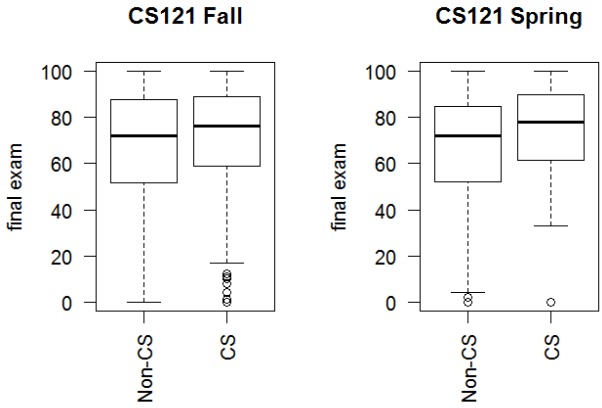


Figure 26: Box plots for CS majors and the rest vs. final exam for CS121 courses.

The following table summarizes the differences in outcomes between the levels of each survey variable.

Table 14: Summary of survey variables vs. final exam scores. Comparisons are listed as greater than if a significance difference between means by t test was found ($p < .05$).

Course	Gender	Class level	Major group	Prev prog. exp.	CS Major
CHEM111	F>M	$1^{st}>2^{nd}$>rest	1>2>3	NA	NA
CS121 Fall	F>M	no sig exam diff.	1>2	no exam sig diff.	CS > rest
CS121 Spring	M>F	1^{st}>rest	1>2	Java>None	CS > rest

1.4 Relevant Work

In this section, we present a review of academic work that is relevant to the studies in this dissertation. The goal is to provide the reader with a context to better understand the system and data as well as the techniques we utilized. This section is organized as follows. First, we briefly describe the major components of a computer-based learning system. Then, we discuss relevant background on strategic thinking. This is relevant to our work because the behavioral patterns we study are considered examples of strategic thinking. Next, we compare and contrast results from previous studies that have relevance to our work. Finally, we present background in the main methodology we utilized in this dissertation.

1.4.1 Computer-based educational systems

The basic elements of a computer-based learning environment are defined in Woolf's book: "Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning" [Woolf, 2008]. She defines the basic components as: Content Model, Pedagogical Model, and Student Model. The Content Model is the structure of the educational content of the system. The Pedagogical Model is the teaching strategy used by the system and/or implicit in the content design, including presentation order, hint and feedback structure. The Student Model represents the current state of learning for a student using the system. This model could consist of a set of beliefs about the level of a student's mastery of the topics being taught. In our

case it is the pattern of learning strategies a student has chosen to follow. In Intelligent Tutoring Systems [Woolf, 2008; Polson, 1988; Chambers, 1983], the system makes pedagogical decisions based on the current state of its student model. In contrast, the OWL system used in our studies does not yet implement strategies on its own, and so the Pedagogical Model is provided by the instructor and the content designer. Our research may provide useful directions for developing student models upon which the OWL LMS may make its own pedagogical decisions.

Our work is related to research in the field of educational psychology, especially the topics of how learners approach learning new information. In terms of computer-based education, there are two main communities of research that are relevant to our work: learning analytics and educational data mining. Learning analytics is the "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs"[Eli, 2011]. Another definition is that Learning Analytics is: "… the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data." [Powell et al., 2002]. The Journal of Learning Analytics is a peer-reviewed research publication of the Society for Learning Analytics Research (http://www.solaresearch.org/).

There are three main, interrelated aspects of Learning Analytics [Eli, 2011]:

1. Provisioning of data—Gathering data from different sources that may be of variable quality, poorly integrated and not designed for accessibility and require the development of a data warehouse.
2. Interpretation and Visualization—Working with practitioners to develop an understanding of how data held on systems can be used to inform the enterprise's activities and presenting information in an accessible and informative way and identification of additional data requirements.

3. Actioning insights—Processes by which practitioners and learners can turn insights into actions within their context.

We are involved in the first two of the main aspects of learning analytics. Fortunately, we have good access to most of our data, which is stored in a relational database. We are interpreting the data by building feature sets and models of student behavior. We rely on visual representations of our data and models to understand the relationships between the covariates and outcomes. We are not yet involved in the third aspect: implementing interventions based on our modeling. We do anticipate that the results of our work would be useful towards designing more informative interfaces for students and instructors, and for system and content designers. Instructors may adapt their teaching pedagogy as a result of our analysis.

Learning analytics and the field of educational data mining, or EDM, are quite similar [Siemens & Baker 2012; Siemens & Long, 2011]. They are both involved in data mining educational data from learning software. The former is historically more involved with commercial learning management systems, while the latter with intelligent tutoring systems in research settings. In this sense, our work would be more comparable to that of the learning analytics and educational psychology communities than educational data mining as we deal with data from a commercial LMS used in an uncontrolled, non-experimental setting.

1.4.2 Strategic Thinking and Engagement

Students who take a more active role in managing their time and energy over the fourteen week semester will be more likely to succeed. This kind of strategic thinking involves the use of planning and organizational skills. Our interests are in discovering and assessing the impact of strategic learning patterns on outcomes. In educational and psychological literature, this strategic aspect of behavior is referred to “meta-cognitive” or “executive function”

[McCormick, 2006; Pressley et al., 1990]. It is the thinking one does about how to accomplish a task and monitoring one's progress while making any necessary changes to keep "on track." This type of thinking is also called "self-regulation". Zimmerman [2000] defines self-regulation as: "self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals". There is research that claims one of the best predictors of academic success appears to be self-regulation and its strategies in educational environments [Zimmerman, 2002; Pintrich & DeGroot, 1990]. Other research shows that successful students use self-regulated learning strategies in online courses [Azevedo et al., 2004; Whipp & Chiarelli, 2004; King et al., 2000].

The level and quality of engagement is a crucial factor in student success. Arroyo et al. [2010] found that three measures of problem solving behavior were adequate to quantify problem solving behavior in a tutoring system that presented a sequence of problems: attempt, time, and help seeking (hint access). The authors categorized student attempts, time, and help seeking by quartiles applied to the distributions of these features. They classified students based on their categorizations in for the three features. They then mapped interventions to these classifications. For example, a student with low attempts but high time and low help seeking is probably not understanding the material and not actively seeking help. This student would require guidance for getting hints. It is interesting to note that the categories of interest represent the extremes in the distributions. Their analysis was based on categorizations that were made relative to the current student population. We also take this approach by calculating cut points on session and interval times in chapter 3. We also calculate variable values for most of our studies on individual chapters or sections as it maintains a control on factors that may affect a population that changes from one semester to another (or even during a semester due to attrition), and content changes that will inevitably occur.

1.4.3 Previous Studies of Student Behavior in LMS Supported Courses

Barber and Sharkey [2012] created models that included student behavior data to predict the likelihood of failure. They used data from three sources: financial aid records, demographic information, and measures of student behavior from LMS- managed courses. Their models, logistic regression and naïve Bayes, achieved a range of predictive accuracy, from 50% to as high as 90%. The most significant features overall were financial aid status and the number of transfer credits. The most significant behavioral feature was the amount of engagement in coursework, measured by points earned plus discussions posts made. Although the scope of this modeling, considering many courses at once, is higher than our study of one course at a time, it is interesting to note that the most significant behavioral feature was a measure of engagement. We attempt to control for the effects of engagement level when we determine the effect of specific behaviors that are likely to co-vary with engagement. One interesting note about this study: the authors intended to add a measure of submission timeliness to their list of variables, but were unable to do so on account of difficulties obtaining accurate data. We look forward to a future study from them on "working late" (our chapter 4 study).

Andergassen et al. [2014] studied the effects of student exam preparation from log data mined from Learn@WU, an LMS used for introductory business, law and economics courses at the Vienna University of Economics and Business. The courses were "blended" courses, which means physical attendance is optional. In particular, they studied how patterns of working time during the exam prep period and the intervals between times affected exam scores. They calculated the total number of distinct days the system was accessed, the intervals between accesses, and the number of solved exercises, a measure of engagement. They correlated these variables (Pearson's r) with exam scores and found r values ranging from approximately .2 to .4 for students with longer intervals between accesses. In Chapter 3 we investigate the pattern of

working sessions over the semester. We determine that a pattern of long intervals between accesses occurs frequently, and that is seems to be negatively associated with exam scores.

A causal study involving a similar environment and population to ours was conducted by Scheines et al. [2005]. Their analysis was done in two studies: 1. What is the effect (pre-test to post-test gains) of lecture attendance versus online learning and 2. What are effective and non-effective learning strategies in a semester long college course? Their study included a total of 650 students over 5 different semesters. The results from study 1 were that online students did as well or better than the Lecture group. The second study of student behaviors is more germane to our work. For this study, 52 students were sampled from two different classes. In their analysis of student behavior, they proposed canonical or "presumed" student behavioral profiles, such as "The Good Online Student". The presumed behavior that maximizes learning is that the student will do all of the work in a timely manner, attend lecture and recitation, and study for quizzes and exams.

The authors used these features to measure behavior: Pretest, Printout, Voluntary Questions Attempted (VolQs), Quiz, Final exam. The "Printout" feature measured how much of the material students printed: the ratio of “Print” button clicks/total number of modules. Previous findings suggested that students benefit from printed copies because they take notes and otherwise annotate the paper copies. This effect was not borne out in this study. Of more interest was the feature "Voluntary Questions Attempted". The presentation material contained embedded "cognitive checks", or questions. The measure was number of questions attempted/total available questions. This feature turned out to be the strongest predictor of quiz and Final exam scores. This parallels our finding about the effectiveness of embedded questions in our electronic textbook. Unfortunately, timeliness was not measured except by whether a student did the material or not. The canonical good student behavior is not explained

by the models, except that the more embedded questions done the better the outcome. We attempt to find measures that will more fully profile “good” and “bad” behavior.

Table 3: Correlations, Means, SDs (N = 52)

	Pre	*Print*	*Vol*	*Quiz*	*Final*	Mean	SD
Pre	1.000					22.2/30	16.6
Print	0.301*	1.000				0.5	0.60
VolQs	-0.258	-0.421*	1.000			0.4	0.30
Quiz	-0.112	-0.419*	0.774*	1.000		0.5	0.20
Final	0.164	-0.259	0.346*	0.399*	1.000	0.75	0.10

Figure 27: Results from Scheines et al. 2005.

This study does apply causal modeling to attempt to discover the effect certain student behavior has on outcomes. The environment was very similar to ours in that the authors studied a semester-long, graded (high stakes) course. However, we are dealing with data obtained in a “real-world” environment, and have no experimental control over student choices, such as whether or not they attend lecture. Pretests are not given in our courses, although they are given in some online courses. Student behavior such as of printing course material was of interest because of the online students in the study. It does not, as the authors acknowledge, have any learning value itself. We use behavioral measures that are commonly tracked by a typical LMS. Because they are obtained without experimental control, we are required to structure our models according to quasi-experimental conditions. One method we employ for doing so is to discover cohorts of statistically similar students by matching and stratification, and then compare outcomes between members who engaged in a treatment condition and those who did not.

Another analysis of student learning strategies is a study in the domain of medical education [Kusurkar et al., 2013]. 383 students from years 2 to 6 of the VU University Medical

Center Amsterdam were invited to participate in the study. Data was collected by electronic questionnaire. The study evaluated the hypothesis that student motivation would positively affect Good Study Strategy (GSS) and study effort, which in turn would positively affect academic performance measured by grade point averages. The variables of interest to this study were student strategies and effort. The instrument used to measure student learning strategies was the Revised Study Process Questionnaire-2 Factors (R-SPQ-2F). The strategies are Deep Strategy, DS, were students look for underlying meaning in the material, and Surface Strategy, SS, which are exemplified by rote memorization of facts. Good study strategies, GSS, were defined as the difference between these two factors:

GSS = DS - SS

Effort was measured by the amount of time spent on studying. This value was obtained by self-report. Positive effects were found for study strategy: a significant positive correlation of .248, supporting the hypothesis. This study is an example of how data obtained by survey combined with features calculated from log data can be used to measure how strategies affect effort and outcomes.

1.4.4 Studies of Procrastination

There is some research on procrastination and its effects on learning outcomes. Procrastination is typically defined as the difference in time between when a student does their assigned work and the due date of the assignment. We refer to this as “working-late”. Moon and Illingworth [2005] studied how the expression of procrastination changes over time and its relationship to outcomes. One of the hypotheses the authors tested was that procrastination and academic performance are negatively related. Their study involved a similar population to

ours. The study participants (N=303) were drawn from introductory psychology courses at a large Midwestern university.

The authors measured procrastination in two ways. The first, which they refer to as "academic procrastination", used an instrument called the Aitken Procrastination Inventory, a self-report inventory, to measure procrastination. The authors refer to the second measure as "behavioral academic procrastination". They used time-stamped data from computer delivered exam assignments to calculate this measure. Each exam included a 1-week window in which students could take the test at their convenience. Behavioral academic procrastination was calculated as the difference between the date the window opened and the date students took the exam, with larger differences indicating more procrastination. Scores for all five test windows ranged from 0 (took the test the same day the test window opened) to 6 (took the test the last day available in the test window). Therefore, the authors collected five measures of behavioral academic procrastination for each student in the study. The outcome measure they used were scores from five exams that were administered during the semester. The demographics of the study population included gender, with 64% female, 36% male, and class level, with 67% freshmen, 19% second year, 14% above.

Inspection of the pattern of correlations (Pearson's r) between behavioral procrastination and test performance indicated that they were negatively related: $r = -0.43$, $p < 0.01$. The more students procrastinated, the lower their grades tended to be. Therefore, their hypothesis mentioned above was supported, and the authors concluded that procrastination behavior was negatively related to test performance throughout the semester.

The use of recorded time-stamped data to measure lateness in this study and the measurement of a time interval to represent procrastination resemble our approach in chapter

4. It was not clear, however; that the authors made use of the demographic data or any other means to adjust for selection bias in students' "choosing" to procrastinate.

In another study, Lakshminarayan et al. [2012] investigated the relationship between procrastination and academic performance, namely exam scores, for a population of 209 undergraduate dental students.

In this study, the authors assessed the student's level of procrastination, or to what degree did they put off doing assigned work until its due date by means of a sixteen-question, prevalidated questionnaire. They then calculated the correlation between the procrastination scores from the survey and final exam scores. They then categorized the survey results into three groups, hi, medium, low, using the 33^{rd} and 66^{th} percentiles of the distribution of procrastination scores as cut points.

The authors found a highly significant negative correlation between the procrastination category and exam scores: $r=-.39$, $p<0.01$ (two-tailed test), indicating that students who showed high procrastination scores performed below average in their academics. They also found that the procrastination value was highly significant in a regression model predicting outcome from age, gender, procrastination, and year of study.

While this study supports our view that working late is negatively correlated with academic performance, it relies on survey data to measure behavior. We have the benefit of being able to measure the actual time of student working activity, thus avoiding the reliance on self-reporting instrumentation.

1.5 Methodology

In this section, we provide the reader with an overview of observational vs experimental designs in educational research. We discuss methods of dealing with the problems of measuring effects when random assignment is not possible.

1.5.1 Observational versus Experimental Studies

This dissertation is a study of the effects certain learning behaviors have on exam scores. It is a purely observational study, as we did not conduct any randomized experiments. Both observational and experimental studies share common aspects. The scenario for both types of study can be described in simplest terms as a population of subjects which is divided into two groups: a treatment and control group. The outcomes for each group are analyzed for a significant difference. The outcome differential indicates the effect of the treatment. This model holds only if we can assume that the outcome is independent of the assignment to the treatment or control groups, the Conditional Independence Assumption (CIA), and if we can assume that subjects with the same attribute values have a positive probability of being in the treatment and control groups, the common support assumption (CSA) [Rubin, 1974]. The CSA ensures that subject attributes such as age, height, etc. do not perfectly predict the outcome.

The main difference between the experimental and observational studies is how the subjects are divided into treatment and control groups. In an experimental study, the subjects are assigned to treatment and control at random, ensuring the CIA holds, while in an observational study the subjects select whether they are in the treatment or control group themselves. The issue with self-selection in the observational study is that there may be an imbalance in certain key aspects of the subjects who choose one group over the other. One or more of these aspects may have an influence on the selection to the treatment group, as well as

an effect on the outcome. This is an example of a factor that “confounds” the analysis of the effect the treatment may have had on the outcome [Rubin, 1974].

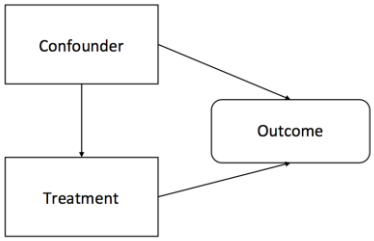


Figure 28: Confounder affects both treatment and outcome.

For example, suppose we are attempting to measure the effect of the book-first pattern on exam scores. Our subjects are students and the treatment is following the book-first pattern of completing all book problems before starting on homework problems. We could design an experiment that randomly assigned students to one of two groups: a treatment group where students follow the book-first pattern or a control group where they do no book problems before starting homework (book-last). We then compare exam scores of the two groups to gauge the effect of the book-first pattern on the exam scores. One way a confounder could affect this study is if it causes students to select the book-first pattern and also affects exam scores. Let’s take previous experience with programming as an example. Suppose that students with previous experience do better on the exam in the programming course than students without experience. Also, suppose that students with previous experience are more likely to follow the book-first pattern. Now, suppose we find that the exam scores for the book-first group is higher than the control group. We attribute the difference to the effect of following the book-first pattern. This is not a correct attribution, however, since the putative effect of book-first is “confounded” by the influence of previous experience.

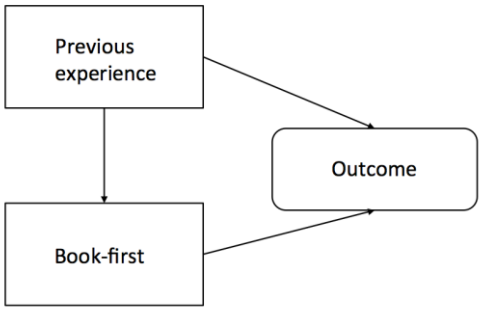


Figure 29: previous experience as a confounder.

In theory, random assignment would have mitigated the effect of previous experience, as there would have been a balance between the number of students with previous experience in both the book-first and control groups. In practice there is no guarantee that this will be the case. Balancing by random assignment requires a sufficiently large number of subjects. Despite this requirement, one major advantage of random assignment is that unobservable confounders are also taken care of. This cannot be done in a quasi-experiment by balancing techniques.

We cannot actually run the randomized experiment described above for several reasons, chiefly because we have no way of assigning students at random to follow a book-first pattern. There is no mechanism in Owl to enforce the book-first pattern on students chosen at random. There are also ethical concerns about fairness and disclosure in a non-experimental academic setting. This means that in our study, students self-select into following the book-first pattern (and how much of the book problems they do before starting homework).

In the case where random assignment is not possible, the remedy is to try to mimic random assignment by selecting the treatment and control groups so that we have a balance on previous experience. One problem with this approach is that we often loose subjects by excluding them to achieve this balance. If many confounding factors have to be accounted for, then the number of subjects can become very small, with a corresponding loss of statistical

power for estimating any effects. This type of analysis where random assignment is not possible but attempts are made to account for confounding effects is referred to as a quasi-experiment [Shadish et al., 2002].

If the confounding factors in the treatment and control groups are balanced in a quasi-experiment, in theory they are essentially the same populations, just as we would assume in a random assignment. If we assume the subjects are the same, then we could say that if subject A had treatment and subject B did not, then, since they are equivalent subjects, any effect from treatment to A is what would have happened to B if she had had the treatment. This reasoning is known as the "counterfactual" argument, which asks what would have happened if the same subject had not been in the treatment group? [Holland, 1986; Rubin, 1974]. Of course, subjects A and B are not truly identical. The only way for that to happen with individual people is if the same person was duplicated and one duplicate had treatment while the other did not. An example of this type of quasi-experiment is studies on identical twins since they share the same genome but vary in their exposure to environmental factors.

1.5.2 Methods of Compensating for Self-selection.

There are methods of controlling for confounding effects in an observational study with self-selection, including simple matching and propensity score matching [Caliendo & Kopeinig, 2008; Morgan & Harding, 2006; Smith & Todd 2005; Dehejia & Wahba, 2002; Heckman et al., 1997; Rosenbaum & Rubin, 1983]. They are designed to deal with the issue of confounding factors by simulating random assignment. The idea is to create a group or groups of subjects for the control or treatment condition that are balanced with regard to all observable factors that may confound the effect of treatment on outcomes. These evenly balanced groups are what we

would have expected from random assignment. Unfortunately, unobservable confounding factors cannot be balanced by matching methods.

1.5.2.2 Simple matching

Simple matching is the process of selecting such groups where the balance of attributes is maintained between the treatment and control groups [Kuehl, 1994]. In contrast to propensity score matching, where the matching is done on propensity scores which are based on the covariates, simple matching deals directly with the covariates. The result of simple matching should be a set of balanced populations. Simple matching may be done manually or by special software.

According to Morgan and Harding [2006], matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger pools of such cases or (2) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted. The first scenario is more relevant to our work. Morgan and Harding state that in the first scenario, matching is a method of strategic subsampling from among treated and control cases. The investigator selects a non-treated control case for each treated case based on the factors that represent the characteristics of individuals, All treated cases and matched control cases are retained, and all non-matched control cases are discarded. Differences in the outcomes are then calculated for treated and matched cases, with the average difference serving as the treatment effect estimate for the group of individuals given the treatment.

There are some issues with creating balanced groups in an observational study with any technique [Morgan & Harding, 2006]. One caveat is that the methods described above do not deal with latent, or unobserved variables. There are likely to be unobserved, confounding factors in the subject pool that are not controlled for. Another is that there may not be enough subjects, a lack of "common support" to make up a balanced group. An argument has to be made that all reasonable factors have been controlled for in order to make any sort of claim for a causal effect. In matching, there is a trade-off between bias and variance. The closer the matches are made, the smaller the bias but the larger the variance of the estimates. Conversely, reducing the variance (by including more observations) in the matches will decrease the quality of the matches and introduce bias.

In our studies, we sometimes have difficulty finding enough subjects for treatment and control groups. One reason is that we are interested in studying the effect of a learning strategy on specific subpopulations, such as students with previous programming experience versus those without experience. Segregating, or stratifying, the population and balancing the values of other possible confounding variables often leaves us with small groups. For example, we may have groups with 15 to 25 subjects. We also may have groups of different sizes. If we dropped some of the data from the larger group, we might bias the results. For these reasons we rely on propensity score matching to adjust for self-selection and to create balanced treatment and control groups. We also manually created groups of specific subpopulations of interest for some studies. In these examples, we use non-parametric tests for effect given the low statistical power due to the smaller sample size.

1.5.2.1 Propensity score matching

Propensity score matching [Rosenbaum, 2002; Rosenbaum & Rubin, 1983], is one method of dealing with self-selection. A propensity score estimates the probability of a subject entering the treatment group given the observed factors. Subjects are grouped according to their propensity scores. The result is that the distribution of (observed) variables is similar between treated and untreated subjects. This strengthens the argument for any effect we observe from treatment being caused by the exposure to treatment and not from the observed confounding variables.

Propensity scores deal with an important problem that arises when there are a large number of variables, both continuous and discrete, to account for. The problem is that there will be very few exact matches on all variable values as the number of variables grow. Comparing subjects on one or more continuous variables is difficult as well. Propensity scoring provides a way to match on a model of the variables of concern rather than on the individual values of those variables. A major assumption is that the propensity scores do not correlate with the outcome. One concern with propensity score matching is that it cannot adjust for unmeasurable effects, works with many variables, and requires larger data sets. Another concern is that the treatment effect estimated from propensity scores can be sensitive to how the propensity score model was specified [Smith & Todd, 2005].

The process of propensity score matching involves several steps: selecting a model for calculating propensity scores, selecting the variables for the model, selecting a matching algorithm, evaluating the results of the matching, and evaluating the average treatment effect.

Propensity scores are typically, but not always, estimated using a logistic regression model, in which treatment status is regressed on observed covariates. The estimated propensity score is the predicted probability of treatment derived from the fitted model. Besides

regression, there are other methods for estimating propensity scores, including the use of tree-based methods [Lee et al., 2010].

There is no general agreement as to which variables to include in a propensity score model. Possible sets of variables for inclusion in the propensity score model include the following: all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (i.e., the potential confounders), and all covariates that affect both treatment assignment and the outcome (i.e., the true confounders). Since the propensity score is defined to be the probability of assignment to treatment, some researchers argue for the inclusion of only those variables that influence the treatment assignment and the outcome. It should also be clear that only variables that are unaffected by the assignment to treatment should be included in the model. To ensure this, variables should be measured before assignment time [Austin, 2014; Heckman et al., 1999].

There are many matching algorithms that may be used for propensity score matching, and for matching in general. The figure below depicts some of the major types of matching algorithms and the parameters that must be specified for each [Sekhon, 2011].

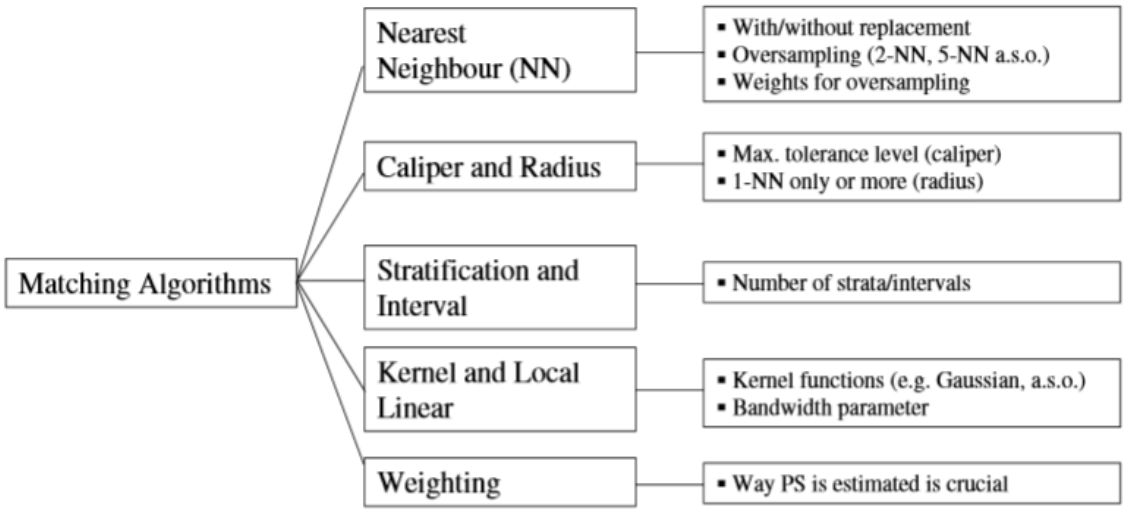


Figure 30: Some matching algorithms and their parameters (from Sekhon 2011).

The simplest matching algorithm is the nearest neighbor (NN) matching. A subject from the control group is chosen as to match with a treatment group subject based on a distance metric, which is the propensity score. NN matching can be done with replacement and without replacement. In the former case, a subject from the control group can be used more than once as a match, whereas in the latter case it is matched only once. Matching with replacement involves a trade-off between bias and variance. If we use replacement, the average quality of matching will increase and the bias will decrease. This is of particular interest with data where the propensity score distribution is very different in the treatment and the control group. For example, if we have a lot of individuals in the treatment group with high propensity scores but few control group individuals with high propensity scores, we can possibly get bad matches as some of the high-score treatment subjects will get matched to low-score control subjects. This can be overcome by allowing replacement, which in turn reduces the number of distinct non-participants used to construct the counterfactual outcome and thereby increases the variance of the estimator [Smith & Todd, 2005].

Stratification and interval matching are another technique used on propensity score matching. The idea of stratification matching is to partition the common support of the propensity score into a set of intervals, called strata, and to calculate the impact within each interval by taking the mean difference in outcomes between treated and control observations. This method is also known as interval matching, blocking and subclassifcation [Rosenbaum & Rubin, 1983]. The key parameter to the stratification process is the number of strata to use in the analysis. Cochrane and Chambers [1965] demonstrated that five subclasses were usually adequate. One way to justify the choice of the number of strata is to check the balance of the propensity score (or the covariates) within each stratum.

Genetic algorithms are also used in matching. A genetic algorithm works by an iterative process of generating new populations of individuals or candidate solutions, evaluating the fitness of each member of the population, selecting the most fit from the current population, and then modifying each member's genome to form a new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population [Mitchell, 1996].

Genetic matching uses a genetic search algorithm to find a set of weights for each covariate such that a version of optimal balance is achieved after matching. Different Matching methods may be passed in to the algorithm. Balance is determined by two univariate tests, paired t-tests for dichotomous variables and a Kolmogorov-Smirnov test for multinomial and continuous variables [Sekhon & Diamond, 2005]. Genetic matching algorithms do not depend on knowing or estimating the propensity score, but the method is improved when a propensity score is incorporated.

1.5.2.2 Matching Evaluation

The goal of matching is to achieve groups with a balanced distribution of the subject's baseline variables. After a matching has been done, the quality of this balance has to be checked in both the control and treatment group. The basic idea of is to compare the balance before and after matching and check if there remain any differences after conditioning on the propensity score. If there are differences, matching on the score was not totally successful and corrective measures may be done, either by modifying the terms in the model used in the estimation of the propensity score or by changing the way matching was done.

There are many ways to evaluate the quality of matching. Three techniques are: 1. compare the difference in means for the treatment and control groups; 2. compare summary statistics based on standardized empirical-QQ plots, used for comparing differences in distributions with regard to location, dispersion, and skewness; and 3. calculate the variance ratio of treatment over control, which equals 1 if there is perfect balance. The Kolmogorov–Smirnov test (KS test) can also be used to evaluate matching. It is a nonparametric test that can be used to compare a sample with a reference probability distribution. The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples [Sekhon, 2011]. A qualitative way to assess matching is graphically, by using data visualization software. Creating a propensity score matching is an iterative process, where the model and matching parameters may be adjusted to achieve the best balance in treatment and control groups [Rosenbaum & Rubin, 1983].

1.5.3 Measuring Effect Size

An effect size is a measure of the magnitude or strength of influence of one factor on another. If factor A has an effect on factor B, then we want to know the size and variability of the effect. For example, if the effect size was 15, we would want to know if every member of A had an effect of 15 or, more likely, a smaller number of A had a lesser effect on B. Both the mean effect size and the variability of that size are useful estimates [Kelly et al., 2012].

The average treatment effect, or ATE, (also referred to as ACE for average causal effect) measures the difference in mean outcomes between subjects assigned to the treatment and control groups. For each subject *i*, the effect of treatment is defined to be $Y_i(1) - Y_i(0)$, and the average treatment effect (ATE) is defined to be $E[Y_i(1) - Y_i(0)]$. A related measure of treatment effect is the average treatment effect for the treated, or ATE. The ATE is defined as $E[Y_i(1) - Y_i(0)|D = 1]$, where $D = 1$ indicates membership the treatment group. The ATE is the average effect of treatment on those subjects who received treatment. In a randomized study, these two measures of treatment effect coincide because, due to randomization, the treated population will not, on average, differ from the overall population.

For our specific subpopulation studies, we use two different types of effect size measurement techniques: Pearson's r and Cohen's d. Pearson's correlation coefficient, r, is often used as a measure of effect size. It measures the degree of linear dependence between two variables by calculating the covariance of the two variables divided by the product of their standard deviations. A value of 0 means no correlation, 1 means a perfect, positive correlation, and -1 means a perfect negative correlation [Ellis, 2010]. In addition to the correlation coefficient, Pearson's r reports a significance level, referred to as a p value, which measures the probability that a particular value of a statistic, the correlation coefficient in this case, could

have been obtained under the null hypothesis, which in this case is that there is a zero true dependence between treatment and outcome.

Cohen assigned the following descriptions for values of r for social science research: 0.1, medium for .3, and large for .5, assigning the terms “large”, “medium”, and “small” to these values is somewhat arbitrary [Cohen, 1992]. In social science and educational studies, a p value of .05 or less is considered significant ($p<.05$). One issue with this value is that it will become significant with large sample sizes regardless of the value of r. We are dealing with small sample sizes.

Cohen’s d reflects a standardized difference between group means. Cohen's d is defined as the difference between two means divided by a standard deviation for the data. It does not report a significance level. A d value of 1, would indicate that the means differ by one standard deviation, and a d value of .5 half a standard deviation. Cohen suggested that $d=0.2$ be considered a “small” effect size, 0.5 a “medium” effect size and 0.8 a “large” effect size. A trivial difference would be a d value of less than 0.2 [Cohen, 1992].

It is important to consider a possible bias in measuring effect sizes on averaged data, as we do frequently. We calculate many of our variables on individual units of assigned work (called sections or chapters in our courses). We aggregate these variables over several chapters by averaging their values. It’s possible for effect sizes to be biased when they are calculated over averaged data (Simpson’s Paradox) [Brand et al., 2011]. We are not averaging results of individual trials, however. A “trial” implies that the value is the result of some analysis. The values we are averaging are measuring the quantity of an occurrence of a specific event.

1.5.4 Matching Methods in Educational Research

The use of matching designs in educational setting has become recognized as a way to deal with non-experimental situations in which self-selection is an issue. Reynolds, et al. [2009] studied the difference in 4-year institution graduation outcomes between students who started in a community college and those who started in a four year institution. They sought to determine what the outcome would have been for students who actually started in a 2-year college, if they had matriculated at a 4-year institution. They used a variety of matching software to select groups for the study. The matching algorithms included simple stratification with matching on individual variables, K nearest neighbor, weighted attributes and propensity scoring. They controlled for gender, income, previous school experience, and other demographic factors. They found that 20% fewer community college students completed 4 year degrees than students who started at 4 year institutions. Their study was an investigation in the use of matching on propensity scores. Our data is simpler, with fewer variables. We adopted basic matching techniques.

Lockwood and DesJardins [2009] investigated the effects on educational attainment of initially attending a 2-year college instead of a 4-year institution, the former being the treated condition and the latter designated as untreated condition. They estimated the effect on educational attainment of attending a 2-year college instead. Outcome measures included retention rates and amount of credits earned. They compared OLS regression to propensity score matching and found a smaller effect for the latter method. This difference was attributed to selection bias which was not adjusted for in the OLS method.

1.5.5 Clustering

Azarnoush et al. [2013] Conducted a study of clustering users of an online educational environment designed to provide resilience training, how to learn skills necessary to cope with the doctoral degree process, for women STEM doctoral students. They used features derived from survey responses and scalar measures of usage. They sought stable clustering that would identify groups of users that described meaningful characteristics to domain experts in order to allow insight about the learning needs of the users in each cluster. The clusters formed were associated with a set of attributes that described characteristics of interest, such as whether they were dissatisfied with some aspects of their study and for what reason: advisor, program, etc. The authors used a random forest [Breiman et al., 2003] to create a dissimilarity matrix to use in their clustering. They found this produced more stable clusters that the Euclidean metric. They also use the Adjusted Rand Index, ARI, [Hubert et al., 1985] to measure cluster stability. We tried the same approach to clustering on frequency vectors, vectors of frequency counts, but found the Euclidean metric superior to the random forest. We plan to incorporate non-numeric data in the future, and will revisit the random forest dissimilarity matrix for clustering). We adopted their means of splitting the data into test and training sets to generate pairs of cluster labels to be compared by the ARI measure.

Perera et al. [2009] use sequential pattern mining to discover patterns that characterized students in an online project management tutoring system. The authors clustered learners according to quantitative indicators of activity and also proposed the use of alphabets to represent sequential patterns of interactions that can distinguish strong from weak groups. They found this method allowed them to create clusters that could be understood by educators, and related to test scores.

Bouchet et al. [2012] performed a cluster analysis on 51 students (using the EM cluster algorithm in Weka (http://www.cs.waikato.ac.nz/ml/weka/)) with 13 variables extracted from their tutoring system's log after the end of each learning session in an ITS. Generally, their variables measured the amount of work attempted, the time spent working, time spent reading, and the number of times a student checked their notes. They found three clusters to be the most frequently obtained number of clusters in 1000 runs. Cluster 1 had high outcomes and spent less time but attempted more material. The negative correlation with outcomes has been frequently observed in our data. Cluster 2 had low outcomes with low time and attempts, and Cluster 3 was intermediate in outcomes, took more time and attempts than 1 or 2.

Having established the clusters, they looked to find any patterns of student actions that were characteristic of students in a cluster when the system prompted them with an automatically generated hint. The authors used a sequence mining approach to finding statistically significantly different patterns. They found that students with prior knowledge with good outcomes "tended to be more compliant with system prompts, using them to validate their progression". On other words, they used the system structure to help them marshal the system resources efficiently.

CHAPTER 2

STUDY OF THE "BOOK-FIRST" STRATEGY

In this study, we evaluate the effectiveness of the book-first pattern, first described in section 1.1.1. First we introduce the research question that motivated our study and why it is of interest for educators and researchers. Then we describe our methodology, results, and discuss our conclusions along with possible confounding effects on our results. We first analyzed the data using a propensity score matching approach, then manually selected subpopulations of interest and estimated the effect of the book-first pattern for the subpopulations. The former approach has the benefit of using all of the data, mitigating selection bias by achieving treatment and control groups with an equal balance of the observed covariates that might affect selection. The downside of this approach is that it requires many steps and assumptions which may affect the validity of the results by introducing bias. The latter approach, where the subpopulations are manually selected, does not involve any intermediate steps. The downside of this approach is that we lose power since the data sets become small. In addition, the latter approach does not adjust for balance in the remaining covariates in the selected subpopulation. The reason for taking two such approaches is that it will strengthen the case for an effect of book-first if there is a positive effect in the results from both methods.

2.1 Introduction

Our main research question is: does interacting with the teachable content, in our case the textbook, before working on homework problems result in better learning outcomes? We call this sequence of working the "book-first" pattern. This question is of interest for several reasons. Most college courses in the STEM disciplines, like the two we studied, CS121 and CHEM111, use textbooks that are designed with a narrative section, often including worked

examples or brief question and answer vignettes, followed by a reinforcing problem set, often called “end of chapter questions” or “homework”. This structure is designed to support the acquisition of new knowledge followed by the reinforcement of this acquired knowledge by applying it to homework style problems. Given that this is the way most STEM material is organized and presented, it is very important to know if students actually follow that sequence as they work through a semester, and if it is an effective sequence. The answers to these questions could have a major impact on the design of textbooks and educational content in general as well as how instructors teach courses.

Following a sequential approach to learning a technical subject has been shown to have a positive effect on outcomes. Lau and Yuen [2009] investigated the effects of gender and learning styles on computer programming performance on secondary school students of age from 14 to 19 participated in this study. Their results indicated that no gender differences in programming performance were found after controlling for the effect of student ability. They also found that students who worked through the material sequentially performed better than random learners.

We have much anecdotal evidence from instructors that many students do not follow the book-first pattern, and that instructors believe they should. If these opinions are justified, then instructors would have a basis to make changes to their pedagogical strategies. They could make stronger suggestions about how students work. They could change the presentation order of material, and motivate students by assigning more weight to grading certain activities.

Evidence for the book-first pattern is also of interest to the learning research community. There are contrary points of view about whether students learn better by following a narrative with tutorial activities and worked examples, or learning by directly engaging in problem solving. Problem Based Learning (PBL) [Hmelo-Silver, 2004; Dods, 1997; DeGrave et al.,

1996] is one term used to refer to the problem solving first approach. The content used in this dissertation is narrative based.

We hypothesize that students who fully interact with the text, as evidenced by the number of "embedded" questions they answer, before they begin homework problems will have significantly higher exam scores than students who interact to a lesser degree. Another hypothesis we evaluated is that students who are relative novices will show a greater effect from following a book-first strategy. In the CS121 courses, we add the condition of no previous programming experience to the novice (first year) group. Novice students in CS121 show a greater effect from following a book-first strategy than students who report having previous Java programming experience (the computer science course is conducted in Java).

2.2 Method Overview

In order to measure the effectiveness of the book-first pattern, we define a variable, BF, which represents a student's level of engagement with the book. Given that the amount of assigned work attempted is correlated with both the BF score and with final exam scores, we will include a measure of engagement in our analysis. This will allow us to adjust for participation in the estimated effect of the BF score on final exam scores. Students who do more of the assigned work will be more likely to do more book questions before they attempt homework and will score higher on the exam. Adjusting for the participation level while allowing the BF score to vary will help isolate any effect on outcome of the BF score alone.

We use PC_HWK (see section 1.3.2), the percentage of assigned homework problems attempted, as a way of adjusting for participation level. This measure is needed because higher participation is correlated with higher exam scores. Engagement with the book is also correlated

with the level of participation, so adjusting for participation is needed to infer an effect from book engagement alone.

In addition to the derived variables described above, we use several demographic variables gathered from course surveys. These variables, such as gender, class level, major, and previous programming experience (computer science only) describe attributes that distinguish subpopulations of interest. These variables also may contribute to a student's choosing to follow a book-first pattern. We use final exam scores as our outcome measure.

In order to estimate the effect a book-first approach has on exam scores, we carried out two kinds of analyses. First, we performed an analysis utilizing propensity score matching to mitigate the selection bias in our non-experimental data (the reader is referred to section 1.5 for relevant background). Second, we manually selected subpopulations of interest and measured the correlation between the BF measure and final exam scores.

2.2.1 Data sets

We pooled data from the two CS121 fall and spring courses. We did this as propensity scoring works better with larger datasets. Merging courses based on the semester maintains a more uniform composition of the student populations, as they vary from fall to spring (section 1.3.2). The following table summarizes these combined datasets.

Table 15: Combined datasets.

CHEM111 Fall (2012)	N=516
CS121 Fall (2012 + 2013)	N=705
CS121 Spring (2013 + 2014)	N=738

Next, we describe our methods for computing the BF and PC_HWK variables. Following that, we describe the survey variables used in this study. Then, we present the propensity score

matching analysis followed by our manually selected subpopulation analysis. We end this chapter with a discussion of the results and our conclusions.

2.2.2 Derived Variables

Students encounter the course material as a sequence of assignments (see section 1.2.1). Each course is organized into units of content which correspond to the main topics taught in the course. These units are referred to as “chapters” in CS121 and “sections” in CHEM111. Regardless of how they are named, each unit consists of two assignments: a textbook and a homework assignment. The textbook assignments contain strategically placed, automatically evaluated “embedded” questions (section 1.2.4) that are assigned for credit. The time stamped data from these textbook problems provide us with a way to track student engagement with the text. Recall that each chapter or section of textbook material is associated with homework problem sets (see section 1.2.1), which are also automatically evaluated and time stamped. In order to represent the book-first pattern, we counted the number of embedded questions answered before the first homework problem was answered. This measure, called BF, is described next.

2.2.2.1 The BF score

The BF measure is the ratio of distinct book questions attempted before the student’s first homework problem attempt to the total number of distinct book questions assigned for a given chapter or section. We calculate a BF score for each student for each chapter or section that contains a book and accompanying homework assignment.

a = distinct book questions attempted before the student’s first homework problem attempt
b = total number of distinct book questions assigned

BF = a/b

The BF score is a value between 0 and 1, inclusive. A student has a BF score of 1 if he attempted all of the available book questions in the chapter before he attempted any of the homework problems. A student with a BF value of 0 has attempted none of the book questions before attempting a homework problem. We calculate a BF score for each student for each *chapter* or *section*. (Note that the terms chapter and section are equivalent: both are units of content where *chapter* refers to CS121 courses and *section* to CHEM111). The following figure depicts how BF is calculated for a student who does all assigned book questions before homework and one who does none of them before homework for a single chapter (or section).

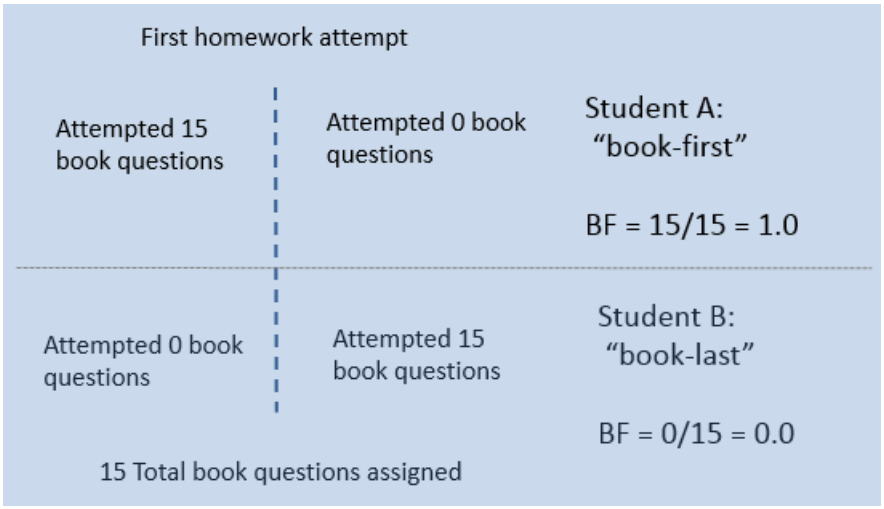


Figure 31: Calculation of BF for two hypothetical students.

To provide the reader with an impression of what the BF scores of students look like, we present the distribution of BF scores for a single unit of content in the following figures. The BF scores shown are fairly typical of other assignments in the courses.

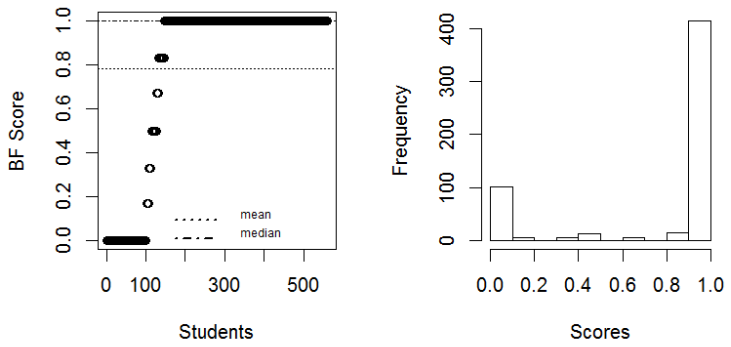


Figure 32: CHEM111 Fall 2012 Section 7.5 BF Scores.

Figure 18 shows BF scores for one section in CHEM111. Comparing this histogram with the one for CS121 below, we see that they are very similar.

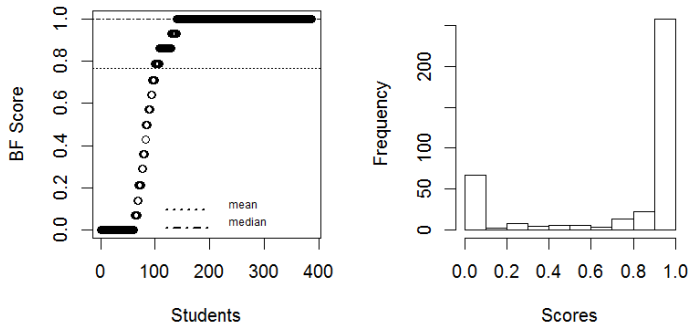


Figure 33: CS121 Fall 2012 Chapter 7 BF Scores.

In both examples, the majority of students are in the BF=1 category, with the next highest group in the BF=0 category. We suspect that some of the BF=0 scores are due to low participation rather than not following the book-first pattern. For example, if a student does not attempt any book questions (but at least one homework problem) then that student's BF score will be 0. The following figure shows the same histogram of BF shown in figure 19 alongside a histogram of BF scores for students who did at least 80% of the assigned homework.

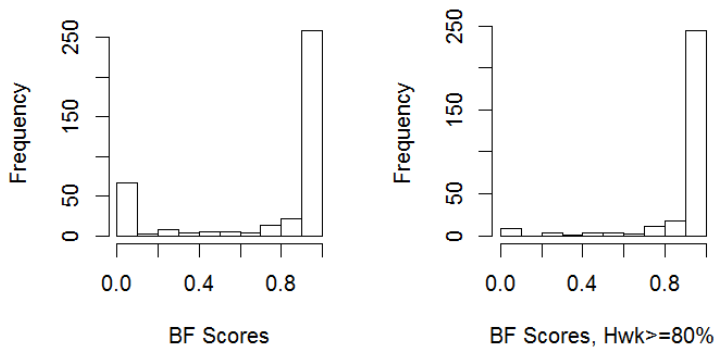


Figure 34: CS121 Fall 2012 BF scores for Ch7 for all students, and for students attempting at least 80% of assigned homework.

The change in BF=0 score frequency proves the point about participation. Since participation is likely correlated with both treatment, the BF score, and outcome, the final exam score, we will take the level of participation into account when we analyze the effect of BF scores on exam scores. We will do this by including a measure of participation, the percent homework attempted, in our set of covariates.

Next, we show the distribution of the BF scores averaged across the entire course. We show this because our analysis will be based on the individual chapter or section BF scores averaged over the entire course, using final exam scores as the outcome measure. There are two main reasons why we did not analyze the data on a chapter (or section) basis. First, we do not have any evaluation of student performance on that level of granularity, i.e. there are no quizzes or exams for each chapter. Second, the midterm exams are cumulative, and do not test specific chapters in isolation.

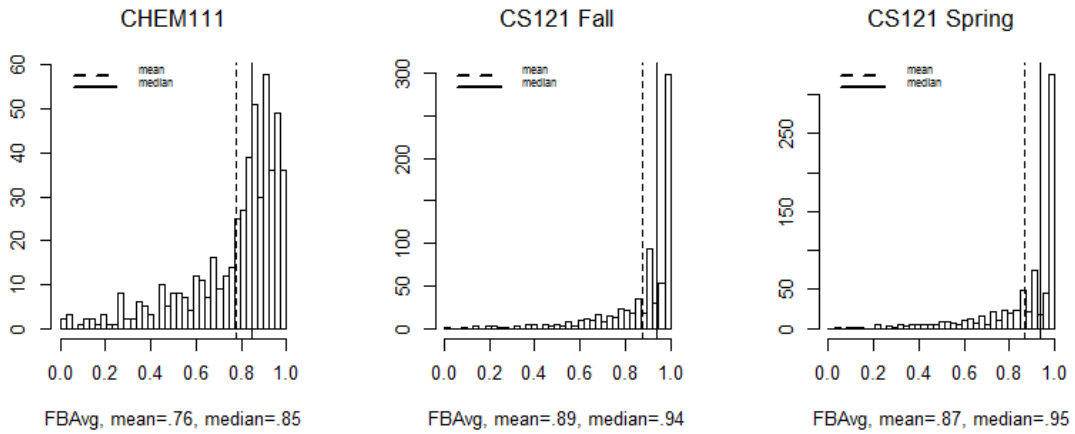


Figure 35: CHEM111 and CS121 BFAvg Scores. The BFAvg scores shown are averages of all individual assignment (chapter or section) BF scores.

The figure above shows the averaged chapter BF scores- BFAvg. From the average score distributions, we see that there are more scores in the 0.8 to 0.9 range in CHEM111 than in CS121. We also see very few students who are doing no book questions before starting homework. This means that we have practically no examples of BF scores of 0.

We were curious if female students followed the book-first pattern more than male students. We plotted the BFAvg distributions for both genders for all three datasets.

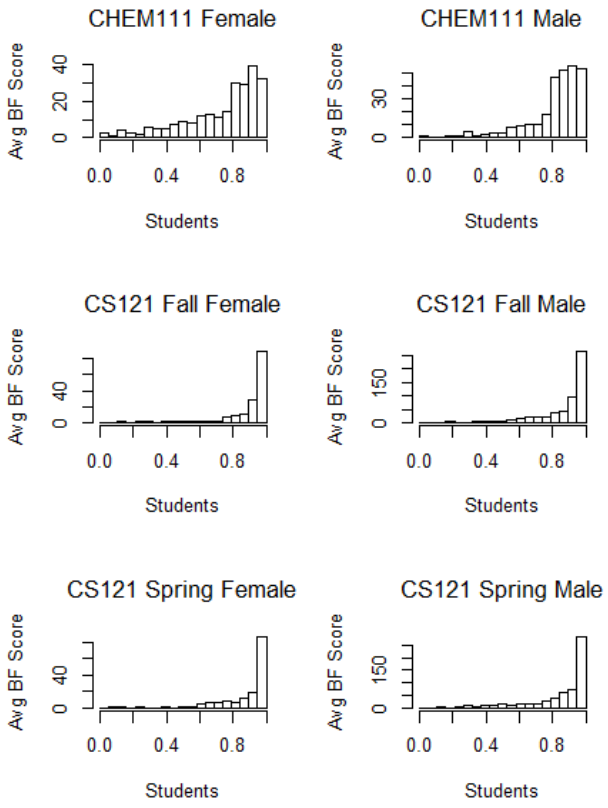


Figure 36: Distributions of BFAvg scores for the three data sets by gender.

In the figure above, we see that female CHEM111 students may follow the book first pattern less than their male counterparts. In the CS121 datasets, we see they are approximately the same behavior.

2.2.2.2 Changes to the BF scores in CS121

There have been several changes to the CS121 course that are worth mentioning because they have affected the way students work, and especially their Book-first related behavior. The course has used the OWL system for managing the iJava textbook and homework since 2007. Until fall 2012, the embedded questions in the text could be presented and evaluated, but could not be assigned with due dates as this part of the OWL system had not been implemented. The instructor would state that the embedded questions were worth a

nominal amount of points, and were due at the end of the semester. We began to see that students who did the embedded questions before they started their homework did better on the exams. A typical example was a difference in 12 points on the final ($p<0.01$) between those who had BF scores of 1, did 100% of the book questions first, vs those with BF scores of 0, did none of the book questions first. The analysis adjusted for the percent of homework participation. In response to these findings, the instructor began to strongly encourage students to do the text questions before homework, citing the results. The instructor tried several methods of motivating students to follow the book-first pattern. One way was to assign 3 points for book questions done before lectures, and 1 point if done at all. This had to be done manually as the OWL system did not have this capability, as mentioned above. The result of these changes was a gradual increase in the number of students following the Book-first strategy. An example of this shift in behavior can be seen in the following graphs.

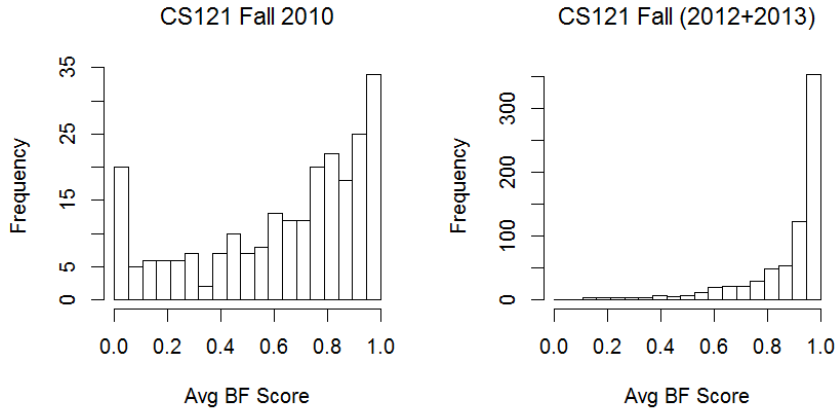


Figure 37: Comparison of frequencies of average BF scores between the fall 2010 and combined fall 2012, fall 2013 CS121 courses. The BF scores shown are averages of all individual assignment BF scores.

The histogram on the left shows that the frequencies of students with average BF scores of less than 1.0, i.e. who did less than 100% of the book questions before starting a homework assignment, are much higher than the for the combined Fall semesters on the right. Notice the

much larger proportion of students have BF scores of 0 in the fall 2010 semester, and the much larger proportion of students with BF scores of 1 in the Fall 2012-2013 semesters. To further illustrate this trend, the table below shows the proportions of extreme BF scores of 0 and 1 for the fall 2010 and fall 2013 courses.

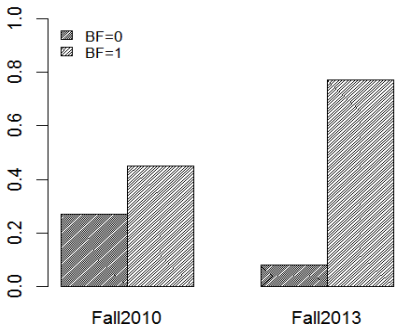


Figure 38: Proportion of BF=0 and BF=1 scores in CS121 courses.

The fact that the embedded questions are now fully assignable and appear with due dates has, in our opinion, made it much easier for students to follow the Book-first strategy. This trend seems to have stabilized in the last four semesters as shown in the following graph.

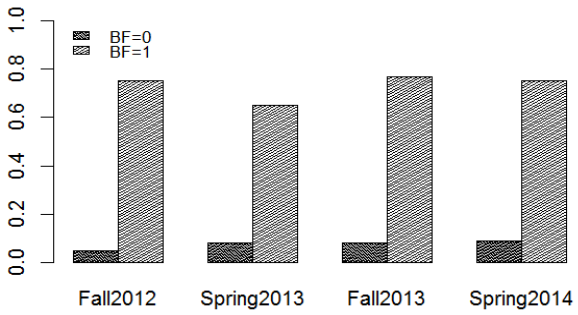


Figure 39: Proportion of BF=0 and BF=1 scores in 4 semesters of CS121 courses.

Also, anecdotally, exams have become more difficult as student performance has improved. While we have evidence for the positive effect of BF scores on exam scores for some semesters before fall 2012, we do not have a longitudinal study that shows the exam proficiency increase and if it was a result of more students following the BF pattern. Another possible explanation is that compliant students are more likely to follow the instructor's advice, and compliant students may tend to do better on exams.

One consequence of the trend towards higher BF scores is that it makes it more difficult to find the effect of the BF strategy on exam scores. If everyone got a BF score of 1 then the variation in outcome would be due to other sources. The proportion of BF=1 vs BF<1 scores for the last four semesters is shown in the following table. The same situation exists for the CHEM111 course.

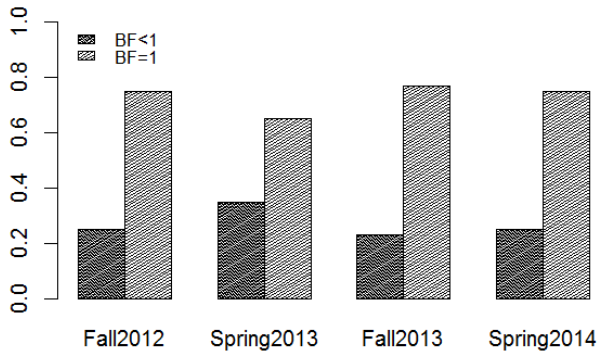


Figure 40: Proportion of BF<1 and BF=1 scores in 4 semesters of CS121 courses.

Given the fact that there are very few students who do none of the embedded questions before their homework, we uncovered an effect for the BF strategy by looking at two groups shown in the figure above: those with BF=1, and those with BF<1.

2.2.3 Survey variables

In addition to the derived variables described above, we gather several, demographic measures from our course surveys. These variables (see section 1.3) describe aspects of students that may have an effect of their outcomes as well as on their selection of following the book-first pattern. For example, do females follow the book-first pattern more than males? Our data suggests this is often the case. Do students with previous programming experience tend to follow the book-first pattern, and does it help them in the same way as students with no experience? The survey data is also helpful in identifying subpopulations of interest. For example, does the book-first pattern benefit first year students who are computer science majors less than non-majors?

2.2.3.1 Survey participation

Survey data is problematic in that it is self-reported, and thus open to bias. Students may not answer truthfully, or may not interpret a question in the way it was intended. Furthermore, by using only students who took the survey (it was optional), we are potentially introducing more bias into our population as the students who participate in the survey may also be more likely to follow the book-first pattern, and to do better on exams. We looked at the population of students who did not take the survey versus the students who did take it with regard to three quantities: their final exam scores, the percentage of homework attempted, and their adoption of the book-first pattern.

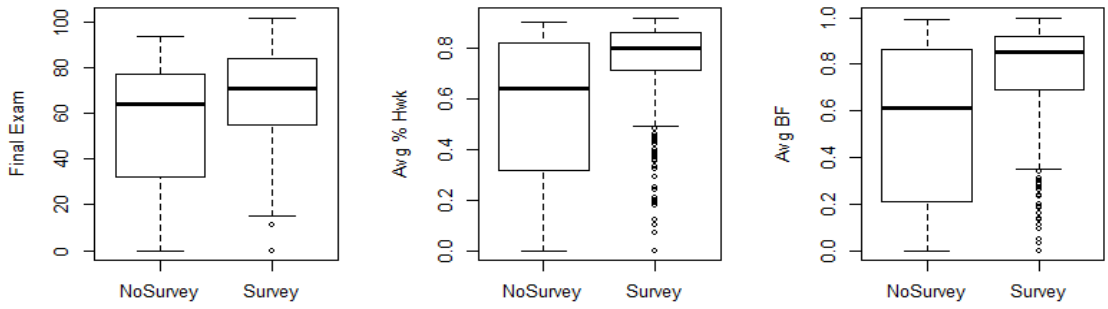


Figure 41: Comparison of three measures for the survey taking population and the non-takers for CHEM111.

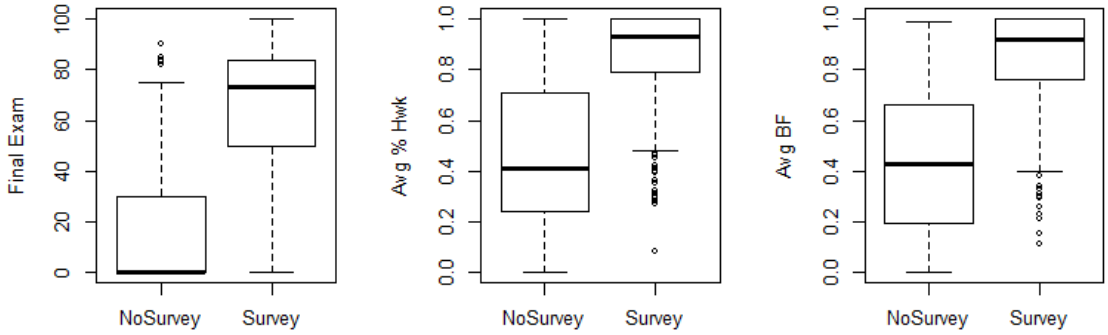


Figure 42: Comparison of three measures for the survey taking population and the non-takers for CS121- all courses combined.

Fall 2012 CHEM111 92% took the survey, CS121 95% took the survey. It is clear that the survey is "selecting" the students who are participating at a high level, which is the majority of the population. This is especially true for the CS121 courses. It is also clear that the non-survey takers suffered in the final exam, most likely due to low participation in the assigned work. It is also interesting to note the correlation between Avg % Hwk and Avg BF. The information presented above shows that we are removing the lowest participating students from our study

by considering only survey respondents. This should not adversely affect our study as the non-survey takers have both low BF and PC homework scores on average.

We now give a summary of all of the variables in the following table.

Table 16: Variables used in the study.

Variable	Levels	Description
PREV	none, some, java	Previous experience. Not used in CHEM111.
GENDER	female, male	Gender
MAJ_GROUP	1,2,3 (CHEM) or 1,2 (CS)	Groups formed from the survey variable MAJOR.
CLASS_LVL	1, 2, 3, 4	1^{st}, 2^{nd}, 3^{rd}, and 4^{th} year students.
BFAvg	[0,1]	The average of BF scores for all course assignments. Used to define the treatment condition.
PCAvg	[0,1]	The average of PC_HWK scores for all course assignments.
FINAL	[0,100]	The outcome measure.

In summary, we have three variable categories in this study: student attributes, treatment condition, and outcome. The student attributes are measured by: GENDER, MAJ_GROUP, CLASS_LVL, PCAvg, and, for CS121, PREV. The treatment condition will be determined by the value of BFAvg. The outcome measure is FINAL, the final exam score.

2.3 Propensity Score Matching

In this study, we estimated the effect of BF scores on final exam scores using the technique of propensity score matching. Refer to chapter one for background information on this method. Propensity score matching is used in non-experimental analyses where selection bias is a concern. The fact that we have a non-experimental situation, where students self-select into following the book-first pattern, makes this technique appropriate for our analysis.

Propensity score analysis requires several steps: 1) the definition of three variables or sets of variables: an outcome measure, a treatment condition, and a set of observable variables, X, that are likely to affect a subject's selection of treatment condition and outcome state, 2)

calculate propensity scores based on a model of treatment condition predicted by the covariates in X, 3) apply a matching algorithm to achieve treated and control groups that are balanced on the distributions of X, 4) calculate the average effect of treatment on outcomes. The output of step 3 is crucial to the success of this technique since selection bias is reduced when the treatment and control groups are evenly matched on the covariate (variables in the set X) distributions, thus simulating a random assignment to treatment, at least on the variables in X. We next describe our method and results for these steps.

2.3.1 Treatment variable definition.

In step one, we used final exam scores as the outcome measure. The treatment condition must be a binary variable. Since we are using the continuous variable BFAvg to measure the degree to which a student followed the book-first pattern, we chose a threshold to define a binary treatment and control condition for CHEM11 and CS121 courses, as shown in the following table.

Table 17: Thresholds for defining treatment and control conditions for CHEM111 and CS121.

	T=1, Treatment	T=0, Control
CHEM111	BFAvg>0.85	BFAvg<=0.85
CS121 (Fall + Spring)	BFAvg>=0.95	BFAvg<0.95

These thresholds are depicted graphically in the histograms below. In each distribution, we have chosen a threshold that segregates the high level BF score averages from lower levels.

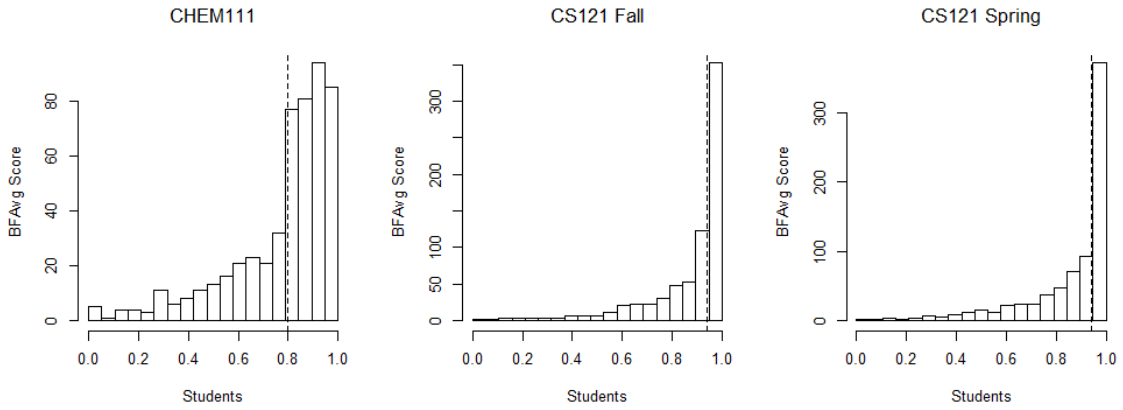


Figure 43: Thresholds for T=1, represented by vertical, dotted lines, based on the distributions of BFAvg scores.

We then defined the set of variables that are likely to affect both a student's selection of following the book-first pattern and exam scores. This set of variables includes gender, previous programming experience (for CS courses), major, class level (section 1.3 describes these variable distributions in detail), and PCAvg, the percent of homework problems attempted as a measure of engagement with the course material to serve as a proxy for the level of overall participation in the assigned work.

2.3.2 Propensity score calculation

In step 2, we calculated propensity scores based on a logistic regression on the variables in the set of observed covariates X. We experimented with several models by adding interaction and higher-order terms. We considered that a more flexible model may help with data that is probably violating the linearity assumption inherent in the regression model, although we do not observe a drastic difference in the distributions of covariates in the treated and non-treated groups. The models we used for propensity score calculation for CHEM111 and CS121 respectively was:

$$T = GENDER + CLASS\_LVL + MAJ\_GROUP + PCAvg + PCAvg^2$$

T = GENDER + CLASS_LVL + MAJ_GROUP + PREV + (MAJ_GROUP * PREV) + PCAvg+ $PCAvg^2$

Since the propensity scores are used in the matching step, we checked their distributions in the treated and control groups to assess the amount of overlap, or common support. Common support ensures that there are representatives from both treatment and control groups over the distribution of all propensity scores. The figure below shows these distributions for the CHEM111 data.

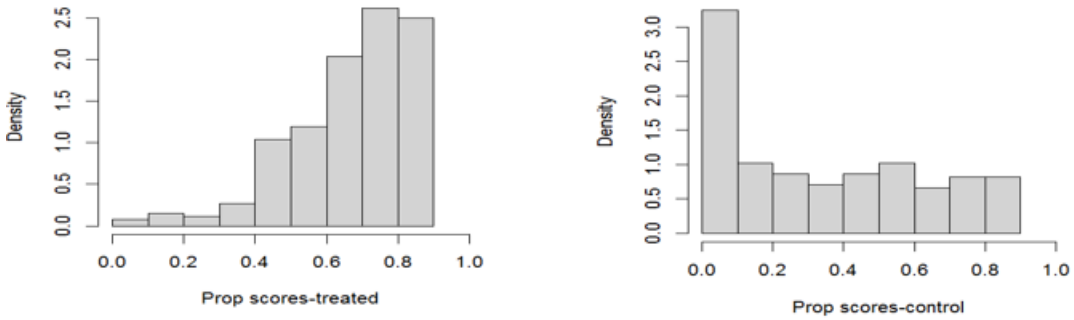


Figure 44: Propensity score distributions for CHEM111 Fall 2012.

From these graphs, we see that there is a fair amount of common support with the exception of the treatment and control groups' lowest scores. In this area there are very few treatment subjects to match with many control subjects. The next figure shows the histogram from the CS data for fall, which is representative of the CS spring distribution as well. We see a similar difference to the CHEM111 histograms in the lowest scores. We can exclude this region in our matching to improve balance.

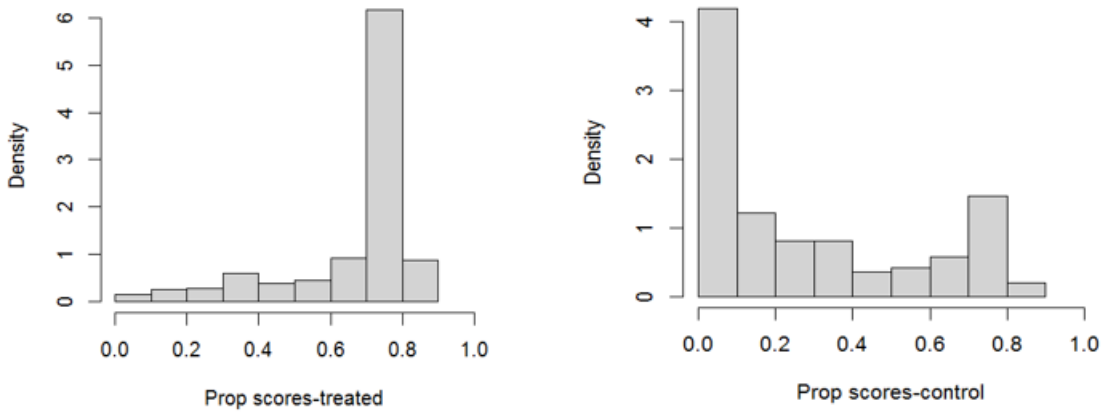


Figure 45: Propensity score distributions for CS121 Fall.

2.3.3 Matching

For step 3, we used two matching techniques using the Match package available in R [Sehkon, 2011]. The first was matching on propensity scores, while the second used a combination of propensity scores and the covariates in X as parameters to a genetic matching algorithm. For the first technique, we used a 1:1 matching with replacement, with the additional function that if one observation in the treatment group matches more than one observation in the control group, the matched dataset will include the multiple matched control observations and the matched data will be weighted to reflect the multiple matches. The second technique used the GenMatch [Sekhon &Diamond, 2005] algorithm (see section 1.5.2.1). We calculated measures of balance on each variable distribution before and after matching to evaluate the success of the matching. These measures included the standard difference of the means, the differences in the variance ratio of the treatment to control groups, summary statistics (mean, median, max diff) on the differences in QQ plots, and the results of a bootstrapped Kolmogorov–Smirnov test, which evaluates the similarity between probability distributions. We

found that the best matching was obtained with the genetic algorithm. An example of the CHEM111 data matching is shown below.

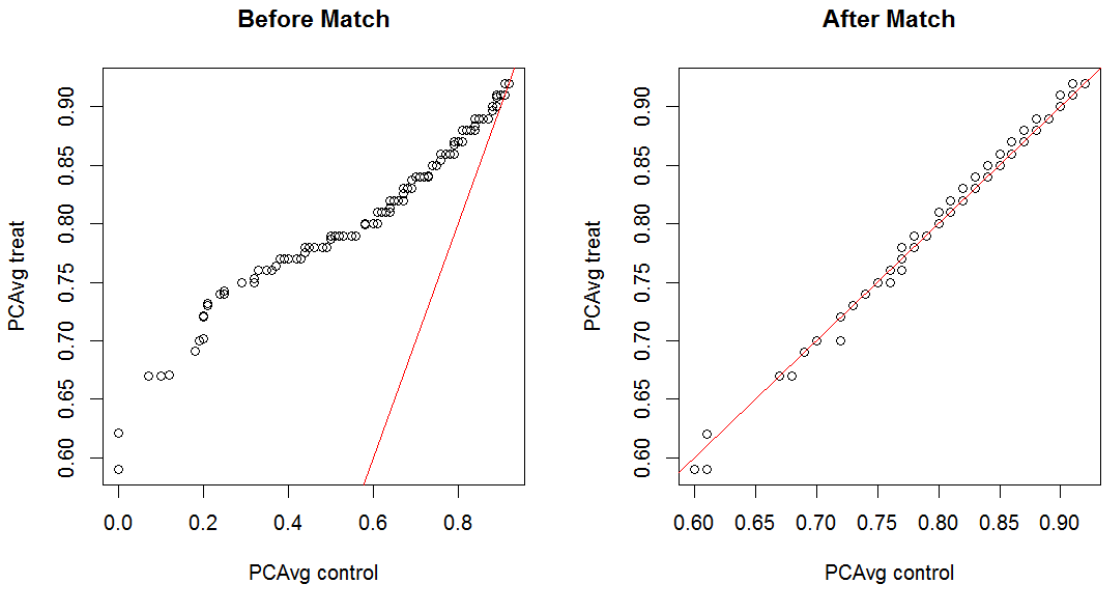


Figure 46: QQ plot for average percent of homework before and after matching for CHEM111 data set.

The results of the matching evaluation produced by the two matching algorithms is partially reported in the tables below. In essence we are comparing the similarity of two distributions of the covariates for the treatment and control groups. We chose to report only the standardized difference in means between the control and treatment groups and the variance ratio for brevity. The best matching would result in a difference of means of zero and a variance ratio of 1.0. Not all model terms are reported below for brevity.

Table 18: Evaluation of matching on CHEM111 data using two matching algorithms. The standardized difference of the means and variance ratios are reported.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	-27.67	-13.95	0.00
CLASS_LVL	- 34.84	-11.57	10.40
MAJOR	-2.130	-9.13	-0.67
PCAvg	302.20	24.47	2.51
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	0.94	0.94	1.00
CLASS_LVL	0.55	0.74	1.17
MAJOR	1.09	1.12	0.98
PCAvg	0.07	0.38	0.98

From this table it is quite apparent that the GenMatch algorithm produced the best overall matching. The standard difference of means was reduced in all cases except for MAJOR, where it actually got worse in the first matching. In the GenMatch column, all except CLASS_LVL were dramatically reduced. The same improvement is seen in the variance ratio, where GENDER is perfectly matched by the genetic algorithm. The other matching evaluation test results were consistent with those in the table above.

The results of the matching evaluation for both the Fall and Spring CS121 data sets produced by the two matching algorithms is partially reported in the tables below. Again, we chose to report only the standardized difference in means between the control and treatment groups and the variance ratio for brevity.

Table 19: Evaluation of matching on CS121 Fall data using two matching algorithms. The standardized difference of the means and variance ratios are reported.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	-27.67	-13.95	0.00
CLASS_LVL	- 34.84	-11.57	10.40
MAJOR	-2.13	-9.13	-0.67
PCAvg	302.20	24.47	2.51
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	0.94	0.94	1.00
CLASS_LVL	0.55	0.74	1.17
MAJOR	1.09	1.12	0.98
PCAvg	0.07	0.38	0.98

Table 20: Evaluation of matching on CS121 Spring data using two matching algorithms. The standardized difference of the means and variance ratios are reported.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	-27.678	-13.95	0.00
CLASS_LVL	- 34.84	-11.57	10.40
MAJOR	-2.13	-9.13	-0.67
PCAvg	302.20	24.47	2.51
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	0.94	0.94	1.00
CLASS_LVL	0.55	0.74	1.17
MAJOR	1.09	1.12	0.98
PCAvg	0.07	0.38	0.98

Once we achieved a reasonable matching, we calculated the treatment effect on outcomes, reported in the next section.

2.3.4 Effect Estimation

We calculated two estimands for treatment effect: the average treatment effect, ATE, and the average treatment effect for the treated, ATT (refer to section 1.5.3). We also computed an adjusted Abadie-Imbens [2006] (A&I) standard error, and a 95% confidence interval. The A&I

error is larger than that produced by a regular standard error calculation in order to take into account the possible introduction of bias as the propensity scores were estimated in a separate step from the calculation of the treatment effect. The following figures and tables describe the results of these calculations for the three data sets.

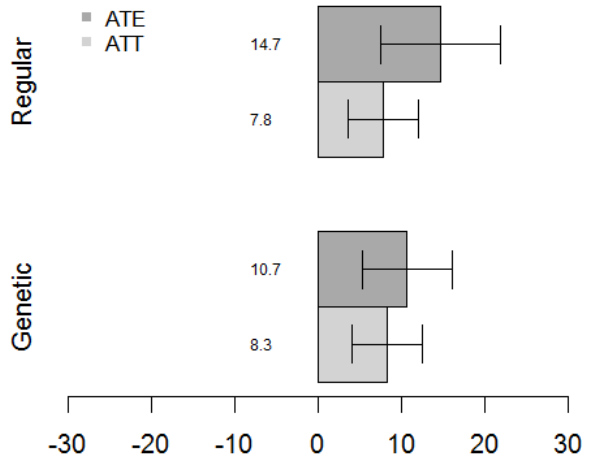


Figure 47: Effect size estimation and 95% confidence intervals (using the A&I computation) for CHEM111 Fall 2012 matched data using the regular and genetic algorithms.

The following table provides the numerical version of the result shown in the figure above.

Table 21: Summary of effect estimands and confidence intervals for the CHEM 111 data set.

Matching	ATE	95% Conf.	ATT	95% Conf.
Regular 1:1 w/repl	14.7	7.6, 21.9	7.8	3.6, 12.0
Genetic	10.7	5.3, 16.0	8.3	4.1, 12.5

From these results we see a positive effect for both matching techniques and for both estimands. The genetic algorithm produced lower estimates than the “regular” matching algorithm. This may be due to the fact that the genetic match quality was better for treatment

and control groups, and that covariate effects were better adjusted for. In both result sets, the ATT effect is lower than the ATE effect, as expected.

The results of effect estimation for the CS121 Fall data set are shown below.

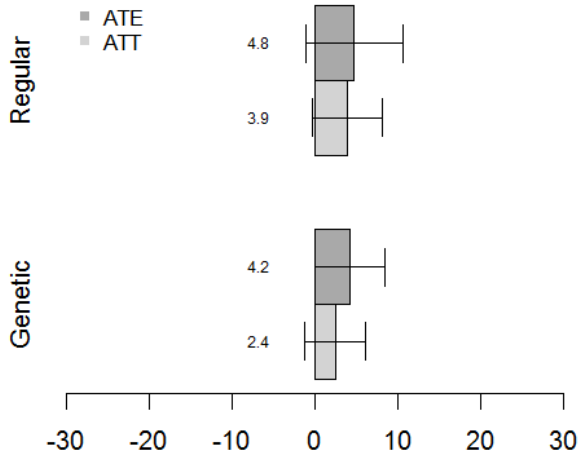


Figure 48: Effect size estimation and 95% confidence intervals (using the A&I computation) for CS121 Fall (2012+2013) matched data using the regular and genetic algorithms.

The following table provides the numerical version of the result shown in the figure above.

Table 22: Summary of effect estimands and confidence intervals for the CS Fall data set.

Matching	ATE	95% Conf.	ATT	95% Conf.
Regular 1:1 w/repl	4.8	-1.3, 10.9	3.9	-0.4, 7.2
Genetic	4.2	0.0, 8.4	2.4	-1.5, 6.3

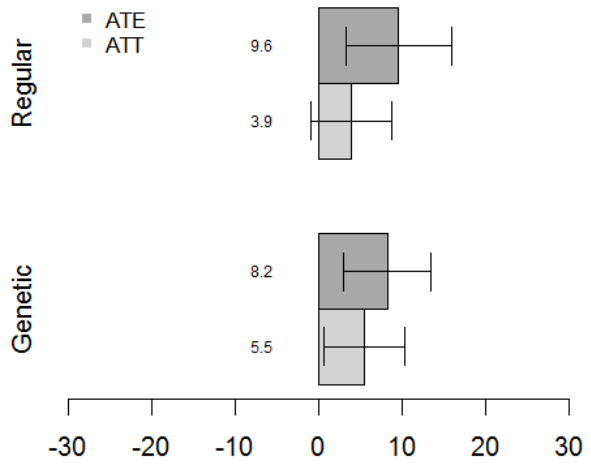


Figure 49: Effect size estimation and 95% confidence intervals (using the A&I computation) for CS121 Spring (2013+2014) matched data using the regular and genetic algorithms.

The following table provides the numerical version of the result shown in the figure above.

Table 23: Summary of effect estimands and confidence intervals for the CS Spring data set.

Matching	ATE	95% Conf.	ATT	95% Conf.
Regular 1:1 w/repl	9.6	3.6, 15.3	3.9	-1.8, 9.5
Genetic	8.2	3.6, 12.9	5.5	0.9, 10.1

The results from the CS121 fall and spring data sets shows a much lower effect size when compared to the Chemistry results. This is likely due to the fact that there are fewer students not following the book-first pattern in the CS121 courses, resulting in fewer counterexamples. There are also fewer assignments in the CS courses. If the content for CHEM111 and CS121 were of similar granularity we might see a more similar distribution of BF scores between the Chemistry and CS courses, with perhaps more similar results. The fact that the effect estimates are all in the positive direction is an encouraging result for the positive effect of the book-first pattern, as measured by the BF score, on final exam scores.

In the next section, we estimate the effect of BF scores on final exam scores for manually selected subpopulations as a check on the results obtained by propensity score matching.

2.4 Analysis of selected subpopulations.

In this section, we study specific subpopulations in isolation from the general population of students in our data sets. In the previous section, we used propensity score matching on the entire population to estimate the effect of BF scores on the outcome variable, final exam scores. That process involved carrying out many steps and making many assumptions. In this section we make as few intervening steps as possible in estimating an effect. We do this as a way of checking the results we obtained from the much more complex analysis above. The main drawback with our approach in this section is that we are losing statistical power since we are examining smaller data sets. Another issue is that we are not attempting to balance all covariates in the subpopulations we select. An imbalance in covariates may lead to biased results, the type of bias we were adjusting for in the previous section. We are, however, avoiding the bias introduced by the methodology used in that section.

Our method here is to manually isolate specific subpopulations of interest and calculate the correlation between the average BF scores and final exam scores. A correlation must exist for there to be a causal link. The results from the previous section make the case for a causal effect as self-selection was adjusted for. In this section, we check for a correlation with far simpler methods. In addition to an analysis of the entire data set, we selected subpopulations of the data according to four hypotheses we have postulated about how certain groups of students will be affected by the book-first pattern. Generally speaking, we believe that novice students are more likely to benefit from following the book-first pattern.

2.4.1 Method

We estimated the effect of following the book-first pattern on the general population of the three datasets used in the previous section: CHEM111, CS121Fall, CS121Spring. This was done to check the results we obtained above. We then assessed the effect of the book-first pattern for four specific subpopulations about which we had developed hypotheses. Our hypotheses regarding the effect of the book-first pattern on the accompanying subpopulations are shown in the following table.

Table 24: List of hypotheses and subpopulations. Note that hypotheses 3 and 4 apply only to CS121 data sets.

	Hypothesis	Subpopulations
1	First year students will show a greater positive correlation between BFAvg and final exam scores than other class levels.	CLASS_LVL=1 vs. CLASS_LVL>1
2	First year female students will show a greater positive correlation between BFAvg and final exam scores than first year male students.	CLASS_LVL=1, GENDER=female vs CLASS_LVL=1, GENDER=male
3	In CS121 course: novice, first year students will show a greater positive correlation between BFAvg and final exam scores than first year students with Java experience.	CLASS_LVL=1, PREV=none vs CLASS_LVL=1, PREV=Java
4	In CS121 course: first year non-computer science majors will show a greater positive correlation between BFAvg and final exam scores than first year computer science majors.	CLASS_LVL=1, CS_MAJOR=0 vs CLASS_LVL=1, CS_MAJOR =1

Before manually selecting the data for each subpopulation, we adjusted for the level of participation, as measured by the variable PCAvg, which is the percentage of assigned homework attempted over the entire course.

Given the fact that participation is strongly associated with both book-first and final exam scores, we selected data with a PCAvg value of 85% or higher for CHEM111 and 90% or higher for the CS121 data sets. These thresholds are depicted in the following figure.

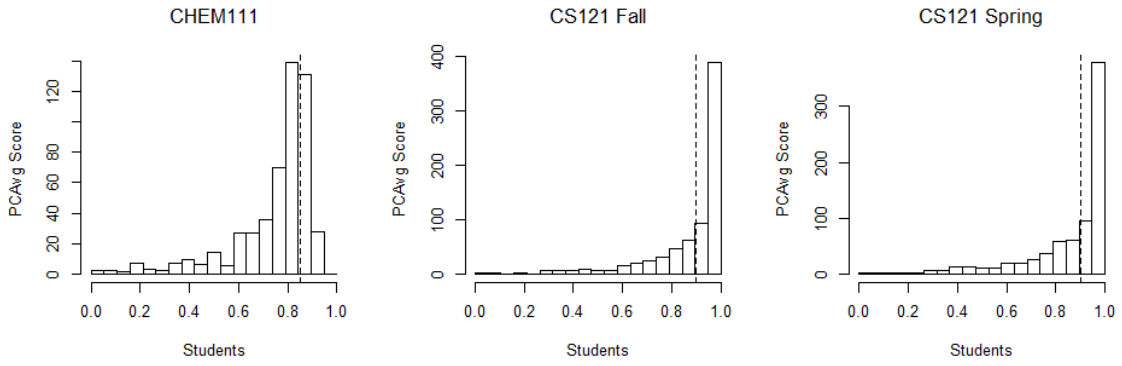


Figure 50: Average percent homework attempted (PCAvg) distributions with 85% and 90% thresholds.

These thresholds capture the highest participating students. Any effect of BF scores on the final exam is more likely to be attributable to the book-first pattern than to participation level.

After selecting the specific subpopulation (with the specified range of PCAvg values), we calculated an effect size with Pearson's correlation coefficient, r, and simple, linear regression. We report on a 95% confidence interval for r and the significance level of the estimate. We also report the regression coefficient, its significance and standard error along with the amount of variance explained by the model (R^2). We also plot a regression line to help visualize the correlation. The simple regression model is: FINAL = Coeff * BFAvg + Constant.

Our regression coefficient is the change in final exam points expected, on average, with the change in one unit of BFAvg. Since these values range from 0 to 1, dividing the coefficient by 10 provides the change in exam scores for a one tenth change in the BFAvg value. The Constant would provide the final exam score with a BFAvg of 0.

2.4.3 Results

The results for the entire data set and for each hypothesis are presented in the following tables and figures. Note: asterisks indicate significant p values: *** 0.001, ** 0.01, * 0.05.

Entire data sets: The results for the entire data sets are presented below, first by regression plots and then in tabular format.

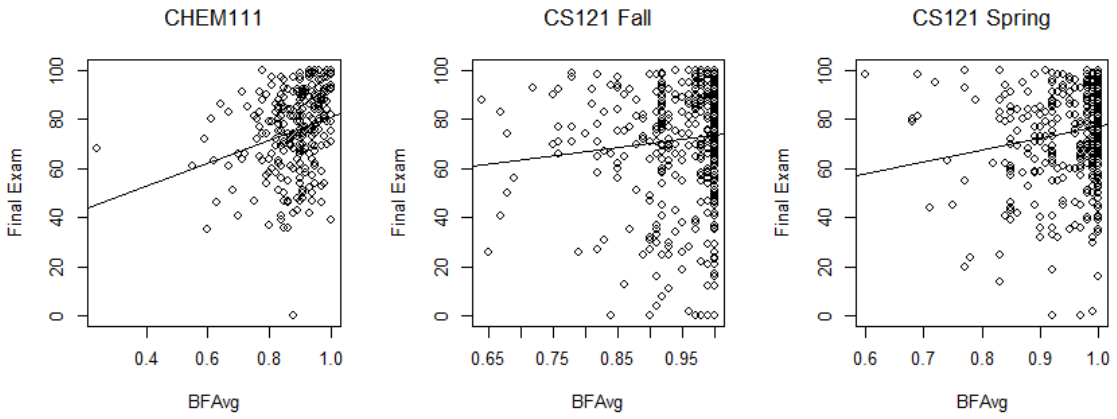


Figure 51: Scatter plots of final exam scores vs. BFAvg with regression lines for CHEM111, CS121 Fall and Spring.

Table 25: Results for CHEM111, CS121 Fall and Spring. (Data selected above the PCAvg thresholds of 85% and 90% respectively).

	N	r	95% conf	Reg coeff	Std Error	R^2
CHEM111	159	.27***	0.15 0.37	46.7***	10.2	.13
CS121 Fall	482	.11*	0.04, 0.18	29.0*	13.2	.01
CS121 Spring	472	.22***	0.13, 0.29	58.6***	11.5	.09

From the results above, we see that the relationship between BFAvg and final exam is small but positive in each dataset. The amount of variation in the data is relatively large, and the regression models have a corresponding low r-squared value. Though the coefficients are significant, meaning there is a correlation with the response, these models would not be useful for prediction purposes. The largest correlation is CHEM111, with CS121 Spring next and CS121

Fall least. This echoes the results from the propensity score analysis. Next we present the results of the subpopulations for the four hypotheses.

Hypothesis 1: First year students will show a greater positive correlation between BFAvg and final exam scores than other class levels.

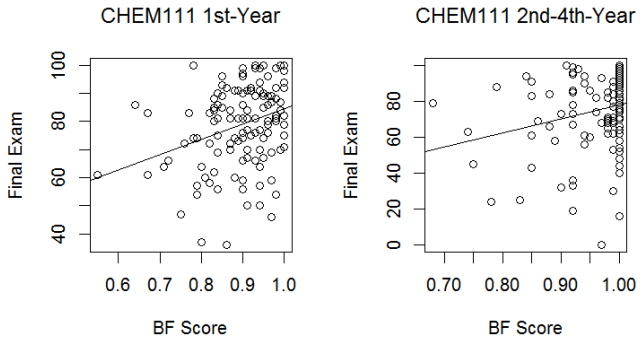


Figure 52: Plots of the data sets with regression lines for CHEM111.

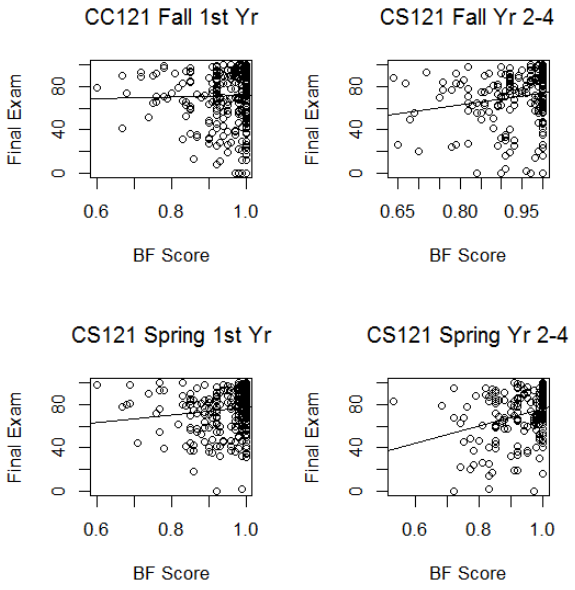


Figure 53: Scatter plots for CS121 first and second year student subpopulations.

Table 26: Hypothesis 1 results.

		N	r	95% conf	Reg. coeff	Std Error	R^2
CHEM111	1st Year	128	.29***	0.16, 0.40	45.2***	10.2	.09
	2nd-4th Year	31	.26	-0.01 0.47	53.0*	28.1	.06
CS121 Fall	1st Year	318	.02	-0.09, 0.12	9.8	10.0	.01
	2nd-4th Year	211	.18*	0.04, 0.30	54.8*	18.2	.04
CS121 Spring	1st Year	325	.13*	0.02, 0.22	34.2*	13.2	.05
	2nd-4th Year	202	.29***	0.17, 0.41	78.5***	18.4	.08

From the results above, we can see that the hypothesis seems to apply to the CHEM111 data but not for the CS121 data, where the opposite seems to be true: the first year groups showed little to no correlation.

Hypothesis 2: First year female students will show a greater positive correlation between BFAvg and final exam scores than first year male students.

Female and male students seem to follow the book-first pattern in a similar manner as evidenced by the histograms in figure 23 in section 2.2.2.1. The correlation results for these subpopulations are presented below.

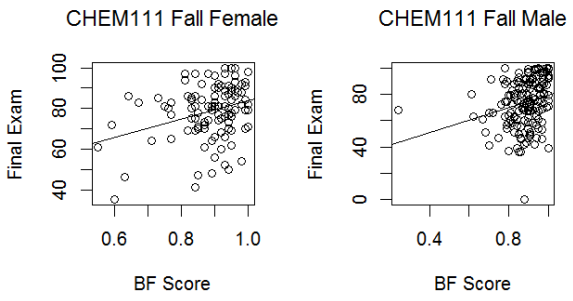


Figure 54: BFAvg vs. Final Exam scores for CHEM111 females and males.

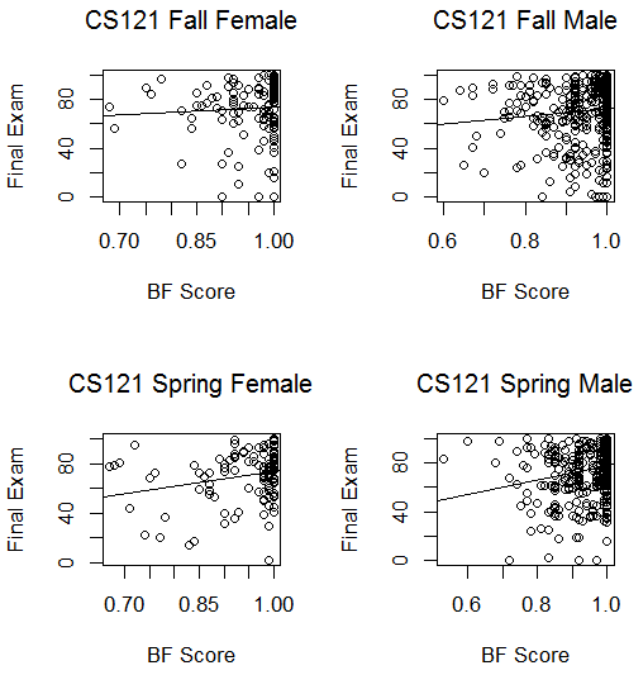


Figure 55: BFAvg vs. Final Exam scores for CS121 females and males.

Table 27: Hypothesis 2 results (First year female and male).

		N	r	95% conf	Reg. coeff	Std Error	R^2
CHEM111	Female	109	.31***	0.12, 0.46	44.2**	12.2	.08
	Male	164	.25***	0.10, 0.38	47.2**	14.1	.07
CS121 Fall	Female	130	.05	-0.12, 0.22	19.0	25.4	.00
	Male	413	.10*	0.00, 0.19	30.8	13.7	.06
CS121 Spring	Female	125	.23**	0.05, 0.39	59.2**	20.3	.06
	Male	407	.20***	0.11, 0.30	57.5***	12.5	.08

These results do not show any significant difference in correlation between male and female students for any dataset. This is consistent with other findings in learning science courses [Lau, 2009; Murphy, 2006].

Hypothesis 3: In CS121 course: novice, first year students will show a greater positive correlation between BFAvg and final exam scores than first year students with Java experience.

Both novice and Java experienced students followed the book-first pattern to a similar degree, as evidenced by the following histograms of the distributions of their average BF scores.

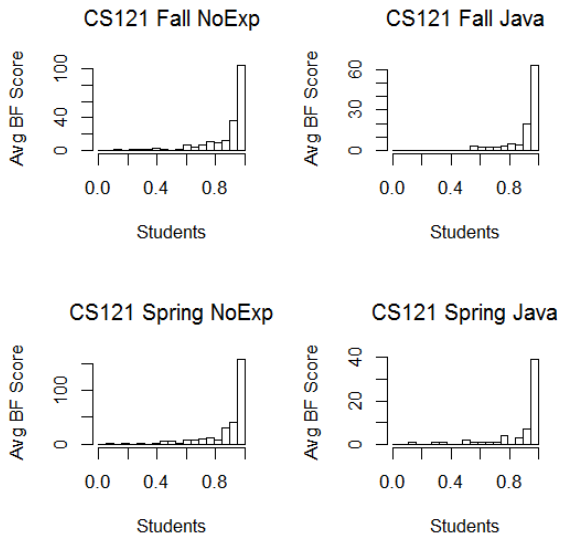


Figure 56: BFAvg distributions for 1st year novice and Java-experienced students in CS121 courses.

The following results show the correlation between their average BF scores and their final exam scores.

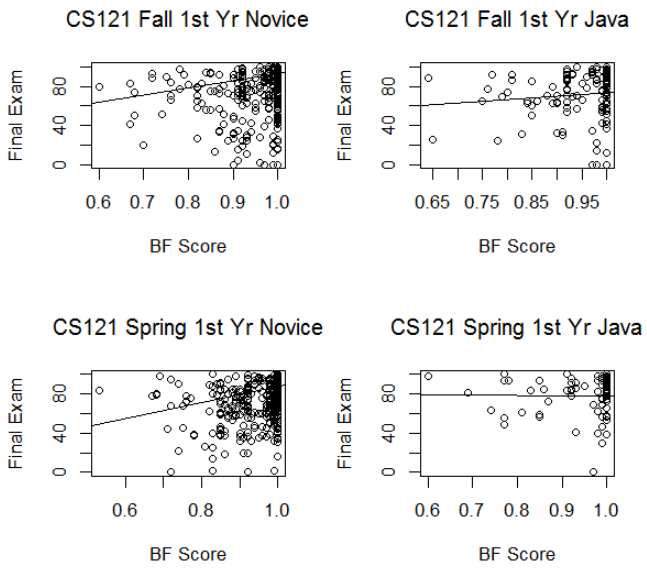


Figure 57: CS121 Fall and Spring data sets.

Table 28: Hypothesis 3 results.

		N	r	95% conf	Reg. coeff	Std Error	R^2
CS121 Fall	Novice	155	.20**	0.11, 0.30	59.0*	20.8	.07
	Java	89	.05	-0.12, 0.22	30.1	15.2	.01
CS121 Spring	Novice	227	.23**	0.06, 0.39	55.3***	13.4	.08
	Java	52	.09	-0.01, 0.19	19.0	30.4	.00

The results show significant correlations for the novice students but not for the Java-experienced students.

We also plotted the final exam distributions for novice versus Java-experienced students based on their level of BFAvg (following the book-first pattern). These distributions, presented below, show that for novice students the mass of the exam distributions are located more rightwards (higher scores) for subpopulations with BF scores of 1 than those with BF scores of less than 1. The Java-experienced distributions show no such difference.

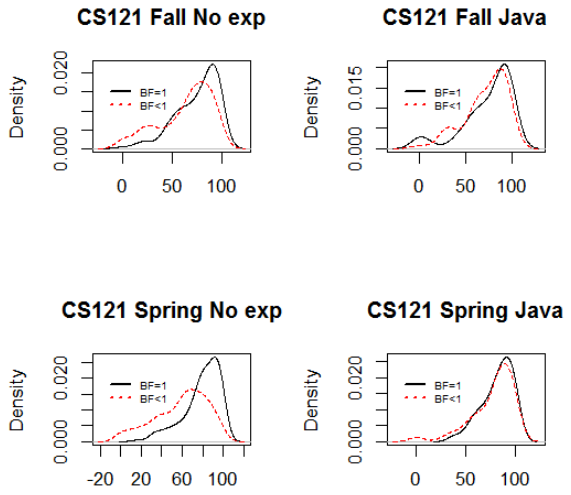


Figure 58: Final exam score density plots for CS121 courses for novice and Java-experienced students with BFAvg of 1 and less than 1.

Hypothesis 4: In CS121 course: first year non-computer science majors will show a greater positive correlation between BFAvg and final exam scores than first year computer science majors.

The following figure depicts the distributions of BFAvg scores for Non-computer science majors versus computer science majors. These distributions show that both subpopulations appear to follow the book-first pattern in a similar manner.

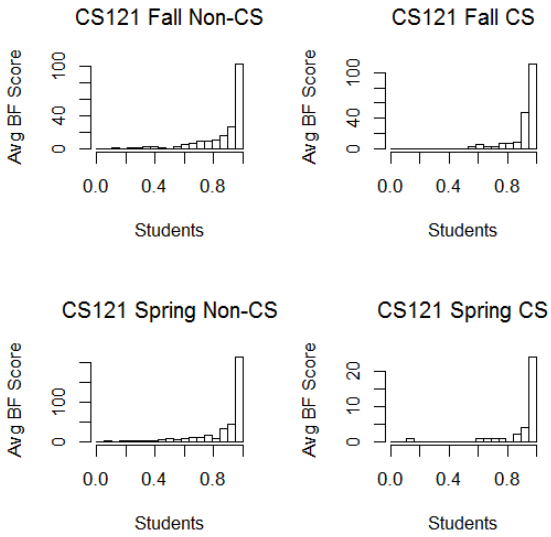


Figure 59: BFAvg distributions for 1st year non CS majors and CS majors in CS121 courses.

The following results show the correlation between their average BF scores and their final exam scores.

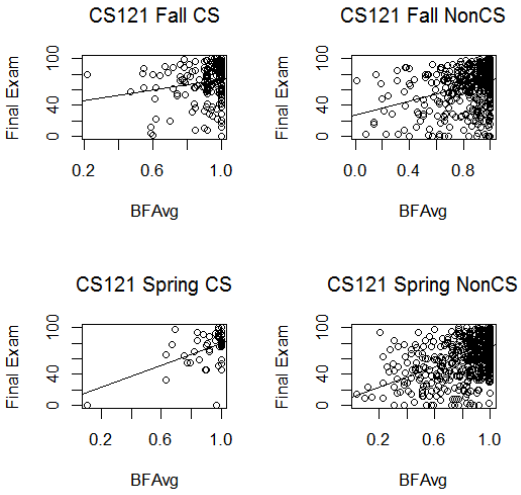


Figure 60: CS121 Fall and Spring data sets.

Table 29: Hypothesis 4 results.

		N	r	95% conf	Reg. coeff	Std Error	R^2
CS121 Fall	CS Major	170	.17*	0.03, 0.29	33.4*	13.5	.04
	Non-CS Major	148	.32***	0.23, 0.39	45.1***	6.4	.12
CS121 Spring	CS Major	30	.48***	0.24, 0.68	71.2***	18.3	.25
	Non-CS Major	295	.49***	0.42, 0.53	65.4***	5.6	.29

From the results above, it is clear that the book-first pattern is less significant for CS majors in the fall data than non-majors. The results for the Spring show strong, positive correlations for both subpopulations regardless of status as a CS major. The following table summarizes the results from the analysis of subpopulations in this section.

Table 30: Summary of results for the four hypotheses in section 2.4. The parentheses in the Results column show a quick summary, where + means the hypothesis was upheld by the result, and – means it was not.

	Hypothesis	Results
1	First year students will show a greater positive correlation between BFAvg and final exam scores than other class levels.	(+,-,-)True for CHEM111, CS121 showed 1st year less than 2nd-4th year.
2	First year female students will show a greater positive correlation between BFAvg and final exam scores than first year male students.	(-,-,-) CHEM111 equal in correlation, CS121 no significant difference, perhaps males slightly more.
3	In CS121 course: novice, first year students will show a greater positive correlation between BFAvg and final exam scores than first year students with Java experience.	(+,+) Novice students show higher, positive correlation for higher BF scores.
4	In CS121 course: first year non-computer science majors will show a greater positive correlation between BFAvg and final exam scores than first year computer science majors.	(+,-) Fall CS has lower correlation while spring shows no significant differences.

2.5 Conclusion and Discussion

From the results of our analyses above, we can conclude that BF scores, on average, seem to have a positive effect on final exam scores. This was shown to be the case in the propensity score matching analysis especially for the CHEM111 data set, and less so for the

CS121 data. This is perhaps due to the fact that the CHEM111 course has more content, which provides more samples of behavior. It is true that the CHEM111 data had more examples of students with high participation and low book-first patterns. This suggests there were more counterexamples to the "treatment" of following the book-first pattern to a high degree. The fact that the CS data showed less of an effect is most probably due to the fact that there were fewer counterexamples of not following the book-first pattern, i.e. most students followed that pattern to a larger degree, making the effect harder to measure. The results of the second analysis for the entire data set, with participation rates (PCAvg) selected for, reinforced the propensity score results as we saw a correlation for the averaged BF scores and final exam scores. The absence of a positive correlation here would cast doubts upon the result for the previous analysis.

The results from section 2.4 were mixed. The main theme of our hypotheses were that novice students would benefit more from the book-first pattern. We did find some evidence for this in hypothesis 1 for CHEM111, and in CS121 for hypothesis 3, relating to programming language experience. The use of first year as a measure of a student's novelty in hypothesis 1 is fairly crude. We assume that first year students have less experience with college-level study habits and in assimilating the amount of material presented at the college level. This may be true about chemistry more than it is for a domain such as learning about programming, which is partially a skill that is mastered through practice.

Hypotheses 2 was not supported by the results. Perhaps there was a slightly more positive correlation for males in the CS fall data, however; we claim that there was no difference between females and males for any of the data sets. We note that the correlations agreed with the results for the whole data set and for the propensity score analysis: CHEM111 showed a higher effect with CS Spring next and then Fall.

Hypothesis 4 was supported by the fall but not the spring data. This is likely due to the disparity in the size of the CS subpopulation in the spring semester. Perhaps those student are also more experienced in some way as well.

2.5.1 Discussion

The main concern of this study is adjusting for influences that would affect a student's choice of following the book-first pattern and final exam scores. We attempted to do this in the first analysis by the method of propensity score matching, and to a lesser extent by manually selecting specific subpopulations for study in the second analysis. Both of these methods rely on the fact that we can observe all of the influences that might confound our claim of causal effect for book-first. In fact, there are many influences that we did not measure. One such influence is a student's ability to learn the material, which we refer to here as aptitude.

The matter of aptitude as a confounding influence is important for our results. Students who are better able to learn this material may also choose to do all of the work, and have high BF scores. They will also do well on the exams. If they did not follow the book-first pattern they would also do well on the exams. We did not include a proxy variable for aptitude in our study. One reason is that we did not have a pretest to measure what we would consider aptitude for the computer science courses. Approximately 95% of the chemistry students have taken chemistry in high school as measured by survey. In both cases, the most likely and easily obtainable proxy would be standardized test math scores. Including this measure in future studies is planned.

A significant assumption we made is that we are successfully adjusting for the degree of participation. Participation is the level of engagement with the assigned material. Hi participation is also a component of good study skills. Our proxy for participation level was the

percentage of homework attempted. We assumed that first year students would have a lower level of previous study skills. Of course, students with higher degree of participation would do the assigned work, and hence follow the book-first pattern. They would also likely score higher on the exam. If that is the case, then our propensity score matching should have adjusted for this as percent homework was balanced in the treatment and control groups. We also assumed that first year students would be more likely to have less experience with study skills. We did not see a large difference in correlation of first year students in section 2.4, hypothesis 1 results. Perhaps prior study skills are more evenly dispersed among the entire population.

The matching methods in section 2.3 involve choices that may affect results. Besides the fact that we can adjust only for observed confounding variables, the type of matching algorithm we used affected the magnitude of effect sizes. The quality of the matching, that is, the balance of covariates, is crucial to the validity of the results. We used a matching algorithm that sampled from the population with replacement to create treatment and control groups. This type of "bootstrapping" is commonly used, but may create bias in the results. The genetic algorithm we utilized may also have internal parameters that add bias. The best way to avoid these problems is to have complete overlap of values for all covariates. In fact, the biggest problem we have in this study is the lack of hi participating students with BF scores of zero, i.e. did not do any book questions before starting homework.

Although there is other teachable content available in the course, such as lecture videos, short tutorial videos, as well as live lecture and discussion sections, we used only textbook access as a proxy to represent a student's access to teachable content. It is also possible that students receive their information from other sources, such as web queries and textbooks, as well as the consultation of other individuals. Access to outside sources is a potential confounding influence on our study in that students may not access the course system

as much if they are doing their learning elsewhere, and the learning we attribute to use of our system was actually due to another source. One way we have of dealing with this type of access is to control for the level of system usage in our study by grouping students according to the amount of work done in the system. While this does not control for the amount of learning done outside the system, it does identify students who are accessing the system to similar degree.

As previously stated, there are many ways that students may learn the material before starting their homework. We are assuming that their recorded book question answering activity will accurately reflect their level of interaction with that material, and serve as a proxy for their commitment to learning the material before doing homework problem solving. Another assumption we make is that the book material is of good quality and that the majority of students would learn from it if they fully engaged with the material. Of course this is not true in every case as students learn in many different ways. Other, non-observable effects are the timing of external events, such as exams, course load, and other commitments to student time and energy. Of course, the innate ability to learn is another factor that we are not able to observe in this study. A major concern is that more able students will choose a Book-first strategy, and that we are really observing ability, not the effect of engaging with the text. A more able student might find the book work easier, or have better time-management skills, and the book work has little or no effect on their outcomes, while students who find the material more difficult may not attempt all of the questions, or may not budget enough time for the book work. Many students view the homework problems as more important as they are worth more points, and may skip some or all of the book questions. These behaviors may change over time. We are aggregating at the course-wide level and may not be accounting for effects of these changes.

The fact that we found a positive effect consistently across several courses and tends to reinforce our confidence that the effect is real. It does not, of course, provide us with the basis for making an absolute causal claim. We plan on future studies where we have access to test scores and other data a priori to our study that can help us measure these qualities. Another way for us to make a causal claim is to manipulate the amount of Book-first work in an experimental setting. We would design an experiment with two groups: those with the current system, and those with a different set of content where the text was integrated with the homework problems. This configuration is actually used by several publishers in the belief that students don't learn best by reading their textbooks, and so they encourage the use of the text as a lookup reference. The key difference between their approach and ours is that their text is static in the sense that there are no embedded questions. It would not make sense to have more questions in a text that is meant to be used as a reference. We, on the other hand, have moved questions, albeit specially designed questions, into the text. Our experiment would involve some reconfiguration so that book usage as a reference was encouraged during homework.

CHAPTER 3

ANALYSIS OF THE "INFREQUENT CONTACT" PATTERN

3.1 Introduction

Although the Book-first strategy may be effective for novice students, there is still variability in their outcomes. Although we recognize that there are many factors that influence students and their outcomes, and that most of those factors are not observable by our study, we investigated the categorization student interaction with the material; we sought patterns of usage that recurred throughout the semester and bore a relationship to outcomes. As previously stated, many instructors, past and present, believe that students do not read their textbook. The courses in our study use an electronic text that contains assignable, embedded questions. Thus, the text now appears more like a homework assignment, and, as shown previously, the number of students who work the embedded questions has increased dramatically. The fact that a student does all of the assigned work says that they are following the right plan to succeed, yet it does not say anything about how they are doing their work. In fact, the same issue applies to homework problems as well. In this part of the dissertation, we study the pattern of how student work in the OWL system: how many times they access the system and for how long.

Students work in the OWL system, and in any computer-based learning system, in time intervals called sessions. The system records the start of a new session when a student accesses an assignment, and records the end of the session when the student logs off, or when a predetermined time interval has elapsed, typically 30 to 60 minutes. We investigated if there were discernable patterns of sessions and between session intervals, and what, if anything, the pattern of sessions and the time intervals between sessions would reveal about outcomes.

We posit two hypotheses: 1. Students who engage for relatively short periods of time between long intervals are likely not engaging with the work in a manner that is beneficial for learning, and 2. Students who engage for long or long sessions with frequent, short intervals in between work sessions will have higher outcomes on average.

3.2 Method

To test these hypotheses, we needed to represent the lengths of time that students worked and the duration of the time intervals between their working sessions. These representations could then be analyzed to discover the patterns of interest and their effect on outcomes measured.

We calculated the length of student sessions as well as the time intervals between their sessions in minutes for each chapter or section. Each module may vary in several aspects: the nature and difficulty of its content as well as when it appears during the semester. Another consideration is that the student population changes over the fourteen week semester as students drop out or otherwise disengage from the course. In many large introductory STEM courses this can amount to a sizable group. Therefore, modules that come later in the course tend to have students who are still actively engaged and more successful as a result. This change in population and content is likely to affect the pattern of work that we observe. For example, the CS121 content becomes more abstract and more challenging in the last half of the semester. Given the changing nature of the college course environment it makes sense to perform our analysis on a per module, or chapter, basis.

For each chapter, we generated a sequence of sessions and intervals between sessions for each student. Each session time and interval time were calculated in minutes. We calculated the duration of a session in the following manner. The start date and time is the beginning of

the duration. The end of the duration is determined by the last recorded activity in the OWL system. If no activity was recorded, the session timed out. We used a somewhat arbitrary length of 20 minutes. This was arrived at from the observation of the data and by consultation with former students and instructor with knowledge of the course. One complication is that, although session start and, sometimes, end times are recorded, it's possible that overlapping sessions might be recorded. For example, a student may start working on an assignment and 15 minutes later starts another assignment. The system will record the start of the first session but the student has not logged off, so no end time is recorded. The second session start time is recorded. The fact that the sessions overlap means we have to consider both sessions as one "logical" session. The figure below depicts the scenario of overlapping sessions.

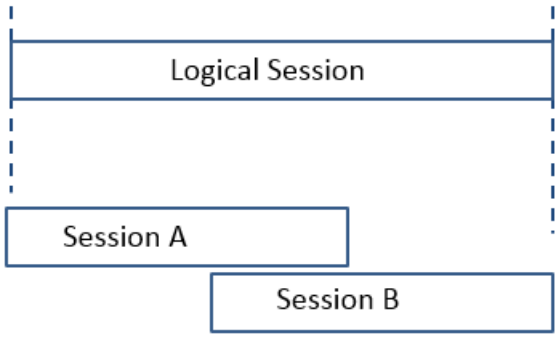


Figure 61: Overlapping sessions.

Once we have the logical sessions and their durations, we can calculate the intervals between sessions easily. We now have a sequence of sessions and intervals, each with its duration in minutes.

3.2.1 Encoding

To better visualize and handle these sequences, we represented sessions and intervals symbolically. For example, a long session was represented by a capital L, a short session by a

capital S. A long interval was represented by a small l and a short interval by a small s. Sessions and intervals that were determined to be neither long nor short were represented by M and m respectively. To determine what constituted long, short, and medium sessions and intervals, we used quartile cut points on the distribution of session and interval duration times for the chapter. The following table summarizes the symbolic assignments based on the time duration of the student sessions and intervals between sessions.

Table 31: Symbolic representation of sessions and intervals.

Sessions	Intervals	Percentile	Typical Session Range	Typical Interval Range
S	s	Lowest 25%	1 to 5 minutes	20 min to 5 hours
M	m	Interquartile range	6 to 59 minutes	5 hours to 80 hours
L	l	Highest 25%	1 to 7 hours	80 hours and up (3.3 days)

An example of a short session is 2 minutes, a medium session 25 minutes to an hour, and a long session can be several hours. Intervals are, of course, much longer and of greater range. We set a minimum interval time of 20 minutes. Any shorter interval is not considered a true break between sessions. Short intervals can be up to 5 hour. Medium sessions can be over three days. A long interval can in theory be until the end of the course. Since we only consider work done before the chapter due date, the longest interval would be for a student who started an assignment very early in the course and then restarted a session for that assignment again right before the due date. The ranges in the table are just examples as the actual values depend on the chapter and population. All chapters are close to these ranges. The following histograms show an example of cut points on the distributions of session duration times for CS121 Fall 2013 Ch7.

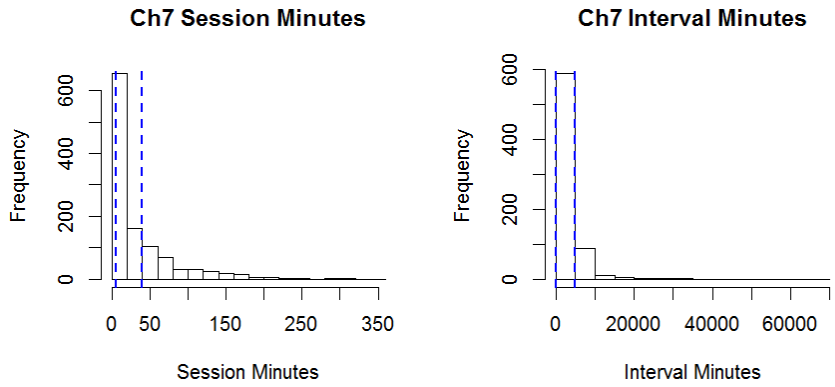


Figure 62: CS121 Fall 2013 Ch7 Session and Interval Durations with S, M, and L cut points.

Example encodings might be:

Student 1: S s S s M
Student 2: S l L m L m M s M m L l L l M

Student 1 has three sessions, the first two are of relatively short duration while the last session is of medium duration. The intervals are both short. This pattern can be described as a sequence of two quick sessions followed by a longer session. Student 2 starts with a short session and a long interval, followed by a long session and medium interval, and so forth. Student 2 has more sessions and of longer duration than student 1, and represents a much different pattern of working.

Consider the following pattern:

Student 3: S l S m S l S m S

This student exhibits the type of pattern we hypothesize as likely to be unhelpful. This student has short sessions followed by long or medium intervals. Perhaps this student is not "staying in touch" with the work enough.

3.2.2 Relationship between patterns and outcomes

In order to test the hypotheses in section 3.1, we have to relate the patterns of sessions and intervals to final exam scores. We defined two simple "patterns", each consisting of one session and interval pair. The following table summarizes these pairs and their putative effect on outcomes.

Table 32: Two pairs of Session-Interval symbols and the predicted sign of the difference between their average exam scores with the population average.

Session/Interval	Representation	Pair avg. - Pop. avg
Short/Long	Sl	negative
Long/short	Ls	positive

We counted the frequency of occurrence of each of these pairs for each student in our data sets. We used two data sets: one for CHEM111 and one for CS121, for which we combined the four semesters. We included data in the Sl group if there was at least one instance of the pattern in the count. For example, if a student ever recorded a Sl pair in their encoding string that student was included in the Short/Long group. In the table below, we see that there were 49 students out of 514 students who had the Sl sequence, and only 22 with the Ls sequence. We included only data with an average percent homework attempted of 70% or higher as a way to adjust for an effect on outcome and pattern group.

Table 33: Size of the groups in both data sets. Only data with PCAvg > .70 included.

Group	CHEM111	CS121
Short/Long Sl	49	120
Long/short Ls	22	55
totals	388	1280

We then compared the Sl and Ls group final exam averages compared to the population averages. Given the large disparity in the data sizes, we performed t tests on samples from

larger distribution. We repeated the tests on 1000 iterations of sampling with replacement. A summary of those comparisons are shown below.

Table 34: Comparison of final exam score distributions between the Short/Long and Long/Short groups and the entire population distributions. The asterisk denotes a significant difference of means as determined by t test, $p<0.05$.

CHEM111	mean	median
Entire population	73.6	76
Short/Long Sl	65.1*	69
Long/short Ls	74.7	77
CS121		
Entire population	67.2	73
Short/Long Sl	62.7	70
Long/short Ls	69.1	74

3.2.2.1 Conclusion

From the tables above we see that the results are not conclusive. While there is a significant difference in the exam scores for the Short/Long group in the CHEM111 course, the CS121 data failed to show a significant difference. There were no significant differences between the Long/Short group and the general population. We conclude that while there seems to be a trend for lower exam scores for the Short/Long group, more work has to be done before we can claim a result.

There are several ways our method can introduce bias in these results. The choice of discretization intervals for performing the encoding are somewhat arbitrary. Considering frequencies of pairs of sessions and intervals alone may not be capturing the intended pattern. Finally, the relative rarity of occurrence of these pairs makes inference more difficult as the sample of positive (for the pattern) examples are low in comparison to the negative examples.

3.3 Further modeling on the encodings

Next, we experimented with combining a bag of words model with clustering techniques in an attempt to more precisely identify patterns in our symbolic encodings.

3.3.1 Bag of words model

At this point, we have a list of encodings that represent session activity for each student. Our goal is to find clusters of these activity strings that are describe meaningful behavior and have some relationship to outcome measures. One way of clustering such encodings is to represent each symbol by its frequency of occurrence in the string. This is the well-known bag of words (BoW) approach used to categorize documents by the topics they describe. In that approach, a document is represented by a vector of the frequency of occurrence for each word that it contains. A dictionary that contains all words in the document set is maintained. The number of words in the dictionary is the size of the vectors that represent the documents in the catalog. If a word does not exist in a document there is a 0 in that word's place in the vector. In this way, documents can be compared to other documents by a distance metric on their representative vectors. Each document vector is compared to a topic vector (made from all of the documents known to describe that topic) to see if that document should be labeled as belonging to that topic. In our case, we do not know the categories we may discover; we are clustering and not classifying, but we will use the BoW representation of our encodings.

Our dictionary, or vocabulary is simple the six letters: S, M, L, s, m, l, so each encoding will be represented by a frequency vector of length six. For example, the encodings of the three student examples above are shown in table 21 below.

Table 35: Frequency vectors for three example encodings.

	S	M	L	s	m	l
Student 1	2	1	0	2	0	0
Student 2	1	3	4	1	3	3
Student 3	5	0	0	0	2	2

It is interesting to note that there are ways of creating an "augmented" BoW model to capture some of the sequential nature of the encodings which the single letter (or word) frequencies miss. This is done in various ways. For example, regular expressions can be generated to be matched on the encodings, or by other sampling means. These techniques are used in video pattern recognition. We plan to explore the augmentation approach in future work.

3.3.2 Clustering

Now that we have frequency vectors calculated we proceed to the clustering task. There are two main parameters to any clustering problem: the number of clusters to use, and the metric to use to calculate similarity (or dissimilarity) between the data. The latter parameter is referred to as a distance metric. We are dealing with numeric data, so we used the Euclidean distance metric. This is the standard geometric way of calculating distance on numerical data. Given that our data is numeric and low-dimensional, we chose the Euclidean metric. We calculated a distance matrix on our data, which is simply an NxN table of distances between the data points, N being the size of the data. We also investigated the use of a dissimilarity matrix generated by a random forest (RF) for comparison with the Euclidean distance matrix, as some research on genomic data has found the RF dissimilarity matrix to be superior to the Euclidean. As described in the section on related work, a random forest is an ensemble of tree classifiers. Trees partition data in a different manner than the Euclidean metric. The details of how the

random forest, normally used for classification, is used for creating the dissimilarity matrix is described in the related work section. We did not find the RF generated matrix to be superior, probably because of the small size and homogenous scale of our data. In the future we plan on implementing an augmented BoW model, which would greatly increase the size of our data. We will revisit using the RF matrix as well as kernel methods.

The other parameter to the clustering problem is how many clusters to use. Most clustering algorithms require the number of clusters as input. We used several means of determining the probable number of clusters for our data: within group sum of squared error, gap statistic, and the Adjusted Rand Index, or ARI. The latter is a measure of cluster stability. We ran these three methods on chapters 7 through 9 in the CS121 Fall 2013 course. We chose these chapters because we believe that the latter half of the course is more challenging and most of the students are going to finish the course. These chapters also span the interval between the midterm exam and the final, and so provide data closer to the final than material before the midterm. We plan on expanding this study to the other courses as well. We next present an example of how we determined the number of clusters for one chapter. These results were quite similar for the other six chapters.

We first ran a k-means clustering algorithm on 1 to 10 clusters, calculating the within groups sum of squares, or wss, value for each cluster. The k means algorithm uses the Euclidean distance metric. The wss is a measure the variability within clusters. As more clusters are used, the wss declines, as each cluster should be smaller, and able to “explain” more of the variability. Of course, overfitting can occur if too many clusters are used. The worst case is that one cluster is used for each data point. In the figure below, the left hand plot was made from synthetic data that was engineered to have three distinct clusters. The inflection point is obvious at 3. This is where the wss is lowest as the data is exactly explained by three clusters.

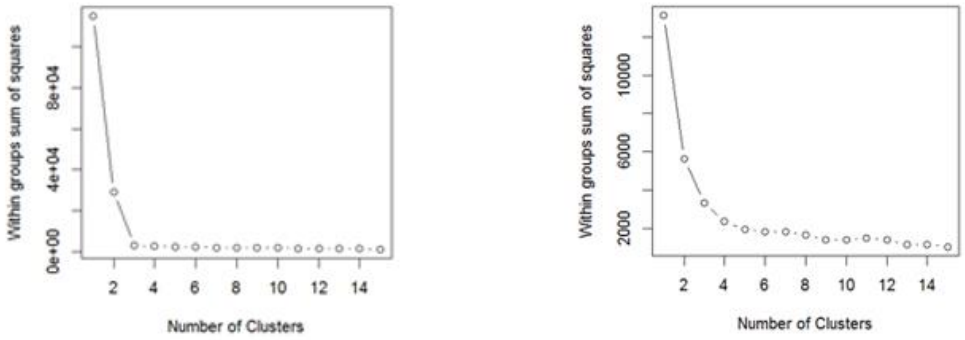


Figure 63: Plots of wss vs number of clusters. The plot on the left demonstrates synthetic data with exactly three clusters. The graph on the right is typical for one of our chapters.

The right hand plot shows a typical result for our data. There is an inflexion point at 4 or 5, after which the error cannot be reduced by adding more clusters. In this case we chose 5 clusters.

Another method of determining the number of clusters is the gap statistic. This algorithm, developed by Tibshirani et al. [2001] uses the wss measure, but calculates a “gap” statistic which is the difference between the wss for a given cluster number and the wss calculated on a randomly sampled distribution from the input data. The largest difference in the range of clusters is the best. The following figure shows a representative plot produced by running a gap statistic calculation on our chapter data. It more or less agrees with the wss plot on the left.

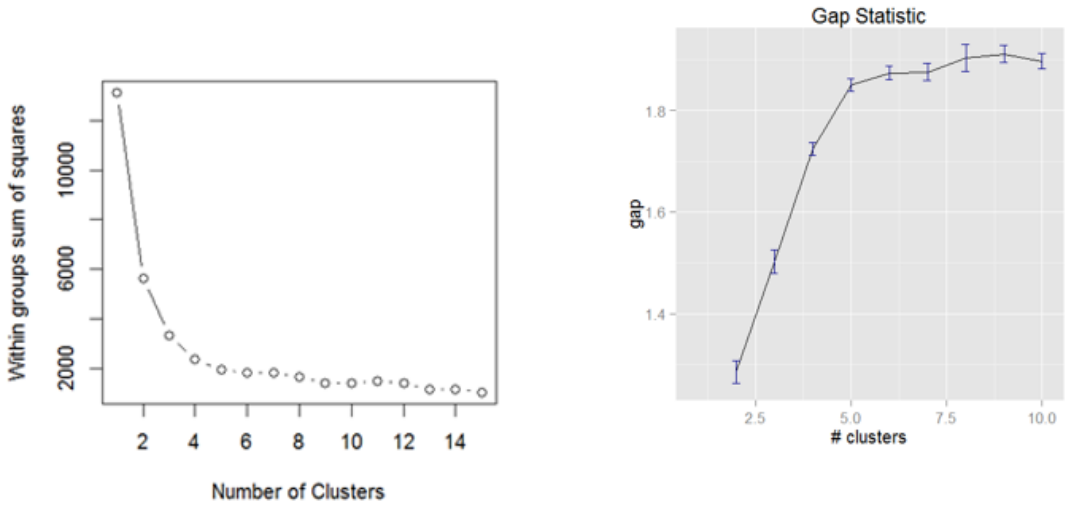


Figure 64: Comparison of wss plot next to the gap statistic plot for our chapter data.

In both graphs we see that the number of clusters to use is 5. In the gap statistic plot, we see that 9 clusters would be more optimal than 5, but the gain is not worth the trouble since we want a reasonable number of clusters to interpret in terms of student behavior patterns. If there really were 9 distinct patterns, we would probably want to consolidate some of them anyway.

Finally, we implemented (in R) an algorithm [Azarnoush, 2013] to calculate the adjusted rand index score (ARI) for each of the proposed cluster numbers. ARI calculates the similarity between two clusterings, with 0 meaning no similarity, and 1 meaning perfect similarity [Rand, 1971]. For each cluster number, we split the data evenly into a test and training set. We clustered each set individually using the PAM (partitioning around medoids) algorithm (described below). We then used a k nearest neighbor algorithm (KNN) with n=10 to generate cluster labels for the test clustering based on the training clustering. The idea is for each cluster label in the training set, find the 10 nearest neighbors (using the distance matrix) in the test set. The mode of the 10 nearest neighbors from the test set is used to form a new set of labels. This

new set is compared to the test set by ARI. This process is repeated 10 times and the distribution of the ARI scores is plotted for each number of clusters.

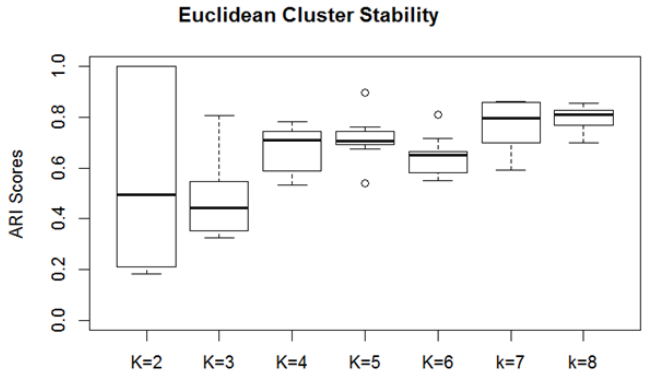


Figure 65: Plot of ARI scores for clusters 2 to 8.

In the figure above, the ARI scores are displayed for 1-8 clusters. This plot shows a measure of cluster stability. The most stable cluster would have the highest average ARI score with the least dispersion. Although cluster 5 does not have the highest mean, we again find it a suitable number for our purposes. In general, we chose to use 5 clusters for each chapter (7-13 in the CS121 fall 2013 course).

Once we decided on the number of clusters, we passed our distance metric to a clustering algorithm. In this case, we used a k-medoid algorithm PAM, for partitioning around the medoid. This is a similar algorithm to k means, except that PAM used medoids, or data points as the centers of clusters. We used the PAM implementation in the R language as it accepts any dissimilarity (or distance) matrix.

3.3.3 Clustering Results

Once we have clustered our data, we inspected the composition of sessions and intervals in each cluster. The following figure shows bar graphs of the compositions of the session and interval durations, encoded as S, M, L and s, m, l for chapter 7.

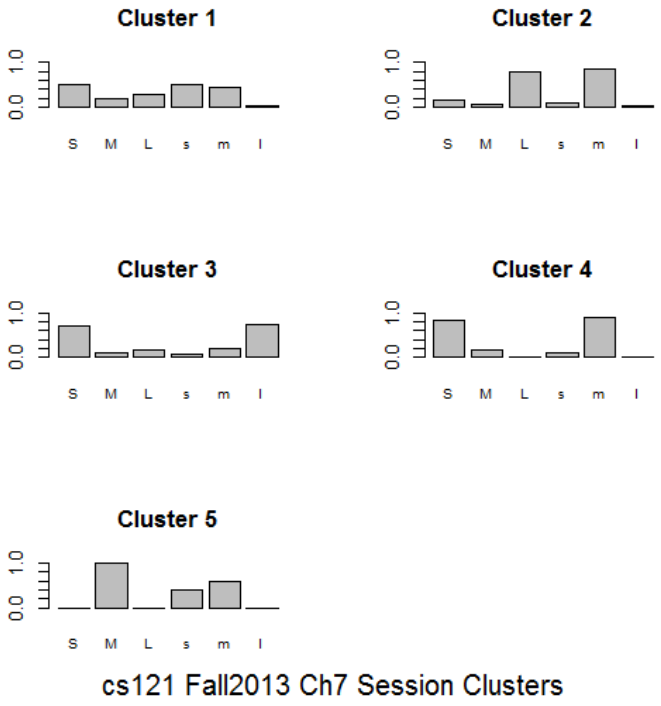


Figure 66: Proportions of session and interval durations for each cluster for chapter 7 data.

We were looking for distinct patterns of session and interval lengths. For example, in the Ch7 graph, cluster 3 has a large proportion of short sessions and long intervals, cluster 2 has a high proportion of long sessions and medium length intervals, while cluster 4 has short sessions with medium length intervals. We refer to these patterns as SL, for short/long, LM, for long medium, and SM, for short/medium. We can see these three patterns in chapters 10, 11, 12 and 13. We do not clearly see these patterns in chapter 8, and we see only the SM pattern in cluster 3 in chapter 9. (The remainder of the pattern graphs are located in appendix A). We speculate

that perhaps a difference in the course content, which is sparser at this point in the course, or some external events may have affected working behavior (The due dates for these chapter assignments are late October/early November). Expanding our analysis to other courses will provide more information. The following table summarizes the patterns we identified in the graphs for chapters 7 to 13.

Table 36: Summary of three identifiable patterns and their cluster labels in the chapter clustering. Note, these patterns do not occur in chapters 8 and 9.

Ch7			Ch10		
pattern	cluster	students	pattern	cluster	students
LM	C2	70	LM	C3	96
SM	C4	53	SM	C2	159
SL	C3	100	SL	C5	58
Ch11			Ch12		
pattern	cluster	students	pattern	cluster	students
LM	C4	105	LM	C5	90
SM	C3	179	SM	C2	160
SL	C2	68	SL	C3	60
Ch13			Note, the SM pattern occurs in chapter 9 in cluster 3. Pattern graphs for chapters 8 to 13 are in appendix A.		
pattern	cluster	students			
LM	C4	118			
SM	C3	55			
SL	C2	45			

The pattern we are interested in is the SL pattern. This pattern represents students who mostly work in short sessions (1 to 5 minutes in length on average), and long intervals, which have a minimum of 3.3 days in length.

3.3.4 Results: Relationship of Patterns to Exam Scores

In order to evaluate the relationship between a working pattern and exam scores, we have to identify a group of students who followed one pattern consistently throughout the chapters we are considering. In our encoding scheme we used the symbols M and m to represent the interquartile range of session times and interval time for each chapter. Therefore, a pattern such as SL will be less frequent than patterns with M or m. This means that there will be fewer students who follow the SL pattern in every chapter compared with other patterns.

We found a group of 10 students who followed the SL pattern in each of chapters 7, 10, 11, 12, and 13. We also found a group of 25 students who followed the LM pattern in each of those chapters. Both groups had similar amounts of homework done, and each had 2 computer science majors. There were no females in the groups. We conducted 100 trials of a permutation test (null hypothesis: means are equivalent) on the exam score distributions of these two groups and found a resulting average p value of .051. The average difference in means was 13.6.

From this result, we would guardedly claim that working exclusively in the SL pattern is associated with lower exam scores, if all other conditions are equal. However, the numbers of students who follow this and other "extreme" patterns is low. We plan to design a study that pools students from other courses so we can work with higher numbers.

3.3.5 Conclusion and Discussion

In conclusion, we have seen that the student session data does show one or more recurring patterns that can be interpreted by instructors, at least for the course we analyzed. The clusters we generated do seem to have some relationship to outcomes, albeit a small one. The SL pattern does appear in several of the chapter data we analyzed, and it seems to be related to lower outcomes. There are many steps in the process of creating the encodings and subsequent clustering. The encoding are based on an arbitrary set of cut points, and the calculation of session and interval durations relies to some degree on an arbitrary time for the estimation of the end of a session. Calculating time is very difficult to do in a web-based system. We also assume that long intervals mean the student is not in touch with the course material. This may not be the case, as students may be working in other ways, or checking in for on campus help.

We realize that more than one pattern could lead to a successful outcome, and that a SL worker may be learning perfectly well. We are not sure if those students stay in the SL pattern throughout the semester or change frequently. One key point is that we have defined an SL pattern as a cluster in which the proportion of short sessions and long intervals are dominant. We do not quantitatively define what these levels are. Furthermore, a cluster in the SL pattern also may have other session and interval types. We are claiming that a student in an SL pattern has more instances of short sessions and long intervals than other students, and when we isolated a group who fit this pattern in five chapters we saw a difference in their outcome scores. There is the possibility that we are observing some other effects that are tied to this pattern. It's also possible that because our bag of words does not capture the sequential nature of the encoded patters a student in the SL pattern could do most of their work in a few medium sessions with medium or short intervals. The fact remains that they are logging in mostly for short sessions and they do have mostly long periods between most of their sessions. We are planning to refine our method to deal with these issues. Please refer to the future work section 5.2 in chapter 5 for a further discussion of our plans.

CHAPTER 4

A STUDY OF THE "WORKING LATE" STRATEGY

4.1 Introduction

In this study, we analyze the effect of working on assigned material close to a due date on outcomes. Students who consistently chose to do the majority of their work close to a due date are following a pattern we refer to as "working late". There are two factors involved in the definition of working late: the "closeness" to the due date the work is being done, and the amount of the assigned work that is being done in this "close" time interval. The following charts illustrate how close to due dates students work in the courses we examined in our study.

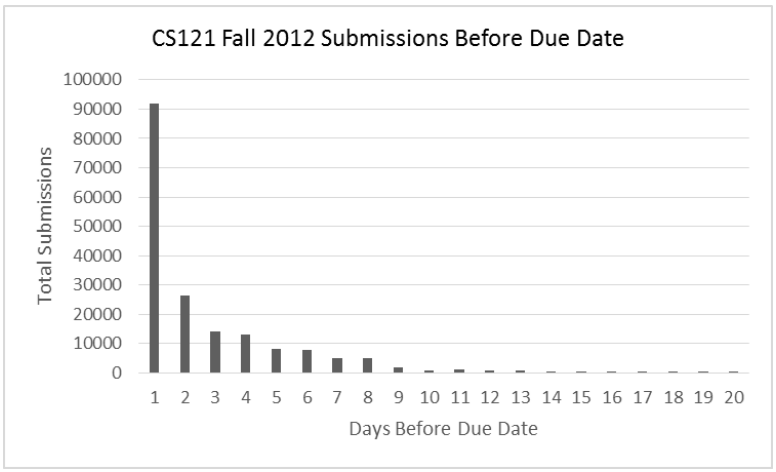


Figure 67: Homework submissions for CS121 Fall 2012 by the day. Day 1 is the day before the due date.

The amount of work done was measured by counting the number of submissions of answers to homework problems within certain time intervals. It's clear that most of the work done by students in these courses is done in the last 24 hours for CS121, and within the last 48 hours for CHEM111. Anecdotally, many instructors will agree that students who work late are not allowing themselves time to get assistance if they need it, and are not benefitting from the

ability to contemplate the topics and skills they are being asked to exercise. The frequency of submissions close to due dates may be “thrashing”: attempts made in desperation without a clear rationale about how to solve the problem. It may also be the case that students who are having a tougher time with the course material are putting themselves at a disadvantage by working late as they are more likely to require some help, but do not have the time to get it. Late workers may also misjudge the time it takes to complete the work. Students tend to work through homework questions sequentially. Usually, the easier questions come first. If a student is working late, she may not have the time left to work on the tougher questions as the end of the assignment, and consequently miss out on an opportunity to apply any acquired skills.

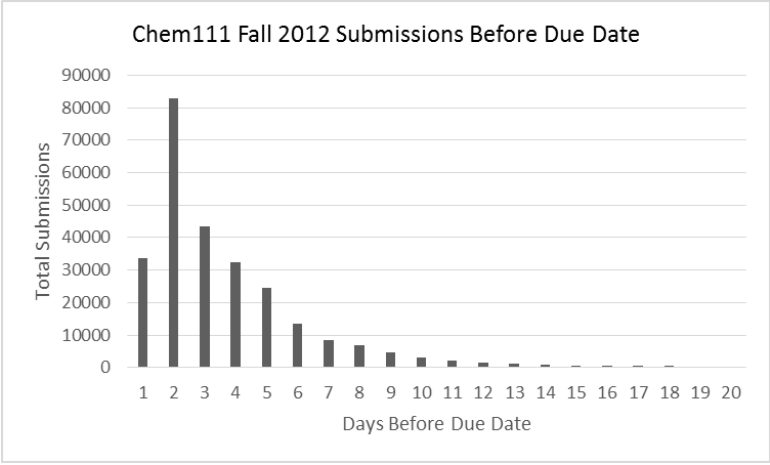


Figure 68: Homework submissions for CHEM111 Fall 2012 by the day. Day 1 is the day before the due date.

The idea that late working is deleterious to learning and subsequent outcomes has a growing research base. In a study of 465 undergraduates, Klassen et al. [2008], found that students who were classified as procrastinators by survey had significantly lower GPAs ($r=-.29$, $p<.01$), higher levels of daily and task procrastination, lower predicted and actual class grades, and lower self-efficacy for self-regulation.

Of course we know that one does not do one's best work when being rushed. In the case of college students in large courses where they do homework problems in many different settings and at any hour of the day, we want to know if they work late and if so how late do they work? In the first part of this study, we investigate the question of the effect on outcomes of working late. In the second part of this study, we investigate the effect of working late on novice students, namely first year students and CS121 students with no programming experience. Novice students may be more affected by working late, or may tend to work late as a result of a lack of time management skills.

4.1.1 Main hypotheses

We hypothesize that students who repeatedly do most of their work close to due dates, i.e. are late workers, will have lower exam scores on average than student who do not follow the working late pattern. We further hypothesize that novice students, first year and, in CS121 courses, students with no previous programming experience, will be more affected by following a working late pattern.

The rest of this study is organized as follows: we begin with an overview of how and when students do their work in our data sets. This is to provide the reader with a context to understand how we measure the working late pattern. Next, we introduce two studies of the effects of working late on outcomes measures of working late using a fixed proportion of work done, and the second calculates the proportion of work done for a specified time before the due date. We then conclude and discuss the results and our ideas for future work.

4.2 Introduction and overview of working late.

First, we present the results of some exploration of the data to provide the reader with an idea of when students are making submissions to their assigned homework. We use homework submissions as a measure of work because it is generally attempted by most students and constitutes the majority of the assigned work in the courses. We utilize time-stamped recordings of student homework question submissions from the OWL system. Each homework assignment in the courses we analyze have due dates associated with them. Homework is assigned on a weekly basis. Each assignment is associated with a unit of instruction, such as a section or chapter (see section 1.2.1 about course structure).

The data sets we used for this study are the same as used for chapters 2 and 3 (see table 14 in section 2.2.1). These data sets are for one semester of CHEM111, and two fall semesters and spring semesters of CS121. We pooled the two fall and spring semesters into two data sets: CS121Fall and CS121Spring.

Using the assignment due date as a reference, we looked at the pattern of homework activity twenty days out from the due date for all homework activity. We plotted the frequencies of submissions for each day until the due date, as the following figure shows. The huge difference between the activity on the last day before the due date, Day 1, and any other day is support for the anecdotal evidence about the number of students who work late. The plots for the other CS121 courses are virtually identical.

We show the same plot for the CHEM111 course as well. Notice that the big peak in the chemistry plot occurs before the last day. Apparently, many of those students are choosing to work earlier than the CS students.

In order to discover how late students are working, we need to get a closer view of what is happening on the day before the due date. The next figures show activity by the hour for the 24 hours before the due date. We see that there is a similar curve in the hour plots from what we see in the daily plots. It does indeed seem that many students are working close to the due date. In fact, it seems that most of the activity happens in the last three hours.

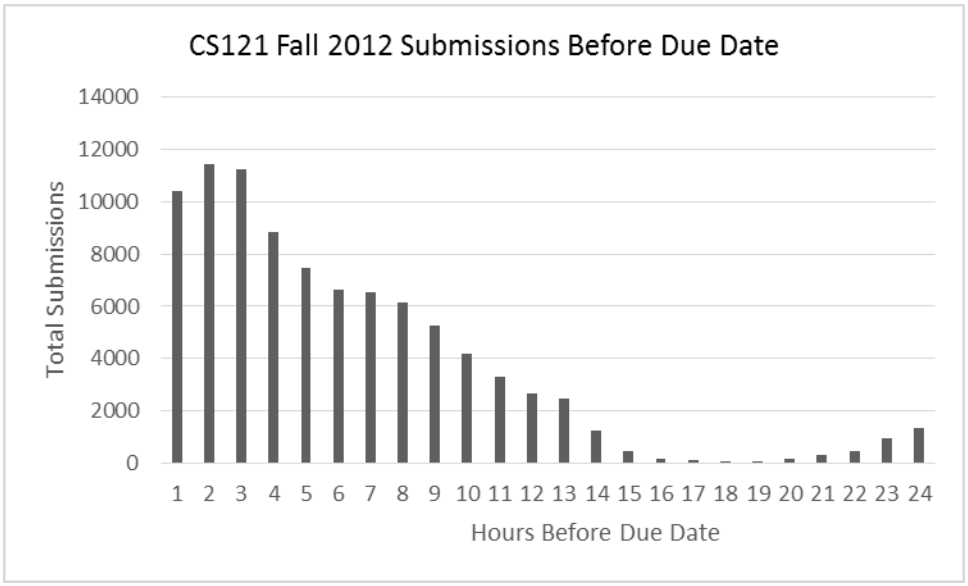


Figure 69: Homework submissions for CS121 Fall 2012 by the hour before the due date.

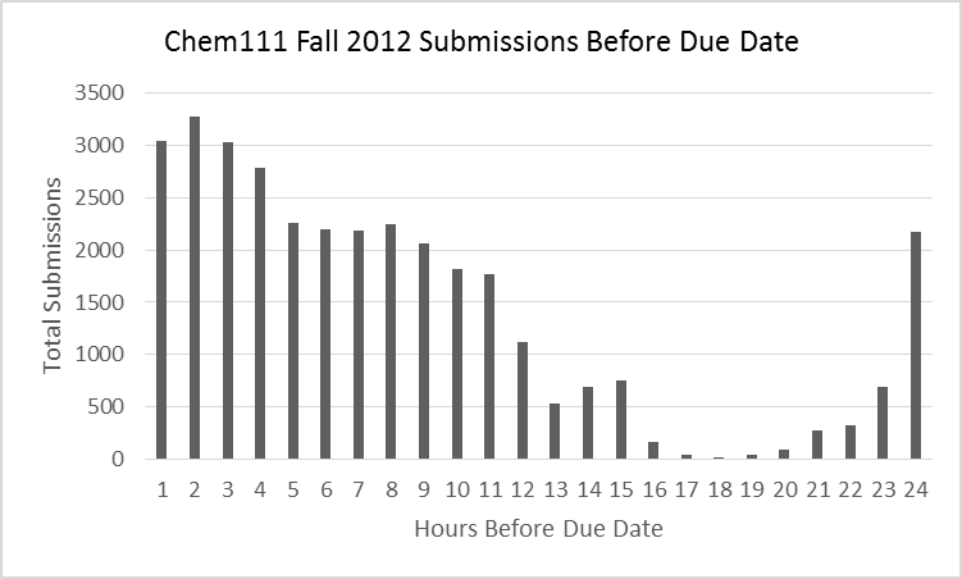


Figure 70: Homework submissions for CHEM111 Fall 2012 by the hour before the due date.

Although the above graphs are very telling about the time of student activity, they don't say anything about how much work is being done by individuals. For example, a student may have a lot of submissions within an hour of the due date, but that may be only a fraction of his total submissions. In addition to defining when we consider work to be late working, we need to know how much of an individual's total output is occurring close to the due date.

4.3 Study 1: Representing the working late pattern with WORK50.

We represented working late by calculating how many hours from an assignment due date a student submitted fifty percent of her homework. We refer to this measure as WORK50. We next describe our method for calculating this variable.

4.3.1 The WORK50 measure.

WORK50 was calculated as follows: we record a time stamp for a student homework submission that is half of all submissions that the student will make before the assignment due

date. Call this time stamp submit50TS. WORK50 is the difference in hours between submit50TS and the assignment due date. These are the steps taken to calculate WORK50 for a student for one homework assignment:

$$index = \ ceiling\left(\frac{Number\ of\ submissions\ before\ due\ date}{2}\right)$$

$$submit50TS = timeStamp_{index}$$

$$WORK50 = diff(dueDate - submit50TS)$$

A WORK50 value is calculated for each assignment in a course. In this study we take the average of WORK50 over all course assignments. We refer to this aggregate value as WORK50Avg. The following figure plots the distribution of WORK50Avg for the three datasets: CHEM111, CS121Fall, and CS121Spring.

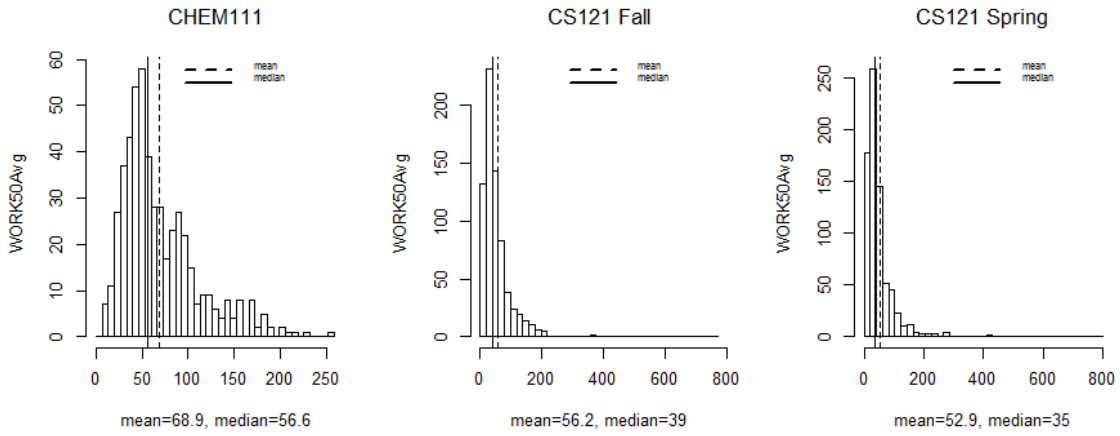


Figure 71: Distribution of when students did 50% of their submissions relative to the assignment dues date including mean and median. The values shown are averages over all course assignments.

It is clear from the plots above that most student do 50% of their work within a few days of the assignment due dates.

Next, we looked at the relationship between percent of assigned homework attempted and WORK50. We hypothesized that students who had a low WORK50, i.e. they submitted 50% of their work close to the due date, would also attempt a lower total amount of homework.

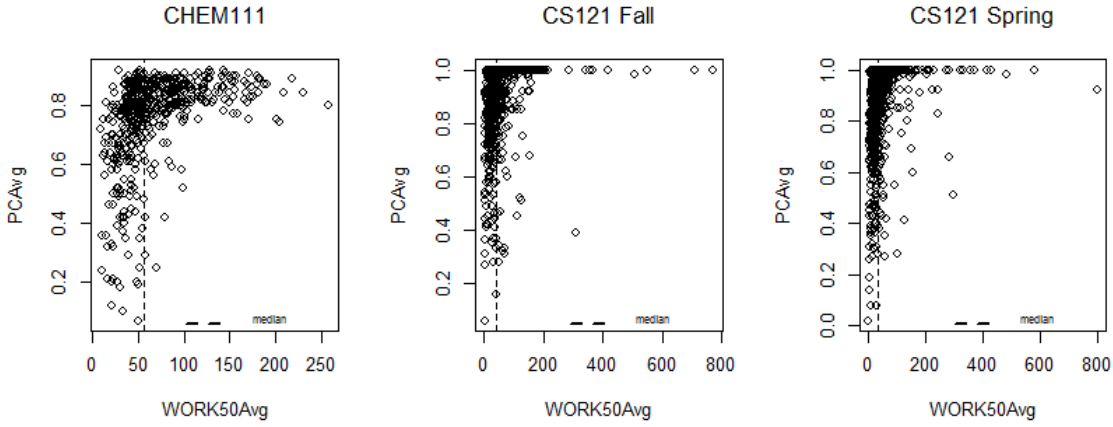


Figure 72: Plots of the percent of assigned homework attempted vs. hours before the due date when 50% of submissions had been made. Medians for WORK50 included.

The plots above do not support our hypothesis. The data points to the left of the median lines do include the majority of lower participants, yet this area is not exclusively populated by lower participants. Another hypothesis was that low WORK50Avg values would be associated with lower exam scores (negatively correlated). We plotted the final exam scores against WORK50Avg to visualize this relationship.

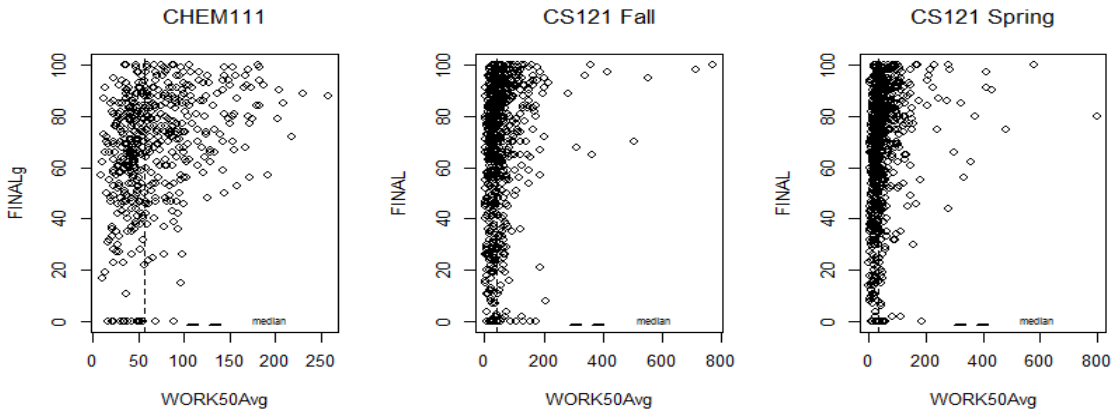


Figure 73: Plots of the percent of final exam scores vs. hours before the due date when 50% of submissions had been made. Medians for WORK50 included.

The plots above do not show some positive correlation (for social and behavioral science small is an r value of .10, medium .30 [Cohen, 1992]). The CHEM111 data does show more of a mass of data points in the region of higher exam scores with higher WORK50Avg values, i.e. students who did half of their work relatively early. A correlation on these data showed some relationship, as shown in the following table.

Table 37: Correlations for final exam scores and WORK50Avg.

	Pearson's r, 95% conf.
CHEM111	.36, [0.24, 0.39]
CS121Fall	.18, [0.11 0.25]
CS121Spring	.23, [0.16 0.23]

4.3.2 Hypothesis

We hypothesize that lower WORK50 values will have a negative effect on final exam scores, adjusting for the total percentage of homework done.

4.3.3 Method

In order to test this hypothesis beyond the results from the correlations calculated in the previous section, we use a matching approach using propensity scores, as we did in Chapter 2. We take this approach because, as with the book-first pattern, the working late pattern is a self-selected behavior, and we want to adjust for the selection bias that is inherent in self-selection.

As stated in Chapter 2: propensity score analysis requires several steps: 1) the definition of three variables or sets of variables: an outcome measure, a treatment condition, and a set of observable variables, say X, that are likely to affect a subject's selection of treatment condition and outcome state, 2) calculate propensity scores based on a model of treatment condition predicted by the covariates in X, 3) apply a matching algorithm to achieve treated and control groups that are balanced on the distributions of X, 4) calculate the average effect of treatment on outcomes. The matching in step 3 is crucial to the success of this technique since selection bias is reduced when the treatment and control groups are evenly matched on the covariate (variables in the set X) distributions, thus simulating a random assignment to treatment, at least on the variables in X.

4.3.3.1 Variable definitions

A summary of the variables used in this analysis is listed in the table below. As in chapter 2, the observed covariates that for the set X mentioned above are: PREV, GENDER, MAJ_GROUP, CLASS_LVL, and PCAvg (These variables are described in detail in Chapter 1).

Table 38: A summary of the variables used in this study.

Variable	Levels	Description
PREV	none, some, java	Previous experience. Not used in CHEM111.
GENDER	female, male	Gender
MAJ_GROUP	1,2,3 (CHEM) or 1,2 (CS)	Groups formed from the survey variable MAJOR.
CLASS_LVL	1, 2, 3, 4	1^{st}, 2^{nd}, 3^{rd}, and 4^{th} year students.
WORK50Avg	[0, +) in units of hours	The average of WORK50 for all course assignments. Used to define the treatment condition.
PCAvg	[0,1]	The average of PC_HWK scores for all course assignments.
FINAL	[0,100]	The outcome measure.

The outcome measure we use is the final exam score. The treatment condition must be a binary valued variable, so we categorize WORK50Avg according to the following value ranges. We selected the lowest 25% (first quartile) of the WORK50Avg distribution as we are interested in the most extreme examples of working late.

Table 39: Categorization of WORK50Avg into treatment and control conditions.

	T=1, Treatment	T=0, Control
CHEM111	WORK50Avg < 40, N=126	WORK50Avg >= 40, N=388
CS121 Fall	WORK50Avg < 23, N=175	WORK50Avg >= 23, N=542
CS121 Spring	WORK50Avg < 22, N=194	WORK50Avg >= 22, N=559

4.3.3.2 Calculation of propensity scores.

We calculated propensity scores based on a logistic regression on the variables in the set of observed covariates: PREV, GENDER, MAJ_GROUP, CLASS_LVL, and PCAvg. The models we used for propensity score calculation for CHEM111 and CS121 respectively was:

$T = GENDER + CLASS\_LVL + MAJ\_GROUP + PCAvg + PCAvg^2$

$T = GENDER + CLASS\_LVL + MAJ\_GROUP + PREV + (MAJ\_GROUP * PREV) + PCAvg + PCAvg^2$

Since the propensity scores are used in the matching step, we checked their distributions in the treated and control groups to assess the amount of overlap, or common

support. Common support ensures that there are representatives from both treatment and control groups over the distribution of all propensity scores. The figure below shows these distributions for the CHEM111 data.

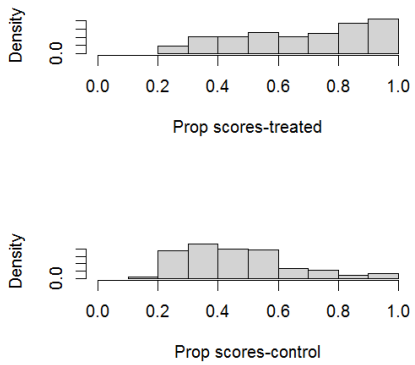


Figure 74: Propensity score distributions for CHEM111.

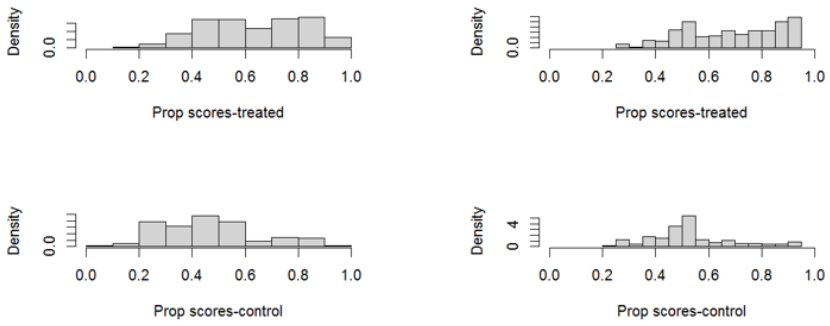


Figure 75: Propensity score distributions for CS121 Fall and CS121 Spring respectively.

The propensity score distributions show fairly good common support except for a general weakness of support for the control groups at the higher end. We can exclude some of this range in the matching phase.

4.3.3.3 Matching.

The results of the matching evaluation produced by the two matching algorithms is partially reported in the tables below. In essence we are comparing the similarity of two

distributions of the covariates used in the propensity score model for the treatment and control groups. For brevity, we chose to report only the standardized difference in means between the control and treatment groups and the variance ratio. The best matching would result in a difference of means of zero and a variance ratio of 1.0. Note that not all of the model terms are included in the tables.

Table 40: Matching statistics for CHEM111.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	-26.98	13.19	5.42
CLASS_LVL	28.26	1.72	4.12
MAJ_GROUP	-17.67	4.73	2.07
PCAvg	-71.48	-2.90	-1.22
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	1.06	0.99	0.99
CLASS_LVL	1.77	1.11	1.17
MAJ_GROUP	0.81	1.03	0.97
PCAvg	4.61	1.16	1.11

Table 41: Selected matching statistics for CS121 Fall.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	19.51	5.67	-1.67
CLASS_LVL	-11.79	1.94	-5.94
MAJ_GROUP	20.35	14.93	0
PCAvg	-51.30	1.06	-1.01
PREV	-34.30	-1.59	-9.87
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	0.78	0.93	1.02
CLASS_LVL	0.79	0.89	0.80
MAJ_GROUP	0.19	0.20	1
PCAvg	1.73	1.05	1.01
PREV	0.82	1.05	0.95

Table 42: Matching statistics for CS121 Spring.

Std mean diff	Before Matching	After Matching	After Matching GenMatch
GENDER	5.64	4.46	-1.97
CLASS_LVL	12.43	1.65	2.54
MAJ_GROUP	24.02	7.21	2.68
PCAvg	-57.70	0.38	-0.78
PREV	-26.88	-6.61	0.73
Var ratio (Tr/Co)	Before Matching	After Matching	After Matching GenMatch
GENDER	0.92	0.94	1.03
CLASS_LVL	1.08	0.99	1.01
MAJ_GROUP	0.94	0.98	.99
PCAvg	2.30	1.01	1.02
PREV	0.73	0.93	1.00

The results above show that the genetic matching algorithm produces better matches on average for most of the covariates. Some matching resulted in slightly worse balance for some covariates. The genetic algorithm does not rely solely on the propensity scores for matching. Perhaps there was a lack of common support or that the covariates are not distributed such that a useful match could be made. In any case, both algorithms improved the covariate balances of the treatment and control groups. The effect sizes were calculated for both types of matching algorithms in the next section.

4.3.3.4 Effect estimation.

The following graph shows the effect estimation for the three data sets including a 95% confidence interval. Both average effect for the treated, ATT, and whole population, ATE, were calculated.

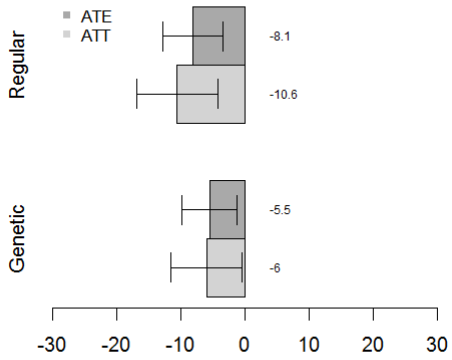


Figure 76: Effect estimations for WORK50Avg for CHEM111.

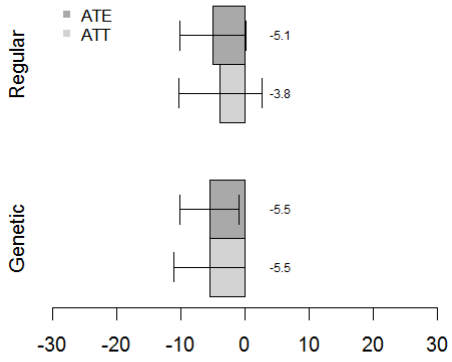


Figure 77: Effect estimations for WORK50Avg for CS121 Fall.

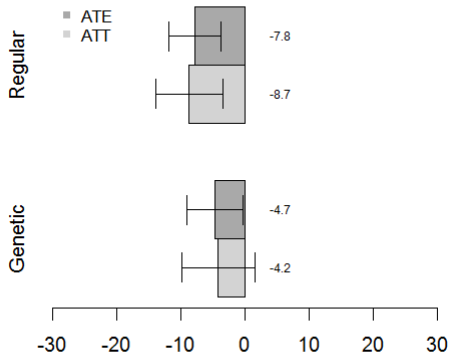


Figure 78: Effect estimations for WORK50Avg for CS121 Spring.

4.3.4 Conclusion and Discussion

The effect estimands, ATE and ATT, are all uniformly negative in direction. The error bars are, for the most part, within the negative range of effect size. The effects for the regular matching are generally larger than the genetic matching. The genetic matches tend to result in closer matches than the "regular" matching algorithm. Perhaps there is less bias in the groups produced by the genetic matching, and that is closer to the true effect size.

It seems that being in the group that does 50% of their work less than 40 hours before due dates in CHEM11 and 24 hours in CS121 will result in around 5 points lower on the final exam. We are estimating an effect on average behavior over a fifteen week semester with one outcome measure at the end. The fact that we see a consistent, negative effect of doing work closer to due dates suggests that this pattern has a detrimental effect, at least for some students on average. We plan to investigate this pattern further by looking at ways to assess the impact of working late on smaller sets of assignments. We are also curious if there are specific subpopulations of students who are more susceptible to negative effects of working late.

4.4 Study 2: The effect of working late on novice students.

Our second study of working late investigates the hypothesis that novice students with low WORK50Avg values, that is, they worked close to the due date often, will have a more negative effect on final exam scores than more experienced students. We define experience in two ways: first, we consider 1st year students as less experienced than 2nd – 4th year students for CHEM111 and CS121, and second, we consider only CS121 students and those with no previous programming experience versus those with some or Java programming experience.

4.4.1 Novices as first year students.

In the first part, we test the hypothesis that first year students will be more affected by working late than second through fourth year students. We used the data from the single CHEM111 course, and, we pooled all four semesters of the CS121 data (fall 2012, fall 2013, spring 2013, spring 2014). We then created two subsets of these data sets, one of first year students only and one of only second through fourth year students. The following table summarizes these groups.

Table 43: Data sets for first year and second through fourth year groups.

Group	CHEM111	CS121
1st year	355	794
2nd-4th year	159	676
totals	514	1470

We then created treatment and control groups based on the categorization of WORK50Avg as described in the previous section. The sizes of these groups are summarized in the following table.

Table 44: Treatment and control group sizes.

		CHEM111	CS121
1st year	Treatment	93	219
	Control	262	575
2nd-4th year	Treatment	60	202
	Control	95	474

To test the hypothesis, we estimated the treatment effect on final exam scores. Support for the hypothesis would be a more negative effect of the treatment on exam scores for the novice group than for the more experienced group.

The working late pattern is a self-selected behavior, and so we have the situation where selection bias could be a problem for the validity of our results. As above, we adopt a propensity score matching method to achieve a balance of the covariates described in the previous section. Of particular interest is the PCAvg variable, the percent homework attempted. This variable is strongly correlated with the final exam score, and the working late pattern, since the more homework you do the more likely it is that you will work late.

We created propensity score models as above, and used the same matching algorithms. The results of the matching on PCAvg is shown below.

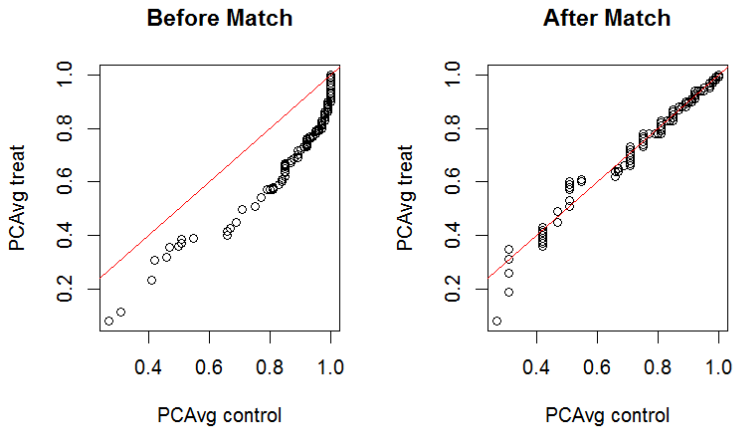


Figure 79: Balance of PCAvg in 1st year (novices) before and after matching in CHEM111.

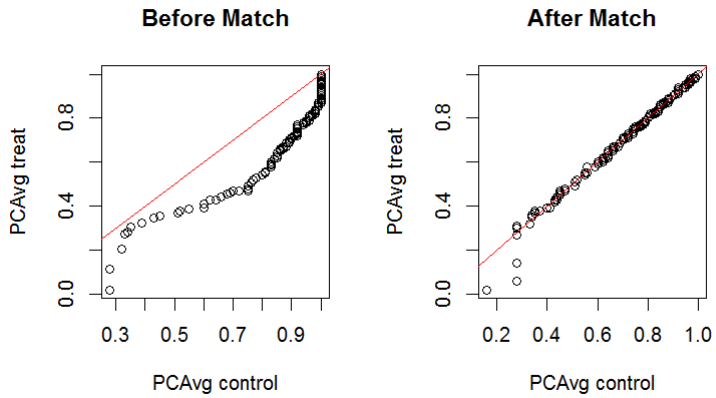


Figure 80: Balance of PCAvg in 2nd-4th year data before and after matching in CS121.

The propensity score distributions common support, and the balance for the other covariates was also checked as in the previous section.

4.4.1.1 Results.

The results for effect estimation on the groups are presented in the following graphs.

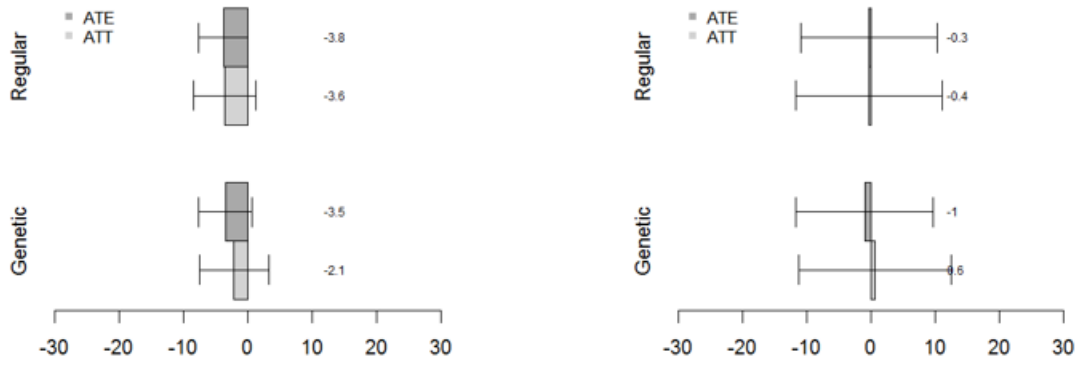


Figure 81: Effect estimation of WORK50 for novices (1st Year) and 2nd-4th Year students respectively for CHEM111.

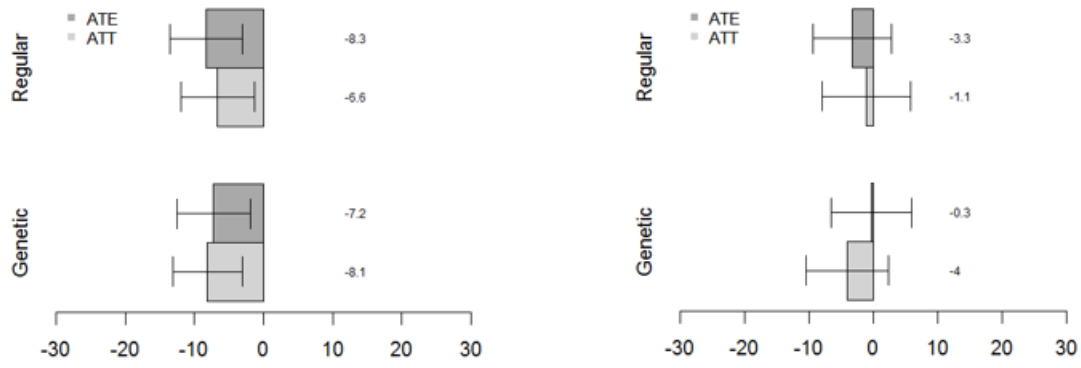


Figure 82: Effect estimation of WORK50 for novices (1st Year) and 2nd-4th Year students respectively for CS121.

4.4.1.2 Conclusion and Discussion

From the results in the graphs above, we see that the effect for the first year students was larger and more negative than that of the other students. The results for CS121 are more evidently negative in both groups, and more negative for the novices. The error bars do indicate that the effect was more variable for the CHEM111 estimates, with the ATT bound crossing into positive territory. The effect on the treated would be hard to interpret as the WORK50Avg is not a binary value. This categorization into treatment and control adds a threat to validity as we chose a somewhat arbitrary threshold- the bottom quartile. Moving this threshold may significantly affect these results to be more positively oriented as we would be including data with less late workers. Another issue is that the sizes of the data sets are unequal. It is also the case that there are more late workers in CS121 than in CHEM111, which may have lent the CS121 results more of an effect. The fact that the results are more negative for the novice groups in both courses lends support for our hypothesis.

4.4.2 Previous programming experience in CS121.

Next, we applied the same method to the CS121 data and used previous programming experience to define novice students. We used only first year students, whom we had defined as novices above. We formed the treatment and control groups for this study using the same categorization of the WORK50Avg variable. The data sizes are summarized below.

Table 45: Data sets and their sizes for the study.

		CS121
No exp.	Treatment	138
	Control	347
Some, Java exp.	Treatment	89
	Control	220

We hypothesized that students with no programming experience would be more negatively affected by working late than those with experience.

4.4.2.1 Results

To test the hypothesis, we again used propensity score matching as above. After matching, we estimated the average effect sizes. The results are plotted below.

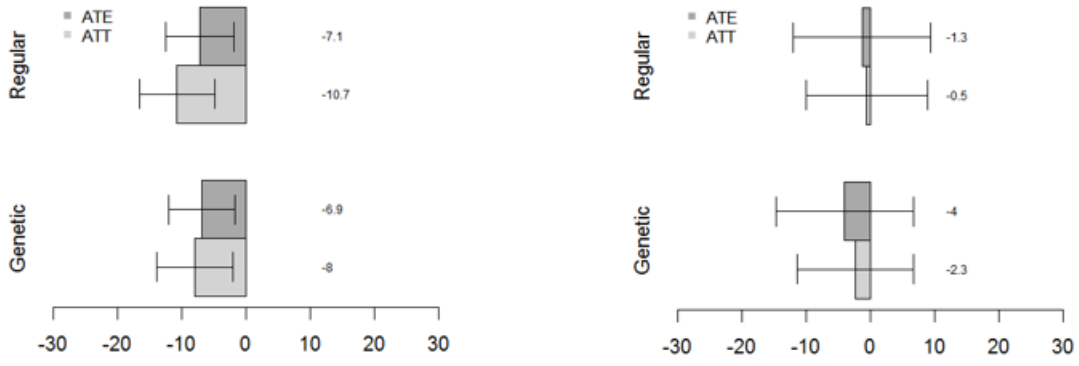


Figure 83: Effect estimation of WORK50 for 1st year novices (no programming experience) and students with some and Java experience respectively for CS121.

4.4.2.2 Conclusion

From the results above, it is clear that the novice group shows a greater negative effect than the experienced group. The high degree of error in the experienced group suggests a high degree of variability in the data. As noted above, there is a disparity in the sizes of the treatment and control groups which may add to the variability in the results. The fact that four semesters of data was used does lend support the view that these results support our hypothesis.

4.5 Chapter 3 Conclusion and Discussion

The results of the propensity score matching are somewhat mixed. In the general study in section 4.3, we do see an average negative effect of the WORK50Avg measure on final exam scores for all courses, though the 95% error bands do approach the zero effect line for CHEM111 and CS121 fall (figures 76 and 77). The error band does cross the zero effect line for the genetic matching result in figure 78. These results show mostly small negative effects but with some uncertainty as the error bands do come close and in one case cross the zero effect line. Based on these small effect estimates, we claim that the direction of an effect is negative, which supports our hypotheses that working late has a negative effect on outcomes, but further study is needed before we can claim our hypothesis to be valid. Previous work has also found a negative correlation between procrastination and grade point average [Lakshminarayan et al., 2012; Klassen et al., 2008; Moon & Illingworth, 2005].

In addition to the general study of WORK50, We also found small but more negative effects for novice students compared with experienced students in sections 4.4.1 and 4.4.2. The 95% error bands do, however, cross the zero effect line three out of four times for the CHEM111

results for novice students (figure 81). The results for CS121 (figures 82 and 83) courses showed a more consistently negative effect for novices in comparison to experienced students, though the latter results had wider error bands. As in the previous paragraph, we cannot conclude that our hypothesis is strongly supported by our results. We do, however, see a more negative direction for novice students that bears further investigation.

There are several threats to the validity of our study's methods. The WORK50 measure captures only submissions to the LMS, and not the exact time that students are working. For example, students may work on solving homework problems early but not submit their answers at that time; waiting until a time closer to the due date to submit. The decision to use 50% as a proportion of student work in the calculation of WORK50 was arbitrary. We could have used 75% or some other value. We also took the average of WORK50 over all assignments in the course, and used the final exam as an outcome measure.

There can be a lot of external factors that could influence student behavior over the semester that we did not adjust for. We are certainly missing an observed variable for ability or aptitude. Additionally, the methods we used in calculating our results involved many steps which may have added bias. We did calculate a larger error band for our results to compensate for such bias, and these error bans did cross the zero effect line several times.

In the future, we plan to study the working late measure on smaller subsets of assignments. We plan to implement short quizzes to use as outcomes for smaller grouping of content.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The purpose of this thesis was to investigate the effect of three patterns of student behavior on outcomes. Book-first, Infrequent-contact, and Working-late. We hypothesized that these patterns describe behavior that has an effect on how students learn in STEM courses at the college and university level where most of the work is done online.

We studied data from five courses in chemistry and computer science. The use of a computer-based learning environment, the OWL system, along with instrumented textbook questions allowed us to gather data necessary to investigate the above mentioned beliefs.

In our first study, we used OWL data to calculate the amount of book work done by students before they attempted their homework problems. We also calculated a measure of total book work regardless of timing relative to homework as a way to control for the total amount of book exposure for each student. We utilized the method of propensity score matching to adjust for self-selection bias by constructing groups of students with balanced distributions of covariates that affect outcomes and behavioral choices. These covariates were the amount of homework done, as well as gender, major, and class level. We also hypothesized that relative novice students would be more affected by these patterns of behavior than more experienced peers. We used propensity score matching to investigate the effect on the book-first and working-late patterns on select subpopulations of interest, such as first year students and students with no previous programming experience in the CS121 course.

Our results showed a significant positive average effect of following the book-first strategy on final exam scores. The effect sizes ranged from about 2 to 14 points. The chemistry course showed the highest effect of the book-first pattern as measured by our BF variable. We tested the difference in outcome distributions as well as effect sizes. We also investigated the difference in the effect of book-first on several subpopulations of students: first year students vs. second through fourth year students, females and males, novice programmers and experienced programmers in CS121, and computer science majors and non-majors. We found no significant difference in effect sizes for gender, but did find that students with programming experience were not affected by the book-first pattern. The results for the other groups were mixed.

In chapter three, we investigated the pattern of working in brief sessions with long intervals between sessions is associated with lower exam scores in the second study. We calculated student work session durations and the durations of the time intervals between their sessions. We then encoded the sessions and intervals according to the distribution of all students' session and interval times. Our encoding was then translated into vectors of encoding symbol frequencies, a bag of words approach. We clustered students on their frequency vectors into one of five clusters per chapter. The clusters showed three recurring patterns in 4 out of the 6 chapters we studied. We evaluated the effect of cluster membership in the pattern with short sessions and long intervals, the SL group, by comparing groups that followed that pattern with those that followed a pattern with long sessions and medium intervals, LM. We found a significant difference between the two groups, with the SL group having a lower exam distribution mean.

In the fourth chapter we investigated the effect of working close to the due date on exam scores. We also studied subpopulations of interest, namely first year students and students with previous programming experience in CS121. We hypothesized that we would find a negative effect of working late on students in general, and that the effect would be increasingly negative for novice students.

We calculated the time difference between when a student submitted half of her work and the assignment due date as a measure of working late. We used propensity score matching to adjust for bias. We did find a negative effect for both chemistry and computer science courses. We also found that first year and novice programmers in CS121 were more negatively affected by working late.

5.2 Future work

The book-first strategy has been adopted by increasingly more students in the CS121 course over the past five years. This is partly due to the use of assignable, embedded questions in the OWLBook. In order to study the effect of the book-first strategy, we had to isolate a relatively small subset of the student population. This means that we have a result with low statistical power. However, the fact that we find an effect consistently across several courses tends to reinforce our confidence that the effect is real. It does not, of course, provide us with the basis for making a causal claim. For that we will have to control for what are now latent factors, such as ability or aptitude. We plan to conduct future studies where we have access to test scores and other data a priori to our study that can help us measure these qualities. Another goal is to use a pretest for such purposes.

Another way for us to make a causal claim would be to manipulate the amount of book-first activity in an experimental setting. Since we are not likely to set up such an experiment on any scale, we rely on a quasi-experimental design where we manipulate the course content with the idea of influencing students to choose not to follow the book-first approach. In the case of the CS121 material, we would create another version of the CS121 course content where the text does not contain embedded questions, and the book chapters are not assigned. This would remove the incentives for a book-first strategy. We have seen that about 40% of the students in CS121 accessed the iJava OWLBook before embedded questions were assignable. Of course, we could just make the embedded questions in the current book optional, except that the book has now been integrated with the homework problems, so that now students access their homework in the book, whereas before the homework problems were accessed via a separate path in OWL. The best design to encourage a problem solving first approach would be to make the book accessible only from homework problems as learning resource links. That way, the homework problems would be the major focus. The links could be made to point to a specific location in the book that is most relevant to the problem at hand. This configuration would result in almost no book first activity. This configuration is actually used by several publishers in the belief that students don't learn best by reading their textbooks, and so they encourage the use of the text as a lookup reference.

Another aspect of student learning we plan to study efficiency. Time and energy are finite resources for students with a full course load. We would like to investigate how a book-first strategy affects time and attempts expended to solve homework problems. We imagine a configuration where one part of the course promotes the book-first strategy alternating with another part that does not. The class could be split into two groups, and the book-non book

sections would be different for each group. We would measure the efficiency of working homework by the time and attempts expended with and without having read the text before.

In chapter 3 we used a bag of words approach to encoding patterns of student sessions and intervals between sessions. We plan to expand this approach in two ways. First, implement an augmented bog of words to capture the sequential information in the strings of sessions and intervals. This can be done in several ways: by sampling sequences of two or more symbols to capture "local" structure, and by generating patterns for regular expressions to be matched on the strings. The other expansion would be to include more detailed events about what happens in a session and when. For example, what would the time interval between an initial page accesses until an embedded question was attempted tell us about a student's method of using the book? Would a pattern of short intervals and many question attempts imply the student is using the text only to answer the questions?

With regard to chapter 3 and our pattern detection methods, we recognize that our bag of words approach is not capturing sequential information in the encodings of session and interval durations. The SL pattern is defined, qualitatively at the moment, as a cluster with mostly short sessions and long intervals. It is possible that students in the SL pattern do all of their work in one or two medium sessions with a medium interval between them. Our current method is too course-grained to capture this. We are planning to refine our method in two ways. First, take into account the amount of submissions done in a session, so if a student does most of their work in a medium session then we would weight their membership in the SL pattern lower than a student who did most of their work in short sessions. Second, implement the augmented bag of words as described in section 3.2.2. This would allow us to capture sequential structure in patterns of sessions and intervals. We would also consider adding more fine-grained information, such as the time interval between a page view and a question

attempt. One issue with timing in a web-based system is the estimation of when some events end. For example, if a session starts and there is no further activity then we must define a somewhat arbitrary cutoff length. We are convinced, however, that this estimation suffices to represent the intention of a student's work.

We also plan to study the extent to which students either move from one pattern to another or stay in the same pattern consistently. We surmise that students who change patterns frequently may not do as well as those who do not. Possibly, moving to an extreme pattern, one with short sessions and long intervals for example, may be a bad indication. Our suspicion that this may be true is due to the results of a previous investigation into lecture attendance. We evaluated the relationship between student outcomes and their answer to a survey question about the extent to which they attended lectures: "always", "sometimes", "never". Controlling for our measures of engagement, such as percent homework done, we found that students who answered "sometimes" had a significantly lower average on the final exam when compared to the other groups. Our best explanation of this result is that students who stick to a plan have an awareness of the notion that one should follow a consistent approach to studying. This assumes that never attending lecture is a conscious decision, and the fact that these student did complete the final and did similar amounts of homework implies that they were making the decision consciously. We plan to look at how session patterns change over time and what consequences this may have for outcomes.

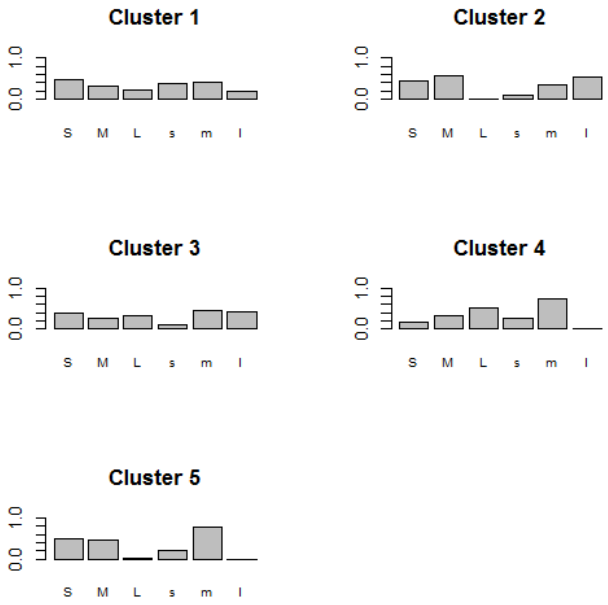
Finally, we plan to apply our working late analysis to future courses to determine if this is a consistent phenomenon, and if so devise methods of uncovering possible causes. In our study, females did not show an effect from late work, but they are a small population, about 29%, in the CS courses, and may not be a representative sample. We were surprised that computer science majors were affected by late work. We may have to look at that

subpopulation in more detail. Perhaps there is a subgroup in these majors that we are not observing.

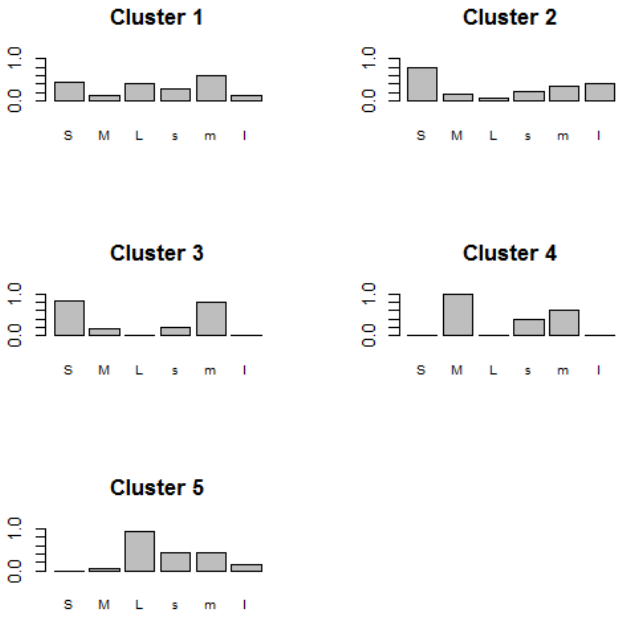
APPENDIX

CHAPTER 3 PATTERN GRAPHS

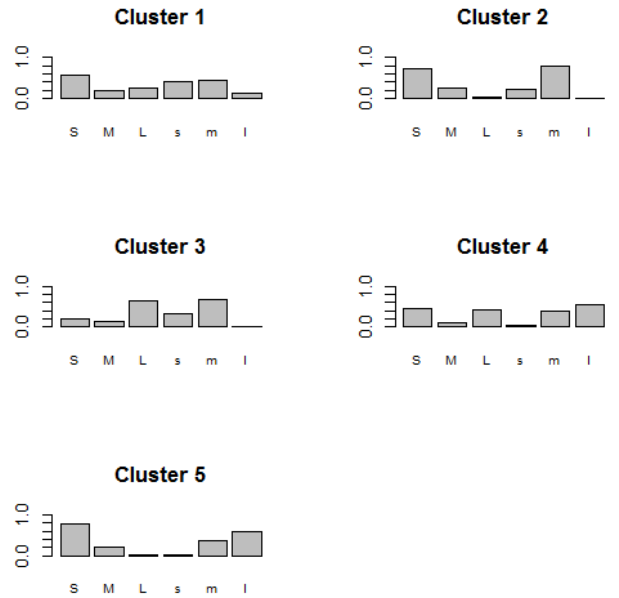
Pattern graphs from Chapter 3 for chapters 8 through 13. (Note that in these charts L=B and l=b).



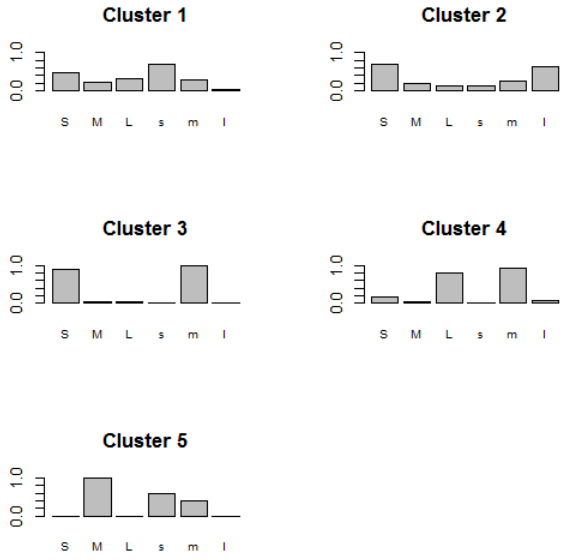
cs121 Fall2013 Ch8 Session Clusters



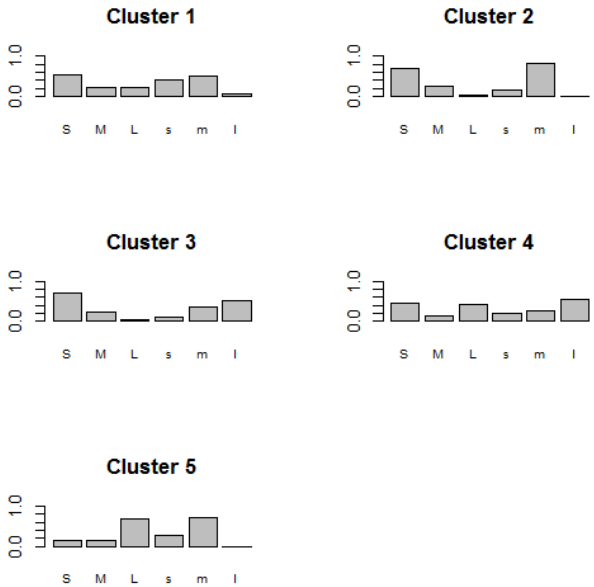
cs121 Fall2013 Ch9 Session Clusters



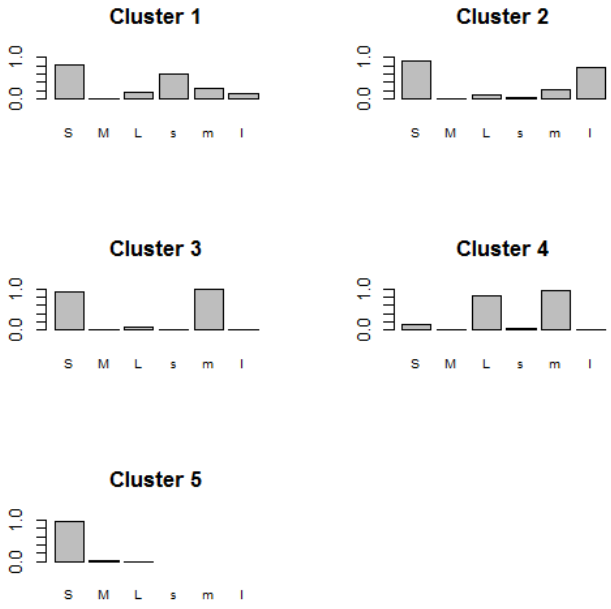
cs121 Fall2013 Ch10 Session Clusters



cs121 Fall2013 Ch11 Session Clusters



cs121 Fall2013 Ch12 Session Clusters



cs121 Fall2013 Ch13 Session Clusters

BIBLIOGRAPHY

Abadie A., Imbens G. (2006). Large Sample Properties of Matching Estimators for Average Treatment Efects. Econometrica, 74, 235-267.

Andergassen, M., Mödritscher, F., Neumann, G. (2014). Practice and Repetition during Exam Preparation in Blended Learning Courses: Correlations with Learning Results. Journal of Learning AnalyticsVol 1, No 1 (2014).

Arroyo, I., Hasmik M., Woolf, B.P. (2010). Effort-based Tutoring: An Empirical Approach to Intelligent Tutoring (2010) In proceeding of: Educational Data Mining. The 3rd International Conference on Educational Data Mining, Pittsburgh, PA, USA, June 11-13

Arroyo, I., Woolf, B.P., Beal, C.R. (2006). Addressing Cognitive Differences and Gender During Problem Solving. International Journal of Technology, Instruction, Cognition and Learning. Vol. 4, pp. 31-63.

Arroyo, I., Beck, J. E., Woolf, B. P., Beal, C. R., Klaus Schultz. (2006). Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. Proceedings of the Fifth International Conference on Intelligent Tutoring Systems. Montreal, Canada. June 2000. pp. 574-583 Springer Verlag.

Austin, P. C. (2014). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research 46.3 (2011): 399–424. PMC. Web. 12 Dec. 2014.

Azarnoush, B., Bekki, J. M., Runger, G. C. , Bernstein, B. L., Atkinson, R. K. (2013). Toward a Framework for Learner Segmentation. 2013. Journal of Educational Data Mining, Volume 5, No 2, 2013.

Azevedo, R., Guthrie, J. T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. Journal of Educational Computing Research, 30(1 & 2), 87–111.

Barber, R., Sharkey, M. (2012). Course Correction: Using Analytics to Predict Course Success. Preceeding of the International Conference on Learning Analytics & Knowledge (LAK), 29 April – 2 May 2012, Vancouver, BC, Canada.

Bettadapura, V., Schindler, G., Ploetz, T., Essa, I. (2013). Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition. 26th IEEE Conference on Computer Vision and Pattern Recognition

Bouchet, F., Kinnebrew, J., Biswas, G., Azevedo, R. (2012). Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. Proceedings of the 5th International Conference on Educational Data Mining 2012, 65-72.

Brand, A., Bradley, M. T., Best L., Stoica, G. (2011). Multiple trials may yield exaggerated effect size estimates. The Journal of General Psychology 138 (1): 1–11. doi:10.1080/00221309.2010.520360.

Chambers, J.; Sprecher, J. (1983). Computer-Assisted Instruction: Its Use in the Classroom. Prentice-Hall Inc.

Cochrane, W., and Chambers, S. (1965). The Planning of Observational Studies of Human Populations," Journal of the Royal Statistical Society, Series A, 128, 234-266.

Cohen, J. (1992). A power primer. Psychological Bulletin 112 (1): 155–159. doi:10.1037/0033-2909.112.1.155. PMID 19565683.

Cotton, K. (1989). Educational time factors, Northwest Regional Educational Laboratory. Retrieved from http://educationnorthwest.org/webfm_send/564.

DeGrave, W. S., Boshuizen, H. P. A., and Schmidt, H. G. (1996). Problem-based learning: Cognitive and metacognitive processes during problem analysis. Instr. Sci. 24:321–341.

Desmarais, M., C., Meshkinfam, P., Gagnon, M. (2006). Learned student models with item to item knowledge structures. User Modeling and User-Adapted Interaction, 16(5):403–434.

Dehejia, R. (2005). Practical propensity score matching: A reply to Smith and Todd. Journal of Econometrics, 125(1–2), 355–364.

Dods, R. F. (1997). An action research study of the effectiveness of problem-based learning in promoting the acquisition and retention of knowledge. J. Educ. Gifted 20: 423–437.

Eli (2011). "Seven Things You Should Know About First Generation Learning Analytics.". EDUCAUSE Learning Initiative Briefing. http://www.educause.edu/library/resources/7-things-you-should-know-about-first-generation-learning-analytics.

Ellis, Paul D. (2010). The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. United Kingdom: Cambridge University Press.

Ellis, R. K. (2009). Field Guide to Learning Management Systems. American Society for Training & Development (ASTD).

Garrison, D., Vaughan, N. (2008). Blended learning in higher education: Framework, principles, and guidelines. San Francisco, CA: John Wiley & Sons.

Garrison, D. R., Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. The Internet and Higher Education, 7, 95–105.

Gray, G., McGuinness, C., Owende, P., Carthy, A. (2014). A Review of Psychometric Data Analysis and Applications in Modelling of Academic Achievement in Tertiary Education. Journal of Learning Analytics, 1(1), 75–106.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. Econometrica, 66, 1017-1098.

Helmreich, J. E., Pruzek, R. M. (2009). PSAgraphics: An R Package to Support Propensity Score Analysis. Journal of Statistical Software, February 2009, Volume 29, Issue 6.

Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn? Educational Psychology Review 16 (3): 235.

Holland, Paul W. (1986). Statistics and Causal Inference. J. Amer. Statist. Assoc. 81 (396): 945–960.

Hubert, L., Arabie, P. (1985). Comparing partitions, Journal of classification, 2.1,193-218.

Imbens G.W., (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics. 2004; 86:4–29.

Johnson,J.,Shum,S.B.,Willis,A.,Bishop,S.,Zamenopoulos,T.,Swithenby,S., Helbing, D. (2012). TheFuturICT education accelerator. The European Physical Journal, Special Topics, 214(1), 215–243.

Kelley, K., Preacher, K. J. (2012). On Effect Size. Psychological Methods 17 (2): 137–152.

King, F.B., Harner M., & Brown S. W. (2000). Self-regulatory behavior influences in distance learning. International Journal of Instructional Media, 27(2), 147–155.

Klassen, R.M., Krawchuk, L.L., & Rajani, S. (2008). Academic procrastination of undergraduates: Low selfefficacy to self-regulate predicts higher levels of procrastination. Contemporary Educational Psychology, 33, 915–931.

Kusurkar,R. A., Ten Cate, Th. J. C., Vos, M. P., Westers, P., and Croiset, G. (2013). How motivation affects academic performance: a structural equation modelling analysis. Advances in Health Sciences Education March 2013, Volume 18, Issue 1, pp 57-69.

Kuehl, R., O. (1994). Statistical Principles of Research Design and Analysis. Wadsworth Publishing Co., Belmont, CA

Lee, B. K., Lessler, J., Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, Volume 29, Issue 3, pages 337–346.

Lakshminarayan, N., Potdar, S., Reddy, S.G. (2012). Relationship Between Procrastination and Academic Performance Among a Group of Undergraduate Dental Students in India. Journal of Dental Education, Volume 77, Number 4.

Lau, W., Yuen, A. (2009). Exploring the effects of gender and learning styles on computer programming performance: implications for programming pedagogy. British journal of educational technology [0007-1013] vol:40 iss:4 pg:696.

Lockwood C. R.,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. Higher Education: Handbook of Theory and Research. The Association for Institutional Research (AIR) and the Association for the Study of Higher Education (ASHE) Volume 24, 2009.

Mackness, J, Mak, S., and Williams, R. (2010). The Ideals and Reality of Participating in a MOOC", Proceedings of the 7th International Conference on Networked Learning.

McCormick, C., B. (2006). Metacognition and Learning. Handbook of Psychology. John Wiley & Sons, Inc.

Meyer, K. A. (2002). Quality in distance education: Focus on on-line learning. In A.J. Kezar (Ed.), ASHE-ERIC Higher Education Report (Vol. 29, pp. 1–134). Jossey-Bass.

Mitchell, M. (1996). An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press.

Moon, S. M., Illingworth, A. J. (2005). Exploring the dynamic nature of procrastination: A latent growth curve analysis of academic procrastination. Personality and Individual Differences 38 (2005) 297–309.

Morgan, S. L., Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. Sociological Methods and Research, 35(1), 3–60.

Muñoz-Merino, Pedro J., Ruipérez-Valiente, José A., Delgado Kloos, Carlos. Inferring higher level learning information from low level data for the Khan Academy platform. (2013) Pages: 112-116, Proceedings of the 3rd International Conference on Learning Analytics and Knowledge.

Murphy, L., Richards, B., McCauley, R., Morrison, B., Westbrook, S., Fossum, T. (2006). Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2006, Houston, Texas, USA, March 3-5

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press

Perera, D., Kay, J., Koprinska, I., Yacef, K., Zaiane, O. R. (2009). Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. In IEEE Transaction on Knowledge and Data Engineering, 21.6, 759-772.

Pintrich, P.R. & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82 (1), 33–40.

Polson, Martha C.; Richardson, J. Jeffrey, eds. (1988). Foundations of Intelligent Tutoring Systems. Lawrence Erlbaum.

Powell, Stephen, and Shiela MacNeil (2002). Instituitional Readiness for Analytics A Briefing Paper. CETIS Analytics Series. JISC CETIS, December 2012.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association (American Statistical Association) 66 (336): 846–850.

Reynolds and DesJardins (2009). Answering Whether Attendance at a Two-Year Institution Results in Differences in Educational Attainment in Higher Education. Handbook of Theory and Research, Volume XXIII Edited by John C. Smart

Roediger H. L. III, Karpicke J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249–255.

Rosenbaum, P.R. (2002). Observational Studies. 2nd edition. Springer-Verlag, New York.

Rosenbaum, P. R., Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, Vol. 70, No. 1. (Apr., 1983), pp. 41-55.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educ. Psychol. 66 (5): 688–701.

Scheines, R., Leinhardt G., Smith, J., Cho, K. (2005) Replacing Lecture with Web-Based Course Materials. Journal of Educational Computing Research, 32 (2005), 1-26.

Sekhon, J. S., Diamond A. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Review of Economics and Statistics, 95(3): 932-945.

Sekhon J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. Journal of Statistical Software, Volume 42, Issue 7.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Siemens, G., & Baker, R.S.J.d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. Proceedings of the Second International Conference on Learning Analytics and Knowledge, Vancouver, Canada, 252–254.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. Educause
Review, 46(5), 30–23.

Smith, J. A., Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, Volume 125, Issues 1–2, March–April 2005, Pages 305–353.

Spirtes, P., Glymour C., Scheines, R. (2000). Causation, Prediction, and Search. 2nd Edition. MIT, Cambridge MA.

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 63, Issue 2, pages 411–423.

Vaughan, N. D. (2010). Blended Learning. In Cleveland-Innes, MF; Garrison, DR. *An Introduction to Distance Education: Understanding Teaching and Learning in a New Era*. Taylor & Francis. p. 165.

Whipp, J. L. & Chiarelli, S. (2004). Self-regulation in a web-based course: A case study. Educational Technology Research and Development, 52(4), 5–22.

Woolf, B. P. (2008). Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning. Published by Elsevier & Morgan Kaufmann.

Yuan, Li, and Stephen Powell (2013). MOOCs and Open Education: Implications for Higher Education White Paper. University of Bolton: CETIS. p.6.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In Boekaerts, M., Pintrich, P., & Zeodmer, M. (Eds.), Handbook of Self-Regulation. Academic Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70.