Gender Matters: Assessing and Addressing the Persistent Gender Gap in Introductory Physics Courses at the University of Michigan

By
Kate Miller

A thesis submitted in partial fulfillment of
the requirements for the degree of

BACHELOR OF SCIENCE WITH HONORS

UNIVERSITY OF MICHIGAN
DEPARTMENT OF PHYISCS

April 5, 2011

Advised by: Professor Timothy McKay

Abstract

Students arrive in class with differing backgrounds. Studying the classroom performance of students with a knowledge of these backgrounds allows us to better understand how a student is transformed by a course of instruction. While we expect that some background factors, such as prior test scores and academic preparation, should influence student success, we are concerned about inappropriate impact that results from uncontrollable factors, such as gender, socio-economic status, and race. In particular, a gender disparity in Science, Technology, Engineering and Mathematics (STEM) fields is nationally recognized and especially prominent in physics. With enrollment and success in introductory college physics courses a crucial factor in this female underrepresentation, our goal is to investigate the gender gap in introductory physics courses at the University of Michigan. We report analysis of data for 48,579 students who have taken introductory physics courses at UM between Fall 1996 and Winter 2010. We clearly detect the presence and persistence of a gendered performance gap in all of these courses in all terms considered. We find that differing mathematical preparation accounts for some of this gender gap, especially in the female dominated life science sequence. The physical science and engineering sequence, which is substantially male dominated, shows a strong gender difference even after differing mathematical preparation is accounted for. Finally, we investigate supplementary study groups as one means of intervention, finding that they do not significantly influence student success. From these conclusions and relevant literature, we provide recommendations of intervention techniques aimed at increasing student success and reducing the gender gap at the University of Michigan.

I. Table of Contents

I. Table of Contents
II. Figures
III. Introduction
- A. Statement
- B. Background
  1. Assessment Instruments
     - *a) Overview of Concept Inventories*
     - *b) Validity of Concept Inventories*
     - *c) Development of Distracters*
     - *d) Format of Questions*
     - *e) Pre-test Influence on Post-test*
     - *f) Influence of the Lack of Grading Incentive*
     - *g) Effectiveness of Individual Questions on the FCI and BEMA*
     - *h) Accepted Validity*
  2. Pedagogy, Coursework and Student Background
     - *a) Comparing Interactive Engagement and Traditional Instruction*
     - *b) Seat Location*
     - *c) The Gender Gap*
     - *d) Influence of Interactive Engagement on the Gender Gap*
     - *e) Cumulative GPA and Grade Predictive Schemes*
     - *f) Additional Findings and Proposed Improvements from Lai*

IV. Methods
- A. Data
- B. Analysis Tools
  1. Kolmogorov-Smirinov Test
  2. Bootstrap Re-sampling

V. Analysis
- A. Gender
  1. The Presence of the Gender Gap
  2. Mathematical Preparation
  3. Female Representation
  4. Instructor Gender
- B. Science Learning Center Study Groups
  1. The Effects of Science Learning Center Study Groups on Students
  2. The Effects of Science Learning Center Study Groups on Male Students
  3. The Effects of Science Learning Center Study Groups on Female Students

VI. Conclusion
VII. Recommendations
VIII. Extensions
IX. Acknowledgments
X. Work Cited
XI. Appendix
- A. Term Codes and Descriptions
- B. External Data Parameters
- C. Internal Data Parameters

II. Figures

Figure 1: Methodological Comparison of Concept Inventories 2
Figure 2: Comparison of FCI Post-Test Scores 2
Figure 3: Signs of Lack of Seriousness 2
Figure 4: Histogram of Average Normalized Gain 2
Figure 5: Average FCI gain by Instruction 2
Figure 6: Students Grouped by Initial Seating Assignment 2
Figure 7: Initial Seating Location and Attendance 2
Figure 8: Initial Seating Location and Course Performance 2
Figure 9: Instructional Approach and Normalized Gain, by Gender (Harvard) 2
Figure 10: Instructional Approach and Normalized Gain, by Gender (Colorado) 2
Figure 11: The Effects of Values Affirmation on Mean Overall Exam Score and Mean Post-test FMCE Score 2
Figure 12: The Effects of Values Affirmation by Gender Stereotype Endorsement 2
Figure 13: Grade Prediction Model for Physics 125 2
Figure 14: Gender Gap in Final Grades 2
Figure 15: Incoming GPA Difference vs. Output Grade Difference by Term 2
Figure 16: Gender gap vs. Gender Probability, by Term 2
Figure 17: Gender Gap in SAT Math Scores 2
Figure 18: Difference in Fractions Receiving B or Better vs. Female Representation 2
Figure 19: Difference in Fractions Receiving Less Than C vs. Female Representation 2
Figure 20: Incoming GPA Difference vs. Output Grade Difference, by Term, for Female Instructors 2
Figure 21: Cumulative GPA 2
Figure 22: Probability of Course Grades 2
Figure 23: Comparison with Predicted Grades 2
Figure 24: Differences in SLC Means 2
Figure 25: Male Cumulative GPA 2
Figure 26: Male Probability of Course Grades 2
Figure 27: Male Comparison with Predicted Grades 2
Figure 28: Differences in Male SLC Means 2
Figure 29: Female Cumulative GPA 2
Figure 30: Female Probability of Course Grades 2
Figure 31: Female Comparison with Predicted Grades 2
Figure 32: Differences in Female SLC Means 2

III. Introduction

A. Statement

Quantitative study of learning environments has potential to reveal driving factors of student success and failure. Exploring correlations between various parameters and student performance can yield revealing trajectories. These correlations are found to be independent of instructor, and therefore universally applicable. While some variables, such as student effort and academic preparation, are expected to affect performance, the influence of other uncontrollable variables, such as gender, race and socio-economic status, present an unjust bias. In particular, we will refer to the gender differences that arise when controlling for incoming factors as the 'gender gap'. Furthermore, Havarti et al. assert,

> …heterogeneity rather than homogeneity will lead to progress in the field by introducing new perspectives. Kenway and Gough (1998) observe that the intellectual potential of females is an untapped source for furthering scientific knowledge. (Hazari, Tai, & Sadler, 2007)

Whether motivated by equity in the classroom or by the evolution of the field of physics, with a further knowledge of what influences effective learning we can then start to answer the question: how can we encourage student success for everyone in the course?

One issue of national concern is the gender disparity in participation in the Science, Technology, Engineering and Math (STEM) fields. In particular, physics has historically been a male-dominated field with few females. According to a recent report from the American Institute of Physics, "During 2003, women earned 22% of the bachelor's degrees in physics and 18% of the PhDs in physics – a record high" (Ivie & Ray, 2005). As compared to other science disciplines, such as Astronomy, this statistic is staggeringly low. While women are grossly underrepresented in physics, the problem does not arise from a lack of female enrollment in introductory physics courses at the high school level. The "pipeline" tracks populations at various points throughout a physics career: high school, bachelor's degree, PhD degree, assistant professor, associate professor and full professor. Examining this pipeline by gender, it becomes clear that enrollment in introductory physics courses at the college level is the main contributing factor for the drop-off in female involvement; in high school, approximately half of physics students are women whereas less than one-fourth of bachelor's degrees in physics are earned by women. After this initial "leak" in the pipeline, women are represented at about the levels we would expect based on past degree production (Ivie & Ray, 2005). Enrollment in an introductory college physics course is crucial. Perhaps it is the culture of physics or the reputation of a course that drives the choice to enroll. The question remains: what is happening to cause this leak?

This thesis explores the presence of the gender gap at the University of Michigan in an effort to understand the nature of this underrepresentation of females in introductory physics courses. We investigate correlations between the gender gap and several parameters that are thought to contribute to this inequity. Particularly, we look at both variables specific to a student's experience in the course as well as background factors that are thought to influence learning, placing an emphasis on incoming GPA, SAT math score, male to female ratio in the classroom and instructor gender. It is the hope that further

understanding will lead to the development of effective intervention techniques and eventually a reduction of the gender gap.

One means of intervention which we explore is the use of study groups led by the University of Michigan's Science Learning Center (SLC). These study groups provide a smaller, more interactive atmosphere of learning. We investigate whether these supplementary sessions can serve as a proactive step towards increasing the likelihood of student success in an introductory physics course. We began this study optimistic that study groups would be found effective in narrowing the gender gap.

Overall, our main goal is to better understand how students learn in introductory physics courses. By exploring correlations between the incoming students and their final course grades a more complete picture of the effects of student identity and pedagogical choices is formed. This information can then be used to inform instruction and provide a basis for recommendations of proactive intervention techniques.

This thesis is unusual from other undergraduate physics theses as the topic is not the inanimate world. Instead, we focus on analyzing human subjects in the context of a social environment. Despite this unique topic, we have tried to approach our data from the perspective of a physicist as we aim to provide quantitative evidence to validate or disregard our hypotheses. The training in making measurements and determining levels of uncertainty is comparable to that of fellow physics undergraduate researchers. It has been our goal to explore how the scientific approaches and sensibilities of a physicist might be brought to bear on an unusual problem, but one of great importance to the physics community.

We begin with a summary of relevant literature, first exploring methods of evaluating student understanding and subsequently exposing the challenges that arise in using these forms of assessment. An additional discussion of literature on the influences of pedagogical decisions and important factors which correlate with student success is provided as motivation for our analysis. After reviewing our compiled data and two analytical tools, the Kolmogorov-Smirinov Test and Bootstrap Re-sampling, which we will use in our analysis, we present our evidence for the presence and persistence of the gender gap in introductory physics courses at the University of Michigan. We also explore the influence of mathematical preparation, female representation and instructor gender on the gender gap. Next, we investigate SLC study groups as a means of intervention, analyzing student success for all students as well as by gender. Finally, we conclude with recommendations of ways to improve student performance and reduce the gender gap in introductory physics courses at the University of Michigan.

B. Background

The pedagogical approaches to introductory physics courses have been, and continue to be, investigated by the physics education research (PER) community. It is important to understand this research in order to build a foundation for further inquiry. Therefore, we will familiarize the reader with basic concepts upon which we will later build.

First, we review assessment instruments to provide a basis for evaluating student success. As with all assessment techniques, careful consideration what of measurements actually imply is essential as data is analyzed. Second, we will summarize the previous research on the effects of student identity upon entering the course. Of particular relevance are gender and cumulative GPA. Finally, an overview of what happens in the physics course itself, including curriculum, structure and overall pedagogy, is provided. The reader will develop a basic knowledge of physics pedagogy and understand how differing instruction techniques can affect student success. Subsequently, we will build on this understanding as we analyze the presence of the gender gap, potential correlating factors, and study group intervention at the University of Michigan.

1. Assessment Instruments

There are many ways to define and evaluate student success. The most widely used method of assessment in the field of PER is a *concept inventory*; a multiple choice conceptual test administered at the beginning and end of a term of instruction. Concept inventories stand out in their design by accounting for disparities in background knowledge, providing a measurement of how much understanding each individual student *gains* throughout the course. This is especially useful when comparing data at different institutions as the gain score attempts to account for student preparation. However, while concept inventories test knowledge that is thought to be what typical students *should* learn, these content objectives do not always align with the objectives of a particular course.

Course grades and student evaluations give another perspective on student understanding and overall experience in the classroom. Course grades primarily reflect exam scores, with less emphasis on several other course elements, while student evaluations are based on the students' personal experiences and opinions about the course. In our analysis, instead of using concept inventories, we will use course grades as a measurement of student performance. Final course grade data is readily available and, in some sense, represents what really matters at the end of a course. Course grades should also be better aligned with course goals since exams are created by the professor of the course. It is also important to recognize that our data is continuous at the University of Michigan, so we need not be concerned with calibrating our data with different institutions. Finally, similar to the gain measured from concept inventories, we can account for differences in student background by using grade prediction schemes. With predicted grades as a reference point, we can determine if a student does worse, better or the same as expected. Like gain scores, this method recognizes differences in student understanding upon entering the course.

We will begin with a brief overview of concept inventories and some of the many factors to consider in using these assessments as an indicator of student success. As we explore the challenges that arise in measuring student understanding, we provide the reader with a basis for later understanding relevant PER literature.

a) Overview of Concept Inventories

Administering tests that probe student understanding of basic physics ideas is a common mode of assessment. Such *concept inventories* are formally defined as, "A multiple-choice instrument designed to

evaluate whether a person has an accurate and working knowledge of a concept or concepts" (Lindell, Peak, & Foster, 2007). These timed assessments often present widely held misconceptions and common-sense alternatives, known as 'distracters', as answer choices. They typically avoid challenging quantitative questions and a calculator is not permitted. Additionally, closed-book and closed-note restrictions ensure an individual's knowledge is indeed being tested.

In particular, a pioneering concept inventory is the Force Concept Inventory (FCI). The FCI is carefully designed to test the basic principles of Newtonian Mechanics, a focus of the traditional first-semester physics curriculum. Topics include kinematics, Newton's laws, the superposition principle, and force analysis. In general, a score of 60% on the FCI is accepted as a reasonable benchmark for an understanding of Newtonian concepts (Hestenes, Wells, & Swackhamer, 1992). By collecting question response data from the FCI, one can begin to determine which topics are well understood. Subsequently, these conclusions can be used to prompt a discussion of potential pedagogical revisions.

Similarly, The Brief Electricity and Magnetism Assessment (BEMA) aims to explore student understanding of the Electricity and Magnetism portion of second-semester physics. Topics covered on the BEMA include electric field, electric potential, magnetic field, basic circuits, and applications of Maxwell's equations. Like the FCI, the BEMA is consciously designed to test a student's basic knowledge of Electricity and Magnetism, exposing weak and strong content areas.

Strong alignment between assessment and course content is key. Besides these two most popular assessments, there are a variety of other comparable concept inventories focusing on different content areas. Such tests include the Astronomy Diagnostic Test (ADT), Conceptual Survey in Electricity and Magnetism (CSEM), Diagnostic Exam Electricity and Magnetism (DEEM), Determining and Interpreting Resistive Electric Circuits Concept Test (DIRECT), Energy and Motion Conceptual Survey (EMCS), Force and Motion Conceptual Evaluation (FMCE), Lunar Phases Concept Inventory (LPCI), Mechanics Baseline Test (MBT), Test of Understanding of Graphs in Kinematics (TUG-K), Wave Concept Inventory (WCI), etc. (Lindell, Peak, & Foster, 2007). It is important to consider which test, if any, is consistent with the content objectives for the course when choosing which assessment to use. Great differences between the material that is being taught and the material that is being assessed may reduce a concept inventory's accuracy in portraying student understanding. In these cases, course grades should instead be used as an assessment of student success.

One way to measure how much a student has learned between the beginning and end of the course is to administer a concept inventory as a pre-test and a post-test. Both times, the same questions are presented in a consistent format. Combining pre-test and post-test scores, a 'gain score' can be calculated:

$$G = \frac{postscore\ \% - prescore\ \%}{100 - prescore\%},$$

where postscore % and prescore % refer to the fraction of correct answers on the post-test and pre-test, respectively. Often, further analysis is conducted by comparing gain scores. This gain score represents how much students have learned between the pre-test and post-test; that is, how much student

understanding has changed (improved, remained constant, or declined) in the specified time period. A typical FCI gain score is on the order of 25%, indicating that a student erased one fourth of their misconceptions, or gained one fourth of what they still needed to learn. Exceptional FCI gain scores are on the order of 45% (Lindell, Peak, & Foster, 2007).

b) Validity of Concept Inventories

While the use of assessments such as the FCI and BEMA present rich results and form a basis for many PER analyses, it is important to deeply consider the accuracy of these assessments. Foremost, one must consider the methodology used in writing a concept inventory. By comparing the similarities and differences of many methodological factors, Rebecca Lindell, Elizabeth Peak and Thomas Foster set out to answer the question, "Are all concept inventories created equal?"

Figure 1: Methodological Comparison of Concept Inventories

	Concept Domain Determined by			Test Specifications				Item Statistics Reported			Field Testing					Validity Studies			Reliability Statistics Reported		
				Basis of Distracters		Distracter Correspondence to Alternate Models					Size			Location							
Instrument	Qualitative Study	Researcher	Existing Literature	Researcher's Understanding	Student understanding	Corresponds	Does Not Correspond	Difficulty	Discrimenation	Concentration Analysis	>500 students	500 - 1000 students	>1000 students	Local	National	Criterion	Construct	Content	Cronbach Alpha	Kuder - Richardson	Point Biserial
ADT[2]	■	■	■	■			■	■	■	■			■		■		■	■			
BEMA		■		■			■	■	■			■		■				■		■	■
CSEM								■	■				■		■			■		■	
DEEM	■	■						■	■		■			■					■	■	
DIRECT		■	■				■	■	■				■		■		■	■			
EMCS		■			■		■	■	■				■		■		■	■	■		
FCI		■		■			■	■					■		■						
FMCE	■		■		■	■							■		■	■	■		■		
LPCI	■		■		■	■		■	■	■			■		■		■	■	■		
MBT		■		■	■		■						■	■			■	■		■	
TUG-K		■	■	■	■		■	■	■			■			■		■	■		■	■
WCI											■			■			■	■			

[1] All analysis based on research reported in original paper or personal communication. Blanks refer to non-reported information.
[2] Questions relating to Lunar Phases and Seasons were based on qualitative investigation conducted by R. Lindell; rest of concept domain determined by researchers.

The results (see Figure 1) indicate all concept inventories are, in fact, *not* created equal (Lindell, Peak, & Foster, 2007). For example, researcher understanding forms the basis for distracters on the FCI with no reported input from student understanding. Conversely, the EMCS only takes student understanding into account when composing distracters, and does not report researchers understanding as a basis of distracter construction. This particular decision will perhaps affect the answer choices presented in the concept inventory, yielding different results.

Overall, there is no consistent approach to the development of these concept inventories, despite their similar analytical applications. Lindell, Peak and Foster urge the reader to consider these varying approaches when deeming a concept inventory an accurate source of data. These same considerations apply to all assessment instruments, including course grades.

c) Development of Distracters

In addition to the lack of consistency in concept inventory development, there are specific concerns about using these assessments as indicators of student understanding. First, there are complications introduced by having distracters as multiple choice answer alternatives. Distracters are included with the intention of determining whether students have overcome common-sense misconceptions, indicating a true understanding of the correct physics explanation. Because of this, distracters must be carefully composed, reflecting typical student misunderstandings, if the question is to be an accurate reflection of a student's grasp on physics concepts.

Dean Zollman and Sanjay Rebello explored the alignment of responses on FCI questions and equivalent open-ended questions with a sample student population of non-majors who generally had some physics background at Kansas State University. After administering the FCI to one randomly chosen group and the same questions in an open-ended format to another randomly chosen group, the open-ended answers were sorted based on naturally occurring categories in the responses. Comparing these answers, it is apparent that misconceptions presented in the multiple choice format differ from the misconceptions that appear in the open-ended format (Zollman & Rebello, 2004). While there is only one right answer, there are many possible wrong answers. It seems that the distracters in the FCI do not necessarily reflect the misconceptions of the students. Therefore, conclusions about student misunderstanding based on the FCI distracters may not be accurate. This is a fundamental limitation of all multiple choice assessments.

Furthermore, researchers presented revised multiple choice questions in which the misconceptions resulting from the open-ended questions replaced irrelevant FCI distracters. Upon comparing the number of students who chose the original FCI distracters verses the revised distracters, the latter tended to dominate. Thus, it can again be concluded that, "an analysis of the incorrect responses to FCI questions may not be an effective way to determine which parts of the students' conceptual understanding are deficient" (Zollman & Rebello, 2004). Furthermore, a caveat is included in the final discussion of this study warning that distracters are ephemeral; misconceptions fluctuate as the students learn physics jargon and confuse content throughout the semester (Zollman & Rebello, 2004).

Considering these two results, it is clear that not all distracters are useful in identifying misconceptions; in fact, most are not accurate. Yet, distracters are still included in concept inventories to fulfill their originally intended purpose: differentiate between a student's true understanding of physics concepts and some common prevailing fallacies.

d) Format of Questions

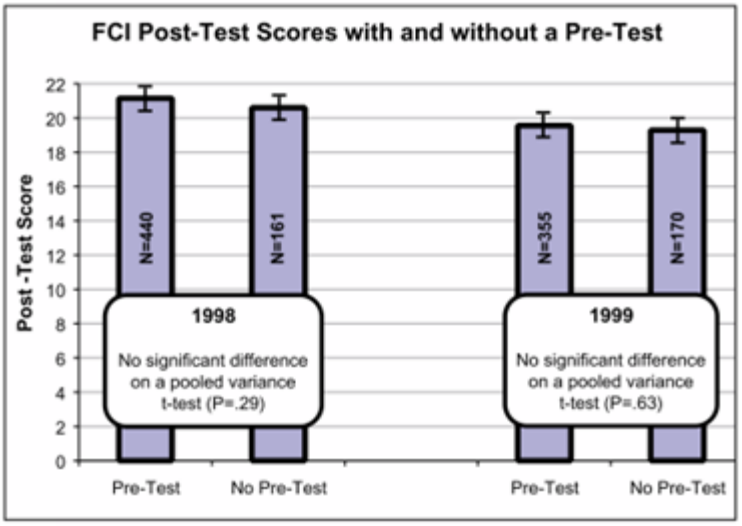
The format of the questions in concept inventories also requires attention. The same question can be presented in a variety of different ways including verbal, graphical, diagrammatic and mathematical/symbolic. Does the format of the questions affect student performance, and thus, affect subsequent conclusions regarding student understanding? David Meltzer administered similar Newton's third-law questions in two different ways: verbal and diagrammatic. He found that correct responses to the verbal format are consistently higher than the diagrammatic format, with incorrect responses following a similarly consistent pattern (Meltzer, 2005). Since verbal and diagrammatic questions produce quite different results, further analysis must bear in mind the implied uncertainty.

Second, Meltzer presented four quizzes, each containing four similar questions. Every quiz was a different representation: verbal, diagrammatic, mathematical/symbolic and graphical. The analysis of the responses to these quizzes revealed generally consistent results among the males, yet a decreased performance among females answering the graphically formatted questions. Thus, the format does not dramatically affect male performance whereas females tend to do worse on questions involving graphical interpretation (Meltzer, 2005). It is apparent that the format of a question on concept inventories affects the accuracy of responses in a gendered manner.

e) Pre-test Influence on Post-test

It is a common concern for all gain score assessments that taking a pre-test and, later, taking the same assessment as a post-test may inflate the post-test scores. Perhaps, the students remember the questions or are desensitized to the format of the assessment. However, this does not prove to be the case. A comparative study at the University of Minnesota, between a control group that takes only a FCI post-test and the experimental group that takes both the FCI pre-test and the post-test, yields no statistical difference (Henderson, 2002).

Figure 2: Comparison of FCI Post-Test Scores



Thus, based on these results (see Figure 2) we see that the pre-test does not bias the post-test scores. This lack of influence is generally accepted among concept inventory administrators.

f) Influence of the Lack of Grading Incentive

With points generally not rewarded for correct answers (but perhaps based on completeness), the question arises: do students take this assessment seriously? To examine this, Charles Henderson administered an ungraded FCI pre-test, followed by ungraded and graded post-tests. With these varying grading incentives, Henderson examined how seriously the students took the assessment. Indicators to conclude a lack of a serious attitude include refusing to take the test, drawing a picture on the Scantron sheet, answering all the same letter, leaving six or more blanks, and other letter patterns in the responses. The results from this experiment (see Figure 3) show that a maximum of 2.8% of students do not take the assessment seriously due to the lack of grading incentive (Henderson, 2002). These students simply do not try to accurately respond to the questions.

Figure 3: Signs of Lack of Seriousness

	Pre-Ungraded	Post-Ungraded	Post-Graded	Maximum % of students not taking test seriously as a result of grading option
	N = 1856	*N* = 524	*N* = 1332	
	1997, 1998, 1999	1999	1997, 1998	
Refuses to take test	0.5%	0.5%	0.0%	0.5%
Draws a picture	0.2%	0.2%	0.0%	0.2%
Answers all A's, B's, etc.	0.0%	0.0%	0.0%	0.0%
Leaves a lot of blanks (six or more)	1.5%	1.5%	0.1%	1.4%
Other Patterns				
ABCDE, EDCBA	0.8%	1.0%	0.5%	0.5%
Six A's, B's, etc.	0.2%	0.2%	0.0%	0.2%
Total	**3.2%**	**3.4%**	**0.6%**	**2.8%**

What about the students who do respond to the questions? Do they put as much effort in when it is ungraded as they do when it is graded? There is, in fact, a small effect due to the grading incentive: about half of a FCI item between the graded test (21.4 ± 0.2) and the ungraded test (20.9 ± 0.2). Once again, this is a relatively small difference which leads to the conclusion that the lack of grading incentive does not significantly affect the resulting data (Henderson, 2002).

g) Effectiveness of Individual Questions on the FCI and BEMA

Furthermore, detailed studies have been performed on the FCI and BEMA analyzing the effectiveness of each individual question. Upon conducting follow-up interviews with test takers and knowledgeable graduate students, it is apparent that some questions contain weak discriminators or confusing diction, which likely affects the analysis of these particular questions (Hestenes, Wells, & Swackhamer, 1992; Ding, Chabay, Sherwood, & Beichner, 2006). In some studies which utilize concept inventories as a main source of data collection, such ineffective questions are omitted in the analysis in order to account for this known inaccuracy.

h) Accepted Validity

Overall, despite the above statistical considerations, concept inventories, particularly the FCI and BEMA, are generally accepted and widely used as valid measurements of student understanding by the PER community. Analysis of gain scores is considered an accurate basis for testing pedagogical efficacy. However, it is important to remember that, with any assessment, uncertainty and biases remain. Whether it is the lack of complete methodology review in the development process, a distracter that does not reflect a student's misconception or the choice of question format, analyses and conclusions must be considered with care. In the end, we must recognize that assessments generally provide data of what is measured, not necessarily what the experimenter *wishes* to measure. Concept inventories record student answers to particular questions. This may be correlated with ultimate student understanding, but they are not perfect measures.

2. Pedagogy, Coursework and Student Background

We next review relevant literature that explores specific factors unique to both a student's identity and the structural set-up of the classroom and how they affect student success. We will begin by exploring internal factors, focusing on the effectiveness of different pedagogical techniques in the physics classroom. We will then shift to external factors that the students bring to the course, with an emphasis on the impact of gender and prior GPA on student performance.

Together, internal and external factors give the reader an understanding of what impacts student success. This basis motivates our research as we hypothesize which factors may affect student performance at the University of Michigan. We will later examine these important parameters in our analysis.

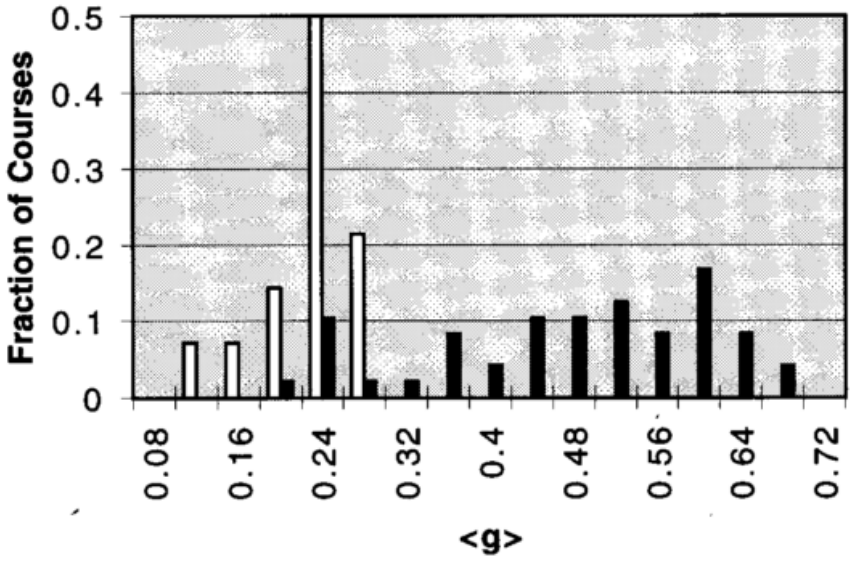
a) Comparing Interactive Engagement and Traditional Instruction

Introductory physics courses generally have a curriculum that is quite uniform as compared to other subjects; however, while the content is consistent, the structure of the activities in the courses varies widely. One basic pedagogical decision is the mode of instruction. Generally, one may view the teaching style in relation to two basic categories: interactive engagement and traditional instruction. Interactive engagement is when students are actively engaged in the class. The key to interactive engagement is student participation; this may include, but is not limited to, the creation of a discussion atmosphere, use of hands-on experiments, and/or a real-time feedback system. On the other hand, traditional instruction refers to the conventional lecture structure in which the teacher lectures to their students. This instructional style invokes little student involvement and aims to teach students primarily via lecture presentations.

Several studies have compared interactive engagement and traditional methods of instruction. In almost all studies, better student understanding occurs in courses that utilize interactive engagement methods. Richard Hake compares gain scores on the FCI in 62 introductory physics courses at Indiana University (N=6542), structured as either interactive or traditional instruction. The results are overwhelmingly in favor of interactive instruction.

Figure 4: Histogram of Average Normalized Gain

White bars correspond to traditionally taught courses whereas black bars correspond to interactive engagement courses.

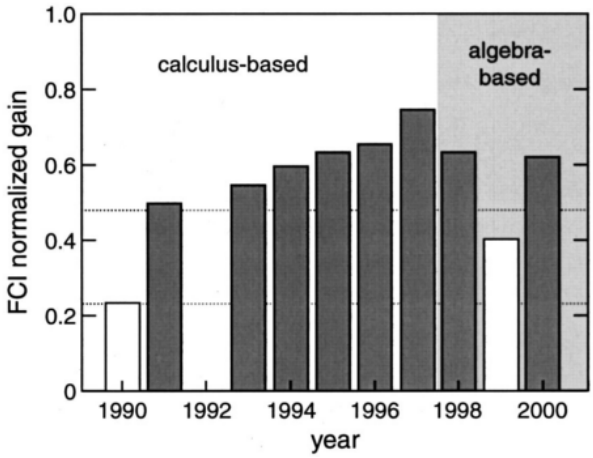


As seen in the above graph (Figure 4), all gain scores for the traditionally taught courses fall into the low-gain region ($g < 0.3$) with an average of $g=0.23\pm0.04$. On the other hand, 85% of the interactive engagement courses fall in the medium-gain range ($0.7 > g \geq 0.3$) and 15% in the low-gain region, with an average of $g=0.48\pm0.14$. While no courses achieved a high-gain ($g \geq 0.7$), the results are conclusive: a higher gain score is achieved in the courses taught with interactive engagement (Hake, 1997).

Similarly, Catherine Crouch and Eric Mazur produce parallel results from introductory physics courses at Harvard University. Here, the term 'peer instruction' refers to modifications to the traditional lecture format which include questions designed to engage students and uncover difficulties with the material. Peer instruction is one widely adopted example of interactive engagement methods. Once again, peer instruction and traditional courses are compared on the basis of FCI gain scores (Crouch & Mazur, 2001).

Figure 5: Average FCI gain by Instruction

The black bars correspond to Peer Instruction courses whereas the white bars correspond to Traditional Instruction courses.



In Figure 5, the upper dotted line represents the average gain score of the interactively taught courses (g=0.48) and the lower dotted line represents the average gain score of the traditionally taught courses (g=0.23). The results are startlingly clear: higher gain scores occur in courses that are interactively structured whereas lower gain scores occur in courses that are traditionally structured.

There is a surprising feature in these results. Crouch and Mazur report exceptionally high gain scores; higher than those seen in any other study. This makes us wonder whether these results are unique to the Harvard environment, and possibly not applicable elsewhere.

b) Seat Location

Another factor to consider particularly in large introductory courses is where a student sits in a large lecture hall relative to the front. Perhaps seat selection itself contributes to student performance. A study at the University of Colorado in Boulder yields interesting correlations pertaining to seat location. Groups of three to four students were randomly assigned a seat location, sitting adjacent to one another, at the beginning of the term. Then, halfway through the term, each group's seating location was switched; the front moved to the back and the back to the front.

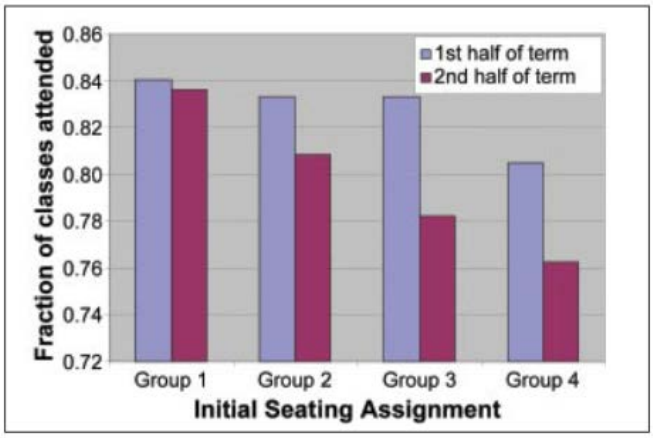
Figure 6: Students Grouped by Initial Seating Assignment

Seating group	Initial seat distance from front of class	# of students	Average GPA (not including this course)
Group 1	0–4 meters	48	2.86 ± 0.11
Group 2	4–6.5 meters	48	2.95 ± 0.09
Group 3	6.5–9 meters	48	2.90 ± 0.08
Group 4	9–12 meters	57	2.89 ± 0.09

As seen Figure 6, each group contained students of essentially the same average GPA (not including the course in the study), suggesting similar student populations.

Despite substantial effort made to engage *all* students (the use of interactive response devices, large fonts and figures on lecture slides and a video camera to project smaller demonstrations), initial seat location was found to have an important impact on outcome: both higher attendance and higher final grades are seen in group 1 (students who initially sat 0-4 meters from the front).

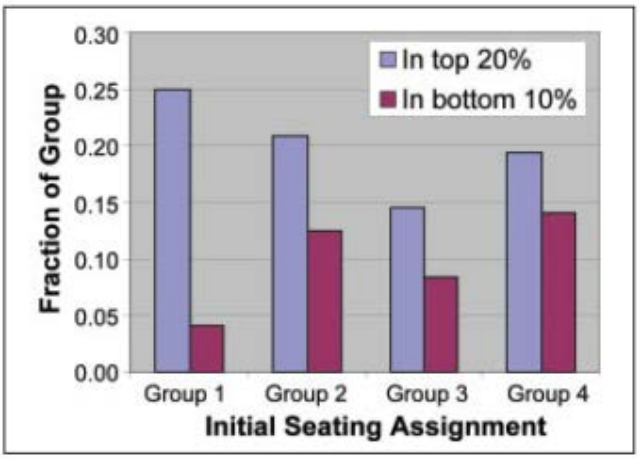
Figure 7: Initial Seating Location and Attendance



Two key trends can be seen in Figure 7: “The further the original seating location is from the front of the classroom, then 1) the lower the average attendance and 2) the larger the drop-off in attendance between the first and second half of the term” (Perkins & Wieman, 2005). In fact, the average attendance of group 1 declined by only 1%, despite being moved to the back of the class halfway through the term.

Figure 8: Initial Seating Location and Course Performance

(Excluding attendance)



Additionally, a significant effect on the total points possible in the course (essentially, the final grade) due to the seat location can be seen (see Figure 8). With 27% of group 1 receiving A's as compared to 18% of group 4, the fraction of A's decreases as original seat location is further from the front. Similarly, 2% of group 1 receives F's compared to 12% of group 4, indicating an increase in F's as original seat location is further from the front.

Furthermore, the effects are different for high and low performers. The performance of students in the top 20% is less correlated to their initial seating assignment whereas the performance of students in the bottom 10% reap positive benefits from initially sitting closer to the front (Perkins & Wieman, 2005).

From this study, it appears that the location of the seat that a student sits in at the beginning of the term indeed affects both prospective attendance throughout the semester and final grade, clearly in favor of those with an initial seat location close to the front of the room, with a more dramatic impact on underperforming students (bottom 10%). These results do not seem to have been tested again in the PER literature.

c) The Gender Gap

In addition to the structure of the course and seat location, gender is an important factor that has been shown to influence student performance. We will first summarize studies that provide evidence for the existence and persistence of the gender gap and then provide an overview of a pedagogical intervention that aims to reduce or eliminate it.

Lauren Kost, Steven Pollock and Noah Finkelstein comment on the presence of the gender gap at the University of Colorado: "…females make up only 25% of the students who enroll in introductory physics and about 15% of the physics majors. Not only is there a gender gap in participation, but there is also a gender gap in performance. Previous studies at CU [University of Colorado], and elsewhere, have

identified difference in males' and females' performances on surveys of conceptual physics. This under-representation and underperformance of females in physics is cause for concern and has led to a variety of studies on the source of the gender gaps in college physics" (Kost, Pollok, & Finkelstein, Characterizing the gender gap in introductory physics , 2009). As reported, there is clear evidence of a gender gap in participation and student performance at the University of Colorado that requires attention.

Additionally, Eric Brewe et al. explore success rates at the diverse Florida International University, sorting students first by their gender independent of ethnicity, and then by ethnicity independent of gender. We will focus on the findings with respect to gender. Comparing FCI scores, it is clear that there exists a gender gap at this institution as well: females (N=115) have an average gain score of $g=25.6 \pm 1.5$ whereas males (N=143) have an average gain score of $g=34.2 \pm 1.4$, a difference in gain of 8.6 in favor of males (Brewe, Sawtelle, Kramer, O'Brien, Rodriguez, & Pamela, 2010).

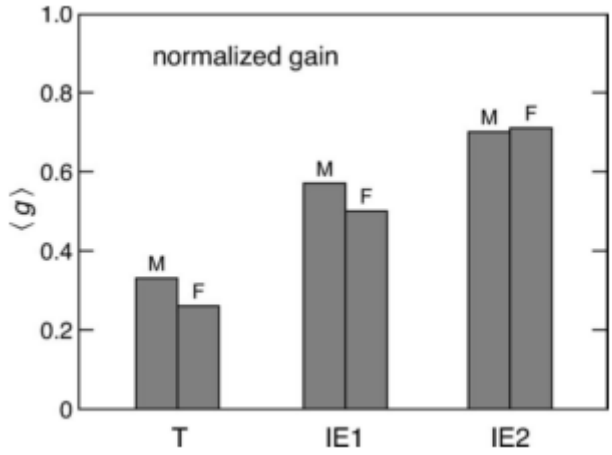
In their conclusion, Brewe et al. hint at an important point: the gender gap may be dependent on the differences in background preparation. However, it is stated, "Gender gaps are not accounted for by precollege preparation alone" (Brewe, Sawtelle, Kramer, O'Brien, Rodriguez, & Pamela, 2010). The question remains: what causes the gender gap?

d) Influence of Interactive Engagement on the Gender Gap

With the gender gap apparent, we now summarize several studies about the influence of interactive engagement on the gender gap. Some studies claim that the use of interactive engagement techniques can severely decrease if not eliminate the gender gap, though subsequent studies have failed to reproduce these results.

Lorenzo, Crouch and Mazur compare three different teaching structures using FCI gain scores at Harvard University. First, the traditionally taught course (T) corresponds to the conventional lecture format. Second, interactive engagement 1 (IE1) is partially interactive; a mix of traditional lectures and engaging discussion. Finally, interactive engagement 2 (IE2) is full interactivity; comparable to the previously discussed interactive engagement pedagogy. With these varying course structures, it is the hope that trends exist between the level of engagement and an increased student understanding for both genders.

Figure 9: Instructional Approach and Normalized Gain, by Gender (Harvard)



As shown in Figure 9, males (M) in the traditionally taught course (T) have higher gain scores than females (F). In contrast, in the highly interactive course (IE2), both females and males have high gain scores. This is consistent with previous studies that demonstrate that high average gain scores occur in interactively taught courses as opposed to traditionally taught courses. Additionally, it becomes clear that females and males have increasingly similar gain scores with the increased use of interactive engagement techniques; a narrowing of the gender gap (Lorenzo, Crouch, & Mazur, 2006).

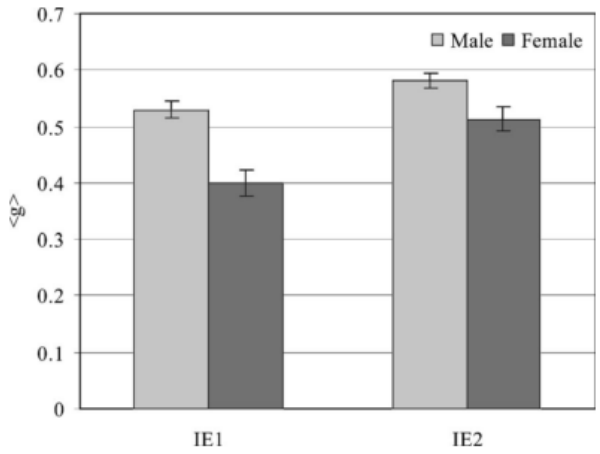
Quantifying this gender gap (the difference between the average gain score of males and the average gain score of females [$\langle S^M \rangle - \langle S^F \rangle$]), the decrease becomes clear: the post-test gender gap of 10% for the traditionally taught course (T) is nearly comparable to the pre-test gender gaps in the interactive course (IE1 and IE2). The post-test gender gaps for IE1 and IE2 are 7.8% and 2.4%, respectively. Perhaps, most significant is that the gender gap for the fully interactive course (IE2) is no longer statistically significant, meaning that males and females are performing at approximately the same level (Lorenzo, Crouch, & Mazur, 2006). Overall, these results are consistent with the claim that the degree of interactive engagement in the classroom is negatively correlated with the presence of a gender gap.

Unfortunately, follow up studies on the effects of interactive engagement have not reproduced these results. Pollock, Finkelstein, and Kost conducted a study comparable to the study above at the University of Colorado at Boulder. However, this study differs from the previous in a few ways: while middle to high-gain scores are seen at both locations, students at the University of Colorado at Boulder start at approximately 30% as opposed to 65% at Harvard and end with scores of approximately 65% as opposed to 85% at Harvard. Additionally, instead of using the FCI, Pollock, Finkelstein, and Kost use the FMCE as the primary assessment instrument. Finally, perhaps the most notable differences the class size is roughly three times larger than Harvard's with a gender ratio (M:F) nearly double (Pollock, Finkelstein, & Kost, 2007).

The results from this study are in firm contrast to the previous results: "While there are individual cases where the gender gap is reduced, on average we see no statistically significant reduction in the gender gap in the IE2 and IE1 classes" (Pollock, Finkelstein, & Kost, 2007). There is an increase in gain scores

with higher levels of interactive engagement in first-semester physics, yet there is still a statistically significant ($p<0.01$) gender gap in both IE1 and IE2 (see Figure 10; Pollock, Finkelstein, & Kost, 2007).

Figure 10: Instructional Approach and Normalized Gain, by Gender (Colorado)



In the case of second-semester physics courses taught with IE2 instruction, a smaller gender gap on the pre-test scores is seen with an overall statistically significantly higher gender gap on the post-test; males make more learning gains than females (Pollock, Finkelstein, & Kost, 2007). These results indicate that the results seen in Lorenzo, Crouch & Mazur's study at Harvard University are not reproduced at the larger University of Colorado at Boulder.

We speculate that these contrasting results are due to the higher male to female ratios at the University of Colorado at Boulder as compared to Harvard University. We hypothesize that the physics stereotype threat, females fearing that they are not expected to do well in physics, is more severe in a male-dominated courses, and will have a greater, negative effect on student performance.

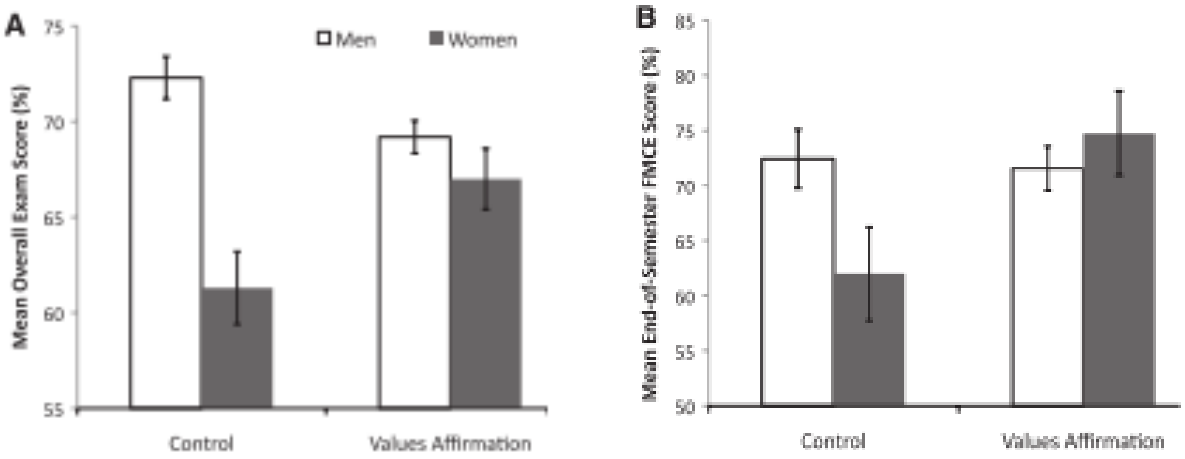
An additional follow-up study conducted by Kost, Pollock, and Finkelstein further explores the source of this persistent gender gap. Modeling the post-test scores using multiple regression, several variables, including high school GPA, combined SAT Math and ACT Math scores, years of physics and/or math preparation in high school, etc., are explored as potential contributing factors.

Using multiple and logistic regression analyses to estimate the impact of student background factors on gain scores, the following are found to be the only parameters to have notable effects on the gender gap: prior physics and math understanding, prior attitudes and beliefs, the semester a student was enrolled, and student gender. The stereotype threat is one example of such an incoming attitude. These results lead to the general conclusion, "We find that the gender gap in conceptual post-test scores is substantially accounted for once prior physics and mathematics understanding and incoming attitudes and beliefs are taken into account" (Kost, Pollock, & Finkelstein, The Persistence of the Gender Gap in Introductory Physics, 2008).

Finally, the most recent article from the Colorado group investigates values affirmation as a means of intervention. At the University of Colorado, up to 60% of the gender gap can be accounted for by previous math and physics preparation (Miyake, Kost-Smith, Finkelstein, Pollock, Cohen, & Ito, 2010). It is the hope that introducing a simple writing exercise about students' values (unrelated to the subject matter of the course), an intervention technique that is widely used in social science disciplines, would further decrease the disparity. Students were divided into two groups: the values affirmation group selected their most important values from a list (relationship with friend, learning, family, etc.) and wrote about why these values are important to them. The control group selected their least important values from a list and wrote why these values might be important to other people. It was predicted that this reflection of self-defining values would reduce the psychological threat of being devalued based on group identity, in this case a female identity. These writing prompts took 10-15 minutes and were administered at several points throughout the semester.

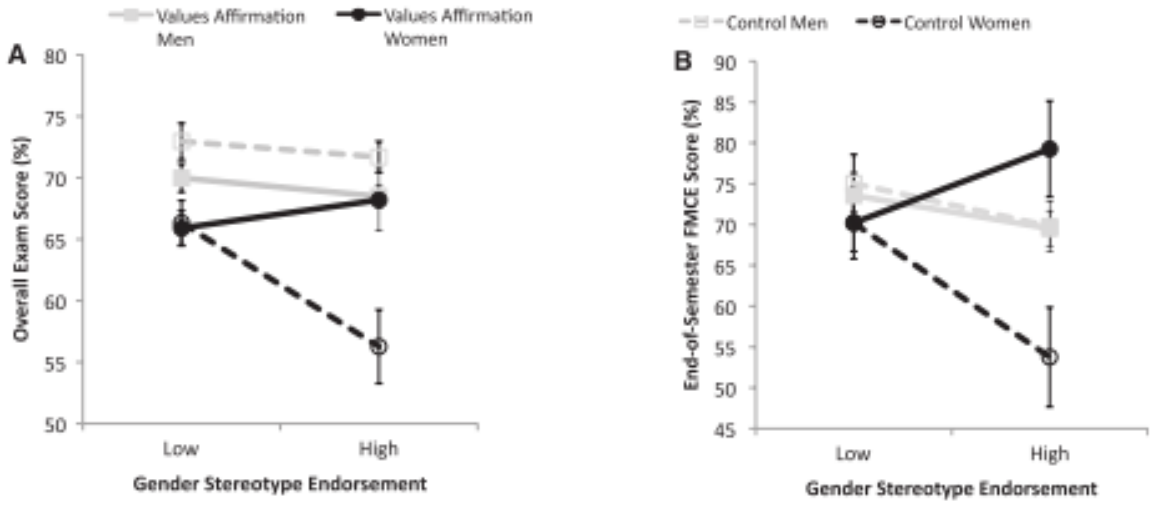
The results are surprising: the gender gap is significantly smaller in both course grades (based 75% on exam scores) and FMCE post-test scores (see Figure 11). Unexpectedly, men in the values affirmation group experienced a decrease in exam scores. However, this is not replicated in post-test FMCE results (Miyake, Kost-Smith, Finkelstein, Pollock, Cohen, & Ito, 2010).

Figure 11: The Effects of Values Affirmation on Mean Overall Exam Score and Mean Post-test FMCE Score



The values affirmation exercise is especially beneficial for women who strongly endorse the stereotype that females are not meant to do physics (see Figure 12; Miyake, Kost-Smith, Finkelstein, Pollock, Cohen, & Ito, 2010):

Figure 12: The Effects of Values Affirmation by Gender Stereotype Endorsement



Overall, this study of values affirmation writing suggests that the psychological threat of identifying with the gender stereotype is a significant contributing factor to the gender gap. These writing prompts provide an effective means of intervention as, "…the affirmation closed the 'residual' gender gap on in-class exam scores by approximately 61% and entirely eliminated the gap on the FCME" (Miyake, Kost-Smith, Finkelstein, Pollock, Cohen, & Ito, 2010).

We are anxious to read follow up studies that present additional findings to confirm or contrast the positive effects seen here. While we expect that there will not be a complete elimination of the gender gap at all institutions, this intervention technique seems to be a simple yet promising way to reduce the magnitude of the gender gap.

e) Cumulative GPA and Grade Predictive Schemes

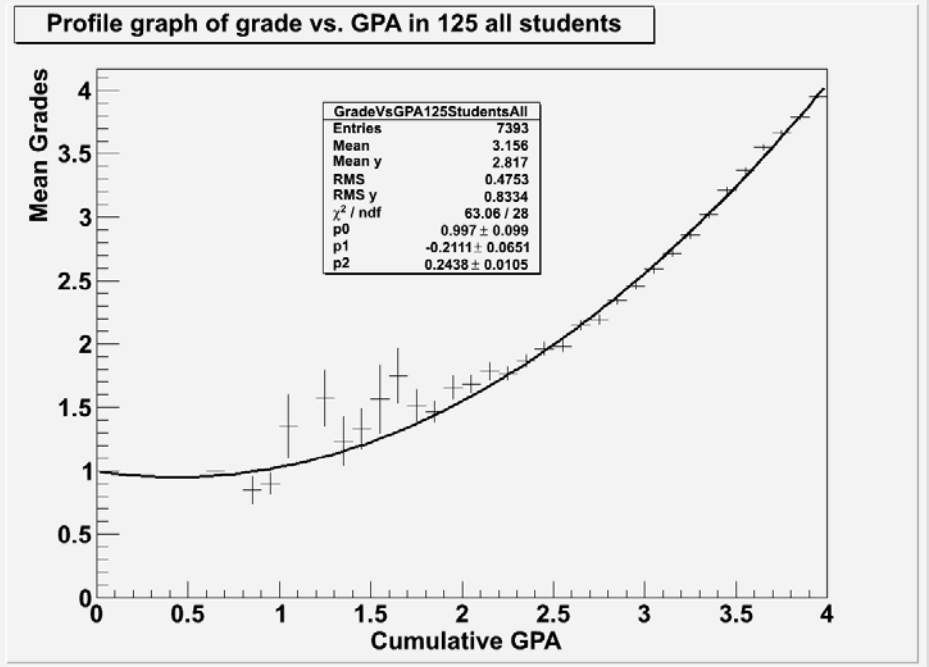
As an alternative, a student's incoming cumulative GPA has been found to be a predictive factor for student success. Scott Freeman et al. at the University of Washington, Seattle, developed prediction schemes for grades in introductory biology courses based on incoming GPA and SAT scores (Freeman, et al., 2007). Motivated by this previous research, University of Michigan undergraduate Laurie Lai, working with Professors McKay, Gerdes, and Evrard, explored correlating factors and ultimately developed grade prediction schemes for introductory physics courses at the University of Michigan. Like Freeman et al., Lai shows that, in addition to gender, a student's cumulative GPA at the beginning of the class greatly predicts their performance. In fact, a student's cumulative GPA is the most correlated parameter to the final grade received in an introductory physics course. Because of this, cumulative GPA can be used to develop grade prediction schemes for the University of Michigan which accurately predict what the student will receive in a course.

Examining 31 terms of data (Winter 1996 through Winter 2008 from the data set we will use for our research), Lai finds that "a student's physics grade tends to be lower than their cumulative GPA" (Lai,

2009). Furthermore, a gender gap can also be seen in introductory physics courses: "…the average grades of females are consistently lower than the average grades of males" (Lai, 2009). Based on this high correlation between cumulative GPA and student performance, Lai developed grade prediction schemes by course and by gender. Plotting course grades vs. incoming cumulative GPA and fitting a quadratic (see Figure 13), equations that predict course grades were developed for each introductory physics courses at the University of Michigan.

Figure 13: Grade Prediction Model for Physics 125

p_0, p_1 and p_2 are variables in the best fit equation of the form $f(x) = p_0 + p_1*x + p_2*x^2$.



These predictions are taken to form a basis for further comparisons as they are highly accurate and largely independent of instructor. We will use these schemes as a point of reference in our analysis. Just as concept inventories use gain scores to account for differences in initial levels of understanding, grade prediction schemes allow us to determine if a student does better than, remains the same as, or does worse *than expected* as we vary a parameter.

f) Additional Findings and Proposed Improvements from Lai

Using these predictive models, Lai noted several additional findings: First, instructor gender does not have a significant effect on student performance. Second, student performance is dependent on the student's year in college; each course favors different years. For example, Physics 126 favors juniors whereas Physics 140 favors sophomores and freshman; students in both of these courses benefit from taking physics earlier in their undergraduate careers. Next, these predictive models are positively correlated with SAT math scores, ACT math scores, and ACT science scores (and therefore also with the ratio of SAT math scores to SAT verbal scores, total SAT scores, and total ACT scores). Finally,

considering students who return to take more semesters of physics, there is a weak correlation between a student's grade in their previous introductory course and their predicted grade. Furthermore, there does not appear to be a significant difference between students who take physics courses successively versus those that take a break (Lai, 2009). From this variety of findings, Lai prompts her reader to take into account additional factors, including year in school, ACT and SAT performance, and previous performance in physics courses, in an attempt to further optimize the grade prediction schemes.

IV. Methods

Building on this understanding of relevant literature, we will begin our own analysis of the gender gap and the effectiveness of supplementary study groups at the University of Michigan. First, we summarize our data set, including a discussion of the content and structure of the courses studied as well as an overview of internal and external parameters that have been compiled. After, we describe two key analytical tools that will be used in our analysis: The Kolmogorov-Smirnov Test and Bootstrap Re-sampling. We then move to our analysis of the gender gap where we provide evidence of the gender gap in all introductory physics courses and in all terms we've studied.

Next, we begin to explore important factors, as suggested by the literature, and their effects on the gender gap. We focus on three parameters: mathematical preparation, female representation in the classroom and instructor gender. With the existent of the gender gap apparent and an initial analysis of potentially correlating factors conducted, we shift our focus to the effects of supplementary study groups on student performance. Starting with the influence of study groups on all students' course grades, we subsequently explore the effects on only males and only females. After summarizing our findings, we provide recommendations, based on PER literature and supported by our results, of ways to decrease the gender gap in introductory physics courses at the University of Michigan.

A. Data

To investigate student performance at the University of Michigan, a data set containing both external and internal parameters is compiled. The external data, referring to student background information, is compiled from the University of Michigan's Registrar's Office. On the other hand, the internal data, referring to the coursework within a given introductory physics class, is collected from the University of Michigan Physics Department and the Science Learning Center. Together, this data provides a broad profile for each student who has taken an introductory physics course at the University of Michigan since Winter 1996. With a total of 39 terms of data collected (Winter 1996 to Fall 2010), we assign a term code to each (see Appendix A).

The University of Michigan offers four main introductory courses: Physics 125, Physics 126, Physics 140 and Physics 240. Physics 125 and 126 are algebra-based Mechanics and Electricity and Magnetism, whereas Physics 140 and 240 are calculus-based Mechanics and Electricity and Magnetism. Generally, non-physics majors enroll in the 125/126 sequence while physics majors and engineers enroll in the 140/240 sequence.

Two new introductory physics courses, Physics 135 and Physics 235, were introduced in Fall 2006 and Fall 2007, respectively. These courses, entitled 'Physics for the Life Sciences', present Mechanics and Electricity and Magnetism material from a biological perspective. The 135/235 sequence was developed for students planning to concentrate in any of the life sciences as well as students planning to pursue a career in medicine, kinesiology, or the health sciences. In Fall 2010 students fitting this description were advised to take Physics 135 instead of Physics 125. The enrollment numbers in our data reflect this shift.

The structure of introductory physics courses until Fall 2010 consisted of bi-weekly, one hour, large lectures with bi-weekly, one hour discussions led by a faculty member other than the lecturing professor. Electronic response units were used in lecture and determined the lecture grade. The discussion grade was based on participation and awarded by the discussion instructor. Homework was completed online using various commercial homework systems. Finally, three multiple choice exams and a multiple choice final exam comprised students' exam grades. The exam scores were heavily weighted (60-80%) in the calculation of final grades.

In Fall 2010, a pedagogical shift to a larger discussion section occurred. The format of introductory courses is presently a bi-weekly, one hour lectures and a bi-weekly, one hour discussions. Discussions are held in the large lecture hall and led by the lecturer. The instructor uses interactive engagement techniques to create a discussion atmosphere. The same electronic response units, online homework, and exam formats are used before and after this shift.

An all inclusive data set, entitled 'SD' for 'student data', is compiled. There are a total of 48,579 students in this structure. The available parameters come from both the external, student background data (see Appendix B) as well as the internal, coursework data (Appendix C).

B. Analysis Tools

We will now provide a brief overview of fundamental analytical techniques. This section will serve as a reference as we utilize these tools in our analysis.

1. Kolmogorov-Smirnov Test

To quantitatively compare two distributions, we use the Kolmogorov-Smirnov Test (K-S Test). This compares the shape and consistency of two distributions by examining how quickly each reaches its whole. From the K-S Test, we obtain two variables: 'PROB' and 'D.' PROB gives the significance level of the K-S statistic. Small values of PROB indicate a small random chance that the cumulative distributions are drawn from the same underlying distributions. D is defined to be the maximum deviation between the two cumulative distributions (Kolmogorov-Smirnov Goodness-of-Fit Test, 2010). We will use the K-S Test as we compare the similarity of two distributions. For example, we will use it to determine the probability that grade distributions observed for males and females are drawn from the same underlying distribution.

2. Bootstrap Re-sampling

Bootstrap re-sampling is a robust way to increase our confidence of measurement uncertainty. Once we determine the mean and uncertainty spread on a set of data, we can re-run the experiment multiple times using a randomized subset of our data sample. Each re-sampling yields a different mean (which fluctuates somewhat more due to the smaller sample sizes) that takes into account the non-Gaussian imperfections which arise in the data. By then normalizing the errors of these re-sampled means to represent the entire data set, we can effectively determine the error on our original mean. This gives us an accurate idea of how much error exists on a mean considering the data set used (Efron & Tibshirani, 1993). Importantly, bootstrap re-sampling is insensitive to assumptions about the distribution. We will use bootstrap re-sampling several times in our analysis. For example, it will allow us to robustly determine the significance of a difference in mean grades between two groups of students.

V. Analysis

A. Gender

With the gender gap in Physics nationally recognized, we want to explore its presence at the University of Michigan. How do females perform in introductory physics courses as compared to males? Based on the literature, we expect males to outperform females. Considering 21 terms of data (not including spring terms in which the number of students is too low for accurate statistics), we find that this is indeed the case. We would like to quantify this difference in order to explicate the gender gap. In what follows, we will examine the gender gap in several different ways.

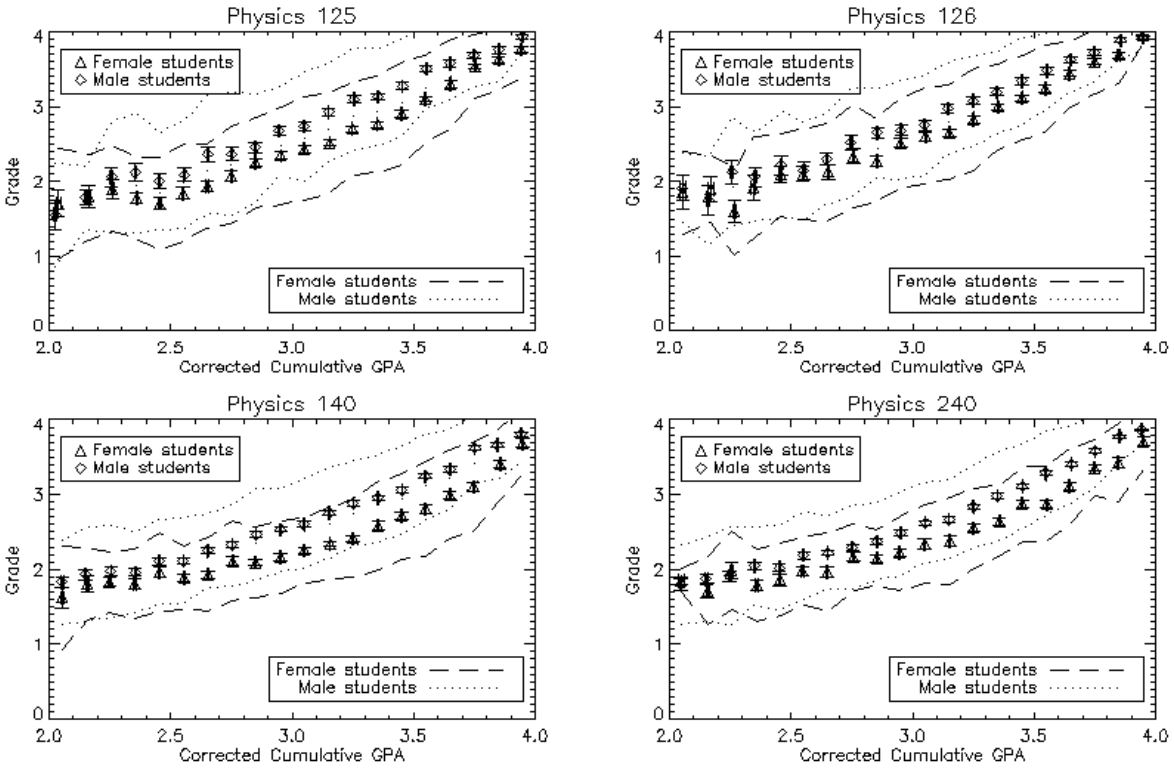
Furthermore, we investigate potential correlating factors which may contribute to the gender gap in these 21 terms. Perhaps it is a disparity in mathematical preparation which leads to this significant gender gap in introductory physics courses. Or perhaps there are socio-psychological factors like stereotype threat affecting performance which have yet to be completely understood.

1. The Presence of the Gender Gap

To study the gender gap in performance, we plot cumulative GPA (at the beginning of the course) vs. final grades for males and females (see Figure 14). Cumulative GPA is chosen as the independent variable because of its predictive value (Lai, 2009; Freeman, et al., 2007). By binning students by prior cumulative GPA, we expect males and females to receive the same grade if there is no gender gap.

Figure 14: Gender Gap in Final Grades

The error bars represent the error on the means (calculated using Bootstrap Re-Sampling) and the dotted lines represent the one-sigma spread in the data.



To the contrary, it is apparent that the mean grades for females are significantly lower than the mean grades for males; males receive about 0.25 of a letter grade higher than females, and this is true for virtually all values of incoming culumative GPA. However, it is important to acknowledge that the grade distributions of male and female students do overlap; there are many cases in which female students outperform male students.

We also recognize that the gender gap is weaker at the extremes of prior GPA: below ~2.25 and above ~3.75. Students who receive a very high or very low grade do so independent of their gender. For the remaining range of GPA (between ~2.25 and ~3.75), it is apparent that gender is a major factor in determining average course grades.

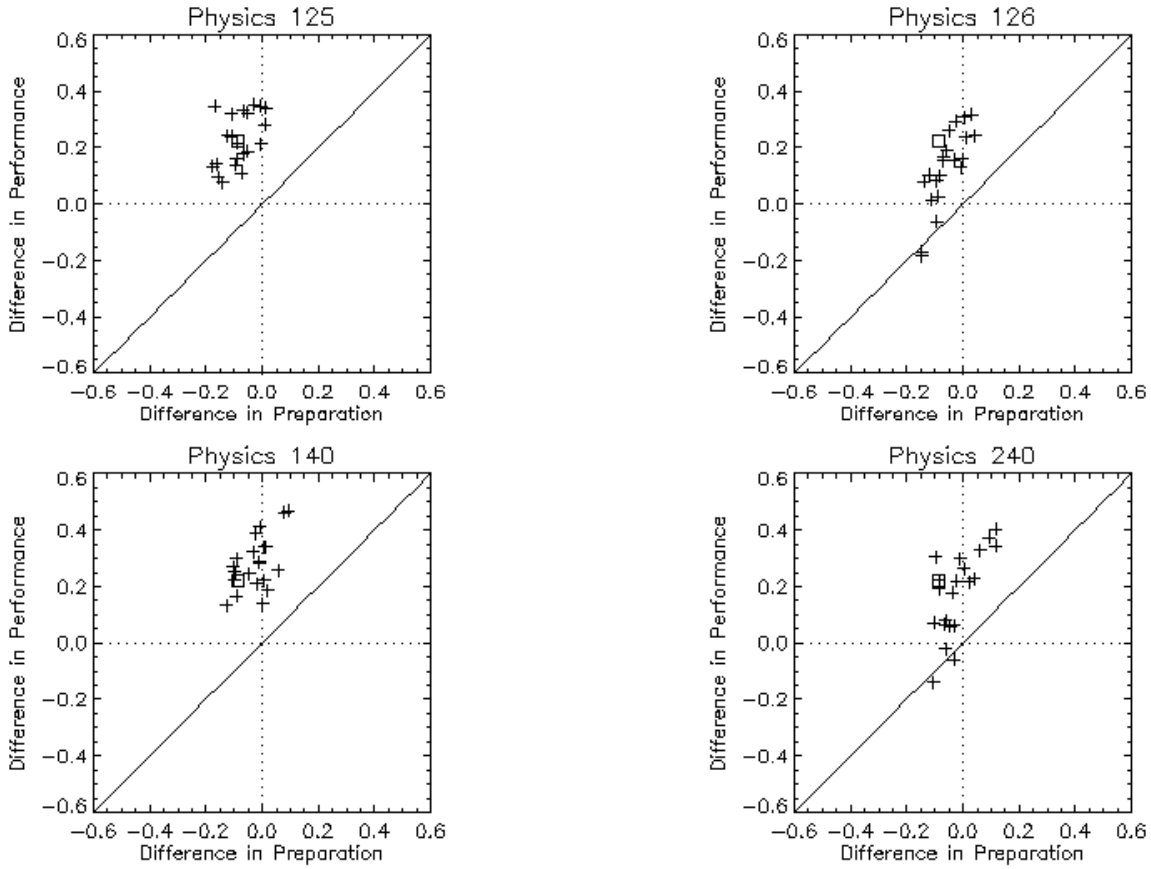
It is important to note that there is a gender gap in all four of the introductory courses, with a more noticeable difference in Physics 140 and 240; females in the 140/240 sequence tend to underperform as compared to their male classmates in a more extreme way than those in the 125/126 sequence. This is an interesting trend considering that the populations of Physics 125 and 126 are female-dominated (56% female, 44% male) whereas the populations of Physics 140 and 240 are male-dominated (76% male, 24% female).

With this evidence of a gender gap in each course for almost all cumulative GPA values, we would like to further explore the relationship between preparation and performance term by term. To accomplish this, we define the difference in preparation to be the difference in incoming GPA. This is calculated by subtracting the mean NGPA (GPA at the beginning of the semester) for female students in one term from the mean NGPA for male students in the same term. Likewise, we defined the difference in performance as the difference in course grades. This is calculated by subtracting the mean grade for female students in one term from the mean grade for male students in the same term. For both of these variables, a positive difference indicates that males have a higher incoming GPA/grade than females and a negative difference indicates that females have a higher incoming GPA/grade than males.

Plotting the difference in preparation against the difference in performance by term for 21 total terms, we will be able to identify general trends between the incoming GPA and output grade, by gender. In particular, if the gender gap were to be explained by incoming GPA differences, then the plotted points would lie along the line y=x indicating a 1 to 1 correlation.

Figure 15: Incoming GPA Difference vs. Output Grade Difference by Term

The plotted line is y=x. The squares represent the overall (Difference in Preparation, Difference in Grade) from sorting independent of term.



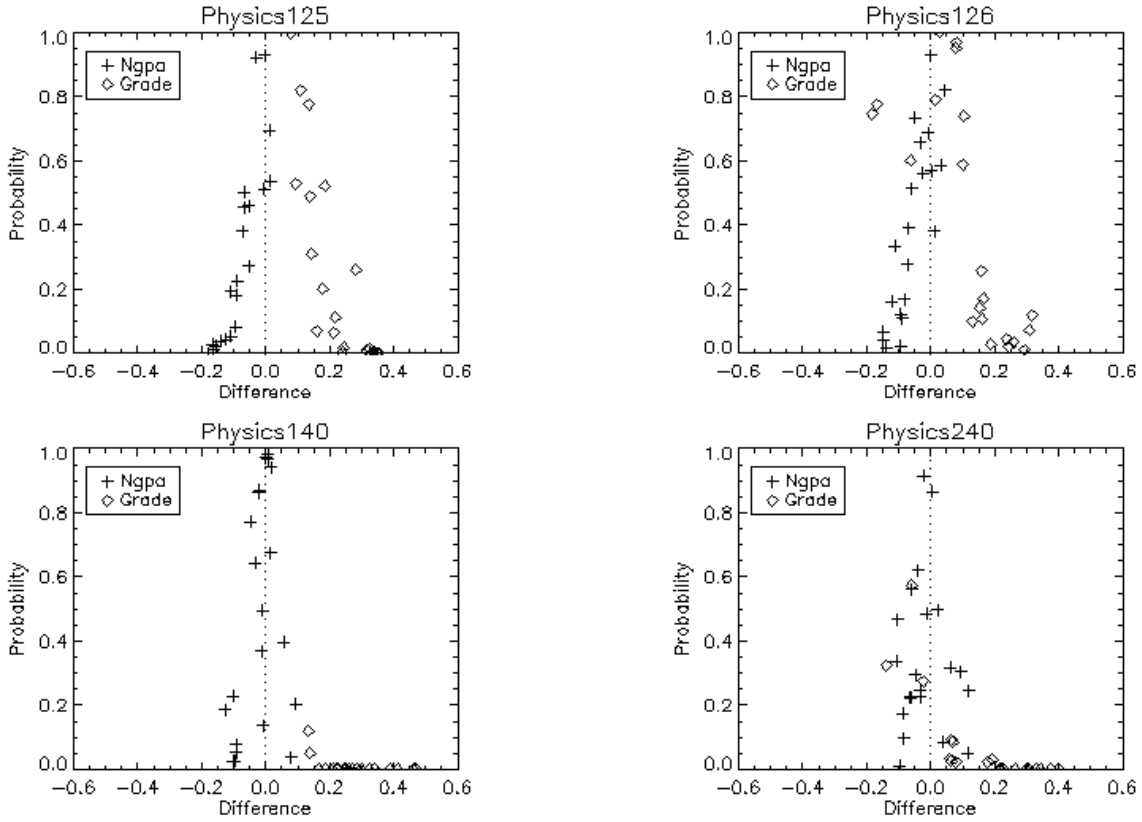
Examining Figure 15, we see that in Physics 125 and 126, the difference in preparation is slightly negative; this is consistent with females having a slightly higher cumulative GPA than males. The difference in performance however is mostly positive, indicating that males receive higher grades than females.

Similarly, in Physics 140 and 240, the difference in preparation is slightly negative (slightly in favor of females), but generally centered on zero, indicating that males' and females' incoming GPA are nearly equal. Once again, we see the majority of difference in performance is positive meaning that males generally receive better grades despite the near equity of incoming GPA. There are a few exceptions in Physics 240 where we observe a negative difference in performance. These are terms in which females received better average grades than males.

Overall, we see that in each class the majority of the plotted points lay above the line y=x confirming that the gender gap cannot be explained by differences in preparation alone. In particular, each term-independent difference (see squares in Figure 15) is at a negative difference in preparation and a positive difference in performance. This is consistent with our previous analyses: although females enter with equal if not slightly higher cumulative GPA, males receive better course grades.

Another way to quantify this gender gap is to compare the incoming GPA and grade distributions for males and females by term. For this, we use the K-S Test to calculate the probability of these gender distributions being different. Recall that small probabilities indicate that the distributions are likely to be different and large probabilities show that the distributions are consistent with one another. We plot the K-S probabilities for both preparation (incoming GPA) and performance (final grades) against the difference in preparation and the difference in performance, respectively, to compare gender distributions. Again, a positive difference indicates that males have a higher incoming GPA/grade than females and a negative difference indicates that females have a higher incoming GPA/grade than males. These plots will show us which gender has more preparation (sign of the difference in incoming GPA) and which gender has higher performance (sign of the difference in grades), as well as the likelihood that the male and female preparation distributions (examining probability values for incoming GPA) and male and female performance distributions are different populations (examining probability values for grades).

Figure 16: Gender gap vs. Gender Probability, by Term



In Figure 16, we see that while incoming GPA gender distributions vary in likelihood of being different from term to term, the grade gender distributions are very different populations (PROB~0) in the majority of terms. With a positive difference in grade distributions, we can confirm that males are receiving higher grades than females.

It is interesting to again note that this effect is weaker in Physics 125 and 126 and stronger in Physics 140 and 240. In the 125/126 sequence, we see varying incoming GPA probability as well as varying final grade probability, with clusters of low grade probability consistent with the above described trend. On the other hand, in the 140/240 sequence, almost all resulting grade probability values are near zero, indicating very different grade distributions despite similar incoming GPA distributions.

With this evidence, we see that there exists a gender gap in the University of Michigan's introductory physics courses. This gender gap persists through each course and all terms considered here. It is important to emphasize that while this gender gap does exist, our analysis also acknowledges the many cases in which a female student outperforms a male student. We merely claim that this event is less likely than the reverse.

With the gender gap apparent, we would like to turn towards an analysis of potential correlating factors. Based on the literature, we have reason to believe that mathematical preparation, female

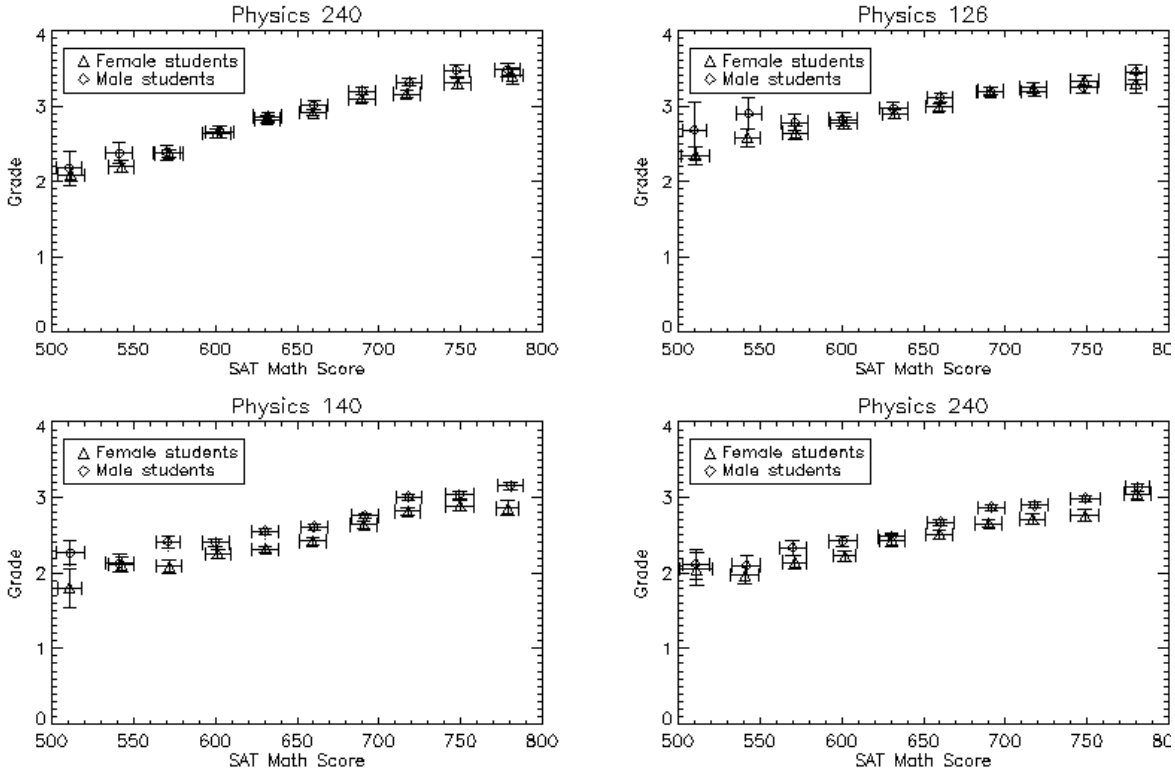
representation in the classroom, and instructor gender, among other factors, may influence the performance of males and females. We would like to explore these parameters to see if they may be contributing to the observed gender gap.

2. Mathematical Preparation

While the level of math required in the course (algebra for Physics 125 and 126 and calculus for Physics 140 and 240) may affect the magnitude of the gender gap as seen above, perhaps the real issue is differences in the level of mathematical preparation females and males bring to the course. We can analyze this preparation by taking the SAT Math score to represent prior mathematical knowledge.

Figure 17: Gender Gap in SAT Math Scores

The error bars represent the error on the means
(calculated by Bootstrap Re-sampling).

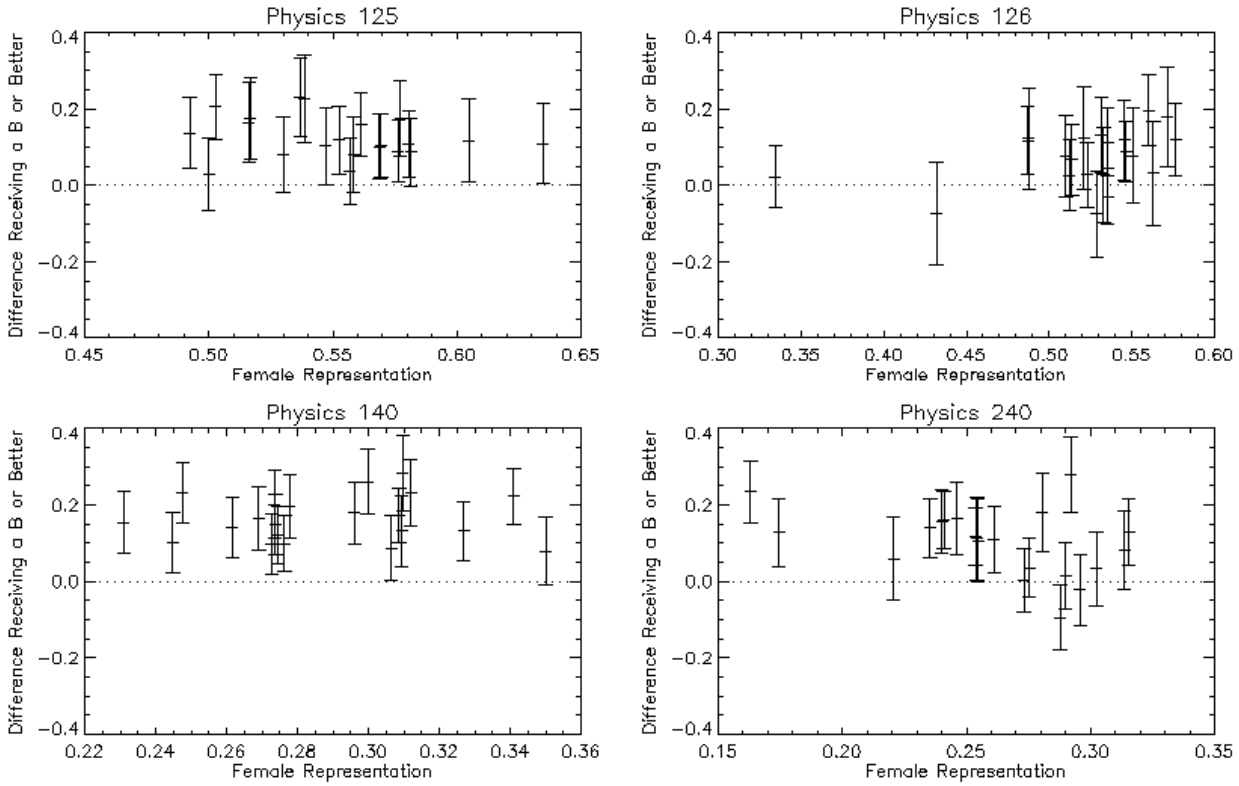


In Figure 17, we see that the gap in final grades between males and females arises at all levels of SAT Math performance for all courses. For example, in Physics 140, even at the highest levels of SAT math performance, there is still a multi-sigma difference in course grades. We note a more dramatic gender gap in Physics 140 and 240, consistent with previous results. Mathematical preparation does not completely account for the gender gap. Therefore, the question remains: what causes the gender gap?

3. Female Representation

In addition to mathematical preparation, we would also like to explore if the number of females as compared to males in the class influences final grades. To do so, we divide the students into two grade categories: those who receive a B or better and those who receive less than a C (failing the course). We take the difference in the fractions of males and females who fit in each category and plot against female representation, that is, the fraction of female students in the course (see Figure 18 and Figure 19). A positive difference indicates more males receive a B or better/C or worse whereas a negative difference indicates more females receive a B or better/C or worse.

Figure 18: Difference in Fractions Receiving B or Better vs. Female Representation



Examining Figure 18, we note that in Physics 125 and 126, 50% or more of the course's population is female. This is in contrast with Physics 140 and 240 where the female representation varies from 16% to 35%. This reiterates the reality that, in general, the 125/126 sequence is slightly female-dominated whereas the 140/240 sequence is male-dominated.

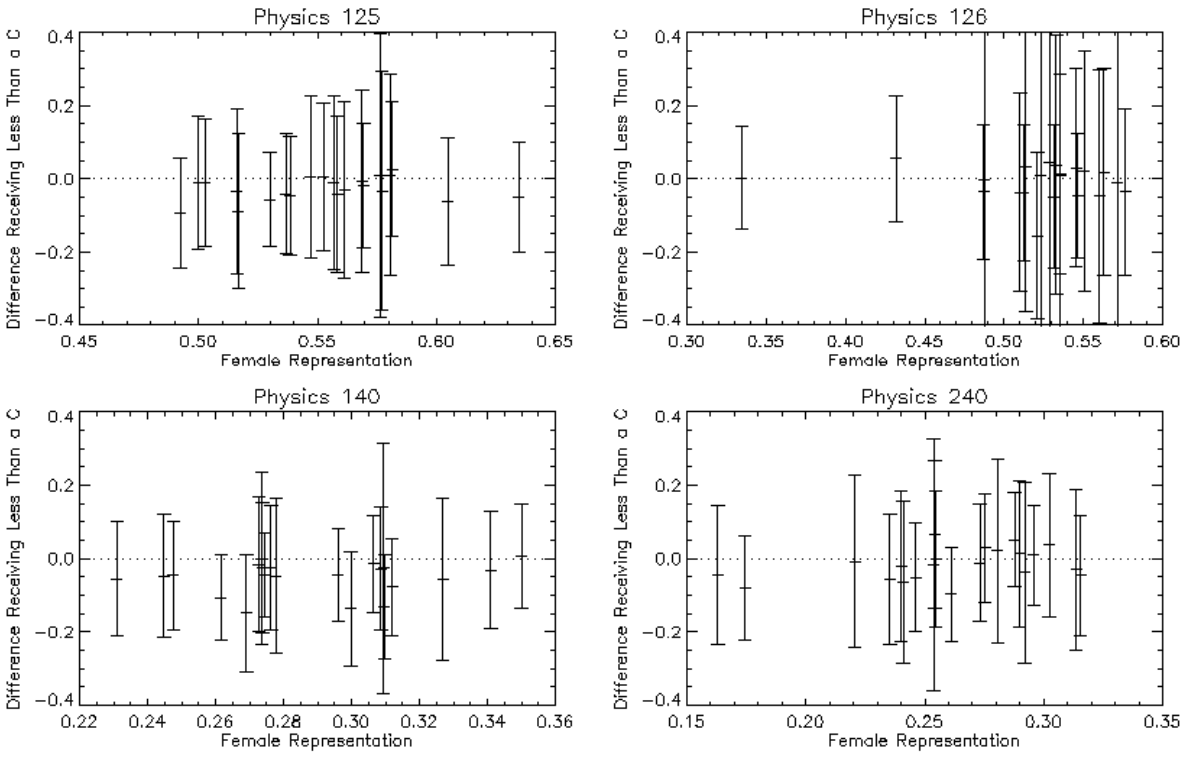
We also observe that, in almost all terms, the difference of the fractions of males and females receiving a B or better is positive. This indicates that it is more likely that males receive a B or better as compared to females. In particular, the strongest effect is seen in Physics 140, with an effect also seen in Physics 240. In Physics 125 and 126, this trend is present, but weaker.

Interestingly, these trends occur in almost every term, independent of the varying female representation when considering each course individually. This leads us to an initial conclusion that the slight variation in representation of females in each independent course does not influence which gender receives a B or better grade; males almost always receive high grades regardless of the gender ratio in that course. Yet, we must recognize that there is little variation of female representation within one course from term to term.

This initial conclusion is in tension with Figure 16. However, we can compare, for example, Physics 125 and Physics 140, two courses which differ greatly in female representation. We then see that more males receive a B or better in Physics 140, where there is a higher male representation, as compared to Physics 125, where there is a higher female representation. We must be cautious about this assertion though—there are many factors, such as content and course structure, which might influence this comparison between courses.

Similarly, we analyze the difference in the fraction of males and females receiving below a C. Once again, plotting this difference against the female representation, that is, the fraction of females in the course, we look for any trends that may arise.

Figure 19: Difference in Fractions Receiving Less Than C vs. Female Representation



Examining Figure 19, we see the same female representation values as in Figure 198, as expected. However, in these plots, we see an overall negative difference in the fractions of males and females receiving less than a C. This means that the fraction of females failing the course is larger than the

fraction of males. This is consistent with the above analysis of B or better students; the fraction of females receiving good grades is lower and the fraction of females receiving poor grades is higher: vice versa for males.

Again, we do not see the development of trends due to the representation of women when considering an individual course. In almost each term, men are receiving better grades and women receiving worse, despite small variations in female representation within that one course. This leads us to conclude that the ratio of male to female populations in each course does not dramatically affect the gender gap. However, as mentioned before, we must recognize the lack of variation in female representation from term to term when examining a course.

A tension arises with this conclusion and Figure 16 as well. Upon comparing Physics 140, with less female representation, and Physics 125, with higher female representation, we see that the effects in Physics 140 are again more extreme. We cautiously assert that this supports our hypothesis that lower female representation negatively influences females' grade performance, while also noting several additional factors which may influence our comparison of these two courses.

4. Instructor Gender

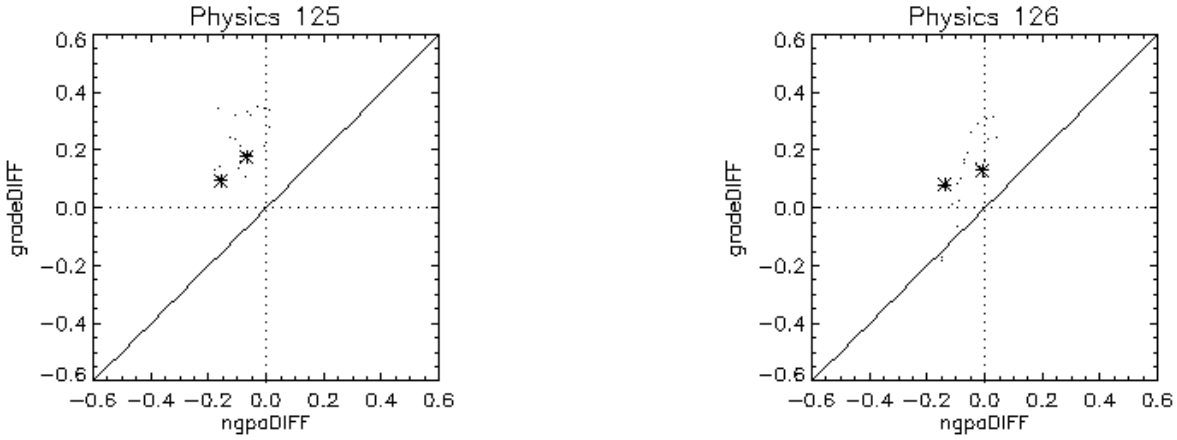
Finally, previous research has led us to believe that instructor gender may play a role in the reduction of the gender gap. In the data set we've compiled, only two female professors taught either Physics 125 or Physics 126 in just four terms of the 84 analyzed course terms:

Fall 2011 (1360): Physics 126
Fall 2005 (1560): Physics 125
Fall 2006 (1610): Physics 125
Winter 2007 (1620): Physics 126

From this alone, it is clear that female representation among professors teaching introductory physics is minimal; less than 5%. And 3 of these 4 terms were actually taught by a lecturer. Below (see Figure 20), we highlight these select terms in a previously analyzed plot (Figure 15):

Figure 20: Incoming GPA Difference vs. Output Grade Difference, by Term, for Female Instructors

** corresponds to the terms in which the instructor was female. The plotted line is y=x.*



In these four terms, there are no major differences in expected incoming GPA and course grade patterns that arise. However, we do notice that the highlighted terms are in the lower half of the grade difference distribution, meaning that the difference in grade by gender, while still in favor of males, is closer to being equal than in other terms. This leads us to the initial conclusion that instructor gender may have a slightly positive effect on the reduction of the gender gap. Unfortunately, there are no other terms in which a female instructor taught an introductory course in our data set. Therefore we are unable to further confirm this conclusion.

B. Science Learning Center Study Groups

With the presence of the gender gap apparent, we now shift our focus to the investigation of different intervention techniques which we hope will improve student performance for both genders as well as decrease the gender gap. Given that female students tend to benefit from interactive engagement techniques, the literature suggests that a potentially effective means of intervention is to utilize small study groups which incorporate such methods. The Science Learning Center (SLC) at the University of Michigan offers supplementary study groups for all introductory physics courses. These groups, which meet 2 hours a week, give students the opportunity to review concepts and work on practice problems with their classmates, under the supervision of an advanced undergraduate.

The parameter SG in our data set (see Appendix C) indicates if a student was signed up for a SLC physics study group, and, if so, for which course. It is important to note that the data available does not account for the level of participation of the students in a SLC study group; we cannot tell if a student consistently attended or even dropped out of the study group, something that may affect the correlation analysis with student performance. SLC study group data overlaps with the larger data structure for only two terms: Fall 2007 and Winter 2008.

1. The Effects of Science Learning Center Study Groups on Students

We would like to investigate the affect SLC study groups have on final course grades by comparing students who participated in a study group to students who did not participate in a study group. To do this, we split the available data into two populations: SLC### and NOSLC###.

SLC### includes students in a SLC study group in a given course ### with a known grade and a known cumulative GPA. This gives a total of 729 students: 207 students for Physics 125, 143 students for Physics 126, 255 students for Physics 140, and 124 students for Physics 240.

On the other hand, NOSLC### includes students not in a SLC study group in a given course ### with a known grade and a known cumulative GPA. This gives a total of 2553 students: 518 students for Physics 125, 465 students for Physics 126, 891 students for Physics 140, and 679 students for Physics 240.

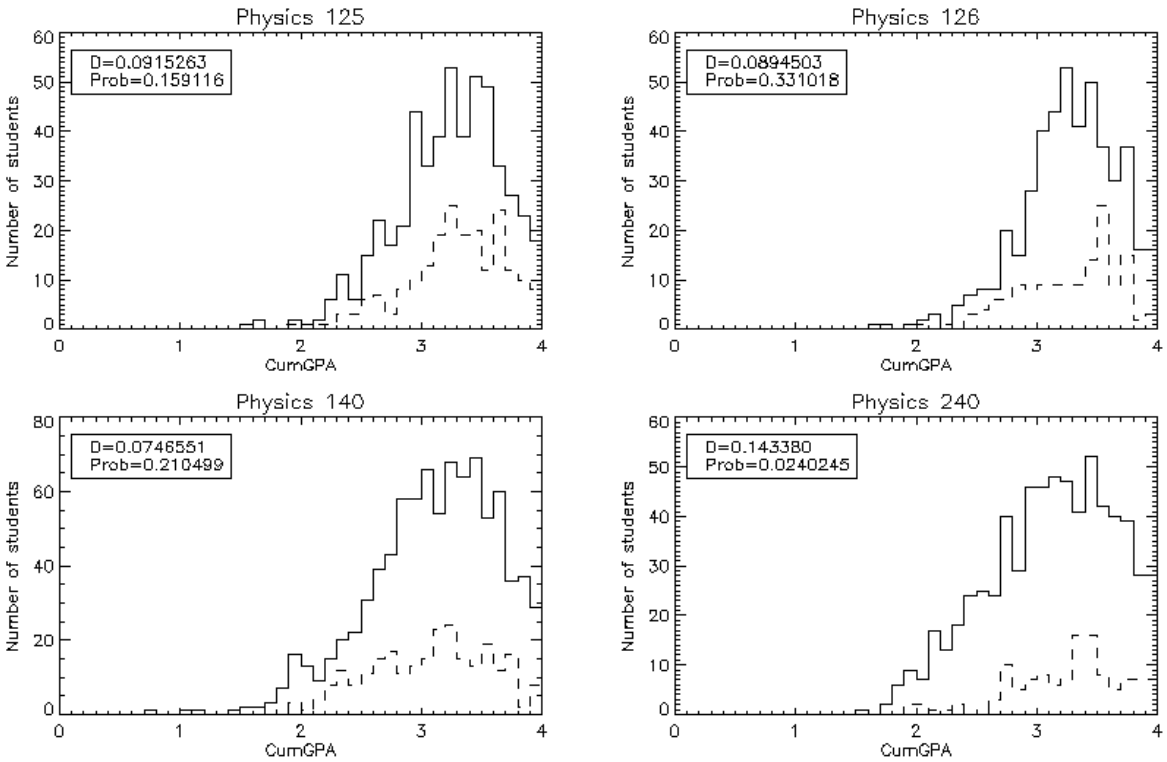
Overall, there are a total of 3,282 students considered in this study.

To begin our analysis we first examine incoming cumulative GPA distributions for both populations, by course. The purpose here is to look for obvious differences in the nature of the students in the SLC and NOSLC cohorts before proceeding to further analysis. Cumulative GPA is chosen as the variable of comparison based on previously investigated correlations between GPA and final course grades (Lai, 2009; Freeman, et al., 2007). Plotting histograms (bin=0.1), we qualitatively compare the shape of the distributions to ensure relatively similar student populations.

Upon first examination of Figure 21, there seems to be little apparent difference.

Figure 21: Cumulative GPA

NOSLC### corresponds to the solid line and SLC### corresponds to the dashed line.

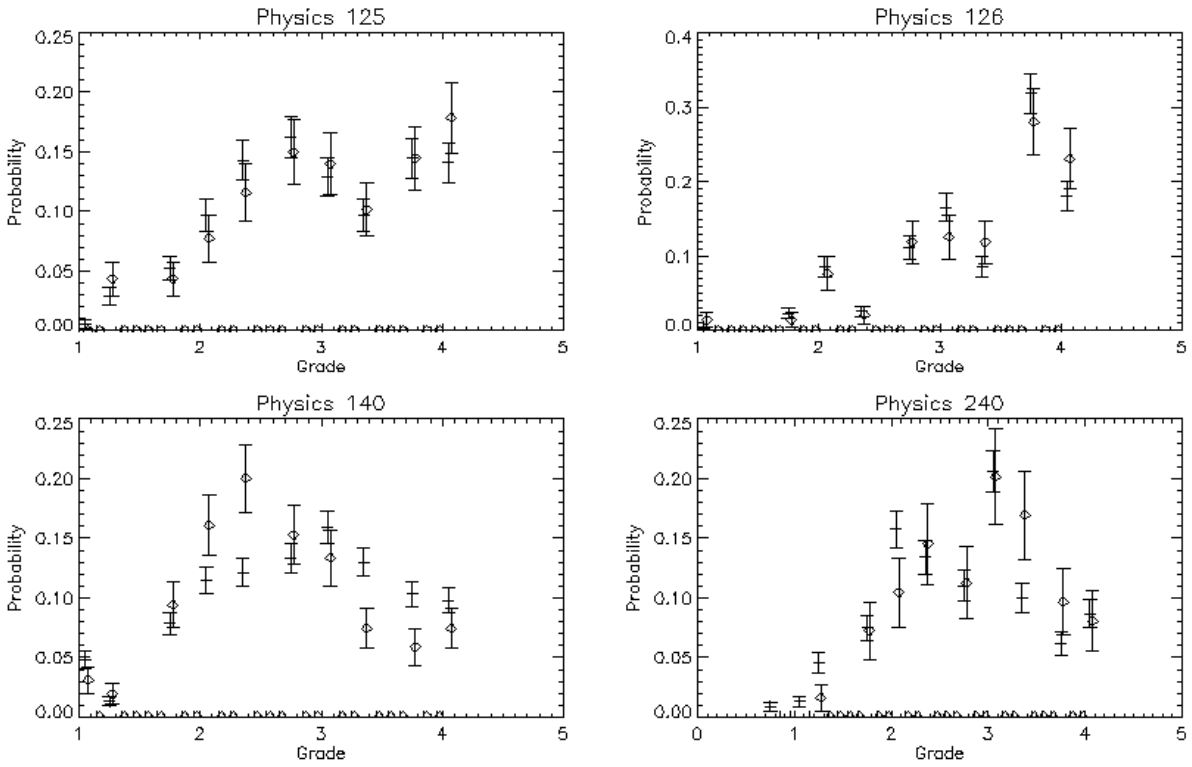


Moreover, by conducting the K-S Test, we can ensure that the distributions are in fact very similar. In particular, the two populations in Physics 240 have the most different cumulative GPA distributions, with PROB=0.024. Physics 125, 126 and 140 are much more likely to be drawn from the same incoming distribution, varying in probability between PROB=0.15 and PROB=0.33.

Second, we investigate the probability of receiving a grade for each population, by course. Plots of course grade versus the probability of receiving that grade are analyzed. This is meant to illuminate terms where students in SLC study groups and students not in SLC study groups have significant differences in the likelihood of receiving a particular grade.

Figure 22: Probability of Course Grades

NOSLC### corresponds to a + and SLC### corresponds to a diamond;
Zero probabilities are due to course grade having discrete values.

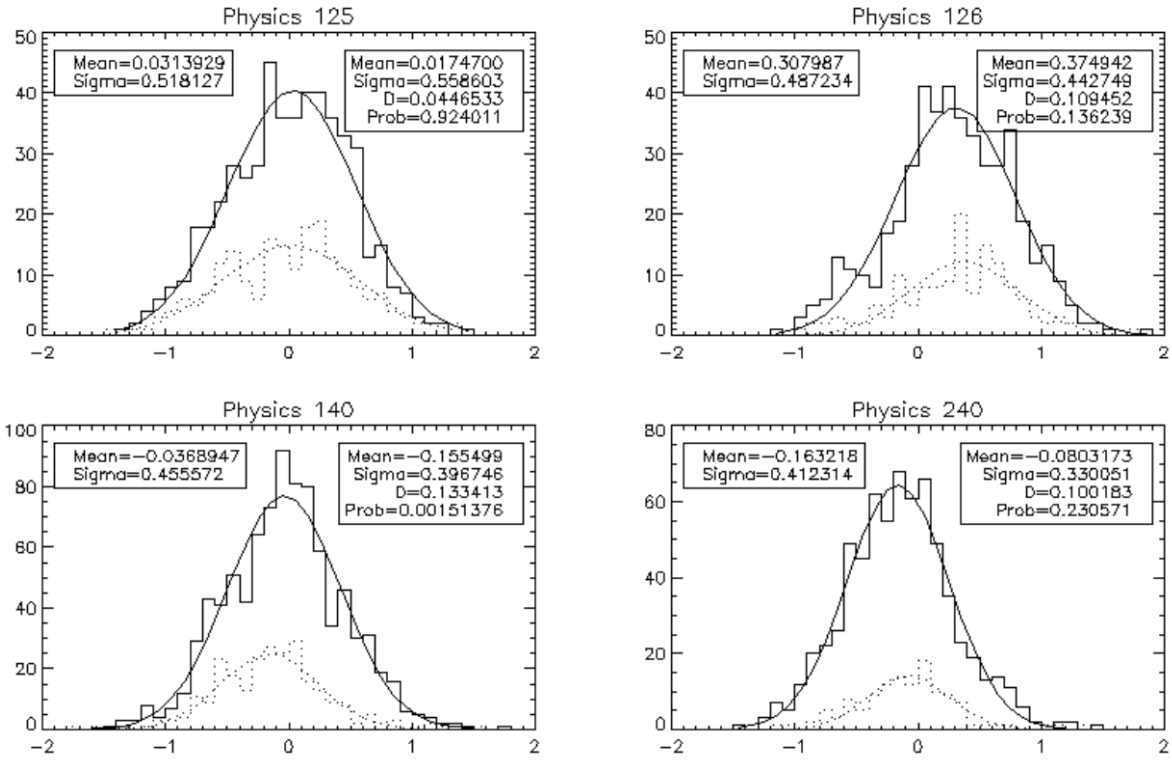


Overall, there does not appear to be a significant difference in final grades except for select grade ranges in Physics 140 and Physics 240. In Physics 140, students in a SLC study group are more likely to receive a ~2.0-2.3 (C range) whereas those not in a SLC study group are more likely to receive a ~3.2-3.75 (B to A- range). Additionally, in Physics 240, those not in a SLC study group are more likely to receive a ~2.2 (C) whereas those in a SLC study group are more likely to receive a ~3.5 (B+/A-). It is important to note that the effects in Physics 240 are smaller than those seen in Physics 140.

Finally, we utilize grade prediction schemes created for these classes as a whole, based on cumulative GPA and defined by course (Lai, 2009). Subtracting the predicted grade from the actual grade, we plot residual values for both populations. Fitting a Gaussian model to these distributions, we can quantitatively compare the mean of the functions as well as the spread. This determines if either group has higher or lower than expected grades. This step is the ultimate indicator of the influence of SLC study groups on student performance.

Figure 23: Comparison with Predicted Grades

NOSLC### corresponds to a solid plot and distribution as well as the upper left legend and SLC### corresponds to a dotted plot and distribution as well as the upper right legend. The K-S Test determines D and Prob as listed in the upper right legend.



Reviewing Figure 23, we can see that the distributions match up fairly well. The differences in means are as follows:

Figure 24: Differences in SLC Means

** indicates a significant probability (Prob<0.05) and*
*** indicates a highly significant probability (Prob<0.01).*

Course	Difference	Prob
Physics 125	0.013 difference in favor of NOSLC125	0.9240
Physics 126	0.066 difference in favor of SLC126	0.1362
Physics 140	0.12 difference in favor of NOSLC140	0.0015**
Physics 240	0.082 difference in favor of SLC240	0.2306

We see that the difference in distributions in Physics 140 is of statistical significance (Prob=0.0015). Physics 140 students who opt to take a SLC study group receive a final grade worse than expected, according to the grade predication scheme used.

Overall, according to these measures, we see little direct impact of SLC study group participation on student performance relative to what is expected. The strongest effect is in Physics 140, where students in SLC groups actually perform worse than expected. There arises an interesting pattern: in first term courses (Physics 125 and 140) students in SLC groups do worse than expected whereas in second term courses (Physics 126 and 240) they do better than expected, though these effects are not highly significant.

2. The Effects of Science Learning Center Study Groups on Males

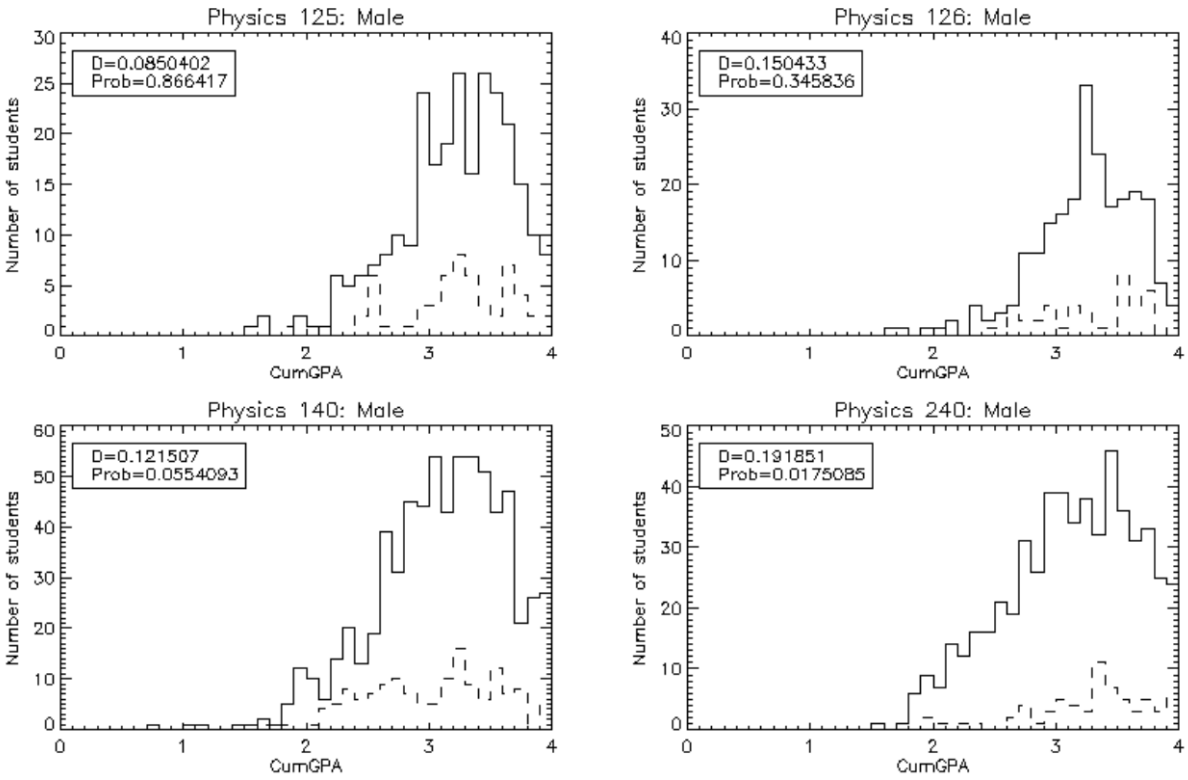
To extend this study, the above analysis is repeated dividing the population by gender: considering only males and then only females. We apply an additional gender restriction to the previously defined SLC### and NOSLC###. It is important to note that the fraction of females choosing to take a SLC study group (34%) is more than double the fraction of males who elect to take a SLC study group (15%).

To begin, we focus on an all male population. The total male sample size is 2,068 students with 315 students participating in SLC study groups (SLCMAL###) (58 students in Physics 125, 44 students in Physics 126, 143 students in Physics 140, and 70 students in Physics 240) whereas 1,753 students did not participate in SLC study groups (NOSLCMAL###) (266 students in Physics 125, 231 students in Physics 126, 695 students in Physics 140, and 561 students in Physics 240).

Upon examining Figure 25, comparing the cumulative GPA distributions of males in a SLC study group and out of a SLC study group, we can again conclude that the incoming populations are very similar for physics 125 and 126, but less so for physics 140 and 240.

Figure 25: Male Cumulative GPA

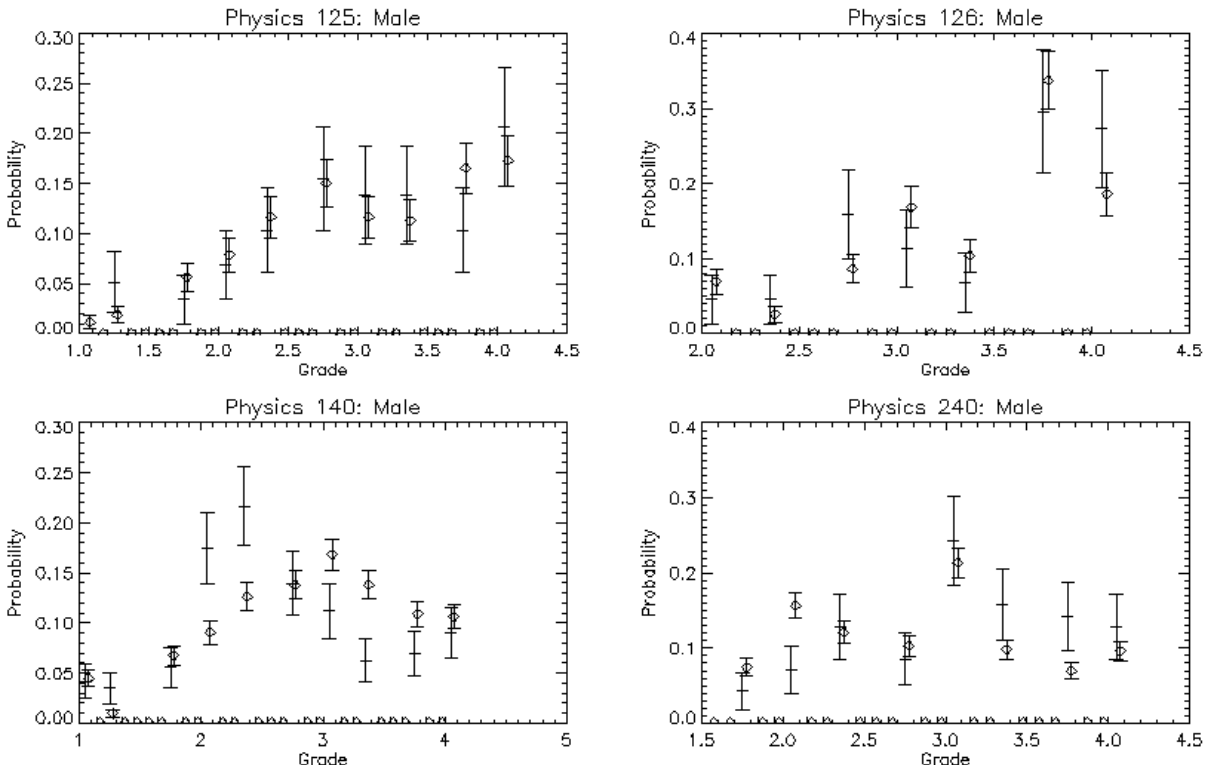
NOSLCMAL### corresponds to the solid line and SLCMAL### corresponds to the dashed line.



Next, we examine the probability of a male student in a SLC study group receiving a particular grade as compared to a male student not in a SLC study group receiving that same grade (see Figure 26):

Figure 26: Male Probability of Course Grades

NOSLCMAL### corresponds to a + and SLCMAL### corresponds to a diamond;
Zero probabilities are due to course grade having discrete values.



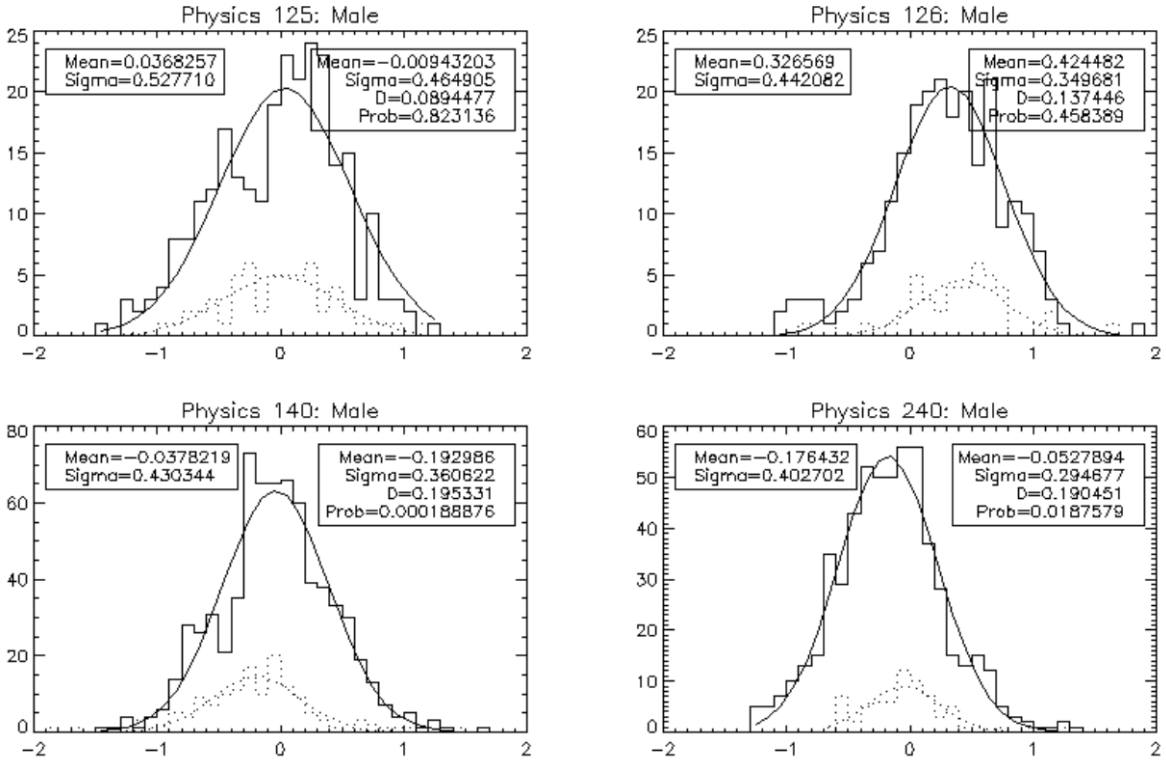
Here, all differences in grade probability for physics 125 and 126 are accounted for. However, in Physics 140, males in a SLC study group are more likely to receive a 2.0 (C) and a ~2.3 (C/C+), and males not in an SLC study group are more likely to receive a 3.0 (B), ~3.3 (B/B+) and ~3.7 (B+/A-); males in a study group do slightly worse. Additionally, in Physics 240, males in a SLC study group are more likely to receive a ~3.7 (B+/A-), and males not in a SLC study group are more likely to receive a 2.0 (C); males in a study group do slightly better.

This implies that there may be a negative effect of SLC study groups on males' grades in first semester, calculus-based physics (Physics 140) with a potential positive effect of SLC study groups on males' grades in second semester, calculus-based physics (Physics 240).

These suspected effects are further explored by comparing the grade distributions to the predicted grade schemes for males in each course (see Figure 27):

Figure 27: Male Comparison with Predicted Grades

NOSLCMAL### corresponds to a solid plot and distribution as well as the upper left legend and SLCMAL### corresponds to a dotted plot and distribution as well as the upper right legend. The K-S Test determines D and Prob as listed in the upper right legend.



The differences in means are as follows:

Figure 28: Differences in Male SLC Means

** indicates a significant probability (Prob<0.05) and*
*** indicates a highly significant probability (Prob<0.01).*

Course	Difference	Prob
Physics 125	0.463 difference in favor of NOSLCMAL125	0.8231
Physics 126	0.098 difference in favor of SLCMAL126	0.4584
Physics 140	0.155 difference in favor of NOSLCMAL140	0.0002**
Physics 240	0.124 difference in favor of SLCMAL240	0.0188*

We see that the differences in Physics 140 (Prob=0.0002) and Physics 240 (Prob=0.019) are of statistical significance. SLC study groups are ineffective at raising the mean of the distribution of course grades for males in Physics 140 but are effective for Physics 240. This is a similar trend to what is seen with the overall population analysis (not split by gender) of SLC study groups being ineffective for Physics 140 students. This is not surprising considering Physics 140 and 240 are male-dominated courses.

3. The Effects of Science Learning Center Study Groups on Females

Similarly, we restrict the sample to an all female population. The total female sample size is 1,211 students with 414 students participating in SLC study groups (SLCFEM###) (149 students in Physics 125, 99 students in Physics 126, 112 students in Physics 140, and 54 in Physics 240) whereas 797 students did not participate in SLC study groups (NOSLCFEM###) (252 students in Physics 125, 233 students in Physics 126, 194 students in Physics 140, and 118 students in physics 240).

Figure 29: Female Cumulative GPA

NOSLCFEM### corresponds to the solid line and SLCFEM### corresponds to the dashed line.

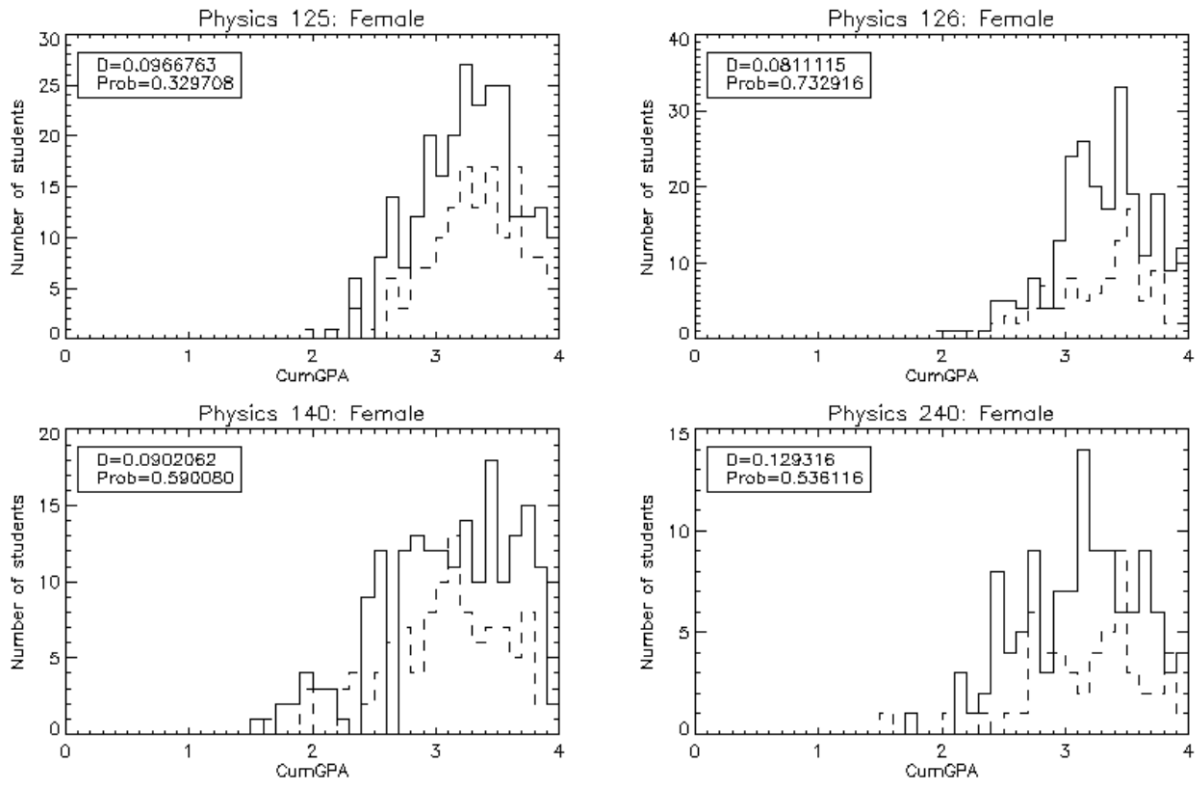
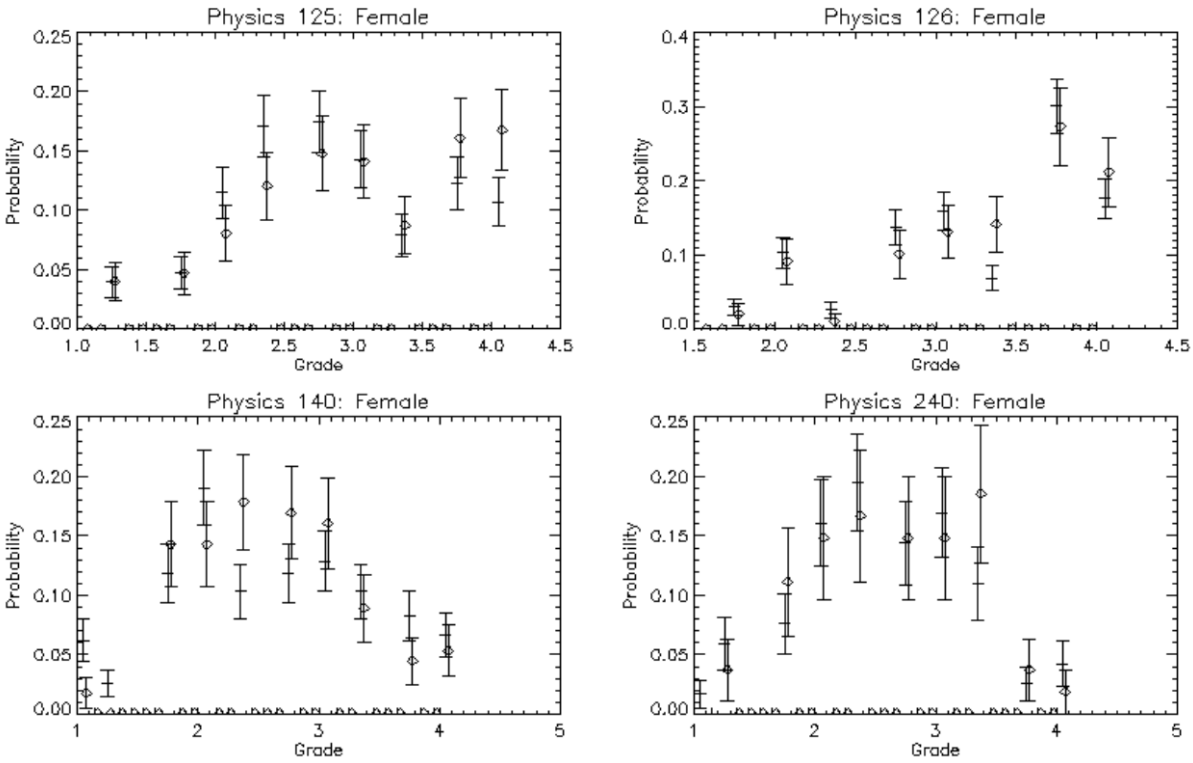


Figure 29 provides evidence that the female sample in a SLC study group and the female sample not in an SLC study groups are very similar. The K-S probability ranges from 0.330 (Physics 125) to 0.733 (Physics 126). It is also clear from this figure that female students are much more likely to participate in study groups.

With these similar populations, we again find the probability of each population receiving a particular grade (see Figure 30).

Figure 30: Female Probability of Course Grades

NOSLCFEM### corresponds to a + and SLCFEM### corresponds to a diamond;
Zero probabilities are due to course grade having discrete values.

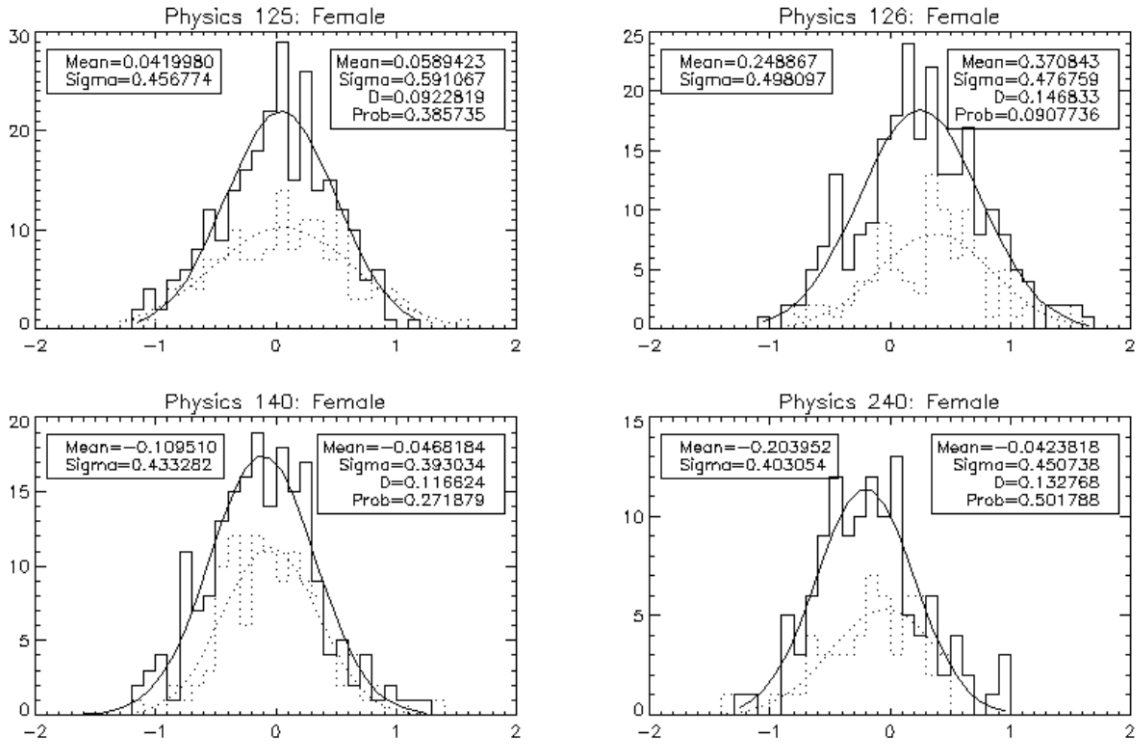


In Physics 125, females in a SLC study group are slightly more likely to receive a 4.0 (A); In Physics 126, females in a SLC study group are slightly more likely to receive a ~3.4 (B+); In Physics140, females in a SLC study group are slightly more likely to receive a ~2.3 (C/C+); In Physics 240, all differences in probabilities are insignificant.

Finally, we again plot and compare the differences in grade distributions to the grade prediction schemes for females in each course (see Figure 31):

Figure 31: Female Comparison with Predicted Grades

NOSLCFEM### corresponds to a solid plot and distribution as well as the upper left legend and SLCFEM### corresponds to a dotted plot and distribution as well as the upper right legend. The K-S Test determines D and Prob as listed in the upper right legend.



The differences in means are as follows:

Figure 32: Differences in Female SLC Means

** indicates a significant probability (Prob<0.05) and*
*** indicates a highly significant probability (Prob<0.01).*

Course	Difference	Prob
Physics 125	0.017 difference in favor of NOSLCFEM125	0.3857
Physics 126	0.122 difference in favor of SLCFEM126	0.0908
Physics 140	0.629 difference in favor of NOSLCFEM140	0.2719
Physics 240	0.162 difference in favor of SLCFEM240	0.5018

While no courses produce statistically significant differences, the greatest difference is seen in Physics 126 (Prob=0.091); females in Physics 126 are slightly helped by SLC study groups.

Overall, considering the analysis conducted with the total population and split by genders, we can conclude that SLC study groups have little significant impact on student performance in introductory physics courses. The pattern of SLC groups negatively impacting first semester courses (Physics 125 and

140) yet positively impacting second semester courses (Physics 126 and 240) as seen in the non-gender split analysis, arises again when considering males and females independently: males in SLC study groups tend to do worse in Physics 140 yet better in Physics 240 with females similarly doing slightly better in Physics 126. This is an effect perhaps worth following up.

VI. Conclusion

Adjusting for incoming GPA differences, we have effectively shown the presence and persistence of a gender gap in student performance in introductory physics courses at the University of Michigan. This gender gap is apparent in every term studied, developing the conclusive trend that male students outperform female students by approximately one fourth of a letter grade. However, we must acknowledge that the distributions do overlap, indicating that there are many cases in which a female student outperforms a male student.

Furthermore, we go on to show that mathematical preparation (SAT math score), does account for a great deal of the gender gap, especially in Physics 125 and 126. Yet, this does not completely explain the magnitude of the gender gap we see, especially in Physics 140 and 240. Investigating other potentially correlating factors, we have shown that the representation of females in a course may affect grade performance slightly—there is a larger gender gap in Physics 140 and 240 where the population is male-dominated. Yet the grade disparity persists despite small variations in female representation in each individual course from term to term. Unfortunately, we cannot vary female representation within a course itself, which would allow us to control for course content, course structure, and other influential factors, so we must be careful about this comparison between different courses. Additionally, we came to the initial conclusion that a female instructor may slightly reduce the magnitude of the gender gap. However, with only four terms in which the instructor was female, we do not yet have the statistics to strongly support this assertion.

Finally, we explore one means of intervention, SLC study groups, in the hope that it would increase grades for all students as well as reduce the gender gap. In our analysis, a trend arises: students in a SLC group for first semester physics (Physics 125 and 140) tend to receive worse than expected grades whereas students in a SLC group for second semester physics (Physics 126 and 240) tend to receive better than expected grades. Again, more extreme effects are seen in Physics 140 and 240. Furthermore, examining a male only population, we see that SLC study groups for Physics 140 negatively impact males' grades whereas SLC study groups for Physics 240 positively impact males' grades. Finally, analyzing a female only sample, we see that SLC study groups have a very small, positive impact on females' grades in Physics 126. These gender specific results are consistent with the general trend described above.

VII. Recommendations

Based on our findings and review of PER literature, we have several recommendations of proactive steps for improving course grades for all students as well as reducing the gender gap in introductory physics courses at the University of Michigan. First, we discuss the disparity in students' academic background, commenting on differences in physics and math preparation. Second, we shift our discussion to focus on

psycho-sociological factors, such as the stereotype threat, self-identification with physics, and understanding the importance of the content. Finally, we conclude with a brief overview of a grant that aims to develop an online system that provides customized, real-time coaching to students. Overall, while we recognize that some variables will inevitably influence student performance, we are motivated by reducing the impact of factors which we feel should not affect student performance, especially gender. It is the hope that these recommendations create more equity in course grades, giving everyone a fair(er) chance of receiving the grade they truly deserve.

Foremost, we have seen that academic preparation dramatically influences the gender gap. While we have shown that the majority of the gender gap is due to differences in mathematical preparation, especially in Physics 125 and 126, the same issue arises when considering previous exposure to physics. Has the student taken physics before? If so, at what level? How long ago? Did they effectively learn the content? One way to acknowledge the difference in physics preparation would be to provide support for struggling students. While supplementary SLC study groups did not prove to be as effective as we had hoped, efforts could be made to refine them to be more successful at raising student grades. Additionally, continued encouragement to use resources such as the physics helproom or attending office hours would likely resolve some difficulties with content. The Physics Department could also utilize their own physics majors by creating a tutoring program with the Society of Physics Students (SPS), Society of Women in Physics (SWIP), and Sigma Pi Sigma. Furthermore, it would be ideal to individually identify with which content areas a student is struggling and provide more in depth explanations and extra practice problems. Compiling and analyzing the online homework data would be a good starting point for identifying a student's difficulties throughout the course.

In addition to differences in physics preparation, we also must recognize the varying levels of math background the students bring to the course. It is essential to 'speak' the 'language' of math in order to fully grasp physics concepts. To encourage a greater understanding of fundamental math techniques while not dedicating valuable course time, supplementary worksheets or even discussion sections could easily be created. The focus would be on the math concepts necessary to understand the physics topics at hand and would be passed out/held the same week as the applicable lecture(s). While discussion sections would come at an additional financial cost, it is an effective way to ensure that students are exposed, at a relevant moment, to the necessary math concepts.

While academic preparation is the main contributor to the gender gap as we have seen, there are also several socio-psychological aspects which have been shown to influence student success. We would like to acknowledge such issues and recommend ways in which a professor could reduce, if not eliminate the gender gap in their course. First, we must acknowledge that stereotype threat exists. Claude Steele has made a compelling argument that the threat of others' negative judgment dramatically affects students, causing them to underperform on academic tests (Steele, 1997). While we believe that science, specifically physics, is not just a male domain, we must *show* that we do not, and students should not, believe this stereotype is true. At the same time, discussion of the stereotype threat is complicated—we must be careful not to unintentionally draw further attention to the issue in order to avoid encouraging propagation. It is also important to recognize that most interventions do not alter the stereotype globally or forever, but rather induce a local and temporary change. Similarly, we should not expect that

these recommended techniques will dramatically change the stereotype or reverse 20 years of reinforcement that students bring to the class. Rather, our goal should be to reduce and eliminate the gender stereotype in one course and one term at a time.

In order to address the stereotype threat effectively, we must consider the self-identification of female students. The goal is to increase self-identification with the field of physics so that student efficacy increases, hopefully resulting in better performance. In this process, however, we must be careful not to alienate male students by focusing too much on the female identity. Some productive ways to encourage this identification is to bring in successful, female, physicists as guest lecturers early in the term. Seeing a female thriving in a physics-related profession makes it easier to imagine oneself in her position. In addition, encouraging involvement in SWIP, Women in Science and Engineering (WISE) and SPS can also increase self-identification with the domain as students gain a socially enjoyable environment as a result of their interest in physics. Finally, simply incorporating female scientists who have made significant contributions to the field of physics, something that is absent from most textbooks, in lecture slides or discussions, reinforces that females do belong in this field.

It is also important for students to understand why they're learning the content. Further understanding the applicability increases effort in the course. Why is this course relevant if you're not planning on winning a Nobel Prize in physics? Of course, the answer varies depending on which course and what student identity is under consideration; for example, telling engineering students that fluid flow occurs in your veins and pre-medicine students that thermal expansion is considered when designing devices would be unproductive. Thus this discussion of applicability must be well thought out and customized to the audience in a specific course in a specific term. In order to do this, incorporating a simple survey question into an early homework assignment asking what prospective majors students are interested in provides a basic understanding of the student population. With this, it is the hope that a shift will occur from viewing introductory physics as 'just another requirement' to something that is interesting and useful.

Finally, it would be ideal to apply the suggestions provided here in a customized, real-time manner. We are currently embarking on the development of an 'ECoach', an online expert/electronic coach which provides personalized feedback and advice based on content difficulties and socio-psychological factors. This will provide a way to reach out to struggling students, isolating particular groups (such as females) while limiting the adverse affects of excluding other groups.

Overall, the gender gap in introductory physics courses needs attention. The Physics Department at the University of Michigan ought to be dedicating time and effort towards the success of their students. A productive way of understanding the existing inequity and beginning to develop intervention techniques would be to encourage other departments on campus to get involved. For example, consulting the College of Engineering, which faces similar issues as related to gender and has shown dedication to the improvement of its introductory courses, would provide a fruitful partnership as we move forward with our efforts.

VIII. Extensions

There are several ways to extend this research. First, it would be interesting to explore correlations between student evaluation data and course performance. This data, already compiled for Winter 1998 through Fall 2009 (see Appendix C), provides insight into the students' attitudes and feelings towards their course experience, the effectiveness of teaching and quality of the instructor. A common claim is that instructors give good grades in order to receive good student evaluations. However, the average grade, controlled by the instructor, remains quite similar term to term. Then why do student evaluations fluctuate? Do the attitudes that arise in these evaluations correlate to subtle gender trends in course performance?

Second, we would like to probe the students' reactions to key events throughout the semester. The first exam, for example, is quite challenging for most students. The literature suggests that males tend to blame external factors for their lower than expected performance, claiming that the exam was too difficult or unfair. On the other hand, females tend to internalize blame, assuming that they could've have studied more, tried harder, and done better. With these expected reactions, first exam scores may be revealing of overall course success (final grade). If the first exam does have this predictive value, expected grades could be determined as early as a month into the course.

Thus far, we have yet to explore persistent enrollment between first and second semester physics. Isolating students who took Physics 125 and 126, 125 and 240, 140 and 240, and 140 and 126, in consecutive terms, we could compare the performance of persistent students to those that take only one semester or two, non-consecutive semesters. With this data, we could also look into the retention rate from semester 1 to semester 2, examining if this differs by gender, race, etc.

For the SLC studies, we could expand this work by including more terms, adding earlier SLC data or later physics data to the original structure. Additionally, we could examine the impact of SLC participation on retention. Are students in SLC study groups more likely to continue to the second term course? We can also examine the impacts of the SLC study groups on subgroups of students, especially divided further by academic level (first, second, third year…).

Finally, while we chose to focus on gender in this thesis, it is important to recognize the several other, non-academic factors that have been shown to influence student performance. Such topics include but are not limited to, race (Steele, 1997), socio-economic status (Raizada & Kishiyama, 2010) and parents' education level. These parameters are also likely to be predictive of student success and therefore should be given adequate attention.

IX. Acknowledgements

With the completion of my senior thesis, I have many people to thank. Foremost, I would like to thank my thesis advisor Dr. Timothy McKay who first exposed me to the PER community and sparked my interest in research at the University of Michigan. During the process of writing my thesis, I gained valuable and indispensable advice, discussions and support from Professor McKay as he guided me through the process without overshadowing my voice.

I also want to thank post-doctoral researcher Brian Nord for the helpful programming tips and many red marks he put on my drafts. His constructive criticism and productive feedback helped me revise my thesis to focus on what I wanted to convey. His perspective of 'the big picture' put me at ease when I found the task overwhelming.

Next, I would like to graciously thank my lab-mate Alex Nguyen for the countless IDL loops she helped me write. Her patience with my beginning programming difficulties and continued support helped me to persist until my thesis was completed.

I would like to acknowledge my closest friends, especially Kathryn Orlando, Salina Halliday, and Holly Gwizdz, who listened to innumerable questions and interpretations of physics education. They have provided much needed encouragement at times when I doubted my abilities.

Finally, continued thanks go to my Mom and Dad for teaching me the value of education and encouraging me to pursue my goals. You instilled in me a desire to think for myself and for that, I am grateful. And to my sister Sarah, for the valuable insight she has passed onto me from her own experiences. Her continued guidance has provided me the support and encouragement I needed to write this thesis. I would also like to acknowledge my brother-in-law, Eddie. Our trips to the shooting range provided me with a much needed stress reliever. Finally, to my baby niece, Sylvia; you do not know it yet, but in the past 5 months you have inspired me to be curious and embrace every experience, while always reminding me where home is.

X. Work Cited

Brewe, E., Sawtelle, V., Kramer, L. H., O'Brien, G. E., Rodriguez, I., & Pamela, P. (2010). Toward equity through participation in Modeling Instruction in introductory university physics. *Physial Review , 6* (1), 1-12.

Crouch, C. H., & Mazur, E. (2001, March 15). Peer Instruction: Ten years of experience and results. *American Journal of Physics* , 970-977.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research , 2* (1), 1-7.

Efron, B., & Tibshirani, R. J. (1993). *Introduction to the Bootstrap.* Boca Raton: Chpman & Hall/CRC.

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., et al. (2007). Prescribed Active Learning Increases Performance in Introducotry Biology. *Life Sciences Education* , 132-139.

Hake, R. R. (1997, May 4). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Baseline* , 64-74.

Hazari, Z., Tai, R. H., & Sadler, P. M. (2007). Gender Differences in Introductory Physics Performance: The Influence of High School Physics Preparation and Affective Factors. *Science Education* , 847-876.

Henderson, C. (2002). Common concerns about the force concept inventory. *Physics Teacher , 40* (Dec.), 542-547.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Phyiscs Teacher , 30*, 141-158.

Ivie, R., & Ray, K. N. (2005). *Women in Physics and Astronomy, 2005.* College Park: American Institude of Physics.

Kolmogorov-Smirnov Goodness-of-Fit Test. (2010, 6 23). Retrieved 3 10, 2011, from Engineering Statistics Handbook: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2008). The Persistence of the Gender Gap in Introductory Physics. *Physics* .

Kost, L. E., Pollok, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics . *Physical Review: Physics Education Research* , 1-14.

Lai, L. (2009). Student Performance in Introductory Physics: A Report on Trends in Physics 125, 126, 140, and 240.

Lindell, R., Peak, E., & Foster, T. (2007). Are They All Created Equal? A Comparison of Different Concept Inventory Development Methodologies. *AIP Conference Proceedings* (pp. 14-17). AIP.

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the Physics classroom. *The American Journal of Physics , 74* (2), 118-122.

Meltzer, D. E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics , 73* (5), 463-478.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the Gender Achievement Gap in COllege Science: A Classroom Study of Values Affirmation. *Science* , 1234-1237.

Perkins, K. K., & Wieman, C. E. (2005). The Surprising Impact of Seat Loction on Student Performance. *Physics Teacher , 43* (1), 30-33.

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics - Physics Education Research* .

Raizada, R. D., & Kishiyama, M. M. (2010). Effects of socioeconomic status on brain development, and how cogntiive neuroscience may contribute to levelling the playing field. *Frontiers in Human Neuroscience* , 1-11.

Steele, C. M. (1997). A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance. *American Psychologist* , 613-629.

Zollman, D. A., & Rebello, N. S. (2004). The effect of distracters on student performance on the force concept inventory. *American Journal of Physics , 72* (1), 116-125.

XI. Appendix

A. Term Codes and Descriptions

Term Code	Term Description	Term Code	Term Description
1070	Winter 1996	1510	Fall 2004
1080	Spring 1996	1520	Winter 2005
1210	Fall 1998	1530	Spring 2005
1220	Winter 1999	1560	Fall 2005
1230	Spring 1999	1570	Winter 2006
1260	Fall 1999	1580	Spring 2006
1270	Winter 2000	1610	Fall 2006
1280	Spring 2000	1620	Winter 2007
1310	Fall 2000	1630	Spring 2007
1320	Winter 2001	1660	Fall 2007
1330	Spring 2001	1670	Winter 2008
1360	Fall 2001	1680	Spring 2008
1370	Winter 2002	1710	Fall 2008
1380	Spring 2002	1720	Winter 2009
1410	Fall 2002	1730	Spring 2009
1420	Winter 2003	1760	Fall 2009
1430	Spring 2003	1770	Winter 2010
1460	Fall 2003	1780	Spring 2010
1470	Winter 2004	1810	Fall 2010
1480	Spring 2004		

B. External Data Parameters

Variable as Listed in SD	Description	Values
UMID UMIDA UMIDS UMIDAGE UMIDP UMIDSLC UMIDC UMIDAD	A system assigned number to uniquely identify a person at U of M.	e.g. 25582542
ACADLEVELB	The academic level of the student at the beginning of the course.	Freshman Sophomore Junior Senior
ACADLEVELE	The academic level of the student at the end of the course.	Freshman Sophomore Junior Senior
CURGPA	A number representing the student's grade point average for the term.	e.g. 2.14300
CUMGPA	A number representing the student's cumulative grade point average at the end of the term.	e.g. 1.73200
NGPA	A number representing the student's cumulative grade point average at the beginning of the term.	e.g. 3.81000
TOTCUMULATIVE	The student's cumulative total of credit hours passed and all transfer credit hours.	e.g. 14
TOTTAKEN	The cumulative number of credit hours taken that count toward the grade point average.	e.g. 16
UNTTAKEN	The number of credit hours that have not yet been taken that will count toward the grade point average.	e.g. 12
UNTTAKENPRGSS	The number of credit hours that are currently being taken that will count toward the grade point average.	e.g. 12
ACTENGL	The score which the student achieved on the English component of the American College Test.	Range: 01-36 e.g. 14

ACTENGLPER	The percentile in which the student placed on the English component of the American College Test.	e.g. 89
ACTREAD	The score which the student achieved on the Reading component of the American College Test.	Range: 01-36 e.g. 25
ACTREADPER	The percentile in which the student placed on the Reading component of the American College Test.	e.g. 67
ACTMATH	The score which the student achieved on the Math component of the American College Test.	Range: 01-36 e.g. 25
ACTMATHPER	The percentile in which the student placed on the Math component of the American College Test.	e.g. 77
ACTSCIRE	The score which the student achieved on the Science component of the American College Test.	Range: 01-36 e.g. 14
ACTSCIREPER	The percentile in which the student placed on the Science component of the American College Test.	e.g. 39
ACTCOMP	The aggregate score which the student achieved for a single administration of the American College Test.	Range: 01-36 e.g. 16
ACTCOMPPER	The percentile in which the student placed on the Composite component of the American College Test.	e.g. 42
SATMATH	The score which the student achieved on the Math component of the Scholastic Aptitude Test.	Range: 200-800 e.g. 650
SATVERB	The score which the student achieved on the Verbal component of the Scholastic Aptitude Test.	Range: 200-800 e.g. 450
SATTOTAL	The aggregate score which the student achieved on one administration of the Scholastic Aptitude Test.	Range: 400-1600 e.g. 1560
GPAHS*	High school grade point average.	e.g. 3.785
RNKHS*	High school ranking percentile.	e.g. 0.90
ESTGROSSIN*	A number representing the Estimated Gross Income of the student's family (in	20=[value unknown] 25=under 25 k

	1000s).	50=under 50k 75=under 75k 100=under 100k 110=over 100k
NUMBERDEPS*	The number of dependants in the student's family.	e.g. 5
SINGLEPARENT*	A number representing the number of parents in the student's family.	800=single parent 900=not single parent
PRNTLVLED*	Parental level of education.	19=[value unknown] 25=[value unknown] 201=[value unknown] 202=[value unknown] 203=[value unknown] 204=[value unknown] 205=[value unknown] 206=[value unknown] 207=[value unknown] 208=[value unknown] 209=[value unknown] 210=[value unknown] 211=[value unknown] 212=[value unknown]
RNKQL*	A number representing the ranking of the student's high school.	1=[value unknown] 2=[value unknown] 3=[value unknown] 4=[value unknown] 5=[value unknown] 32=[value unknown] 89=[value unknown] 92=[value unknown] 96=[value unknown] 98=[value unknown]
AGE	The age of the student.	e.g. 18.2055
GENDER	The gender of the student.	'F' = female 'M' = male
CITIZENSHIP	The citizenship of the student.	e.g. 'United States'

*This parameter available only for Spring 2008 through Fall 2010 (terms 1680-1810)

(Lai, 2009); Most definitions copied or paraphrased from http://www.mais.umich.edu/reporting/download/dwsrdict.doc

It is important to note that the variable 'CUMGPA' does include the grade from the physics course of the term. 'NGPA', the cumulative GPA which excludes the physics course grade, differs from 'CUMGPA' in that resulting correlations are slightly shifted, with a more dramatic effect seen in students with fewer accumulated credit hours towards their cumulative GPA.

C. Internal Data Parameters

Variable as Listed in SD	Description	Values
UMID UMIDA UMIDS UMIDAGE UMIDP UMIDSLC UMIDC UMIDAD	A system assigned number to uniquely identify a person at U of M.	e.g. 255811047
TERM TERMCODE	A code representing the administrative time period within which students are billed and statistics are accumulated.	See Appendix A
TERMDESCRIPTION	A description of the time period within which students are billed and statistics are accumulated.	See Appendix A
COURSE COURSEAGE COURSEC	The number of the physics course, as listed in the LSA course guide.	125 = Physics 125 126 = Physics 126 140 = Physics 140 240 = Physics 240
SECTION DISCSECT	The number of the student's discussion section.	e.g. 4
LECTSECT	The number of the student's lecture section.	e.g. 1
LETTERGRADE LETTERGRADEC	The letter grade assigned to the student at the end of the term.	e.g. 'B'
INCLUDEINGPA	The description of grade inclusion in CUMGPA.	'Y' = the letter grade for the course is included in CUMGPA 'N' = the letter grade for the course is not included in CUMGPA
GRADE	A value representing the grade the student received in the course.	e.g. 3.75000
SG	A description of the student's involvement with a Science Learning Center study group.	' '= no participation in a Science Learning Center study group. 'PHY###' = participation in a Science Learning Center study group for the course Physics ### e.g. 'PHY125' = participation in a Science Learning Center study group for Physics 125

HOMEWORK	A fraction representing a student's grade on the homework portion of the course.	0.00000-1.00000 e.g. 0.850000
LECTURE	A fraction representing a student's grade on the lecture portion of the course.	0.00000-1.00000 e.g. 0.670000
DISCUSSION	A fraction representing a student's grade on the discussion portion of the course.	0.00000-1.00000 e.g. 0.950000
EXAM1	A fraction representing a student's grade on exam 1.	0.00000-1.00000 e.g. 0.650000
EXAM2	A fraction representing a student's grade on exam 2.	0.00000-1.00000 e.g. 0.675000
EXAM3	A fraction representing a student's grade on exam 3.	0.00000-1.00000 e.g. 0.520000
FINALEXAM	A fraction representing a student's grade on the final exam.	0.00000-1.00000 e.g. 0.7750000
TOTALOTHER	A fraction representing a student's grade on other elements of the course. This may include extra credit opportunities.	0.00000-1.00000 e.g. 0.760000
TOTAL	A fraction representing the total points the student earned in the course. This is calculated by adding HOMEWORK, LECTURE, DISCUSSION, EXAM1, EXAM2, EXAM3, FINALEXAM, and TOTALOTHER.	0.00000-1.00000 e.g. 0.842500
BASESUM	The total possible points in the course.	e.g. 100.000
PERCENT	A percent which represents the student's final percentage in the course. This is calculated by dividing TOTAL by BASESUM.	e.g. 84.25%
INSTRUCTORNAME**	Name of the instructor.	e.g. Timothy McKay
INSTRUCTORGENDER**	Gender of the instructor.	0=male 1=female
NRESPONDING**	Number of students who took the student evaluation survey.	e.g. 55

NENROLLED **	Number of students enrolled in a course.	e.g. 140
PERCENT RESPONDING**	Percent of students who responded to the student evaluation survey. This is calculated by diving NRESPONDING by NENROLLED.	e.g. 0.40
Q1**	A number representing the response to the first question on the student evaluation survey: "This was an excellent course."	5=strongly agree 4=agree 3=neutral 2=disagree 1=strongly disagree
Q2**	A number representing the response to the second question on the student evaluation survey: "The instructor was an excellent instructor."	5=strongly agree 4=agree 3=neutral 2=disagree 1=strongly disagree
Q3**	A number representing the response to the third question on the student evaluation survey: "I learned a lot from this course."	5=strongly agree 4=agree 3=neutral 2=disagree 1=strongly disagree
Q4**	A number representing the response to the fourth question on the student evaluation survey: "I had a strong desire to take this course."	5=strongly agree 4=agree 3=neutral 2=disagree 1=strongly disagree

**This parameter available only for Winter 1998 through Fall 2009 (terms 1170-1760)

(Lai, 2009); Most definitions copied or paraphrased from
http://www.mais.umich.edu/reporting/download/dwsrdict.doc