FACTORS THAT INFLUENCE STUDENT PERFORMANCE IN PHYSICS

by

MICHAEL ALAN GREENE

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2018

Acknowledgements

My sincerest gratitude goes to my patient and loving wife who supported me throughout the entire duration of my graduate school career, and will continue to do so long after. I also owe my motivation and emotional strength to the support of my mother, father, and sister. Additionally, the UTA physics faculty, staff, and graduate students have provided the most memorable and fulfilling college experience I have ever known. Furthermore, I am grateful for the guidance, leadership, and insight provided by my award-winning mentor Ramon Lopez.

Finally, I would like to acknowledge the Graduate Assistance in Areas of National Need (GAANN) Department of Education grants P200A130247 and P200A090284 for their financial support.

April 4, 2018

Abstract

FACTORS THAT INFLUENCE STUDENT PERFORMANCE IN PHYSICS

Michael Alan Greene, Ph.D.

The University of Texas at Arlington, 2018

Supervising Professor: Ramon E. Lopez

This study is composed of several projects in the field of physics education research. First, Investigations into flipped physics classrooms at the senior and graduate level show that students uniformly agree that the active learning techniques (peer instruction, group problem solving, etc.) were beneficial to their learning regardless of their feelings toward the flipped format of the class. One possible impact of flipping graduate-level physics courses is an increase in pass rates on the relevant section of the departmental graduate qualifying exam. Next, in the calculus-based introductory physics course, a novel statistical methodology called nonlinear casual resource analysis was used to construct predictive models of student performance based on widely-accepted factors that influence performance in physics courses. The prediction efficiency of the model was compared to traditional linear and nonlinear regression models using dichotomous forecasting, and results from the two different approaches come to similar conclusions. Finally, certain measures of cognitive ability (scientific reasoning, mental rotation ability, mathematics proficiency etc.) and affective beliefs were studied longitudinally in students enrolled in PHYS 1443 from semester to semester. Some of those factors were also studied “vertically” as they were measured in higher level physics courses including a graduate course. Trends in the descriptive statistics are reported for UTA students as well as comparisons to TCU and Yale University students.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Physics Education Research (PER)

The field of education has been under reform ever since Joseph Rice went on a tour of United States schools in 1893 and documented that students across the nation were simply being told what to say and how to say it (Rice 1893). Even today, the belief that teaching is an art form, and that the scientific method cannot be made to fit inside of the classroom, permeates the minds of people today (Beichner 2000). There has been a strong desire to make informed decisions and policies using primary data generated from within schools (Halverson 2007). This data-directed drive further supported the need for a scientific thought process when conducting education research, and is partly responsible for the recognition of a subfield of physics known as Physics Education Research (PER).

It is unclear when exactly PER as a subfield emerged, but there have been formal academic papers published regarding PER since the 1970’s, and regular PER conferences since 1997 (Barthelemy 2015). In 1999, the American Physical Society released a statement supporting PER as a subfield of physics, calling practitioners to be held to the same standards and rigors of research as any other subfield (Research in Physics Education, 1999). Beichner describes “basic” PER as a fundamental attempt to understand what students are thinking when they are learning physics, and “applied” PER when instructors actively modify their teaching methods as a result of the information and theories generated by foundational research (Beichner 2000). The ultimate goal of PER, like any other subfield of physics, is to develop testable hypotheses, implement innovative research methods, and use the results to improve existing theories or challenge what is generally accepted. For this reason, PER specialists are often found inside of physics departments, and depending on the nature of the work, they often collaborate with faculty members of other departments such as the

College of Education or the College of Psychology. Graduate students who are conducting physics education research at top-tier research institutions will possess the quantitative and qualitative skills needed to perform well in scientific and industrial careers, just like their peers conducting research in more traditional subfields of physics. They will also have training and knowledge in the education field, allowing them to find employment at a university where there is strong emphasis on teaching (Beichner 2000).

Relevant Landmarks in Physics Education

One of the most influential early members of the PER community was Lillian McDermott. The work of Dr. McDermott identified the need to address an issue which demonstrated that conceptual misunderstandings of physical topics persisted even after formal instruction in physics (McDermott 1984). Over the next several years, a significant amount of research in student's conceptual understanding lead to the development of an instrument which was meant to probe a student's beliefs on physics topics, known as the Force Concept Inventory or FCI (Hestenes et al. 1992). The widespread usage of the FCI lead to the famous study by Hake (1998) who discovered that classrooms using active learning techniques, discussed later in the *Active Learning* section, lead to greater learning gains than their peers in more traditional classrooms. Since then, other concept inventories have been developed to probe understanding of other concepts such as electromagnetic theory and thermodynamics. Concept inventories have even been developed for other STEM disciplines such as chemistry and biology. An exhaustive list of concept inventories as of 2008 can be found in a list compiled by Libarkin (2008).

In addition to research in student understanding, there has also been a substantial amount of work done related to student learning and knowledge. Hammer (2000) defined resources as mental images or cognitive storage that needs to be activated at the right moment in order to understand some physical process. Work in this area was based on valid psychological theories of constructing knowledge, and has led to many researchers and education practitioners to re-evaluate how they approach student understanding using a more cognitive framework (Elby 2010).

Because of the influence of these researchers and educators, the word “performance” and “resources” can take on many meanings in the context of PER. Performance can refer to how well a student can solve a quantitative problem, how a student handles being confronted with a misconception, or how well a student scores on a standardized test. Resources could mean what a student is thinking during a problem, or what concepts they may need to understand before moving on to a new topic. In this research project, I will define performance and resources in a different way which will be made explicitly clear in the relevant section.

Applied Physics Education Research – The Evolution of the Classroom

The Traditional Classroom

A picture of a “traditional” classroom might include a hall full of silent students seated while taking notes from a distinguished individual who may be considered an expert in the subject being taught. The teacher at the front of the room conveys information by reading from handwritten notes, writing on a chalkboard, or using some multimedia such as electronic text slides with embedded graphics and videos. This teaching method is considered traditional because it has been used since the Middle Ages at medieval universities to study all disciplines of the arts including law and medicine, and is still being widely used today across national and international universities (Rait 2012).

Although traditional lecturing does have its advantages, such as being able to deliver a large amount of information during a short amount of time, this method of teaching is being criticized for its inability to cater to students with various learning styles, its assumption that all students are learning at the same pace and keeping up, and the lack of assessment of formative student comprehension (Bonwell 1996). Many misconceptions can be formed by this method, and those misconceptions often persist after completion of the course, and are extremely resistant to traditional instruction techniques (McDermott, 2001). Teachers and their schools have been using integrated technology to address some of those issues, such as multimedia presentations like showing YouTube videos in class embedded in PowerPoint slides with eye-catching graphics, but still students are not retaining the presented information. Additionally, minority student don’t usually have the background knowledge that many professors assume during a lecture, and will not be able to make connections with the “real-world” content being presented (Paul 2015).

Active Learning

Education practitioners have different interpretations of active learning, but it is generally accepted in the education community that active learning is an instructional method that involves and engages the student in his or her own learning (Prince 2004). Hake (1998) found that using active learning techniques increased student understanding of historically difficult physical concepts (measured by the FCI) and increased the problem solving ability of the students, compared to traditional instruction techniques. Additionally, Freeman et al. (2014) conducted a meta-analysis of 225 studies that compared student performance in undergraduate science, technology, engineering, and mathematics (STEM) courses using traditional lecturing versus active learning methods. Their results show that there was a significant different in examination scores, concept inventories, and failure rates. The results are compelling - students who were using active learning techniques scored higher on assessments, and were less likely to fail the course (Freeman et al. 2014).

A physics classroom that is using active learning methods might have students who are working in small groups of three or four with each person contributing to the solution of a problem, or one student solving a problem from the previous night's homework on the board while his or her peers give feedback and critiques the work. The instructor would not be standing at the front delivering information for an hour, but instead be asking deep questions and allowing students time to think about the answer, share and defend their answers, and let other students discuss their answers. Alternatively, the instructor might pose a multiple choice question, and let students perform a think-pair-share activity where they come up with their own answer, then in small groups discuss their answers until the group comes to a consensus. Active learning spans a continuum of complexity ranging from the most simple of tasks which would be just allowing time during a lecture to pause for reflection and internalization, or as complicated as an inquiry lab where students design and implement an experiment (Active Learning Continuum 2016).

The Flipped Classroom

Performing all of the student-centered activities, such as small group discussions and independent research projects, would undoubtedly deduct from the amount of time teachers would be able to spend lecturing, and thus decrease the amount of material presented in a semester. King (1993) urged instructors to consider using class time to "construct" knowledge rather than just transmit information. Since then, educators have been experimenting with the "flipped" or inverted classroom.

The flipped classroom takes events that traditionally take place inside the classroom and moves outside and into the home, and vice versa (Lage 2000). In general, educators will deliver information to the students outside of class time using various multimedia methods such as recorded in-class lectures from previous years, homemade video lectures, or electronic whiteboard apps with voiceovers, all of which completely or partially replace the in-class lecture. In the classroom, students use a variety of active learning techniques to clarify anything from the videos, solve examples or homework problems, and discuss the concepts from the videos. Freeman et al. (1998) found that there was a significant difference in learning gains when comparing across the size of the classrooms; the smaller the class, the greater the benefit of implementing active learning techniques. However the classes in most of the studies still reported positive learning gains when compared to a traditional lecture (Freeman 2014). This suggests that any size class can be flipped, but there may be some differences and adjustments to be made in meeting the needs of all students.

Flipping the classroom draws the attention and curiosity of instructors who want their students to truly master the content and perform at their best. Fulton (2012) listed several advantages to flipping the classroom: students move and learn at their own pace; solving problems in class allows teachers better insight into student difficulties and misconceptions; classroom time is used more effectively and creatively; teachers using this method see greater interest, engagement, and performance; differentiated multisensory instruction is supported by learning theory; and the usage of technology is flexible and appropriate for “21st century learning”.

The flipped classroom is not without its problems however. Selecting a video hosting service is often intimidating for instructors who aren’t comfortable with today’s fast-changing technology, and many tenured professors do not want to put in the work required to create videos when they could just use the slides they’ve perfected over the years (Herreid 2013). Technology has made it easier than ever to host your own websites where you can uploaded your own videos, and many colleges allot bandwidth and storage space to faculty and students to host academic media. Additionally, a growing number of free and paid services make it easy to record your personal electronic device's screen (such as tablets and phones) and allow you to record voiceovers. The flipped classroom also asks students to put in time at home watching videos, and since most college students don't do assigned readings, why should instructors expect students to watch lecture videos? Detailed more in Chapter 2, our study and other flipped classroom studies suggest that when the class is structured in such a way that using the videos are an essential component to success in the course, students will do what it takes to be successful. In my opinion, I compare the student who doesn't watch the assigned videos to the student who skips class or sleeps during lectures. That type of student is not likely to pass the class regardless of the structure of the course, and there will always be students who are simply unwilling to put in the necessary work to be successful in the course.

Introduction to Modeling Student Performance

Traditional Quantitative Methods Used in PER

When conducting research in more traditional fields of physics such as high-energy or condensed matter, there are often well defined and universally agreed-upon sets of physical properties associated with the entities being studied, and standard techniques for measuring them (Ding 2012). In contrast, quantitative investigations in PER almost always attempt to measure the non-physical characteristics of human beings such as conceptual understanding, scientific reasoning skills, and attitudes and beliefs (Ding 2012). Furthermore, the PER community has not yet reached a consensus on the definitions of these constructs, and there is no universally accepted tool or methodology for measuring them. For this reason, it is commonly understood that most PER studies involving quantitative methods will have validity and reliability within the project, but may be difficult to reproduce and generalize. Nevertheless, there are still a number of fundamental statistical techniques for making inferences and descriptions of the data measured, such as using correlative statistics, single and multiple regressions, and hypothesis testing, which are common among all scientific studies, including PER.

Linear Regression Models

One specific application of correlational statistics is simple linear regression modeling. This powerful and well-studied technique involves identifying a relationship between two variables by plotting the independent variable on the x-axis and the dependent variable on the y-axis. In general, it can be used to develop a prediction based on empirical data, or it can be used to describe how well a linear equation describes the variance in the data. When working with data collected from human subjects, it's expected that the variance among the results is high due to the complicated nature of human behavior. Simple linear regression models can be quite useful when the sample size is large enough to account for this large variation, and has been used extensively in scientific research, economic and finance industries, and business applications.

Nearly every real-world phenomenon depends on several variables at once instead of just one. Linear regression can be extended to include multiple variables acting simultaneously contributing to the same dependent variable; this analysis is known as multiple regression.

Hypothesis Testing

When a researcher wants to compare the means or variance of two or more distinct but similar data samples, such as the means of final exam scores from students are two different universities, the researcher may engage in a statistical methodology known as hypothesis testing. Hypothesis testing always includes testing a null hypothesis against an alternative hypothesis. In the example of final exam scores, the null hypothesis would be "There is no difference between the means of the scores", and the alternative hypothesis could be “University A has a different mean than University B.” Before making this type of inference, the researchers must be careful in choosing a significance level which must be evident before concluding which hypothesis to accept or reject. In PER, the standard significance level is *p ≤ 5%* (Ding 2012).

The primary hypothesis testing statistic used in this study is the Welch’s t-test. The t-statistic can be calculated by using the following equation:

$$t = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}} \qquad (1)$$

Where t is the t-statistic, $\overline{X_1}$ is the mean of sample 1, $\overline{X_2}$ is the mean of sample 2, s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, and N is the respective sample size. The t-statistic, in conjunction with the degrees of freedom of the combined sample, df (rounded down), can be used to look up a p-value.

$$df = \frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 v_1}+\frac{s_2^4}{N_2^2 v_2}} \qquad (2)$$

Where $v_1 = N_1 - 1$, the degrees of freedom of the respective sample. All t-tests and p-values reported in this study use the Welch’s test, which is a modification of the student’s t-test which is more appropriate when the sample sizes are different and if the sample variance is different from the population variance. More information about hypothesis testing and Welch’s method can be found in any modern statistic handbook or Welch’s original paper (Welch 1947). Furthermore, the null hypothesis for all studies in this report, unless otherwise stated, is that there is no difference in the mean values.

Non Traditional Statistical Methods

General Systems Performance Theory

In 1986, professor of electrical engineering and bioengineering Dr. George Kondraske founded the Human Performance Institute at The University of Texas at

Arlington. In this research group, he and his interdisciplinary team have focused on modeling how human beings interact with a task from the perspective of performance (Kondraske 2011). In his 30 years of studying how humans interact with a task in specific circumstances, such as driving a car or performing a surgery, Dr. Kondraske found that no concise body of knowledge exists for a generalized theoretical treatment of system performance, and that most performance models are often specific in their applications (Kondraske 2011).

A system is something that comprises many small parts. A complicated system has many different parts, but understanding how each part functions individually lessens the complexity of the system. Kondraske (2011) points out that human being is a complicated and complex system, due to the high unpredictability of human behavior emergence. Emergence is the phenomenon where the whole system behaves in a way that is different or greater than the sum of the individual parts. Defined by Kondraske (2011), performance is what a system is capable of doing, whereas behavior is what the system actually does. Performance is a central value of today's data-driven decisions, and is fundamental for human and artificial systems. Despite the importance of modeling and predicting performance, most models have been based on specific applications in narrow domains, and generalizations have often been of secondary interest to researchers in that field (Kondraske 2011).

Kondraske (1988) developed such a theoretical framework that can model a complex system's performance by using a nonlinear combination of factors that might influence the performance of the system. He calls the task at hand the high level task (HLT), and the individual factors that influence the overall performance a basic performance resource (BPR). The performance prediction methodology uses Nonlinear Causal Resource Analysis (NCRA), which combines each BPR in such a way that allows a researcher to ask, “What is the least amount of a given performance resource required to support a given degree of higher level task performance?” (Kondraske 2011). This is fundamentally different than using traditional correlative statistics to determine relationships between BPRs and the HLT.

To explain NCRA and highlight the special nature of system performance and performance variables, consider data of the type in Figure 1. This type of distribution is explained by GSPT’s economic resource threshold idea and the logical combination of multiple BPRs at play when a system executes a high level task (Kondraske 2011). Points in the upper left region are particularly informative. For example, one can have excellent visual acuity (one BPR) and not fly a plane well (HLT) if one also has low visual information processing speed (a different BPR). The resource demand function (RDF) represents the result of a task analysis telling us how much of the BPR is required to support a given amount of HLT performance (Kondraske 2011). Note that the RDF is derived from data reflecting BPR availability. Thus, the amount required (or demanded) is inferred from data representing the amount of resource available.

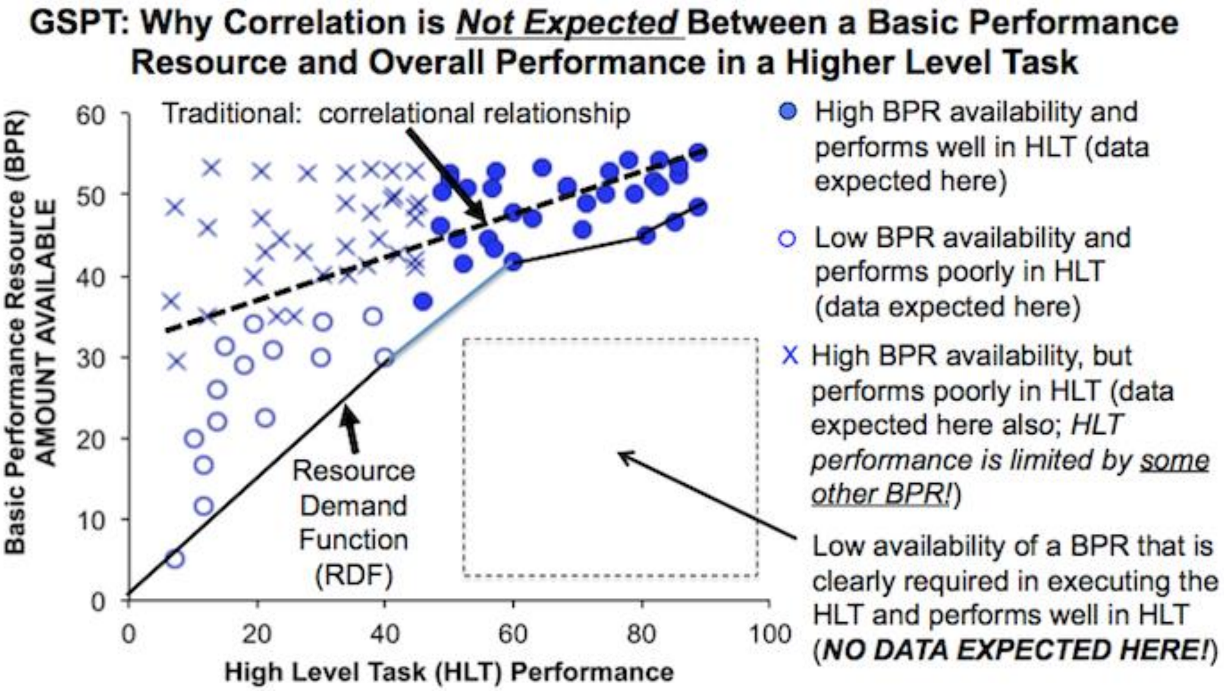


Figure 1 – Resource Demand Function (lower boundary of data points), representing the minimum amount of the BPR required to support a given level of HLT performance (Kondraske 2011).

Prediction Efficiency

When using any predictive model, whether it comes from linear regression or GSPT, it's important to evaluate the efficiency of the model's predictions, since the value of a prediction is measured by how well it helps a decision maker obtain some benefit. Two metrics to interpret the prediction efficiency of the model are distance from the actual result to the predicted result, and a method used by the meteorological community known as dichotomous forecasting. The meteorological community has developed an extensive set of tools for measuring the accuracy of predictions, which we adapted for our purposes (Stanski 1989). The simplest category of prediction is for events that have a "YES" or "NO" outcome, e.g. "this student fails the class" or "this student gets an 'A' in physics".

Analysis of such dichotomous predictions begins with a contingency table, shown in Table 1, which accounts for the four possible combinations of YES/NO events for predictions and observations.

	Observe YES	**Observe NO**	**Total**
Predict YES	Hit (H)	False Alarm (F)	PY=H+F
Predict NO	Miss (M)	Correct Negative (N)	PN=M+N
Total	OY=H+M	ON=F+N	T=PY+PN=OY+ON

Table 1 – Contingency Table.

In Table 1, H is the number of hits, F is the number of false alarms, M is the number of misses, and N is the number of correct negatives. A hit represents a correct prediction (student X was predicted to get an 'A' and that occurred) a false alarm is an incorrect prediction (student X was predicted to get an 'A' but did not). A miss represents an event which did occur which was not predicted, while a correct negative represents a NO event occurring with a correct forecast.

Table 1 can be used to calculate several measures that assess the model's ability to predict correctly, such as Accuracy (A), A=(H+N)/T, which is a simple measure of the fraction of the correct predictions. It ranges from 0 to 1 with 1 being a perfect score. It is fairly intuitive to use, but the results can be misleading since it is heavily biased by the most common situation of correct forecasting of NO events. Model Bias (B) B= (H+F)/ (H+M) compares the predicted frequency of YES events to the observed frequency of YES events. It ranges from 0 to infinity with 1 being a perfect score. It indicates whether the model has a tendency to under forecast (<1) or over forecast (>1) events. It provides no measurement of how well these forecasts correspond to the observations.

Probability of detection (POD) POD= H/(H+M) measures the fraction of observed YES events that were correctly predicted. It ranges from 0 to 1 with 1 being a perfect score. POD is good for rare events, but it can be artificially improved with more YES predictions to increase hits. It should be used in conjunction with the false alarm ratio (FAR) FAR= F/(H+F) which measures the fraction of predicted YES events that did not occur. It ranges from 0 to 1 with 0 being a perfect score. There are many additional metrics not described here.

While these metrics for model prediction efficiency were developed for meteorology, and have also been used for space weather prediction (Lopez 2007), as long as a prediction can be expressed dichotomously to create a contingency table, they can be used. In fact it may be possible to "tune" a model to improve a specific kind of result. For example, the multiple regression models could be compared to find the best balance of variables that provide a large predictive correlational coefficient as well as a large adjusted correlational coefficient. But the models could likewise be tuned to maximize, for example, the POD and the FAR, thus producing a trustworthy model that could make reliable predictions. Similarly, the NCRA model could be adjusted to maximize prediction efficiency by modifying RDFs.

Chapter 2

Investigation of flipped classrooms in the upper division and graduate level physics courses

Statement of the Problem and Purpose of Research

Since the 2000's, thousands of classrooms, then tens of thousands, have been flipped across the nation (Goodwin 2013). Many of the flipped classrooms involve middle school and high school, but flipped classrooms have been found to be successful in large introductory college courses such as physics and chemistry as well (Deslauriers 2011, Flynn 2015). The body of available knowledge diminishes as you progress into higher education and graduate school. The purpose of this project was to implement a flipped classroom in an upper-level physics course and a graduate-level physics course. This project is important because it would represent an application of an existing framework of the flipped classroom model to a population of students that have not been widely studied under this framework.

Research Questions

The flipped classroom was implemented in two upper-level physics courses at UTA, advanced mechanics and classical mechanics. Advanced mechanics is traditionally taken during an undergraduate physics major's junior or senior year. Classical mechanics is traditionally taken in the first semester of graduate school, and effectively picks up where advanced mechanics leaves off. We would like to answer the following questions:

1. To what extent is the flipped classroom appropriate for upper-level physics undergraduate and graduate students?
2. What impact does the flipped classroom have on graduate student performance on the qualifying exam?

3. Do advanced undergraduate or graduate students prefer one instructional method over another?

Course Description

For both the graduate classical mechanics and the undergraduate advanced mechanics, the instructor recorded three or four video lectures, 7-8 minute each, in which the instructor narrates sections from the current chapter in the textbook while writing on a tablet-based whiteboard app. Students watching the videos would hear the narration and see the writing as if it was live. These videos were recorded on and uploaded to a website that was made specifically for these type of videos called Educreations (www.educreations.com). Students will create a free account and log in and view the videos before coming to class. Educreations offers view tracking, so the instructor can see how many times the videos have been viewed. The instructor of this flipped class also required the students to take notes while watching the videos, just as many students take notes during a lecture. During the videos, the instructor will ask the viewer to pause the video periodically, in order to think about the answer to a conceptual question. For example, the instructor can pose a multiple-choice problem, ask the viewer to pause the video and think, and then give the answer and explanation shortly after when the viewer resumes. This is similar to Peer Instruction (e.g., Mazur, 1997), except that students do not have anyone to talk to while thinking about the conceptual question, unless they happen to view the lecture with another student. Additionally, the instructor can ask students to show steps in deriving expressions that the author of the textbook skips, or that the instructor intentionally leaves out of the videos. This allows for immediate feedback to the student on comprehension, and some interactivity between the viewer and the video at the pace of the viewer.

In class, there is an expectation that students will have watched the videos before coming to class, so that they have some understanding of the material. Students are also expected to have notes written for the videos, as well as an attempt that was made on their own to answer any homework problems assigned. The instructor expects students to make an attempt at a homework problem, even if they only get “half way”, so that they have something to contribute to the group discussion. During class, the instructor will divide students into groups of three or four, and have each group discuss one of the problems that are to be completed. When the group comes to a consensus of what they think is the correct answer, a volunteer from the group will explain it to the class. At this time, other groups may offer a different solution or some critique, and together will come to the correct answer. The instructor may need to facilitate the discussion or intervene if the direction is heading towards a misunderstanding. This process is repeated until all homework problems have been solved, and all students have communicated their level of understanding.

At the end of class, students will fill out an index card with their names and any question they still have that have not been answered, or write a request for additional clarification of a certain concept. These index cards allow the instructor to see what the students are still not understanding, and the instructor will incorporate these questions into the next session’s recorded lecture videos. The ability for an instructor to determine a student’s understanding is a crucial and elusive element of any educational program (Roehl 2013). In traditional courses, many instructors rely on exams given infrequently to make this measurement, but in a flipped classroom, the index card and class discussions allow an instructor to make these measurements much more frequently.

Methodologies

This project was a case study and used a combination of surveys and comparative design methodologies.

Survey Design

A survey of 24 Likert-type items was developed to understand a student's work habits and perceptions of a flipped classroom. Students would rank whether they strongly agreed, somewhat agreed, felt neutral about, somewhat disagreed, or strongly disagreed with a statement. Student responses were converted to a numerical value by equating a strongly disagree with a 1, somewhat disagree with a 2, neutral with a 3, somewhat agree with a 4, and strongly agree with a 5. This way, the higher the number, the more students agree with the statement. Statements were organized in four categories: *affective* - e.g., "I enjoyed this flipped classroom." *participatory* - e.g., "I watched every video before class.", *cognitive* - e.g., "Solving problems in class was helpful for my understanding of the topics.", and *procedural* - e.g., "While watching the videos, I took notes on paper." Average statement responses were compared to the neutral response, to determine if students overall agreed or disagreed with the statement. In other words, the null hypothesis is that students have a neutral opinion on particular items, and the alternative hypothesis (two-tailed) is that students either agree or disagree overall on those items.

Comparative Design

To measure the effects of the flipped classroom with graduate students, we will compare pass rates on the UTA qualifying exam from a time period before flipping the class and after flipping. Information about the qualifying exam will be discussed later in this chapter on page 38.

To measure the effects of the flipped classroom with upper-level undergraduates, the survey asks students about their perceptions of the class and the extent to which the active learning techniques and videos were helpful in their performance in the course and in their understanding of the material. Data such as final grades will not be used to make inferences of the flipped classrooms because at the time of the study, there was only one section of each type of course, and thus no “traditional instruction” grades to compare to.

Results

Undergraduate Population

The results from the survey given to the undergraduate students is shown below in Table 2. A few important results can be concluded from the results of the survey:

- Students watched nearly every video that was required
- Students set specific time for them to watch the videos, usually on the weekdays
- When there were multiple videos posted, most students would watch them over the span of one to three days
- While watching the videos, students took notes on paper, but didn’t necessarily follow along in the book
- Students often would pause, rewind, and rewatch videos at their own pace, and it was beneficial to have that option
- Before an exam, some students would rewatch videos
- Student enjoyment of the class was neutral. Some students liked it and would take additional flipped classes, and some disliked it and would not like any more flipped classes.
- Compared to traditional lecture courses, some students felt they learned less during this flipped class
- Nearly every student felt that the active learning components of the course (peer discussions, small groups problem solving) was beneficial to them

Statement	N	Mean	StDev	p
I watched every lecture video before class	12	4.42	0.52	0.000*
I watched most of the lecture videos, but skipped a few	12	2.08	1.62	0.076
I watched the lecture videos as soon as they were posted	12	2.75	1.06	0.429
I dedicated specific times in the week for watching the videos	12	3.58	1.08	0.089
I watched the lecture videos only on the weekends	12	1.92	0.90	0.002*
I watched the group of posted lecture videos all in one session	12	3.50	1.31	0.214
I spread out watching the lecture videos over more than three days	12	2.17	1.27	0.044*
While watching the videos, I took notes on paper	12	4.92	0.29	0.000*
While watching the videos, I also followed along in the textbook	12	2.25	1.55	0.121
While watching the videos, I paused the videos to stop and think	12	4.33	0.65	0.000*
While watching the videos, I would frequently go back a few minutes to listen again	12	4.08	0.90	0.002*
After watching the entire video, I would rewatch it	12	2.00	1.13	0.011*
Before an exam, I rewatched some of the lecture videos	12	3.75	1.55	0.121
Watching the videos more than once helped me understand some topics.	12	3.42	1.17	0.241
I enjoyed this flipped classroom	12	2.75	1.36	0.536
I learned less from this flipped class compared to what I learn in traditionally taught classrooms	12	3.33	1.23	0.368
I learned more from this flipped class compared to what I learn in traditionally taught classrooms	12	2.42	0.90	0.046*
I would enjoy taking additional “flipped” courses	12	2.75	1.22	0.491
Solving problems in class was helpful for my understanding of the topics	12	4.33	0.99	0.001*
Working in groups during class helped me understand the material better than working on my own	12	3.67	1.23	0.087
I enjoyed discussing conceptual questions in class that were asked in the lecture videos	12	4.17	1.03	0.002*
The additional discussions and clarifications in the classroom sessions were important to understanding the material	12	4.33	1.16	0.002*
The recorded lecture videos were helpful for my understanding of the topics	5	4.20	0.37	0.033*

Watching the recorded lecture videos helped me understand the material better than a traditional lecture in class	5	2.80	1.30	0.749

Table 2 – Undergraduate students’ responses to the end of course survey. The null hypothesis is the neutral response, a mean of exactly three.

* indicates the value is significant at the 95% confidence interval.

Graduate Students

The results from the survey given to the graduate students in classical mechanics is shown below in Table 3. A few important results can be concluded from the results of the survey:

- Students watched most of the videos that were required, but skipped several
- Students did not set specific time for them to watch the videos, but watched more videos during the weekdays than the weekends
- When there were multiple videos posted, some students would watch them all at once, and some would spread them out over several days
- While watching the videos, students took notes on paper, and most followed along in the book at the same time
- Students often would pause, rewind, and rewatch videos at their own pace, and it was beneficial to have that option
- Before an exam, a majority of students would rewatch videos
- Student enjoyment of the class was neutral. Some students liked it and would take additional flipped classes, and some disliked it and would not like any more flipped classes.
- Compared to traditional lecture courses, most students felt that they learned the same amount of material as a traditionally taught course

- Nearly every student felt that the active learning components of the course (peer discussions, small groups problem solving) was beneficial to them

Statement	N	Mean	StDev	p
I watched every lecture video before class	12	4.00	1.04	0.007*
I watched most of the lecture videos, but skipped a few	12	3.17	1.64	0.732
I watched the lecture videos as soon as they were posted	12	3.08	1.56	0.857
I dedicated specific times in the week for watching the videos	12	3.17	1.80	0.754
I watched the lecture videos only on the weekends	12	2.33	1.23	0.087
I watched the group of posted lecture videos all in one session	12	2.83	1.40	0.689
I spread out watching the lecture videos over more than three days	12	2.92	1.44	0.845
While watching the videos, I took notes on paper	12	4.17	1.12	0.004*
While watching the videos, I also followed along in the textbook	11	3.82	1.25	0.055
While watching the videos, I paused the videos to stop and think	12	4.67	0.49	0.000*
While watching the videos, I would frequently go back a few minutes to listen again	12	4.67	0.65	0.000*
After watching the entire video, I would rewatch it	12	3.42	1.17	0.241
Before an exam, I rewatched some of the lecture videos	12	4.00	0.60	0.000*
Watching the videos more than once helped me understand some topics.	12	3.75	1.06	0.032*
I enjoyed this flipped classroom	11	3.09	1.22	0.810
I learned less from this flipped class compared to what I learn in traditionally taught classrooms	11	2.91	1.14	0.796
I learned more from this flipped class compared to what I learn in traditionally taught classrooms	11	3.00	1.00	1.000
I would enjoy taking additional “flipped” courses	11	3.09	1.22	0.810
Solving problems in class was helpful for my understanding of the topics	11	4.36	0.67	0.000*
Working in groups during class helped me understand the material better than working on my own	11	3.64	1.43	0.172
I enjoyed discussing conceptual questions in class that were asked in the lecture videos	11	4.18	0.75	0.000*

The additional discussions and clarifications in the classroom sessions were important to understanding the material	11	4.36	0.92	0.001*

Table 3 – Graduate students' responses to the end of course survey. The null hypothesis is the neutral response, a mean of exactly three.

* indicates the value is significant at the 95% confidence interval.

Conclusions and Discussion

There are similarities and differences between the responses to the flipped classroom of the graduate and undergraduate students:

- Both groups watched a majority of the videos that were assigned
- Undergraduate students had more structured video-watching habits, whereas graduate students were more sporadic
- Graduate students used the textbook while watching the videos more than undergraduates
- Having the ability to pause, rewind, and replay videos was beneficial to both groups
- Both groups rewatched several videos before an exam

Although the opinions of the flipped classroom were neutral overall for both graduates and undergraduates. There are clearly three groups of students: one group of students who preferred the flipped format, ones who did not like the flipped structure, and a neutral group, as seen in Figure 2.

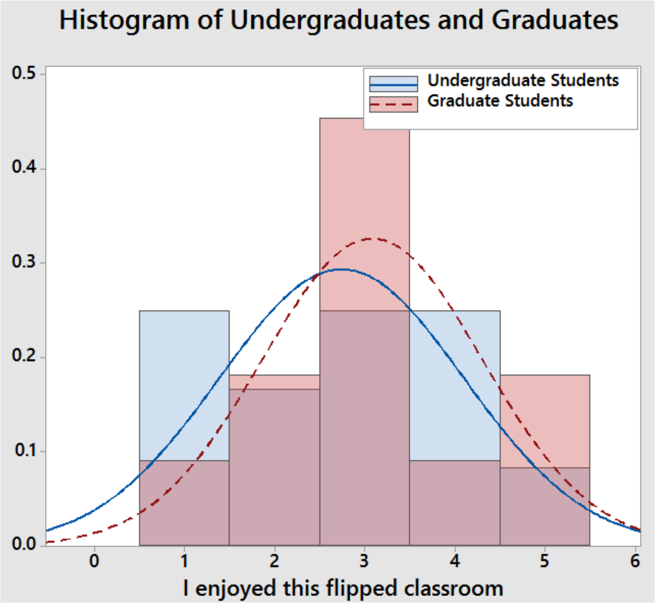


Figure 2 – Histogram of student response to “I enjoyed this flipped classroom”. A score of 1 is strongly disagree, 3 is neutral, and 5 is strongly agree.

However, a majority of students form both groups reported that discussing the lecture videos in class, doing homework in a group, and solving problems in small groups was beneficial to their understanding of the material. This is an important conclusion because it is more evidence for the use of active learning techniques even at the upper division and graduate level. Benefits of active learning are reported regardless of how a student felt about the format of the class. Student attitudes towards the flipped classroom may, however, be a predictor for their perception of how much they learned in the class, as seen in Figure 3. A study by Sarah Zappe suggested that engineering students overall were neutral about the flipped lectures, but would prefer to flip some and keep some traditional (Zappe et al. 2009).

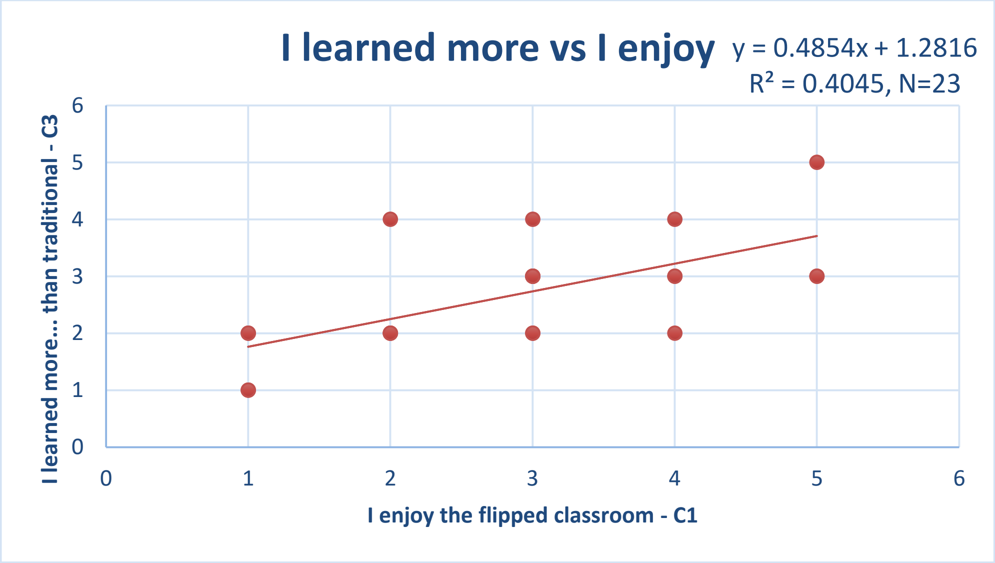


Figure 3 – Correlation between perceived learning and enjoyment of the flipped course.

The Qualifying Exam

The qualifying exam consists of four three-hour exams, one in each of four core subjects in physics: classical mechanics, statistical mechanics, quantum mechanics, and electricity & magnetism. When new graduate students begin their physics PhD program at UTA, they must pass all parts of a qualifying exam by the end of the third long semester (summer does not count) to remain in the program. Students are given three opportunities to take the exam: once at the start the semester they arrive, again during the next semester, and a third and final attempt. The exam taken right after beginning graduate school does not count against the student if the student fails, but if the student passes, the exam is considered passed. Typically, students arrive in the fall and take their first “bonus” exams, then after taking two of the four core courses, they take any unpassed exams again in the spring. Finally, after taking the other two core courses and studying over the summer, students take any remaining unpassed exams in the fall.

Each exam is composed of two “short” problems that are at the introductory physics level, and two “long” problems that are at the upper division/very beginning graduate level. Topics that would be encountered only at the advanced graduate level (Hamilton-Jacobi theory, for example) are not included in the exams. The exams are constructed each semester by committees, which also grade that particular exam and report the score to the Graduate Studies Committee. A passing grade on an exam is 60% or better. If a student does not pass an exam by the third long semester, buts does score between 40%-59%, the students qualifies to take an oral exam on this topic if a faculty member so petitions. This represents a “last ditch” effort for a student to pass a remaining subject and stay in the Ph.D. program. Typically about 10%-20% of students find themselves in this situation, but most pass. Over the past few years only 5% of students have been dismissed from the program for failure to pass the qualifying exams.

Exact historical passing rates of the UTA qualifying exam in classical mechanics is not publically or institutionally available, nor could it be de-identified to comply with IRB protocol. However, it is known that the passing rate for classical mechanics was less than 100% before implementing the flipped classroom, and there was no significant difference between the passing rates in classical mechanics and the other subjects. After implementing the flipped graduate level classical mechanics in the fall of 2014, every student who has taken the flipped version (the only one offered) has passed the classical mechanics section of the qualifying exam – a 100% pass rate. No other section of the qualifying exam has this pass rate.

Although the content of the classical mechanics qualifying exam questions has remained consistent before and after the flip (the committees tend to recycle popular problems while adding one or two new ones), there is not enough evidence at this time to say that this increase in pass rates is because of the flipped classroom model or the inclusion of active learning techniques. However, the circumstantial evidence points in this direction since the change in the pass rate was coincident with the adoption of active learning in the graduate class. No other graduate class uses these techniques, and their pass rates have remained essentially unchanged.

Chapter 3

Predictive models of student performance in introductory calculus-based mechanics using traditional statistical tools

Statement of the Problem and Purpose of Research

Single and multiple llinear regression discussed on page 21 provides a useful tool for determining a relationship, if any, between one or more independent variables and a particular outcome. The purpose of this project is to determine the extent to which a regression model founded on empirical student data has the capacity to provide accurate and consistent predictions of a future set of students' performance in an introductory physics course. Such a predictive model would be incredibly useful for educators to identify students who may be at risk for failing the course due to a lack of performance resources such as scientific thinking skills, math ability, reading ability, etc. Administrators may also be interested in such a prediction to either improve the college selection process, placement of students in appropriate courses, or even budgeting for future scholarships.

Research Questions

The first question that must be addressed is to identify several factors that influence a student's performance in physics which can also be measured readily in the beginning of the semester. Collecting data as soon as possible is imperative because the model will be used to predict the students' performance based upon their abilities coming in the course. After collecting such primary data, the next step is to develop several linear and nonlinear models to form a relationship between these factors and the students' performance. The final research question is to determine the prediction efficiency of the models. Discussed previously on page 25, the model can be considered successful by reporting a high accuracy and probability of detection, and a low false alarm ratio.

Methodology

To address the first research questions, a list of factors that are known to influence performance in physics was generated. Each factor must be well supported in peer-reviewed publications and should be readily measured via standardized assessment or survey. This step was accomplished by several brainstorming sessions, followed by a literature review. The literature review then leads to other possible factors, which returns back to a discussion session among the PER group members[1].

Next, an appropriate assessment must be identified or developed for each relevant factor. This was done by categorizing the factor as something that could be collected from a survey, from a paper-and-pencil test, or from the physics department records. Much of the literature associated with each factor presented a section which described the best way to assess or quantize that factor. For example, a student's mental rotation ability can be assessed by the Mental Rotation Test (Vandenberg and Kuse, 1978); or a student's mathematics ability inferred from the student's SAT Math score can be provided by the department of physics. Other factors that are demographic or attitudinal can be quantized by means of a survey.

To develop the single and multiple linear regression models, the quantized factors are plotted on the x-axis and the student's final grade, a measure of their performance, is plotted on the y-axis. Minitab® Statistical Software[2] and Microsoft Excel

[1] Thank you to Amanda Benson, Colby Hair, and Amanda Horton for their work contributing to this project.

[2] MINITAB® and all other trademarks and logos for the Company's products and services are the exclusive property of Minitab Inc. All other marks referenced remain the property of their respective owners. See minitab.com for more information.

(2013) were used to determine the linear and nonlinear equations which represent the relationships between the dependent and independent variables, as well as conducting hypothesis testing on the coefficients of the resulting equations.

Once the models are developed, they can be used to predict the performance of students taking that physics course the next semester by measuring the aforementioned factors and plugging them into the respective equation. The result of that operation will be what is called the “predicted grade” for that individual. The predicted grade will then be compared to the final grades actually earned in the course at the end of the semester using the metrics described earlier on page 25.

Results

Factors that influence student performance in physics

Below is a table of factors identified, a reference to the literature citing the factor’s important or effect on the performance of the students in physics, and the method used to collect the data.

Factor	Metric	How it’s measured	Citation
Spatial ability (mental rotation ability)	MRT[1] score	In-class assessment	Wai 2009
Scientific Reasoning	CTSR[2] score	In-class assessment	Lawson 1978
English /reading ability	SAT Reading score, primary language spoken	Transcript, in-class survey	Koch 1995
Basic math knowledge	SAT Math score, number of math courses taken	Transcript, in-class survey	Meltzer 2002

Basic physics knowledge	Number of physics courses taken in high school or college	In-class survey	Sadler 2001
Academic preparedness	Grade point average (GPA)	Transcript	Kuncel 2002
Initial self-assessment of science ability	Student-reported score on Likert-scale	In-class survey	Ajzen 2002
Initial self-assessment of math ability	Student-reported score on Likert-scale	In-class survey	Ajzen 2002
Time spent on homework	Hours logged in to online homework service	Reported from online service	Keith 1982
Hours worked at job	Student reported	In-class survey	Trockel 2000
Hours slept per night	Student reported	In-class survey	Trockel 2000

Table 4 – List of factors that influence student performance in physics.

[1] Mental Rotation Test

[2] Classroom Test of Scientific Reasoning

The Mental Rotation Test (MRT) is a multiple choice assessment in which students select two 2-d representation of 3-d objects out of a group of four objects which are identical to a given object with the only difference being that they are presented at a different angle in 3-d space (Vandenburg 1978). It was selected as the best assessment for spatial ability because of its widely accepted reliability, accuracy, and availability. The MRT has a total of 20 questions between two parts. Each part must be completed in 3 minutes, with a one minute break in between. The Classroom Test of Scientific

Reasoning (CTSR) is a multiple choice assessment in which students select an answer which they think is the correct conclusion, outcome of an experiment, or explanation of a phenomenon, as well as indicate why they think that is their answer is correct (Lawson 1978). This assessment was also selected for its common usage in the field of PER as a reliable way to quantify students' scientific reasoning ability. Questions include topics about conservation of mass, probability, and experiment design.

Single-variable linear regression

As an example of a linear equation with regression, Figure 4 below shows the relationship between a student's SAT math score and his or her performance in the course, constructed from four consecutive long semesters.

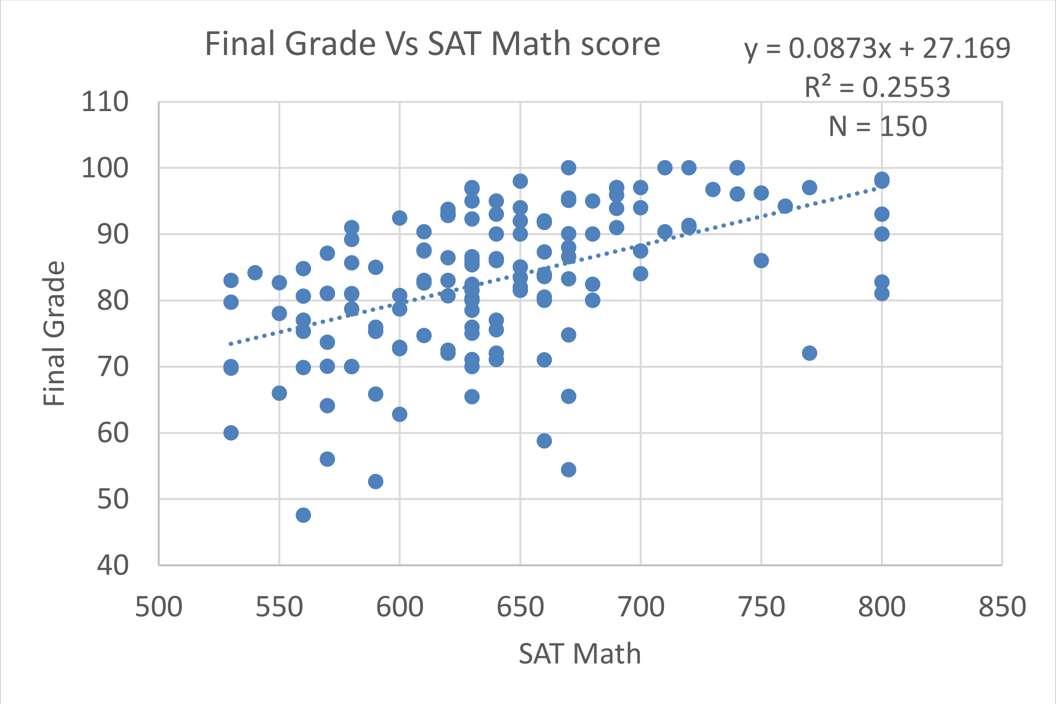


Figure 4 – A linear regression model of students' SAT score and their performance in physics.

The equation in the top right corner represents a linear equation which theoretically represents how much the dependent variable will change given a change in the independent variable. The coefficient of the dependent variable can be statistically different or similar to a value (such as zero), and so can the intercept. The equation is accompanied by a calculated regression correlation coefficient, R^2 , which represents how well the equation accounts for the variability in the data. A value of zero would mean the equation does not model accurately the data at all, and a value of one would mean that the equation models the data perfectly. To determine if the values of the slope or intercept are statistically different from zero, a two-tailed t-test is used with the standard error of the associated value. The t-test computes a t-value, which is then converted to a p-value based on the number of degrees of freedom of the measurement. In PER, researchers typically use a confidence interval of 95%, which means that a p-value less than or equal to 0.05 represents a significant difference from the chosen value, and that the difference between the values is due to something other than random chance.

Single-variable linear models

The following table summarizes the regression coefficients (the slope of the line), intercept, corresponding p-values, correlation coefficients (R^2), and number of data points in that set (N) of the factors found in Table 4. The equations were constructed from data collected from three consecutive long semesters beginning fall 2015 until fall 2016.

Factor	Slope	p-value	Intercept	p-value	R^2	N
Spatial ability	0.696	<0.001*	72.85	<0.001*	0.0670	229
Scientific Reasoning	1.859	<0.001*	67.06	<0.001*	0.1167	219
English/reading ability	0.0402	0.0840	60.60	<0.001*	0.0433	70
Basic math knowledge (SAT Math)	0.0933	<0.001*	24.40	0.020*	0.3107	80

Basic physics knowledge	2.008	0.006*	79.64	<0.001*	0.0425	180
Academic preparedness (GPA)	10.91	<0.001*	47.43	<0.001*	0.3027	150
Initial self-assessment of science ability	2.538	<0.001*	73.97	<0.001*	0.0729	180
Initial self-assessment of math ability	2.843	0.004*	70.62	<0.001*	0.0455	179
Time spent on homework	0.00746	0.018*	75.76	<0.001*	0.0644	86
Hours worked at job	-1.082	<0.001*	84.075	<0.001*	0.0671	179
Hours slept per night	2.478	0.001*	65.83	<0.001*	0.0651	181

* indicates significance at 95% confidence interval

Table 5 – A summary of linear regression equations, correlation coefficients, and hypothesis testing of coefficients.

In general, the linear equations with the largest value of R^2 have the strongest correlation correlation between the independent and dependent variables. According to * indicates significance at 95% confidence interval

Table 5, the factors which have the largest correlation coefficients are the student's score on the mathematics portion of the SAT, his or her GPA at the beginning of the semester, and his or her score on the classroom test of scientific reasoning. These results that they may be among the most influential quantitative factors to influence student performance.

Prediction efficiency of single-variable linear models

Using the models from page **Error! Bookmark not defined.**, students enrolled in the spring 2017 and fall 2017 physics 1443 course were given the same tests and surveys as the previous semesters' students, and a predicted final grade was computed by substituting the quantified factor back into the model. This value was compared to their actual earned final grade, so Table 1 was used to determine the number of hits, misses, etc. in order to determine prediction efficiency of the models. The question being asked was either “Will the student get an ‘A’? (greater than or equal to 90% final grade)” or “Will the student fail the course? (less than 70% final grade)”

The tables of such prediction metrics are presented below:

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Spatial ability	0	4	0	51	0.927	0.000	0.000	N/A
Scientific Reasoning	0	4	0	49	0.925	0.000	0.000	N/A
English/reading ability	0	4	0	23	0.852	0.000	0.000	N/A
Basic math knowledge (SAT Math)	0	3	0	31	0.912	0.000	0.000	N/A
Basic physics knowledge	0	4	0	40	0.909	0.000	0.000	N/A
Academic preparedness (GPA)	3	3	1	49	0.929	0.667	0.500	0.250
Initial self-assessment of science ability	0	4	0	40	0.909	0.000	0.000	N/A
Initial self-assessment of math ability	0	4	0	40	0.909	0.000	0.000	N/A
Hours worked at job	0	4	0	39	0.907	0.000	0.000	N/A

Hours slept per night	0	4	0	40	0.909	0.000	0.000	N/A

Table 6 – Prediction efficiency for single-variable linear regression models; “Will the student fail the course?” – Spring 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Spatial ability	0	15	0	40	0.727	0.000	0.000	N/A
Scientific Reasoning	0	14	0	39	0.736	0.000	0.000	N/A
English/reading ability	1	8	0	18	0.704	0.111	0.111	0.000
Basic math knowledge (SAT Math)	4	7	1	22	0.765	0.455	0.364	0.200
Basic physics knowledge	0	13	0	31	0.705	0.000	0.000	N/A
Academic preparedness (GPA)	3	12	1	40	0.768	0.267	0.200	0.250
Initial self-assessment of science ability	0	13	0	31	0.705	0.000	0.000	N/A
Initial self-assessment of math ability	0	13	0	31	0.705	0.000	0.000	N/A
Hours worked at job	0	12	0	31	0.721	0.000	0.000	N/A

Hours slept per night	0	13	0	31	0.705	0.000	0.000	N/A

Table 7 - Prediction efficiency for single-variable linear regression models; “Will the student get an A?” – Spring 2017

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Spatial ability	0	15	0	70	0.824	0.000	0.000	N/A
Scientific Reasoning	0	14	2	68	0.800	0.133	0.000	N/A
English/reading ability	0	3	0	24	0.889	0.000	0.000	N/A
Basic math knowledge (SAT Math)	0	4	0	26	0.867	0.000	0.000	N/A
Basic physics knowledge	0	11	0	53	0.828	0.000	0.000	N/A
Academic preparedness (GPA)	3	10	3	37	0.755	0.462	0.231	0.500
Initial self-assessment of science ability	0	11	0	53	0.828	0.000	0.000	N/A
Initial self-assessment of math ability	0	11	0	53	0.828	0.000	0.000	N/A
Hours worked at job	0	11	0	52	0.825	0.000	0.000	N/A

Hours slept per night	0	11	0	52	0.825	0.000	0.000	N/A

Table 8 - Prediction efficiency for single-variable linear regression models; “Will the student fail the course?” – Fall 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Spatial ability	0	15	0	70	0.824	0.000	0.000	N/A
Scientific Reasoning	0	15	0	70	0.824	0.000	0.000	N/A
English/reading ability	0	5	0	22	0.815	0.000	0.000	N/A
Basic math knowledge (SAT Math)	2	3	0	25	0.900	0.400	0.400	0.000
Basic physics knowledge	0	9	0	55	0.859	0.000	0.000	N/A
Academic preparedness (GPA)	1	6	1	45	0.868	0.286	0.143	0.500
Initial self-assessment of science ability	0	9	0	55	0.859	0.000	0.000	N/A
Initial self-assessment of math ability	0	9	0	55	0.859	0.000	0.000	N/A
Hours worked at job	0	9	0	54	0.857	0.000	0.000	N/A

Hours slept per night	0	8	1	54	0.857	0.125	0.000	1.000

Table 9 - Prediction efficiency for single-variable linear regression models; “Will the student get an A?” – Fall 2017

Multi-variable linear regression

Simple (single variable) regression is a powerful tool for making predictions for values that are within the range of the data set form which the equations are generated. Data collected from human subjects and projects which measure the performance of humans are almost never modelled by one single regression model. When multiple factors are acting simultaneously and contributing to the dependent variable being measured (performance), it’s possible to achieve higher regression coefficients and therefore higher prediction efficiency by constructing multi-variable linear regression equations. These equations linearly combine two or more factors in such a way that each factor will have its own coefficient of slope, but will produce a single regression coefficient which behaves the same way a single variable regression equation does. There are nCr ways to linearly combine n factors with r terms, and since a multiple regression has at least two variables, the total number of possible linear combinations is

$$\sum_{r=2}^{n} nCr$$

Where n is the number of factors measured, and r is the number of terms in the regression equation (not counting any constants), and C is the standard formula for combinatorics $(\frac{n!}{r!(n-r)!})$. When $n = 10$, this means 1013 possible linear multiple regression equations. The combinations only include equations which have no cross terms, and no order or exponent greater than one. In principal, it’s possible to write a program that will compute all of these equations with the corresponding R^2 values, and one could then identify the maximum value and the corresponding equation. Minitab already has such capabilities, so that equation can be determined readily. Additionally, one could also use only the factors which were found to have significant p-values and include those in another model as well.

Multivariable linear models

Below are several equations which were either constructed using Minitab multiple regression capabilities, or by choosing factors which have the highest linear regression correlations, or factors which have significant p-values, and various combinations of those. The equations will be named according to the order in which they are presented, and then followed by a table with similar statistics to the above tables.

Equation I

Equation I was constructed by including every factor above in a single equation:

$$\begin{aligned} Fnl\_Grd \;=\; 30.4 + 0.345\,MRT + 0.620\,CTSR - 0.0156\,SAT\_R + 0.0522\,SAT\_M \\ - 1.49\,Sleep\_Hrs - 0.475\,Work\_Hrs + 1.24\,HS\_Phys \\ - 1.04\,Phys\_Bfr + 0.59\,Math\_Bfr + 9.36\,GPA\_UTA \end{aligned}$$

Factor	Description	p-value	N	R^2
Overall			33	0.6294
MRT	Mental rotation ability	0.341		
CTSR	Scientific reasoning ability	0.509		
Sat_R	SAT Reading score	0.580		
SAT_M	SAT Math score	0.091		
Sleep_Hrs	Hours slept per day	0.383		
Work_Hrs	Hours worked per day	0.399		

HS_Phys	Number of physics courses taken in high school	0.329		
Phys_Bfr	Self-assessed ability in physics at beginning of course	0.545		
Math_Bfr	Self-assessed ability in math at beginning of course	0.791		
GPA_UTA	Incoming UTA GPA	0.008*		

Table 10 – Multi-variable linear regression Equation I and its correlation statistics

Equation II

Equation II was constructed by taking an individual's mental rotation ability, scientific reasoning ability, and SAT math scores as the only three factors. These were the factors that were originally hypothesized to have a significant impact on a student's performance in physics early on in this research project:

$$Fnl_{Grd} = 22.6\ +\ 0.096\,MRT\ +\ 0.221\,CTSR\ +\ 0.0911\,SAT\_M$$

Factor	Description	p-value	N	R^2
Overall			66	0.2834
MRT	Mental rotation ability	0.737		
CTSR	Scientific reasoning ability	0.729		
Sat_M	SAT Math score	<0.001*		

Table 11 – Multi-variable linear regression Equation II and its correlation statistics

Equation III

Equation III was constructed by taking the top three factors with the largest R^2 values from Table 5:

$$Fnl_{Grd} = 26.4 + 0.579\, CTSR + 0.0342\, SAT\_M + 9.36 GPA\_UTA$$

Factor	Description	p-value	N	R^2
Overall			45	0.4585
CTSR	Scientific reasoning ability	0.278		
Sat_M	SAT Math score	0.133		
GPA_UTA	Incoming UTA GPA	<0.001*		

Table 12 – Multi-variable linear regression Equation III and its correlation statistics

Equation IV

Equation IV was constructed by taking the top two factors with the largest R^2 values from Table 5:

$$Fnl_{Grd} = 25.9 + 0.0342\, SAT\_M + 9.36 GPA\_UTA$$

Factor	Description	p-value	N	R^2
Overall			56	0.4611
Sat_M	SAT Math score	<0.001*		
GPA_UTA	Incoming UTA GPA	0.007*		

Table 13 – Multi-variable linear regression Equation IV and its correlation statistics

Equation V

Finally, Equation V was constructed by taking the two attitudinal factors from Table 5:

$$Fnl_{Grd} = 65.49 + 2.269\, Phys_{Bfr} + 2.316\, Math_{Bfr}$$

Factor	Description	p-value	N	R^2
Overall			179	0.1022

Phys_Bfr	Self-assessed ability in physics at beginning of course	0.001		
Math_Bfr	Self-assessed ability in math at beginning of course	0.017		

Table 14 – Multi-variable linear regression for Equation V and its correlation statistics

Prediction efficiency of multi-variable linear models

With the above equations, we can perform a prediction efficiency analysis in the same way described on page 47. The tables below summarize the results:

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation I	1	1	0	14	0.938	0.500	0.500	0.000
Equation II	0	3	0	27	0.900	0.000	0.000	N/A
Equation III	3	0	0	27	1.000	1.000	1.000	0.000
Equation IV	3	0	1	30	0.971	1.333	1.000	0.250
Equation V	0	4	0	40	0.909	0.000	0.000	N/A

Table 15 – Prediction efficiency for multivariate linear regression models; “Will the student fail the course?” – Spring 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation I	0	5	0	10	0.688	0.167	0.167	0.000

Equation II	3	5	1	20	0.767	0.444	0.333	0.250
Equation III	4	5	0	21	0.833	0.444	0.444	0.000
Equation IV	0	11	0	23	0.676	0.000	0.000	N/A
Equation V	0	13	0	31	0.705	0.000	0.000	N/A

Table 16 - Prediction efficiency for multivariate linear regression models; “Will the student get an ‘A’?” – Spring 2017

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation I	0	1	0	5	0.833	0.000	0.000	N/A
Equation II	0	4	0	25	0.862	0.000	0.000	N/A
Equation III	0	3	2	9	0.642	0.667	0.000	1.000
Equation IV	0	3	3	8	0.571	1.000	0.000	1.000
Equation V	0	11	0	53	0.828	0.000	0.000	N/A

Table 17 - Prediction efficiency for multivariate linear regression models; “Will the student fail the course?” – Fall 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation I	0	0	0	6	1.000	0.000	0.000	N/A
Equation II	2	2	0	25	0.931	0.500	0.500	0.000
Equation III	0	1	0	13	0.929	0.000	0.000	N/A
Equation IV	0	1	0	13	0.929	0.000	0.000	N/A
Equation V	0	9	0	55	0.859	0.000	0.000	N/A

Table 18 - Prediction efficiency for multivariate linear regression models; “Will the student get an ‘A’?” – Fall 2017

Multi-variable non-linear regression

To truly construct as many possible regression equations as possible, we can allow for our multiple regression models to also include higher order terms and cross terms. In theory, there is no limit to the order for which a model could be constructed, for example perhaps the performance could be modelled as GPA to the power of 9 plus SAT math to the power of 50. To limit the computational time and produce a result that is within the scope of this project, I will constrain the model to only include up to quadratic terms and cross-terms. This means that at most one factor will either be multiplied with itself or one other factor. This choice allows for:

$$\sum_{r=1}^{n} jCr$$

combinations where $j = n^2,\ n$ is the number of factors measured, and r is the number of terms in the regression equation (not counting any constants), and C is the standard formula for combinatorics $(\frac{j!}{r!(j-r)!})$. When $n = 10$, this means over 10^{12} possible nonlinear multiple regression equations.

Additionally, nonlinear regression models do not give reliable R^2 values because there are assumptions about linear models that are not true for nonlinear models (Spiess 2010). Furthermore, it’s often not possible to report p-values for the coefficients of nonlinear regression equations. For linear equations, the null hypothesis is that a coefficient of zero implies that additional amounts of a predictive factor have no influence on the final result. For nonlinear equations, however, that predictive factor may show up in several different terms, and thus a single coefficient being zero will not imply that a particular factor changing will have no effect on the result (Frost 2014). For these reasons, we choose to only include a small number of factors which were shown to have a significant correlation to performance, and will rely on the prediction efficiency of the model to evaluate its effectiveness.

Multi-variable non-linear models

Below are equations which were either constructed using Minitab multiple regression capabilities, or by choosing factors which have the highest linear regression correlations, or factors which have significant p-values, and various combinations of those.

Equation VI

Equation VI was constructed by using two cognitive factors and the two highest linear R^2 factors (N=44; R^2=0.7735):

$$\begin{aligned} Fnl\_Grd = {} & -73.2 + 0.42\, MRT + 5.90\, CTSR + 0.581\, SAT\_M - 50.5\, GPA\_UTA \\ & - 0.1939\, MRT * MRT - 0.000703\, SAT\_M * SAT\_M + 0.2800\, MRT \\ & * CTSR + 0.674\, MRT * GPA\_UTA - 2.673\, GPA\_UTA * CTSR \\ & + 0.1124\, SAT\_M * GPA\_UTA \end{aligned}$$

Equation VII

Equation VII was constructed by using only the two highest linear R^2 factors (N=56; R^2=0.5205):

$$Fnl\_Grd = -172.4 + 0.651\, SAT\_M + 8.02\, GPA\_UTA - 0.000450\, SAT\_M * SAT\_M$$

Equation VIII

Equation VIII was constructed by using two behavioral factors and two factors related to self efficacy (N=177; R^2=0.2136):

$$\begin{aligned} Fnl_{Grd} = {} & 37.9 + 6.08\, Sleep_{Hrs} + 0.712\, Work_{Hrs} + 7.48\, Phys_{Bfr} - 0.34\, Math_{Bfr} \\ & - 0.218\, Work_{Hrs} * Work_{Hrs} - 1.220\, Sleep\_Hrs * Phys\_Bfr \\ & + 0.644\, Phys\_Bfr * Math\_Bfr \end{aligned}$$

Prediction efficiency of multi-variable nonlinear models

With the above equations, we can perform a prediction efficiency analysis in the same way as the preceding sections. The tables below summarize the results:

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation VI	2	2	3	24	0.839	1.250	0.500	0.600
Equation VII	1	2	0	31	0.941	0.333	0.333	0.000
Equation VIII	0	4	0	39	0.907	0.000	0.000	N/A

Table 19 – Prediction efficiency for multivariate nonlinear regression models; “Will the student fail the course?” – Spring 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation VI	1	8	0	22	0.742	0.111	0.111	0.000
Equation VII	6	5	1	22	0.824	0.636	0.545	0.143
Equation VIII	0	12	0	31	0.721	0.000	0.000	N/A

Table 20 - Prediction efficiency for multivariate nonlinear regression models; “Will the student get an ‘A’?” – Spring 2017

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation VI	0	3	2	9	0.642	0.667	0.000	1.000

Equation VII	0	3	1	10	0.714	0.333	0.000	1.000
Equation VIII	0	11	2	49	0.790	0.182	0.000	1.000

Table 21 - Prediction efficiency for multivariate nonlinear regression models; “Will the student fail the course?” – Fall 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
Equation VI	0	1	1	12	0.857	1.000	0.000	1.000
Equation VII	0	1	1	13	0.857	1.000	0.000	1.000
Equation VIII	0	8	1	53	0.855	0.125	0.000	1.000

Table 22 - Prediction efficiency for multivariate nonlinear regression models; “Will the student get an ‘A’?” – Fall 2017

Conclusions and Discussion

The goal of this study is to determine the extent to which a model derived from empirical data is useful for modelling student performance in a calculus-based introductory mechanics course, and predicting the academic performance of students in subsequent semesters. For this study, a model is considered to be successful if the metrics used to evaluate the prediction efficiency indicate a high accuracy, high probability of detection (POD), and a low false alarm rate (FAR). Since failing the course is a relatively “rare” event, it’s important to understand what a false alarm might mean for the practical application of this type of work. If an educator asks, “Will the student fail the course?”, it would be better to have a slightly higher FAR because it would be more beneficial to the student to have extra attention given to that individual student who might have passed anyway, rather than the student who might actually fail the course on their own be “ignored”, which would be a “miss” according to our metrics. For these reasons, the models with a higher probability of detection should be given more weight when deciding which model to select as the best, since it contains information about the number of misses (see page 25). Furthermore, the best models should give a consistent prediction from semester to semester, and have similar prediction metrics.

Single Variable Linear Models

With the above preferences in mind, the single variable linear model that produced a finite, non-zero FAR and POD with respect to failing the course was the equation relating a student’s performance with that individual’s incoming grade point average (GPA), indicated by Table 6 and Table 8. The relatively high POD and low FAR in one semester show that this model may be useful for selecting individuals who may not have a strong academic background, and thus may need more personalized attention.

On the other side of the coin, both the single linear models using GPA and SAT math scores were able to consistently give finite, non-zero POD and FAR when it comes to identifying students who may be capable of earning an “A” grade for the course, indicated by Table 7 and Table 9. Interestingly, the SAT math score had a higher POD in the fall than in the spring. One hypothesis is that students in the fall semester have just recently taken the SAT, which is meant to be a predictor of first-year college success, while students taking mechanics in the spring may be less traditional students who may have taken the SAT more than one year ago.

Multivariate Linear Models

The multivariate linear equation which produces the highest POD and lowest FAR for identifying at-risk students by far was Equation III (page 54) with a POD of 1.000, and a false alarm of 0.000, shown in Table 15. This means that out of the 3 students that failed the course that semester (who also had information about their CTSR, GPA, and SAT scores), this particular equation was able to identify all three of them as students at risk to fail the course. Unfortunately, this same equation failed completely the next semester, unable to identify a single at-risk student from Table 17. Further investigation is needed to determine whether different models could be used for different semesters.

With respect to getting an “A”, Equation II (page 54) was the only model to give a consistent finite non-zero POD and FAR over two semesters, although the probability of detection was about 33% in the spring but rose up 50% in the fall, with zero false alarms, shown in Table 16 and Table 18.

Looking at Table 10 through Table 14, we find many coefficients which are not significantly different from zero. In fact, typically only one or two factors are among the set are significantly different from zero. However, the overall correlation coefficient (R^2) increases to a value which is above any individual single linear regression correlation coefficient. This may be due to multicollinearity, which is a condition where some of the factors used in the model correlate to other factors. This can have the effect of making any single variable seem insignificant even when it is expected to be significant. Measurements of multicollinearity can be performed to determine if this effect is influencing the results by computing the variance inflation factor (VIF) for each variable. Further research is needed to determine the extent to which the variables are collinear.

Multivariate Nonlinear Models

There were no nonlinear multivariate models that consistently predicted any failing individual or ‘A’ student in both semesters, especially in the fall. In the spring, however, Equation VI (page 59) shown relatively high POD and low FAR (Table 19 and

Table 20), and may be a possible candidate for further study and refining. Equation VI represented a relationship between SAT math scores and GPA.

Final Remarks

Predictive models are only useful when they allow someone to make a decision. The choice to use any single model to evaluate a possible course of action should be heavily considerate of the impact that decision could have. In this context, admitting or denying a student admission to a university or to a course based on the predicted grade of any single equation would be unethical. However, all of the models presented in this chapter that were somewhat consistent from semester to semester were the ones which included information about their mathematics ability (SAT) and their academic preparedness (GPA). This finding is consistent with today’s body of physics education research and may serve as a starting place for additional research projects in the space of predictive analysis.

Chapter 4

Predictive modelling of student performance in introductory calculus-based mechanics using non-traditional statistical tools

Statement of the Problem and Purpose of Research

The previous chapter used traditional statistical methodology to construct predictive models of academic performance. It may be possible, however, that there is such a complicated interaction among the variables identified in Table 4 that the use of any regression model will not be able to accurately and reliably predict an individual's future performance. Modelling and predicting human behavior as they interact with a task has been the study of Dr. George Kondraske for over 30 years (see page 22), and in this study we will explore the extent to which a theory that considers the human as a system of resources can be used to develop a predictive model of academic performance: General Systems Performance Theory (GSPT).

Background on General Systems Performance Theory

General Systems Performance Theory (GSPT) is based on the idea that an entity may have emergent behavior. Emergent behavior is when a system exhibits more behaviors than the individual parts which make up the system (the resources) can display on their own. For example the brain is composed of about 75% water and the rest is a combination of atoms, molecules, and chemicals (the resources), and yet the brain serves as the primary controller of the nervous system and gives humans consciousness (emergent behavior). In our context, we model the student as a system composed of several resources attempting to perform the task of being successful in a physics course. In this study, we define the basic performance resources (BPR) as the factors which influence a student performance in physics shown in Table 4, and the high-level task (HLT) as passing a physics course.

The basic idea of GSPT is that the system can only function as well as its weakest part. That is, if one critical resource is missing or inadequate then the system will not be able to accomplish the task with a given level of satisfaction. The relationship between a single resource and the task performance is given by a resource demand function (RDF) which represents the amount of a particular resource needed in order to accomplish a specific level of performance. For example, the image in Figure 1 shows a scatterplot of a resource (on the y-axis) and the performance on a task (on the x-axis). The RDF is a function which defines the lower boundary, and points that lie along the RDF represent the minimum amount of that resource to achieve that level of success on the task.

Once the RDF has been determined for a number of independent resources, one can begin to use the RDFs as predictive functions. Inverting the function then substituting values of the quantified resource into the new equation produces a predicted level of performance. Doing this for all factors will produce a range of predictive task performance. Taking the lowest value in this range will serve as the final predicted level of performance because it means that value came from an RDF in which the individual had the lowest amount of resource, and therefore, that resource is the individual’s limiting factor of success. This analysis of using RDFs to predict future performance is calls Nonlinear Causal Research Analysis (NCRA).

Motivations

The motivation for a project such as this comes from two events. The first event was a research project conducted by a former UTA PhD candidate, Alfonso Hinojosa. The bulk of his work relied on traditional statistical analysis, but one part of his dissertation included using NCRA to predict the SAT scores of 30 high school AP physics students (Hinojosa 2015). **Error! Reference source not found.** shows the result of this prediction.

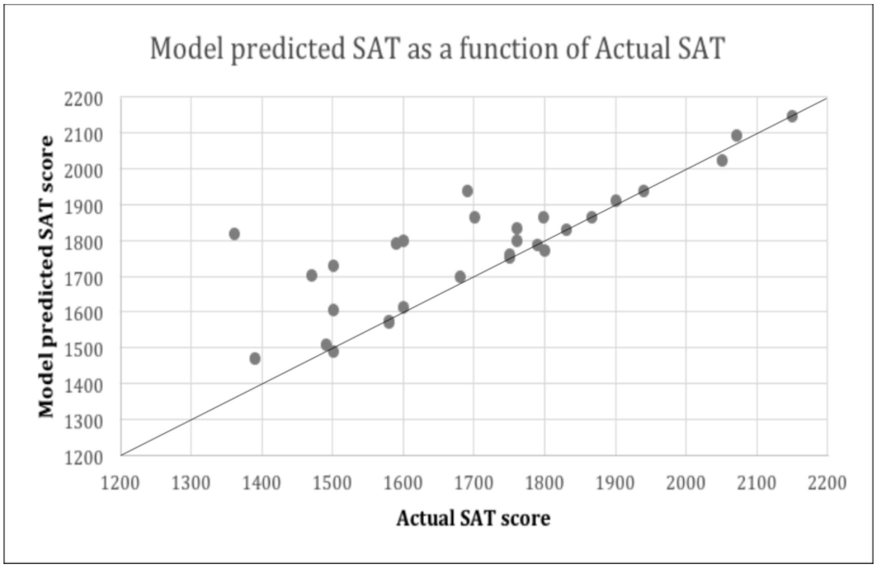


Figure 5 - Model predicted SAT score as a function of actual SAT scores (Hinojosa 2015)

The basic performance resources used were science ability (grades in high school physics and the score on the state science assessment, the TAKS – Texas Assessment of Knowledge and Skills), math ability (math grades and TAKS/Math), English language ability (TAKS/ELA), scientific reasoning ability (measured by Lawson's test), and spatial ability (measured by the MRT). As can be seen, the model predictions are close to the real values, with 20 of the 30 predicted scores within 60 points of the actual scores (the standard error for SAT according to the College Board). But the model is post hoc and researchers only had the SAT scores for the thirty students (Hinojosa 2015).

The second motivating event was a study conducted again by Hinojosa (2015) and others as part of his dissertation which compared the performance in STEM courses of high school students enrolled in a magnet school versus a non-magnet school, and their relationships to math ability, scores on standardized assessments, and other metrics. One of the major conclusions was that many of the factors studied (math ability, English ability, grades in STEM courses, mental rotation ability, etc.) had statically significant correlations to the TAKS standardized assessments and physics courses. However, many of those same factors were not significantly correlated to the non-magnet students, with the exception of mathematics ability.

For example, the statistically significant correlations of the non-magnet students' class grades with the CTSR score disappear, as does the correlation of the MRT with the physics grade. A possible explanation for the non-magnet students' performance is that the primary cognitive resource that determines the grade is the student's math ability. If the non-magnet students have a lower ability to use math in their classes, then math would be the primary limiting factor in their performance in those classes and eclipse other factors. But for a magnet student, with stronger math ability, the limiting performance factor might be spatial ability, scientific reasoning skill, English language ability, or science content knowledge because the requisite threshold math ability exists (Hinojosa 2015).

Research Questions

Inspirited by the accuracy of the post-hoc SAT score prediction from the NCRA (Figure 5) and working with the hypothesis of a threshold effect where you must possess a minimum amount of resource to perform at a particular level on a task before being limited by some other resource, we will investigate the following research questions:

1. To what extent can a nonlinear causal resource analysis (NCRA) be used to predict academic performance of physics students?

2. How does the prediction efficiency of the NCRA compare to that of traditional statistical prediction metrics?

Methodology

To answer the above research questions, I used the same empirical data set described in Chapter 3, that is, three semesters of PHYS 1443, a calculus based introductory physics course. Each factor (with the exception of some, discussed later) identified from Table 4 is considered a basic performance resource (BPR), and a scatter plot is made with the BPR on the y-axis and the final grade (the high level task (HLT)) on the x-axis. The resource demand function (RDF) is then generated by constructing a piecewise function based on the lower boundary of all the data, shown below:

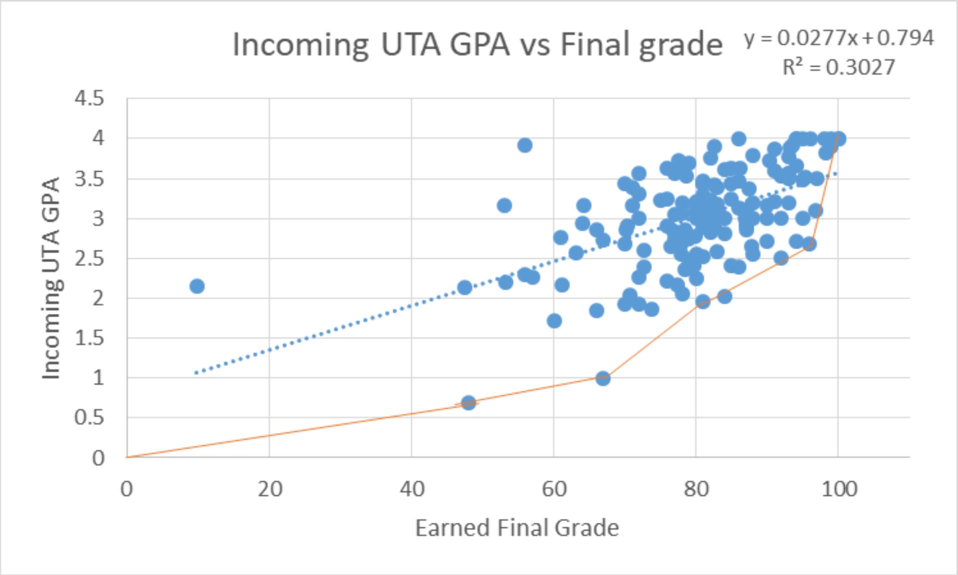


Figure 6 – A scatterplot of GPA vs Final Grade. The dashed blue line is a traditional linear regression line, and the solid red line is the RDF.

Once the RDF’s have been created for all the BPRs, the RDFs are then inverted, and the next semester of quantified BPRs can be inputted. The result will be a number which is the predicted performance based on the amount of BPR the individual has. To construct the predictive models, and conduct NCRA, each individual student will have one BPR which produces the lowest value of performance. That lowest value is the student’s predicted level of performance (final grade), and is associated with the limited performance resource.

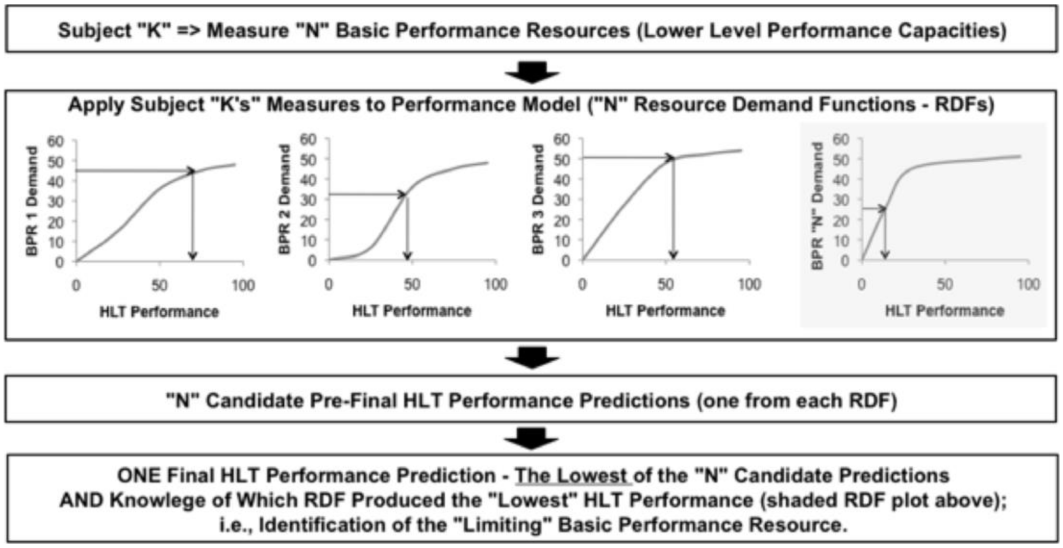

Subject "K" => Measure "N" Basic Performance Resources (Lower Level Performance Capacities)
Apply Subject "K's" Measures to Performance Model ("N" Resource Demand Functions - RDFs)
BPR 1 Demand
BPR 2 Demand
BPR 3 Demand
BPR "N" Demand
0
10
20
30
40
50
60
0
50
100
HLT Performance
"N" Candidate Pre-Final HLT Performance Predictions (one from each RDF)
ONE Final HLT Performance Prediction - The Lowest of the "N" Candidate Predictions
AND Knowlege of Which RDF Produced the "Lowest" HLT Performance (shaded RDF plot above);
i.e., Identification of the "Limiting" Basic Performance Resource.

Figure 7 – An NCRA approach to determine the predicted final grade (Kondraske 2011).

Results

After generating plots like Figure 6 for each BPR, the following RDF’s were constructed. Time spent on homework was not included because we switched to a different homework service that does not report the time spent on the assignment, and SAT reading was also not included because of a high number of outliers. Due to the nature of GSPT requiring that “more” of a quantity is “better”, the number of hours worked in a day is subtracted from 24, and thus becomes number of hours “not worked” during the day.

Resource	Equation	Domain
Spatial ability	y = 1.667E-05*x + 0	for x < 60
	y = 0.049*x - 2.99	for 60 ≤ x < 80
	y = 0.140*x - 10.220	for 80 ≤ x < 87.13
	y = 0.233*x -18.3	for 87.13 ≤ x ≤ 100
Scientific Reasoning ability	y = 0.0173*x + 0	for x < 57.67
	y = 3.53E-05*x +0.997	for 57.67 ≤ x < 85.92
	y = 0.141*x - 11.1	for 85.92 ≤ x < 93
	y = 0.428*x - 37.8	for 93 ≤ x ≤ 100
Basic math knowledge (SAT Math)	y = 8.83*x + 0	for x < 60
	y = 4.34E-05*x + 529	for 60 ≤ x < 83
	y = 6.24*x + 11.3	for 83 ≤ x < 91
	y = 8.33*x - 178	for 91 ≤ x < 97
	y = 13.3*x - 663	for 97 ≤ x ≤ 100
Basic physics knowledge	y = 1.10E-05*x + 0	for x < 90.977
	y = 0.111*x - 10.1	for 90.977 ≤ x ≤ 100
Academic preparedness (GPA)	y = 0.0143*x + 0	for x < 48
	y = 0.0164*x - 0.100	for 48 ≤ x < 67

	y = 0.0683*x -3.58	for 67 ≤ x < 81
	y = 0.0482*x -1.95	for 81 ≤ x < 95.93
	y = 0.325*x -28.5	for 95.93 ≤ x ≤ 100
Initial self-assessment of science ability	y = 0.0115*x + 0	for x < 86.86
	y = 0.0824*x - 6.16	for 86.86 ≤ x < 99
	y = x - 97.0	for 99 ≤ x ≤ 100
Initial self-assessment of math ability	y = 0.0136*x + 0	for x < 73.3
	y = 0.0483*x - 2.54	for 73.3 ≤ x < 94
	y = 0.167*x - 13. 7	for 94 ≤ x ≤ 100
Hours not worked at job (/day)	y = 0.263*x + 0	for x < 57
	y = 6.667E-05*x + 15.0	for 57 ≤ x < 72
	y = 0.0473*x + 11.6	for 72 ≤ x < 93.1
	y = 0.385*x - 19.8	for 93.1 ≤ x < 97
	y = 2.17*x - 193	for 97 ≤ x ≤ 100
Hours slept per night	y = 0.0455*x + 0	for x < 66
	y = 0.0588*x - 0.882	for 66 ≤ x < 83
	y = 0.0625*x - 1.19	for 83 ≤ x < 99
	y = x - 94	for 99 ≤ x ≤ 100

Table 23 – Piecewise linear RDF for each basic performance resource. The x-value is an expected level of performance, and the y-value is the minimum amount of that resource needed.

Using the same RDFs for both semesters, student data from Spring 2017 and Fall 2017 were inputted to produce the predicted final grade for each resource. The lowest value of these functions was taken as the final predicted grade, and it was compared to the actual earned grade. The same dichotomous prediction efficiency metrics was used as previous chapters, and the following figures and tables summarize the prediction efficiency:

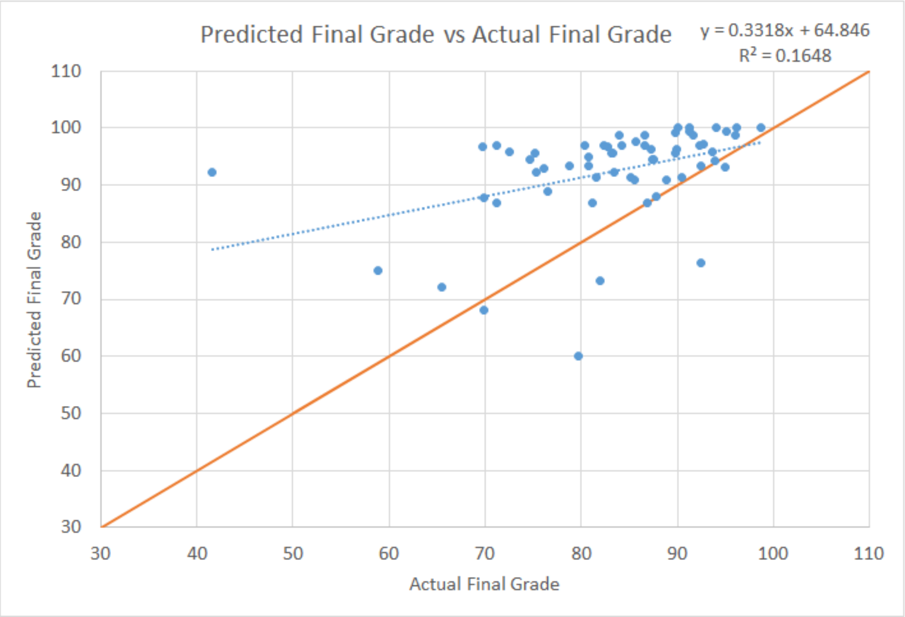


Figure 8 – A graph of spring 2017 students' predicted final grades versus actual final grades. The dashed line represents the linear trendline, and the solid line has been added for reference and has a slope of 1, representing a perfect prediction (N=60).

In Figure 8, the solid line with a slope of 1 was manually added to indicate a “perfect” prediction. If a student is above the line, that individual got a lower grade than what was predicted, if a student is below the line, he or she scored above the predicted grade. The dashed line represents the correlational trend in the predicted vs. actual grade. The following figure shows the predictions for the following 2017 semester.

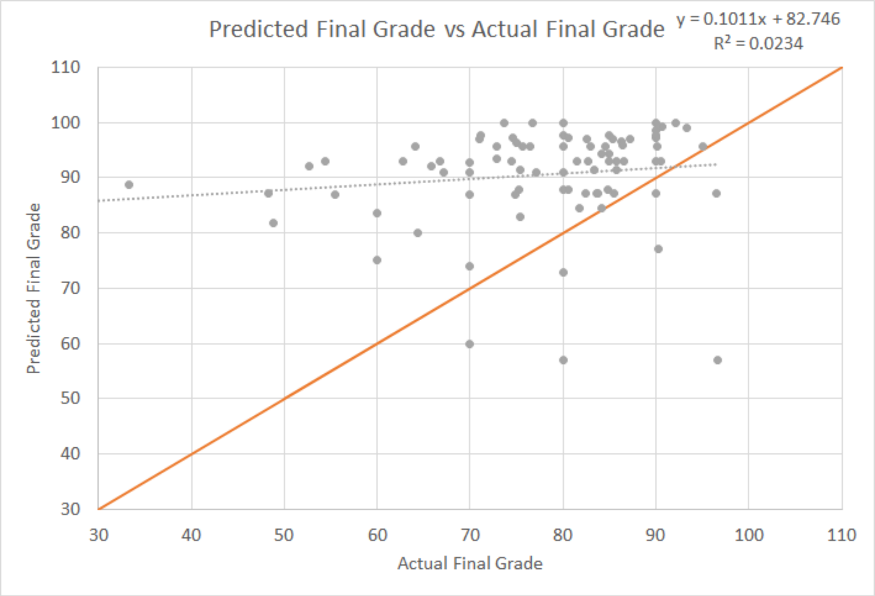


Figure 9 - A graph of fall 2017 students’ predicted final grades versus actual final grades.

Finally, the two semesters combined:

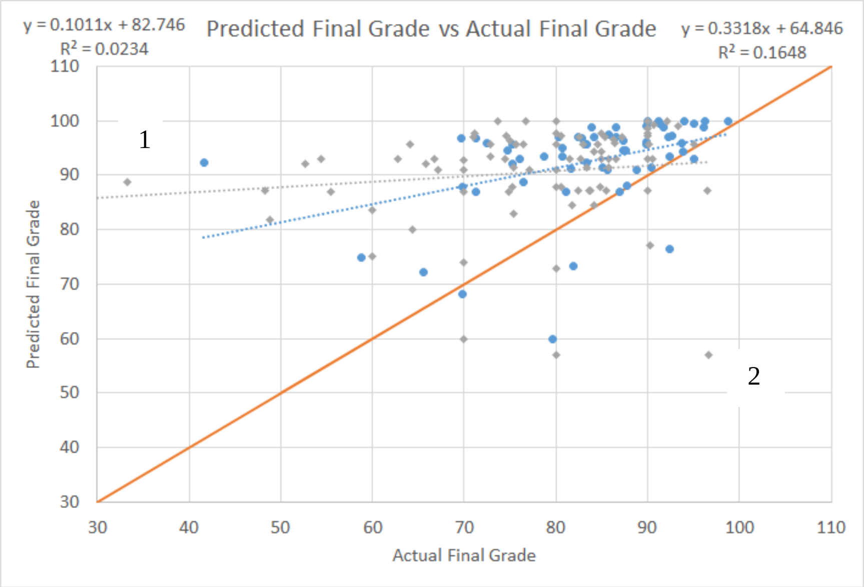


Figure 10 – A graph of spring (dots) and fall (diamonds) 2017 students' predicted final grades versus actual final grades.

In Figure 10, there are two possible outliers identified. To be an outlier on this graph means that the individual scored very differently from the predicted outcome. Through the lens of GSPT, outliers above the solid line, such as student 1, means that they performed worse than predicted, and that they possibly were limited by the lack of resource which was not included in the construction of the NCRA. In other words, there was some factor which was not included in the model that was holding him or her back. It may be possible in the future to interview students like these to get a better understanding of the cause of this poor performance. Student 2 on the other hand is far below the line, which means that this individual performed well despite being predicted to have poor performance due to low BPRs. The first run of this analysis produced several outliers below the line, and a majority of them were due to low SAT reading scores. In each case, the next limiting factor would have been SAT math score, and each of their resulting predicted grades would have been closer to their final grade. Student number 2 was reported to be limited by the number of hours worked at a job per day, and without that factor, he or she would have been predicted to earn an A (which that individual actually did earn).

Prediction Efficiency

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
NCRA	1	5	1	53	0.900	0.333	0.167	0.500

Table 24 – Prediction efficiency for NCRA; “Will the student fail the course?” – Spring 2017

Will the student get an ‘A’?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
NCRA	16	1	32	11	0.450	2.824	0.941	0.667

Table 25 - Prediction efficiency for NCRA; “Will the student get an ‘A’?” – Spring 2017

Will the student fail the course?								
Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
NCRA	0	15	3	67	0.788	0.200	0.000	1.000

Table 26 – Prediction efficiency for NCRA; “Will the student fail the course?” – Fall 2017

Will the student get an ‘A’?

Model Used	Hits	Misses	False Alarm	Correct Negative	Accuracy	Model Bias	Probability of detection	False Alarm Ratio
NCRA	11	4	45	25	0.424	3.733	0.733	0.804

Table 27 - Prediction efficiency for NCRA; “Will the student get an ‘A’?” – Fall 2017

Conclusions and Discussion

General Systems Performance Theory can be used as a framework for modelling human performance on a task. In this study, the student is considered a system of resources which are drawn upon to complete the high level task. The basic performance resources are the factors that influence performance in physics, and the task is passing a physics course. By quantifying and measuring the resources of incoming students, it can be compared to the resource level and performance of previous students in order to construct a predictive model. This type of Nonlinear Casual Resource Analysis (NCRA) approach has never been used in the context of an introductory physics course sequence, and thus presents a unique approach to identifying at-risk or high-performing students.

With regards to the NCRA model, Table 24 through Table 27 show inconsistent or unreliable prediction efficiency. Although the false alarm ratio and probability of detection are relatively low in the spring for detecting at-risk students (Table 24), it does not give the same results in the subsequent semester (Table 26). The high probability of detection for identifying “A” students is consistent and promising (Table 25 and Table 27), but the high false alarm ratio and model bias means that the model in general tends to over predict students’ performance, and may not be the most trustworthy.

One of the benefits to using NCRA over traditional statistical methodology is the fact that the NCRA models will report which resource is the limiting factor for a particular individual. The top three most common limiting resources for the spring was GPA, SAT Math scores, and mental rotation ability. The top two most common limiting resources for the fall was GPA and, surprisingly, hours worked. However, the overwhelming majority of predicted values for hours worked were above 95%, so the third most common limited resource was actually the SAT math scores, followed by mental rotation ability.

One very interesting finding is that the NCRA model actual produces similar trends of prediction efficiency metrics. That is, in both traditional linear regression and in NCRA, there were models which had consistent prediction for “A”’ students, relatively promising results in the spring for identifying at-risk students, and the models which had the biggest influence were those related to GPA, SAT scores, and mental rotation ability. Another benefit to using NCRA is that we are able to identify which factor may be most limiting to a student’s performance. This could allow a researcher to develop a personalized intervention for that individual and thus improve his or her predicted performance. The downside is that if a student has a poor single-point measurement, such as SAT reading score or MRT score, then the predicted performance may show an exceedingly low score. This is the case with the outliers from an initial version of this analysis, in which three additional students were all limited by SAT reading scores in such a way that they were considered extreme outliers in Figure 10. It’s entirely possible that these individuals did not perform well on the reading portion of the SAT due to distraction, lack of sleep, or any other number of reasons and therefore was not a true reflection of their ability. That fact that a majority of these outliers are due to this particular factor supports the hypothesis that this particular factor should not be included in this model, and was therefore left out. Additionally, a majority of the twelve students below the solid orange line from Figure 10 all were limited by a single-points measurement such as number of hours worked, mental rotation ability, or their confidence in science or mathematics, with only one student being limited by their GPA, which is not typically considered a single-point measurement.

Although the conclusion of this study did not lead to profound predictive capability of physics students, a novel application of statistical methodology that comes to a similar conclusion as traditional statistical methods means that the use of GSPT and NCRA may have place in an educational context, but will require further refinement of the approach or the model, and may serve as the beginning of future research projects.

Chapter 5

Trends in cognitive ability and affective measures among courses and institutions

Statement of the Problem and Purpose of Research

When developing predictive models of student performance in Chapter 3 and Chapter 4, a question that arises is whether or not students in the fall and spring semesters are the same, or at least considered to be independent samples of the same population. If so, then we should expect that the same predictive models can apply to both semesters. If not, then it will likely be that two different models are needed for two different populations. Investigating trends in cognitive ability (mental rotation ability, scientific reasoning, SAT scores, etc.) and affective measures (self-efficacy, motivations, etc.) over a period of successive semesters will allow professors and instructor to have quantitative evidence of their students’ ability and thinking, and can allow researchers to determine if students are similar enough to warrant using the same predictive models.

Furthermore, comparing these measures from one course to another may allow insight into the effect of staying in a STEM program may have on their cognitive ability and attitudes towards physics as they progress from an introductory physics course all the way to graduate school. Finally, comparing these metrics across institutions may reveal interesting patterns in the effect an institution may have on their student body such as one institution filtering out individuals or another providing a larger growth in cognitive ability.

Research Questions

In this study, we will attempt to answer the following research questions:

1. Do factors that influence performance in physics have consistent correlation with performance from semester to semester within a course?

2. What are the differences in the descriptive statistics of some of these factors among the same course in different institutions?
3. How do these factors change as physics students grow and take additional physics courses?

Methodology

To answer these research questions, we administered a variety of assessments and surveys either in-class or online to students enrolled in physics courses at The University of Texas at Arlington (UTA), Texas Christian University (TCU), and Yale University (Yale). The assessments primarily included the Mental Rotation Test (MRT), the Classroom Test of Scientific Reasoning (CTSR), and a select number of questions from the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich 1991). The MRT and CTSR were described on page 43. The MSLQ is a questionnaire in which students respond to various statements using a Likert scale. Students were also given a survey which asked various questions related to demographics, academic preparedness, attitudes towards the course, self-efficacy in math and science ability, and other academic habits. This survey was also selected because of its proven reliability and availability.

The collaboration efforts began in spring 2017 with Texas Christian University when a lecturer in the physics department at both UTA and TCU, Dr. Fajer Jaafari, suggested that we also collect data from TCU students for comparison for the upcoming fall semester. Meanwhile at a GIREP conference in Dublin during the summer of 2017, my research advisor Ramon Lopez met Dr. Claudia De Grandi, a Helmsley Postdoctoral

Teaching Scholar in Physics at Yale University, and began to discuss the extent to which Yale could collaborate with UTA during the upcoming fall as well.

To answer research question 1, we gave the same assessments as before to students enrolled in PHYS 1443 at UTA, the same students from Chapter 3 and 4, and recorded their responses to each test and survey item. We are then able to track the means, standard deviations, correlations and other statistical information from semester to semester. To answer research question 2, a collaboration was formed among UTA, TCU, and Yale to give the MRT, CTSR, and MSLQ to their respective calculus-based introductory physics courses. All parties involved collected data under the same UTA Institutional Review Board (IRB) protocol, and all documentation (physical or electronic) including consent forms are kept secure following UTA IRB guidelines. To answer research question 3, the MRT and CTSR was administered to students enrolled in freshman-, junior-, and senior-level physics courses at UTA, such as modern physics and advanced mechanics, as well as to graduate students.

Results

Trends in factors that influence performance in physics in a single course.

Answering research question 1, from fall 2015 to fall 2017, eight factors from Table 4 were individually correlated to the students' final grade in PHYS 1443, the calculus based introductory physics course at UTA. Thus for each factor in each semester, a simple linear regression equation was calculated in the same was as described in the "

Single-variable linear regression" section. The following diagrams show the slope of the regression equation on the x-axis, and the y-intercept of the equation on the y-axis. The errors bars in both the x- and the y-direction represent the standard error in the slope and y-intercept respectively. The standard error is an estimation of the standard deviation of the entire population from which the sample was drawn, and is equal to the standard deviation of the sample divided by the square root of the sample size. Hypothesis testing can be used to determine the statistical differences of these quantities, because overlapping standard error bars may indicate that the quantities are not significantly different, and if we computer the standard deviation from the standard error, we can then determine if the values are statistically different.

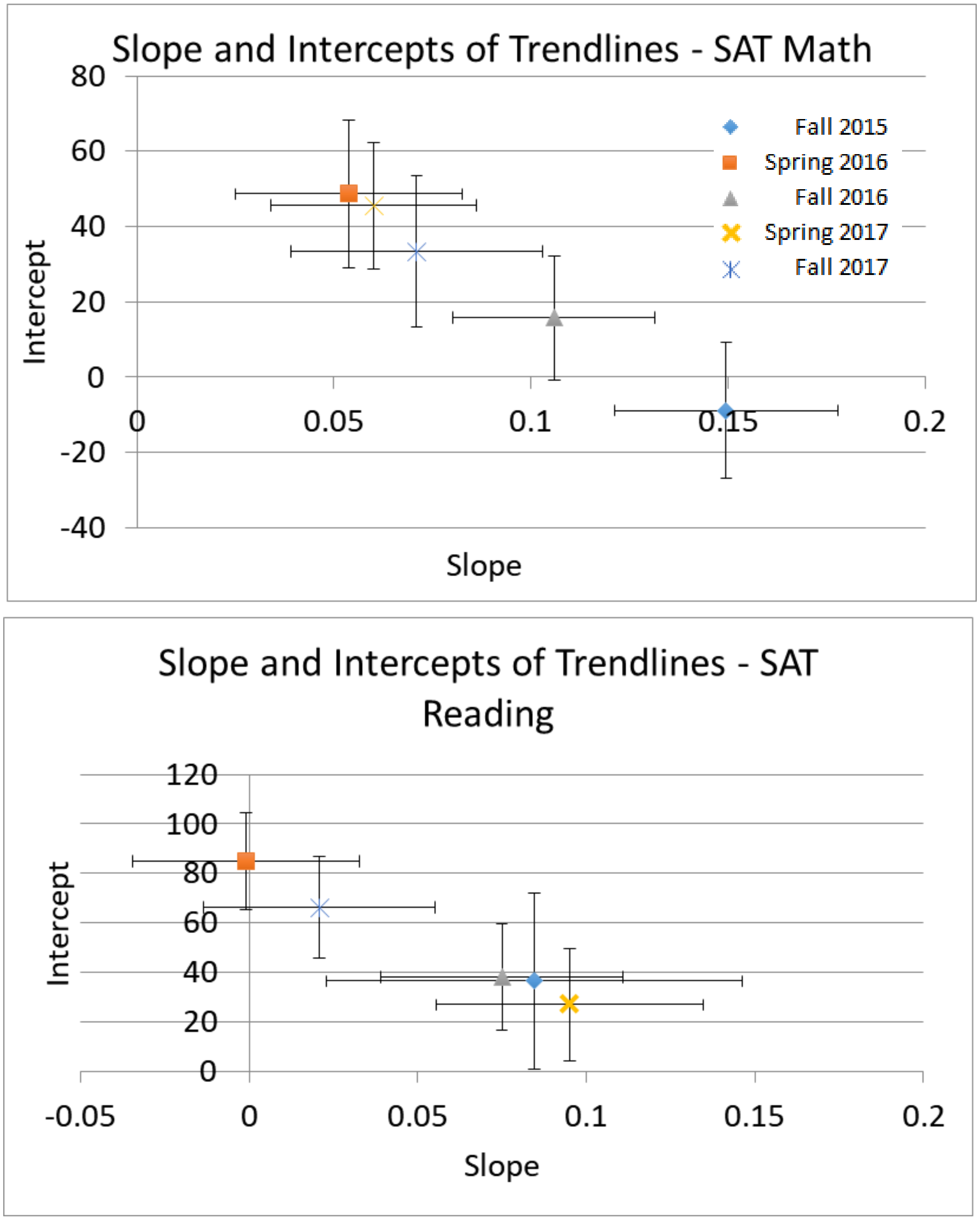
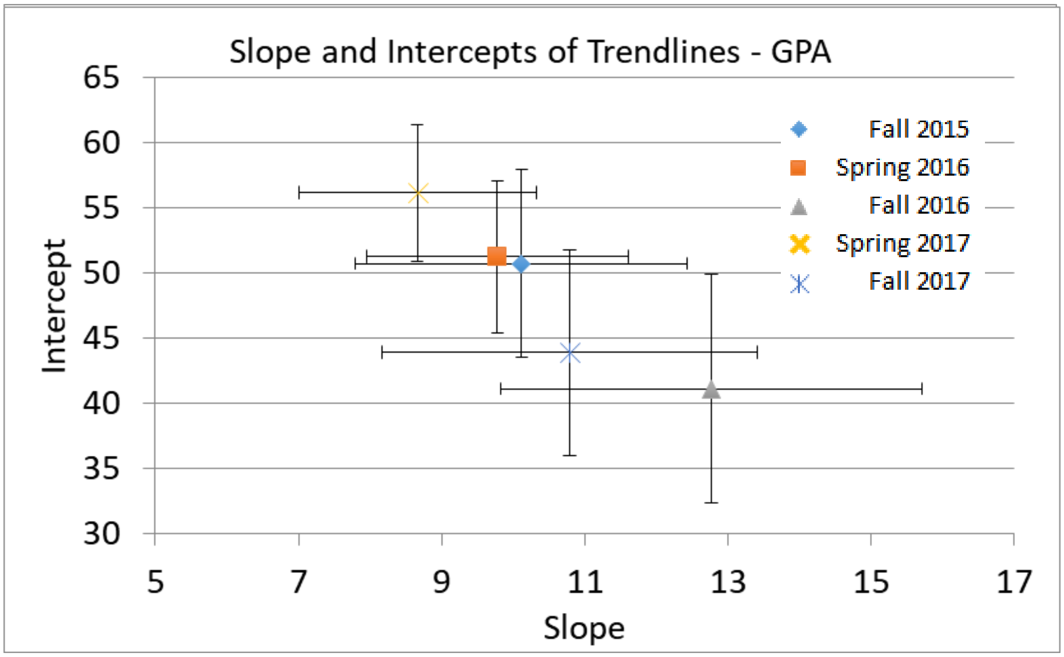


Figure 11 – A representation of single linear regression equations in successive semesters for Final Grade vs SAT Math (top) and Reading (bottom)



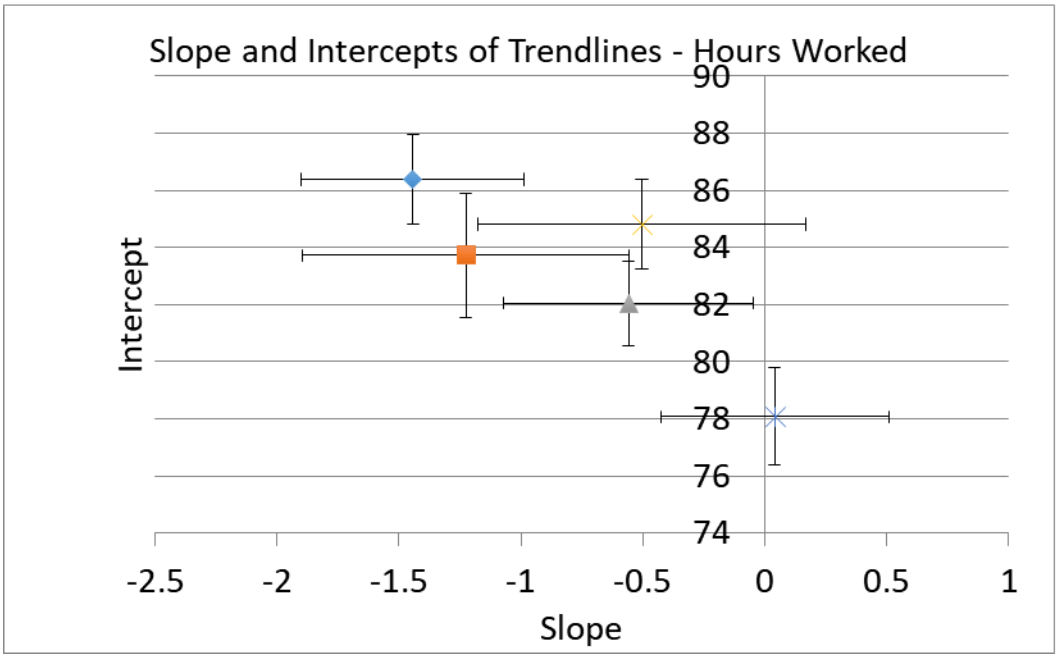


Figure 12 – A representation of single linear regression equations in successive semesters for Final Grade vs GPA (top) and Hours Worked (bottom)

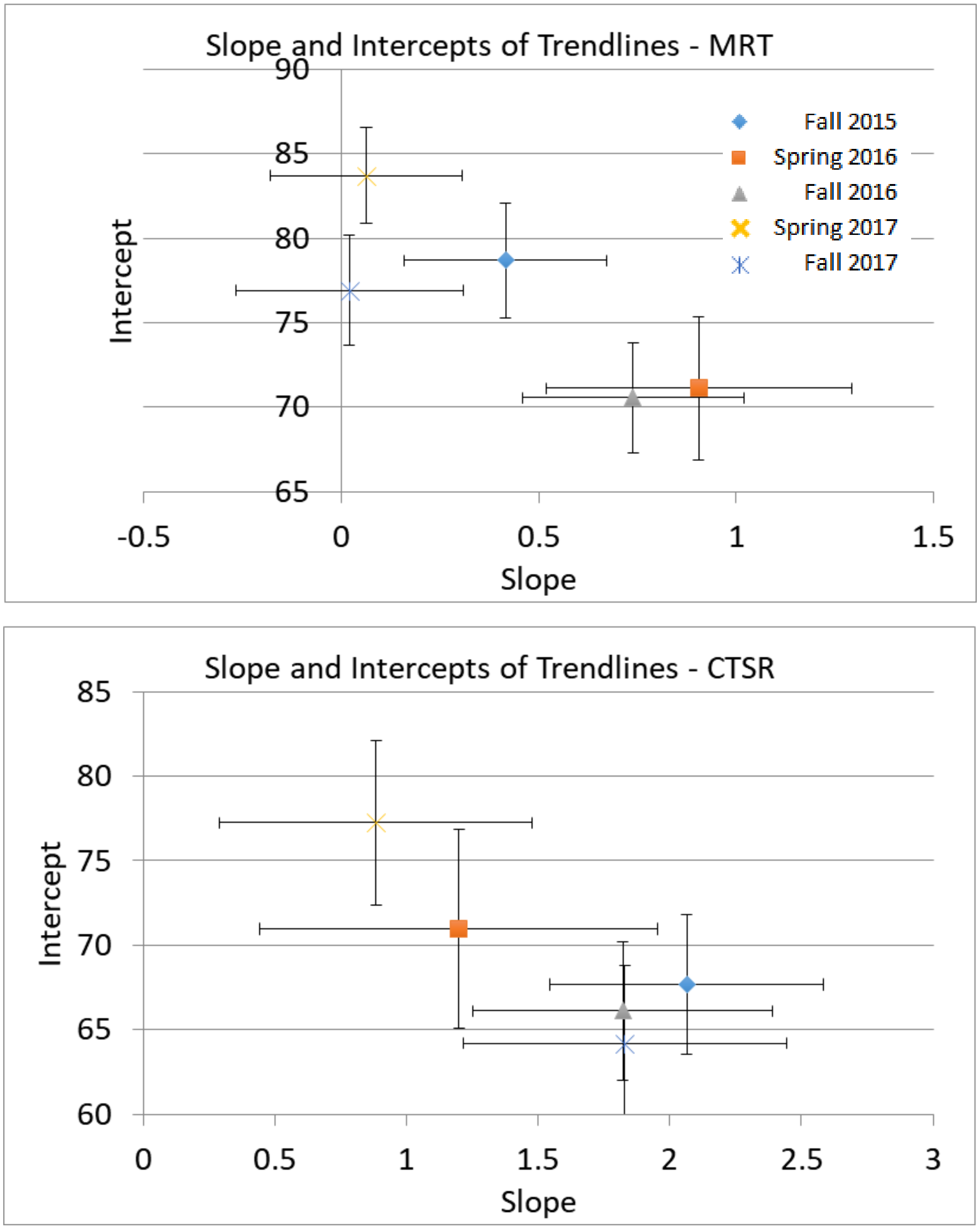
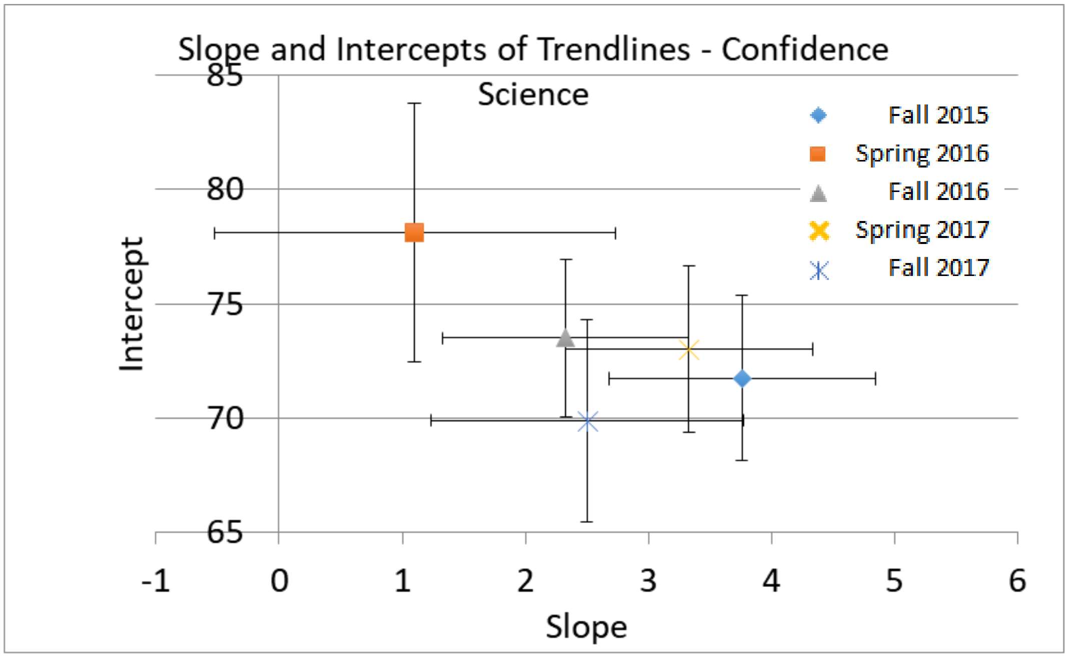


Figure 13 – A representation of single linear regression equations in successive semesters for Final Grade vs MRT (top) and CTSR (bottom)



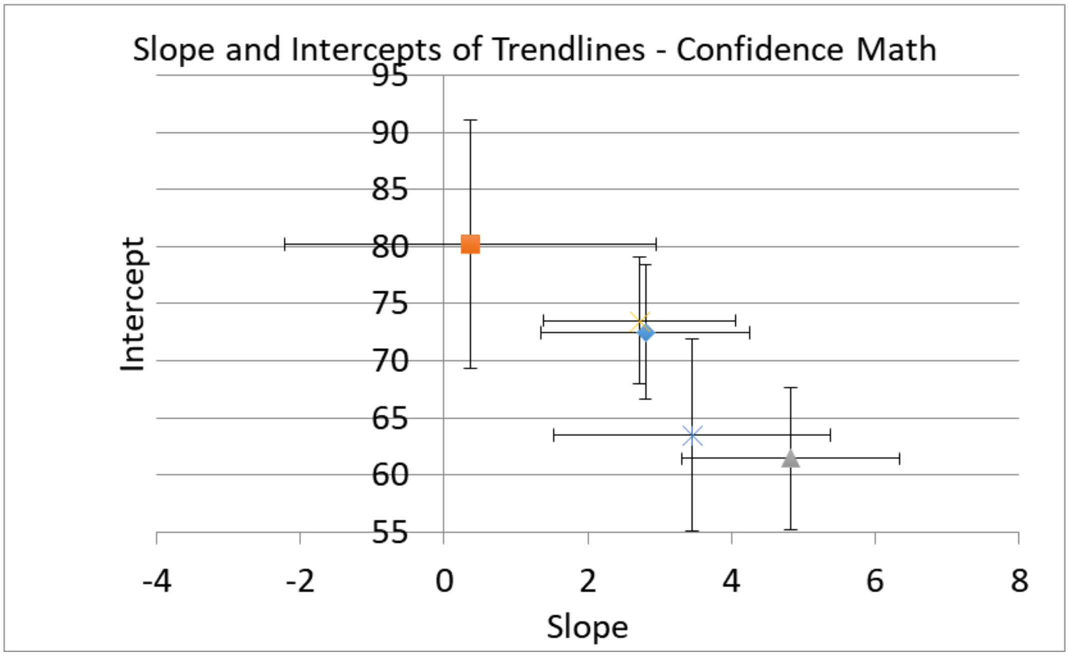


Figure 14 – A representation of single linear regression equations in successive semesters for Final Grade vs Confidence in Physics (top) and Math (bottom)

Figure 11 through Figure 14 show how various factors correlate to performance in the course throughout several semesters. Using a one-way ANOVA test, there were no significant differences in the slope or intercept of any factor. However, a visual inspection of the above figures shows that one or both spring semesters is usually on one extreme end of the figure in either slope, intercept, or both. In addition to the trends in the correlational metrics, below are the trends in the means and standard deviations of the same factors above from fall 2015 to fall 2017.

Factor	Metric	Fall 2015	Spr 2016	Fall 2016	Spr 2017	Fall 2017
SAT Math	Mean	635.52	674.17	640.00	640.00	629.68
	St.Dev	56.23	71.32	70.47	54.62	57.47
	N	29	24	30	36	31
SAT Read	Mean	576.96	588.18	605.71	580.36	592.14
	St.Dev	42.26	75.63	67.08	64.89	60.27
	N	23	22	28	28	28
GPA	Mean	3.07	3.09	2.96	3.04	2.94
	St.Dev	0.56	0.76	0.56	0.72	0.67
	N	44	50	61	62	53
MRT	Mean	12.09	9.79	10.50	10.75	10.21
	St.Dev	5.24	4.45	5.01	4.88	5.16
	N	76	52	112	60	85
CTSR	Mean	7.55	7.38	6.62**	7.84**	7.08
	St.Dev	2.38	2.54	2.50	2.02	2.23
	N	76	50	104	58	86

Hr Worked	Mean	1.99	1.65	1.71	1.05	2.21
	St.Dev	2.80	2.84	2.31	2.11	2.91
	N	67	39	73	43	63
Conf. Sci	Mean	3.13	3.26	3.23	3.43	3.34
	St.Dev	1.17	1.21	1.13	1.25	1.03
	N	67	39	74	44	64
Conf. Math	Mean	3.94	4.16	4.07	4.02	4.28
	St.Dev	0.92	0.79	0.73	1.00	0.68
	N	67	38	74	44	64

Table 28 – Trends in the means and standard deviations among successive semesters of eight factors that influence performance in physics

None of the means or standard deviations of the factors show any significant difference at the 95% CI with the exception of the CSTR mean between fall 2016 and spring 2017. This result somewhat agrees with Figure 11 through Figure 14, which show that the overall correlation to the performance in the course can vary to a small degree from semester to semester, with the exception of the spring semesters being different to a somewhat larger degree.

Multi-institutional differences in cognitive ability and affective beliefs of introductory physics students

Answering research question 2, the Mental Rotation Test (MRT) and Classroom Test of Scientific Reasoning (CTSR) were administered along with selected question from the Motivated Strategies of Learning Questionnaire (MSLQ) to students enrolled in UTA’s PHYS 1443, TCU’s PHYS 20474, and Yale’s PHYS 180. The section tested at TCU also included a small group of honors students enrolled in the same course, these students will be put into a separate group. These three courses are calculus-based physics courses covering Newtonian mechanics taken primarily by STEM majors. The following table summarizes the results from the fall 2017 semester. There will be some data missing because Yale did not administer the MRT to their students.

Factor	Metric	UTA	TCU / Honors	Yale
CTSR	Mean	7.09	7.39 / 9.44	9.95
	St Dev	2.25	1.88 / 1.88	1.49
	N	85	18 / 9	39
MRT	Mean	10.21	12.72 / 15.11	---
	St Dev	5.16	4.05 / 3.48	---
	N	85	18 / 9	---

Table 29 – Descriptive statistics of CTSR and MRT scores of physics students at three institutions in fall 2017

With respect to the CTSR, Yale students scored significantly different (higher) scores than UTA ($p<0.001$) and TCU regular ($p<0.001$), but not significantly higher than TCU honors ($p=0.383$). TCU honors students' mean is also significantly different from the TCU regular students' means ($p=0.0131$), and is significantly higher than UTA students' means ($p=0.0033$). TCU regular and UTA means were not significantly different ($p=0.599$).

With respect to the MRT, TCU honors students did have significantly higher means from UTA students ($p=0.003$), but TCU regular students did not have statistically different means from the honors ($p=0.144$) or UTA students ($p=0.0554$).

The MSLQ comprises 31 statements, and students report the extent to which they agree with the statement using a Likert scale where a 1 is strongly disagree, 4 is neutral, and 7 is strongly agree. There were 14 statements which showed significant difference in the response of UTA students and Yale students. The p-value associated with this difference will be labeled as p_YU. Furthermore, there response could be significantly different from 4 (the neutral response), and the p-values associated with this difference will be p_Y and p_U for Yale and UTA students respectively. The following table summarizes the significant responses:

Statement	Yale Mean	UTA Mean	Yale St. Dev	UTA St. Dev	Yale N	UTA N	p_YU	p_Y	p_U
5	3.98	4.87	1.563	1.361	105	70	0.000	0.901	0.000
6	3.762	4.814	1.638	1.618	105	70	0.000	0.139	0.000
11	4.21	5.17	1.759	1.633	105	70	0.000	0.205	0.000
15	4.067	4.886	1.607	1.565	105	70	0.001	0.672	0.000
16	4.667	5.214	1.567	1.667	105	70	0.031	0.000	0.000
17	4.314	5.029	1.695	1.711	105	70	0.007	0.060	0.000
20	4.267	5.171	1.416	1.383	105	70	0.000	0.056	0.000
21	4.429	5.343	1.44	1.35	105	70	0.000	0.003	0.000
24	3.952	4.543	1.583	1.63	105	70	0.019	0.759	0.007
26	4.562	5.214	1.664	1.559	105	70	0.009	0.001	0.000
28	4.429	3.829	1.885	1.849	105	70	0.039	0.022	0.441
29	4.505	5.1	1.395	1.571	105	70	0.011	0.000	0.000
30	4.413	4.886	1.909	1.915	105	70	0.013	0.445	0.000
31	4.505	5.271	1.395	1.424	105	70	0.001	0.000	0.000

Table 30 – A table of significantly different responses to MSLQ statements between UTA and Yale students in fall 2017. p_YU is the p-value associated with the difference between Yale and UTA students, and the other p-values are the different from 4, the neutral response.

Statement Number	Statement
5	I believe I will receive an excellent grade in this class
6	I'm certain I can understand the most difficult material presented in the readings for this course
11	The most important thing for me right now is improving my overall grade point average, so my main concern in this class is getting a good grade
15	I'm confident I can understand the most complex material presented by the instructor in this course

16	In a class like this, I prefer course material that arouses my curiosity, even if it is difficult to learn
17	I am very interested in the content area of this course
20	I'm confident I can do an excellent job on the assignments and tests in this course
21	I expect to do well in this class
24	When I have the opportunity in this class, I choose course assignments that I can learn from even if they don't guarantee a good grade
26	I like the subject matter of this course
28	I feel my heart beating fast when I take an exam
29	I'm certain I can master the skills being taught in this class
30	I want to do well in this class because it is important to show my ability to my family, friends, employer, or others
31	Considering the difficulty of this course, the teacher, and my skills, I think I will do well in this class

Table 31 – A key for reading Table 30

The first statement that showed a difference in response was “I believe I will receive an excellent grade in this class”. Yale students, on average, responded neutrally, while UTA students tends to agree with the statement more. In fact, UTA students consistently tended to agree more often than Yale students with nearly every statement except for “I feel my heart beating fast when I take an exam” in which the average response was neutral for UTA, whereas most Yale students agreed with the statement.

Changes in the factors as students continue in the program

In previous research by former graduate students, it was shown that scores on the MRT and CTSR can be influenced by either direct experimental intervention or just by taking additional physics courses (Ximena 2011). In this study we administered the MRT and CTSR to the introductory physics course, junior- and senior-level physics courses, and graduate-level courses in various semesters. The courses involved were PHYS 1443, the freshman level introductory calculus based mechanics course; PHYS 3313, a junior-level introductory course into modern physics; PHYS 3446, a junior-level particle physics course; PHYS 4319, a senior-level advance mechanics course, and PHYS 5306, classical mechanics, a graduate-level course. For 1443, data was averaged over five semesters (fall 2015 to spring 2018); for 3313, data was collected in spring 2018; for 3446, data was collected in spring 2016; for 4319, data was collected from spring 2015 and 2016; and for 5306, data was collected in fall 2015. The following table summarizes the means, standard deviations, and number of students for each class for the MRT and CTSR.

Test	**Metric**	**1443**	**3313**	**3446**	**4319**	**5306**
MRT	Mean	10.78	11.13	9.33	10.29	10.75
	St Dev	5.04	5.14	4.33	4.85	5.97
	N	402	40	12	28	8
CTSR	Mean	7.21*	8.71*	8.17+	8.50*+	8.00
	St Dev	2.36	2.32	1.20	2.20	2.07
	N	391	42	12	28	7

Table 32 – Descriptive statistics of MRT and CTSR scores from various undergraduate physics courses up to a graduate level course

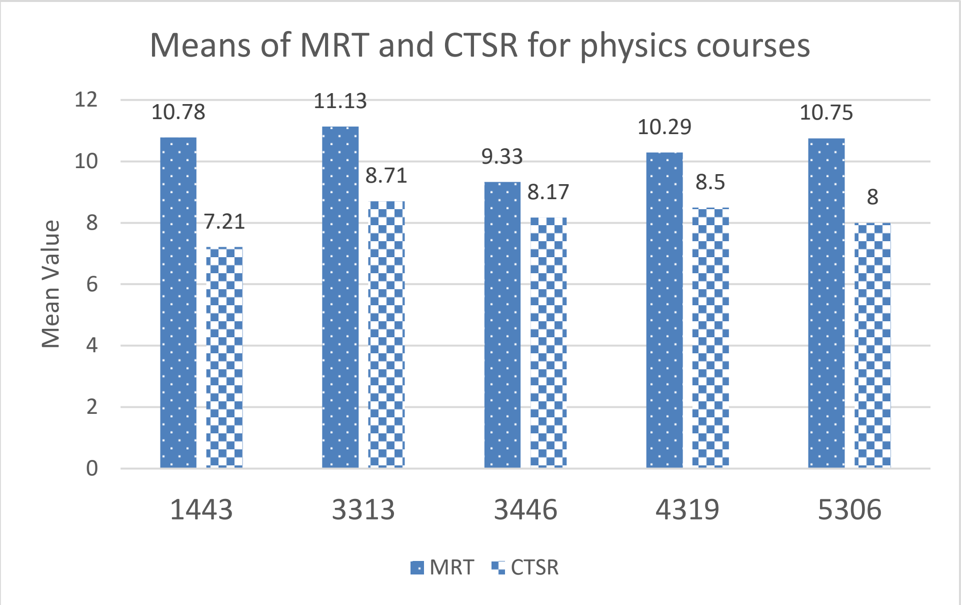


Figure 15 - A visual representation of the mean MRT and CTSR scores from various physics courses

In Table 32, there is no significant difference in the mean MRT scores among any course. This is likely due to the fact that most of the sample sizes presented are small, and are thus highly influence by extreme values. On the CTSR, the freshman-level course, 1443, had the lowest mean overall. This value, however, was only significantly different from 3313 (p=0.0007) and 4319 (p=0.00172), indicted by the * symbol. Interestingly, the next significant difference is between 3446 and 4319, the junior and senior level courses, indicated by the + sign (p<0.001). Strangely, the graduate student scores were not significantly different from any other courses.

Conclusions and Discussion

The goal of this study was to determine the trends in various factors that influence performance in physics over successive semesters and as a student advances in the program. These trends include the changes in the means, standard deviations, and correlations with performance. Furthermore, studying these trends lead to forming collaborations with other universities and investigating how measurements of these factors compare among the other students.

Trends in factors that influence performance in physics in a single course

Figure 11 through Figure 14 show the trends in the linear correlations with each factor to the performance in the course. Although there were no significant differences in the slope or intercept at the 95% CI, something about the spring semesters seem to affect the correlations among factors. With the only exception of number of hours worked per day, a spring semester can be found at either the lowest or highest value of slope or intercept. Because of this result, I would caution against drawing generalized conclusions from any intervention relying on linear regressions from a single semester, because there seems to be an effect where correlations are somewhat different in slope, intercept, and R^2 values (see Appendix B) from spring to fall semesters.

Overall, it does appear that the descriptive statistics of the factors identified remain consistent with each semester. Although most factors described did not have significantly different means and standard deviations, there was a significantly higher average CTSR score in the spring 2016 semester than the fall 2015 (Table 28). An increase in the mean values that does not significantly affect the correlations implies that there may be a more complicated relationship between that factor (and others) to performance in the course which is not well suited to be defined by regression. From a GSPT standpoint, it may seem that the threshold for using scientific reasoning is already being met in both the spring and fall semesters, and that the threshold for "passing" performance falls below the lowest mean CTSR score value. In other words, if the mean CTSR values continue to increase or decrease somewhat, it isn't likely that the correlations will change because there will be some other factor which is more influential on performance.

Multi-institutional differences in cognitive ability and affective beliefs of introductory physics students

When it comes to scientific reasoning, STEM major students enrolled in the same type of course, an introductory calculus based physics course, showed much different ability. In Table 29, Yale and TCU honors students scored significantly higher than their peers at TCU (regular) and UTA. The physics course taught at TCU was the same course, but was composed of honors and regular students. Additionally, regular TCU students were not significantly different from UTA students. One explanation for this finding is that the demographics of the respective institutions are different. UTA is a public, Hispanic-serving institution, which means at least 25% of enrolled students are Hispanic, and 15.8% are African American. TCU and Yale are private universities, and enroll 11% Hispanic students each, with 4.8% and 10% African American students respectively. (https://oir.yale.edu/sites/default/files/factsheet_2016-17.pdf) (http://www.ir.tcu.edu/factbooks/2016/student_data.asp). The demographics of the study body of a school will affect the demographics of the courses being taught, but in the physics classroom specifically, minorities are still underrepresented in general. Another possible reason for the difference in scores is that students at Yale and TCU are already more likely to have been selected based on their academic performance. This is also supported by the fact that TUC honors students scored higher on average than the regular students, but doesn’t necessarily explain how regular TCU physics students performed the same as UTA physics students on average.

With regard to the MRT, TCU honors students again performed on average better than UTA students, but this time not significantly better than their “regular” peers.

Like scientific reasoning ability, individuals who were selected to join the honors college may have been selected on the basis of prior academic performance. This supports the idea that mental rotation ability and scientific reasoning ability may be influential to academic performance, or at least should be considered in the process of determining academic ability or performance. Since Yale did not participate in the MRT, no comparison can be made for those students.

The MSLQ was given to Yale and UTA students, and Table 30 and Table 31 show the results of statements that had a significant difference in levels of agreement. In general, UTA students tended to agree more to statements reflecting self-confidence and high expectations than Yale students who were more neutral. This could reflect a difference in mindset and patterns of thinking for students enrolled in physics courses at these institutions. It may be possible that students attending UTA expect to enjoy their classes and they expect that that they will learn what they need to know in order to be successful, whereas Yale students may be tougher on themselves and expect much more difficult experiences. One interesting exception to this trend is that Yale students tended to agree with the statement “I feel my heart beating fast when I take an exam” more than UTA students. This may reflect a more stressful testing environment at Yale where high academic achievement is required just to be considered for admission.

Changes in the factors as students continue in the program

Previous work in studying trends in cognitive ability have shown that mental rotation ability increases from high school to college (Hinojosa 2015) and that taking engineering or physics courses also tend to increase such ability (Cid 2011). It is then natural to study the progression of mental rotation ability and scientific reasoning skill from freshman-level physics courses up to graduate-level courses. When doing so, it’s important to keep in mind that these students are not the same students that move up, but if the upper-division students follow the same trends as reported in this chapter, then it's safe to assume that each course is a fair representation of the course for any semester. It's also important to note that the findings in this chapter were from MRT and CTSR exams administered early in the semester. It has been shown that there is often improvement in the scores during a test/retest type of intervention, and out data matches up to previous work's "pre" scores. The results in Table 32, visualized in Figure 15, show no significant differences in the mean values of MRT scores among students from PHYS 1443, the introductory physics course, up until PHYS 5306, a graduate course in classical mechanics. There was, however, a significant increase in scientific reasoning ability from 1443 to PHYS 3313, a junior-level physics course in modern physics. The next increase in scientific reasoning then comes from the junior-level particle physics, 3446, to the senior-level advanced mechanics, 4319, which is the last physics course a senior might take. It seems logical that scientific reasoning skills increase as you take more physics courses, but it may also be possible that students who were already scientifically minded persisted through the program and were more capable of taking on the additional challenged associated with higher level physics courses. This claim can be tested by tracking individual students from 1443 onward. This analysis was completed to a small extent by looking at students who participated in the study from 2015 and are now juniors in modern physics by 2018, but there were too few individuals to make any meaningful conclusions.

Chapter 6

Summary and future research direction

This dissertation represents my contribution to the field of physics education research (PER). As classrooms change and technology improves, there is an increasing trend to incorporate primary data into not only PER research methodology but also pedagogy and classroom management. The challenges of PER arise because humans beings to not behave like protons in a vacuum; we are all unique and have complicated motivations, expectations, and backgrounds. This doesn't mean that quantitative research can't exist, it just means that one must be more careful when interpreting the results of such a project. I believe that the strength of my contribution lies with the consistent statistical methodology used throughout it, and that it may be used again in the future as a link in the chain of another project.

There were four major parts to this study. The first was determining the impact of flipped classrooms on upper-division and graduate students. The second and third parts involved using primary, empirical data to developed traditional and nontraditional predictive models of academic performance, and to determine the prediction efficiency of such models. Finally, an investigation of how factors that influence performance in physics change not just from semester to semester, but also from course to course as students progress in their major, and compare them to physics students at other institutions.

In the flipped classroom, the major finding was that students found the active component of the course (small group discussions, group problem solving, notecard feedback, etc.) was uniformly seen as beneficial to their learning, regardless of their feelings toward the flipped class itself. Some students reported that they looked forward to taking more flipped classes, and some did not enjoy it, but seniors and graduate students like both considered the active, in-class portions to continue more to their understanding than traditional lectures. Furthermore, pass rates on the graduate qualifying exam increased to 100% for the section which was taught in a flipped and active format, while other traditionally taught sections remained the same. Future research in the space of studying pass rate of the qualifying exam or other measures of graduate student success, would likely involve developing a concept inventory or another assessment and investigate if the increase in pass rate was due to increased knowledge retention due to active learning classes. Focusing more on the seniors, one might be interested in what students do after graduating, and if performing well in a flipped classroom influences their decisions to pursue graduate school.

Next, developing a reliable and robust predictive models of student performance would be an absolute revolution in the education industry. Identifying students at risk of failing the course would allow educational professionals to target such individuals with interventions or support to help them succeed. I don’t believe that any model will be 100% efficient, but I believe finding the right combination of known factors may be a step in the right direction. To find this combination, factors that are known to influence performance in physics courses can be correlated to such performance. This type of investigation isn’t new by any definition, and complicated human behaviors make this analysis that much more difficult, but it allows us to draw a baseline for trying more exotic statistical and quantitative methodologies. Although the linear and nonlinear models developed in this project didn’t produce consistent predictions in general, there are some that show promise, and perhaps just need to be refined and given more time to make more predictions. Additionally, quantitative methodology from GSPT arrived at similar conclusions to the traditional methodology, which does inspire me to continue working on GPST models in an educational environment. Future work in both of these approaches include identifying more factors that may influence performance in physics courses, expanding into other physics courses, developing more robust RDFs, and perhaps even considering two (or more) different models for different semesters. So much information can be learned from empirical sets of data that it is possible that a human can't identify certain patterns which may exist, so I do have plans to apply principals of machine learning and deep learning to uncover potentially hidden patterns in the data.

Finally, monitoring trends in student ability is of great importance to education and school administrators because it can offer one unique perspective of student learning or achievement. I believe that most teachers can get a “feel” for how well their students have been prepared for the course after only a few weeks, so one motivation for making comparisons from semester to semester of the same course is to provide quantitative records of the class demographics, cognitive ability, and affective beliefs. I must also assert that this analysis should be done by an independent party, as not to bias the instructor. In this study, we found that the means and correlations to performance of various factors remained consistent from one semester to another. Comparing ability among different courses, however, can give insight to how the students are growing mentally, or even what skills are required to make it to the senior year of a physics program. This study shows that there are significant increases in scientific reasoning ability as students move up in the program from some courses to another, but there may eventually be an upper limit which is reached around the junior year.

Institutional differences in physics students may be of interest to recruiters, marketers, and any other individuals who have a stake in comparing student bodies. In this study, it was found that, across similar calculus based introductory physics courses for STEM majors, Yale and TCU honors students shows higher scientific reasoning ability than UTA and TCU regular students. Yale students, however, seem to have different expectations of success than UTA students, and may be under more pressure to be academically successful. Future research in the area of longitudinal, vertical, or institutional comparisons will always benefit from having larger datasets, longer time periods, and additional intuitions. At a joint meeting of APS and Zone 13 of the SPS where some of these findings were presented, it was suggested that universities from state-wide conferences could submit their own data for a project entirely dedicated to these institutional studies. Our collaboration with Yale and TCU could even evolve to have those places collect vertical data on these factors from different courses, and make multi-intuitional vertical comparisons. The next logical step would then be to continuously collect this data over several semesters, to make an extremely comprehensive multi-institutional longitudinal and vertical study of trends in factors that influence performance in physics.

Appendix A

Surveys

Flipped Classroom Survey 1

Throughout the semester, you engaged in what's known as a "flipped classroom" where you review lecture notes and videos before coming to class, and during class you worked on solving problems and deepening your understanding. Please read the following statements and indicate the extent to which you agree with the statement by putting a checkmark in the appropriate box.
Your responses will be kept confidential and will not be read until after final grades have been officially posted.

Name: ________________________________

Statement	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
I watched every lecture video before class					
I watched most of the lecture videos, but skipped a few					
I watched the lecture videos as soon as they were posted					
I dedicated specific times in the week for watching the videos					
I watched the lecture videos only on the weekends					
I watched the group of posted lecture videos all in one session					
I spread out watching the lecture videos over more than three days					

Statement	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
While watching the videos, I took notes on paper					
While watching the videos, I also followed along in the textbook					
While watching the videos, I paused the videos to stop and think					
While watching the videos, I would frequently go back a few minutes to listen again					
After watching the entire video, I would rewatch it within one hour					

Before an exam, I rewatched some of the lecture videos					
Watching the videos more than once helped me understand some topics.					

Statement	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
I enjoyed this flipped classroom					
I learned less from this flipped class compared to what I learn in traditionally taught classrooms					
I learned more from this flipped class compared to what I learn in traditionally taught classrooms					
I would enjoy taking additional “flipped” courses					

Statement	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
Solving problems in class was helpful for my understanding of the topics					
Working in groups during class helped me understand the material better than working on my own					
The additional discussions and clarifications in the classroom sessions were important to understanding the material					
I enjoyed discussing conceptual questions in class that were asked in the lecture videos					

Flipped Classroom Survey 2

The recorded lecture videos were helpful for my understanding of the topics.
1 – Strongly Disagree
2 – Disagree
3 – Neutral
4 – Agree
5 – Strongly Agree

Watching the recorded lecture videos helped me understand the material better than a traditional lecture in class.
1 – Strongly Disagree
2 – Disagree
3 – Neutral
4 – Agree
5 – Strongly Agree

Student demographics and academic preparedness survey

Demographic, Academic Preparedness, and Course Opinions Survey
This survey is only to be completed by participants who have read and signed the informed consent document for:
IRB Protocol Number: **2016-0122**
IRB Protocol Title: *Using general systems performance theory to predict student success in introductory physics*
Participant name: (please print)____________________________________
Your name will only be used to verify your consent to participate in this research project, and then removed and replaced by a code known only to the principal investigator. When completed, please turn in this survey to the folder provided. This folder will be sealed and not opened until after final grades have been issued. Therefore, this survey will have no impact on any part of your grade.
For the following questions, please answer honestly and accurately. Each question should be considered voluntary and optional, you are not required to answer every question. If you cannot remember the answer to a question, or are choosing not to answer a question, you may leave it blank and skip it.

1. What is your age? ________

2. What is your gender? (please circle one)
   a) Male
   b) Female
   c) Other
   d) Prefer not to answer

3. What is your primary language? (please circle one)
   a) English
   b) Spanish
   c) Chinese
   d) Korean
   e) Hindi
   f) Vietnamese
   g) Thai
   h) Other
   i) Prefer not to answer

4. What is the highest degree or level of school you have completed? (please circle one)
   a) Some high school, no diploma
   b) High school graduate, diploma or the equivalent (for example: GED)
   c) Trade/technical/vocational training
   d) Some college credit, no degree
   e) Associate degree
   f) Bachelor's degree
   g) Other
   h) Prefer not to answer

5. Approximately how many inches is the longest hair on your head? ______

6. On average, how many hours of sleep do you get per night? ______

7. How many hours do you spend per day at a job off campus? ______

8. Do you live on, or within walking distance of, campus? (please circle one) Yes
   No

9. On average, how many hours per day do you spend traveling to and from school?
   ______

10. Approximately how many hours per day do you spend on *non-academic* activities such as hobbies and being with friends and family? ______

11. Subtracting time spent sleeping, working a job, walking to and from school, and engaging in non-academic activities, how much time is *available to you* per day to spend on academic activities outside of the classroom such as doing homework, reading textbooks, studying for tests, etc.? ______

12. Approximately how much of the time available to you do you *actually* spend on academic activities per day? ______

13. How many hours per week do you spend in a structured supplemental academic environment (not the classroom) such as Supplemental Instruction, Peer Tutoring, Tutoring Clinic, etc. ______

14. How many hours per week do you spend studying one subject? ______

15. On average, how many of those hours are spent with one or more other people?
    ______

16. How many physics courses did you take in high school? ______
17. Beyond Algebra II, how many math courses did you take in high school? ______
18. How many other physics courses have you taken in college? (do not include this course) ______

19. How many math courses have you taken in college? (include any you are currently enrolled in) ______
20. Have you completed a precalculus course? (circle one) YES NO
21. Have you completed a calculus I course? (circle one) YES NO
22. Have you completed a calculus II course? (circle one) YES NO
23. Have you completed a calculus III course? (circle one) YES NO

If you can remember the following information, please provide it. Remember that responding to any question should be considered voluntary and optional. Every response to this survey will be kept confidential and anonymous.

24. High school GPA: ________ out of ________

25. SAT Score: ________ out of ________

26. ACT Score: ________ out of ________

27. Have you ever received a grade of D or F in a college level mathematics course? (circle one)

YES NO

28. Have you ever received a grade of D or F in a college level science course? (circle one)

YES NO

29. Have you ever withdrawn from a college level mathematics course due to academic reasons? (circle one) YES NO

30. Have you ever withdrawn from a college level science course due to academic reasons? (circle one) YES NO

Please mark the box which best reflects your opinion on the following statements

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I had a good understanding of physical concepts before taking this class					
I have a good understanding of physical concepts after taking this class					
Before taking this class, I was confident in my mathematical ability					
After taking this class, I am confident in my mathematical ability					
I was very interested in taking this class before the start of the semester					

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Class time was used effectively					
Because of my textbook, I was able to understand most physical concepts					
Because of my instructor, I was able to understand most physical concepts					
Attending classes is essential for understanding physics					
My instructor played an important role in my understanding of physics					
I mainly taught myself physics					

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Taking this class has improved my mathematical ability					
I can apply concepts from this course to the real world					
Taking this course improved my ability to analyze questions and problems inside the classroom					
Because of this class, I am able to make logical connections between abstract ideas					
I find myself applying physical concepts to situations outside of classes					
Taking this course improved my ability to analyze problems in the real world					
After taking this class, I better understand the world around me					
After taking this class, I can identify underlying physical concepts behind everyday phenomenon					

I think more scientifically having taken this class					

In a few sentences, please describe your reasons for taking this course, and your motivations for being successful in this course:

____

____

____

____

____

Appendix B

Data for Figure 11 through Figure 14

SAT Math	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	0.1495	0.0283	<0.001	-8.9	18.1	0.626	50.77%	28
Spring 2016	0.0538	0.0289	0.077	48.7	19.6	0.022	14.80%	21
Fall 2016	0.1059	0.0256	<0.001	15.8	16.5	0.346	38.79%	28
Spring 2017	0.0601	0.0261	0.028	45.6	16.8	0.011	14.20%	33
Fall 2017	0.071	0.032	0.035	33.4	20.2	0.109	14.92%	30
SAT Reading	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	0.0845	0.0617	0.185	36.5	35.7	0.318	8.20%	22
Spring 2016	-0.001	0.0335	0.976	84.9	19.7	<0.001	0.01%	19
Fall 2016	0.075	0.0359	0.047	38.1	21.7	0.091	14.85%	26
Spring 2017	0.0951	0.0395	0.024	26.9	22.8	0.248	18.84%	26
Fall 2017	0.0208	0.0344	0.552	66.3	20.5	0.003	1.43%	27
GPA	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	10.12	2.32	<0.001	50.72	7.22	<0.001	31.26%	43
Spring 2016	9.78	1.83	<0.001	51.25	5.83	<0.001	39.37%	45
Fall 2016	12.77	2.94	<0.001	41.11	8.77	<0.001	24.60%	59
Spring 2017	8.66	1.66	<0.001	56.16	5.25	<0.001	33.45%	55
Fall 2017	10.79	2.62	<0.001	43.9	7.9	<0.001	25.40%	52
MRT	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	0.416	0.257	0.109	78.68	3.38	<0.001	3.43%	75
Spring 2016	0.905	0.387	0.024	71.12	4.23	<0.001	11.27%	44
Fall 2016	0.739	0.28	0.01	70.55	3.28	<0.001	6.15%	107
Spring 2017	0.063	0.244	0.799	83.7	2.83	<0.001	0.12%	54
Fall 2017	0.02	0.288	0.945	76.91	3.24	<0.001	0.01%	83
CTSR	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	2.064	0.519	<0.001	67.67	4.11	<0.001	17.62%	75
Spring 2016	1.199	0.755	0.12	70.96	5.86	<0.001	5.79%	42
Fall 2016	1.821	0.571	0.002	66.1	4.08	<0.001	9.40%	99
Spring 2017	0.884	0.594	0.143	77.25	4.85	<0.001	4.16%	52
Fall 2017	1.829	0.614	0.004	64.2	4.56	<0.001	9.77%	84

Hours Worked	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	-1.443	0.457	0.002	86.38	1.56	<0.001	13.33%	66
Spring 2016	-1.226	0.668	0.074	83.72	2.17	<0.001	8.35%	38
Fall 2016	-0.559	0.513	0.279	82.05	1.47	<0.001	1.65%	72
Spring 2017	-0.505	0.672	0.456	84.81	1.57	<0.001	1.36%	42
Fall 2017	0.041	0.467	0.93	78.08	1.7	<0.001	0.01%	63
Confidence Science	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	3.76	1.08	0.001	71.73	3.61	<0.001	15.67%	66
Spring 2016	1.1	1.63	0.504	78.1	5.66	<0.001	1.22%	38
Fall 2016	2.33	1	0.023	73.49	3.43	<0.001	6.98%	73
Spring 2017	3.33	1	0.002	73.02	3.66	<0.001	20.77%	43
Fall 2017	2.5	1.27	0.053	69.9	4.43	<0.001	5.90%	64
Confidence Math	**Slope**	**SE Slope**	**p-value**	**Intercept**	**SE Intercept**	**p-value**	**R^2**	**N**
Fall 2015	2.8	1.45	0.059	72.48	5.88	<0.001	5.39%	66
Spring 2016	0.37	2.58	0.886	80.2	10.9	<0.001	0.06%	37
Fall 2016	4.82	1.51	0.002	61.43	6.24	<0.001	12.38%	73
Spring 2017	2.72	1.34	0.049	73.52	5.56	<0.001	8.89%	43
Fall 2017	3.45	1.93	0.079	63.49	8.37	<0.001	4.89%	64

References

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology, 32*(4), 665-683.

APS Office of Public Affairs. *Research in physics education.* Retrieved October 26, 2016, from

Barthelemy, R. S., Van Dusen, B., & Henderson, C. (2015). Physics education research: A research subfield of physics with gender parity. *Physical Review Special Topics-Physics Education Research, 11*(2), 020107.

Beichner, R. J. (2009). An introduction to physics education research. *Getting Started in Per, 2*(1), 1-25.

Bonwell, C. C. (1996). Enhancing the lecture: Revitalizing a traditional format. *New Directions for Teaching and Learning, 1996*(67), 31-44.

Cid, X. C. C. (2011). *Investigations in the impact of visual cognition and spatial ability of student comprehension of physics and space science* Physics.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*(6031), 862-864.

Ding, L., Liu, X., & Harper, K. (2012). Getting started with quantitative methods in physics education research. *Getting Started in PER—Reviews in PER, Edited by C.Henderson and KA Harper (American Association of Physics Teachers, College Park, MD, 2012), Http://Www.Per-Central.Org/Items/Detail.Cfm,*

Elby, A., & Hammer, D. (2010). Epistemological resources and framing: A cognitive framework for helping teachers interpret and respond to their students' epistemologies. *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice,* , 409-434.

Flynn, A. B. (2015). Structure and evaluation of flipped chemistry courses: Organic & spectroscopy, large and small, first to third year, english and french. *Chemistry Education Research and Practice, 16*(2), 198-211.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics.*Proceedings of the National Academy of Sciences, 111*(23), 8410-8415.

Frost, J. (2014). *Why are there no P values for the variables in nonlinear regression?*http://blog.minitab.com/blog/adventures-in-statistics-2/why-are-there-no-p-values-for-the-variables-in-nonlinear-regression

Fulton, K. (2012). Upside down and inside out: Flip your classroom to improve student learning. *Learning & Leading with Technology, 39*(8), 12-17.

Goodwin, B., & Miller, K. (2013). Evidence on flipped classrooms is still coming in. *Educational Leadership, 70*(6), 78-80.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2007). The new instructional leadership: Creating data-driven instructional systems in school. *Journal of School Leadership, 17*(2), 159.

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics, 68*(S1), S59.

Hawkins, C. A., Smith, M. L., Hawkins, I., Raymond C, & Grant, D. (2005). The relationships among hours employed, perceived work interference, and grades as reported by undergraduate social work students. *Journal of Social Work Education, 41*(1), 13-27.

Herreid, C. F., & Schiller, N. A. (2013). Case studies and the flipped classroom. *Journal of College Science Teaching, 42*(5), 62-66.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158.

Hinojosa, A. J. (2015). *Investigations on the impact of spatial ability and scientific reasoning of student comprehension in physics, state assessment tests, and STEM courses* The University of Texas at Arlington.

Keith, T. Z. (1982). Time spent on homework and high school grades: A large-sample path analysis. *Journal of Educational Psychology, 74*(2), 248.

Koch, A., & Eckstein, S. G. (1995). Skills needed for reading comprehension of physics texts and their relation to problem-solving ability. *Journal of Research in Science Teaching, 32*(6), 613-628.

Kondraske, G. V. (1988). Experimental evaluation of an elemental resource model for human performance. Paper presented at the *Engineering in Medicine and Biology Society, 1988. Proceedings of the Annual International Conference of the IEEE,* pp. 1612-1613.

Kondraske, G. V. (2011). General systems performance theory and its application to understanding complex system performance. *Information Knowledge Systems Management, 10*(1-4), 235-259.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science,* Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education, 31*(1), 30-43.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11-24.

Libarkin, J. (2008). Concept inventories in higher education science. Paper presented at the *BOSE Conf,*

Lopez, R. E., Hernandez, S., Wiltberger, M., Huang, C., Kepko, E. L., Spence, H., et al. (2007). Predicting magnetopause crossings at geosynchronous orbit during the halloween storms.*Space Weather, 5*(1)

Mazur, E. (1997). Peer instruction: Getting students to think in class. Paper presented at the *AIP Conference Proceedings, , 399.* (1) pp. 981-988.

McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today, 37*, 24-32.

McDermott, L. C. (2001). Oersted medal lecture 2001:“Physics education research—the key to student learning”. *American Journal of Physics, 69*(11), 1127-1137.

Meltzer, D. E. (2002). The relationship between mathematics preparation and conceptual learning gains in physics: A possible “hidden variable” in diagnostic pretest scores. *American Journal of Physics, 70*(12), 1259-1268.

Meyers, C., & Jones, T. B. (1993). *Promoting active learning. strategies for the college classroom.* ERIC.

Paul, A. M. (2015). Are college lectures unfair? *The New York Times, 9*, 12.

Pintrich, P. R. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ).

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223-231.

Rait, R. S. (2012). *Life in the medieval university* Cambridge University Press.

Rice, J. M. (1893). *The public-school system of the united states* New York: Century.

Roehl, A., Reddy, S. L., & Shannon, G. J. (2013). The flipped classroom: An opportunity to engage millennial students through active learning. *Journal of Family and Consumer Sciences, 105*(2), 44.

Sadler, P. M., & Tai, R. H. (2001). Success in introductory college physics: The role of high school preparation. *Science Education, 85*(2), 111-136.

Spiess, A., & Neumeyer, N. (2010). An evaluation of R 2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A monte carlo approach. *BMC Pharmacology, 10*(1), 6.

Stanski, H. R., Wilson, L. J., & Burrows, W. R. (1989). Survey of common verification methods in meteorology. *World Weather Watch Tech, 358*(Technical Report No. 8), 114.

Trockel, M. T., Barnes, M. D., & Egget, D. L. (2000). Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors.*Journal of American College Health, 49*(3), 125-131.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47*(2), 599-604.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817.

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2010). Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: A 25-year longitudinal study. *Journal of Educational Psychology, 102*(4), 860.

Welch, B. L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35.

Zappe, S., Leicht, R., Messner, J., Litzinger, T., & Lee, H. W. (2009). Flipping" the classroom to explore active learning in a large undergraduate course. Paper presented at the *American Society for Engineering Education.*

Biographical Information

Michael A Greene was born in Bradenton, Florida and received his Bachelors of Science in Science Education with a concentration in Physics from Florida Institute of Technology in Melbourne, Florida in 2011. He taught at a nearby nationally-recognized college-prep high school for three years until moving to pursue his PhD in physics at The University of Texas at Arlington.

Michael is interested in an academic career in Physics Education Research, but is also interested in applying the quantitative and qualitative statistical skills he developed at UTA into the field of data science or finance, especially if it’s in an educational context.