2023

Methods for Analyzing Physics Student Retention and Physics Curricula

John Darrell Hansen
jdh0079@mix.wvu.edu

Methods for Analyzing Physics Student Retention and Physics Curricula

John D. Hansen

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Physics

John Stewart, Ph.D., Chair
Gay Stewart, Ph.D.
Edward Flagg, Ph.D.
Matthew Campbell, Ph.D.

Department of Physics and Astronomy

Morgantown, West Virginia
2023

Keywords: physics, education, conceptual inventories, retention, Bayesian Network, curriculum

ABSTRACT

Methods for Analyzing Physics Student Retention and Physics Curricula

John D. Hansen

Retention of students in college has been a concern of academic institutions for many years. In the last two decades, the focus on student retention in STEM fields has intensified. The current graduation rate of students in science, technology, engineering and mathematics (STEM) fields is well below that required to fill the projected need of STEM professionals. The work presented in this dissertation investigates the problem of student retention in physics programs. Four studies were performed. The first identifies the relationships between student retention and pre-college and early-college academic factors at an eastern U.S. university using logistic regression and Bayesian networks. The second uses Bayesian networks to predict the outcomes of physics course grades, using prior physics and math course grades as evidence, to assist academic advisors and physics departments as they help students progress through their physics curriculum. The third part investigates the complexity of physics curricula at 60 U.S. institutions using Curricular Analytics and compares the differences in complexity of programs with different national rankings. The final part evaluates a common physics conceptual assessment to determine the structure of knowledge the assessment measures; assessments that accurately measure student knowledge in physics are essential in designing courses and programs that successfully train future STEM professionals.

ACKNOWLEDGMENTS

Firstly, I would like to acknowledge and thank Dr. John Stewart, the chair of my PhD committee. He has worked tirelessly to help me grow as a student, researcher, and teacher. You are the consummate example of a professional academic, and working with you has been a pleasure.

I would also like to recognize the other members of my committee: Gay Stewart, Matthew Campbell, and Ned Flagg. Thank you for your time and efforts to make me a better academic. From letters of recommendation to lectures on introductory Quantum Mechanics, I am very grateful for all you have done for me.

Although not directly involved in this work, I would like to recognize my siblings. I look up to each of you, and to say that you have not influenced me and my work would be a lie. Thank you, ya' filthy animals!

I would like to thank my parents for their support and encouragement, and three decades of lessons in determination, resiliency, faithfulness, and love. You are the most influential people in my life. Thank you.

Finally, I would like to thank my wife, Kaitlin. You moved 2,000 miles away from your family to follow me on this adventure. You have worked part-time, all while caring for our children full-time, to help support us. You have lifted me up every time I have been discouraged. I can not express how grateful I am for you. I love you.

"Now therefore, *O God*, strengthen my hands." Nehemiah 6:9 KJV

Contents

1 Introduction to Physics Education Research **1**
- 1.1 Introduction . . . 2
- 1.2 Conceptual Understanding . . . 3
  - 1.2.1 Conceptual Inventories . . . 3
  - 1.2.2 Normalized Gain . . . 4
- 1.3 Research-Based Instructional Strategies . . . 6
  - 1.3.1 Lecture-based Strategies . . . 6
  - 1.3.2 Recitation-based Strategies . . . 7
  - 1.3.3 Laboratory-based Strategies . . . 7
  - 1.3.4 Classroom Environment-Based Strategies . . . 8
  - 1.3.5 General Instructional Materials . . . 9

2 Student Retention and Educational Data Mining **11**
- 2.1 Retention . . . 12
  - 2.1.1 Physics retention . . . 12
  - 2.1.2 General college retention . . . 14
  - 2.1.3 STEM retention . . . 15
  - 2.1.4 Physics course success . . . 16
- 2.2 Educational Data Mining . . . 18

3 Statistical Methods **20**
- 3.1 Descriptive Statistics . . . 21
  - 3.1.1 Measures of Central Tendency . . . 21
  - 3.1.2 Variability . . . 22
- 3.2 Inferential Statistics . . . 23
  - 3.2.1 Hypothesis Testing . . . 24
  - 3.2.2 Effect Size . . . 25
  - 3.2.3 Error . . . 26
  - 3.2.4 Beyond Significance Testing . . . 26
  - 3.2.5 Boot-strapping . . . 27
- 3.3 Regression Analysis . . . 28
  - 3.3.1 Linear Regression . . . 28
  - 3.3.2 Logistic Regression . . . 28
- 3.4 Factor Analysis . . . 29
  - 3.4.1 Exploratory Factor Analysis . . . 29

3.4.2 Confirmatory Factor Analysis . . . 31
3.5 Machine Learning . . . 31
3.5.1 Supervised Learning . . . 32

4 Exploring the Retention of Physics Students 36
4.1 Introduction . . . 37
4.1.1 Research Questions . . . 37
4.1.2 Results of prior research . . . 38
4.2 Methods . . . 39
4.2.1 Sample . . . 39
4.2.2 Variables . . . 40
4.2.3 Statistical and Graphical Methods . . . 41
4.3 Results . . . 41
4.3.1 Descriptive Analysis . . . 41
4.3.2 Visualizing Retention . . . 44
4.3.3 Survival Analysis . . . 46
4.3.4 Logistic regression . . . 52
4.3.5 Decision Trees . . . 58
4.3.6 Traversing the course network . . . 63
4.4 Discussion . . . 66
4.5 Implications . . . 70
4.6 Limitations and Future Work . . . 71
4.7 Conclusions . . . 71

5 Examining the Conditional Probabilities of Physics Student Retention with Bayesian Networks 73
5.1 Introduction . . . 74
5.1.1 Research Question . . . 75
5.1.2 Bayes' Theorem . . . 76
5.1.3 Bayesian Networks . . . 77
5.1.4 Prior Studies of Bayesian Networks in Retention . . . 80
5.2 Methods . . . 82
5.2.1 Sample . . . 82
5.2.2 Building Bayesian Networks . . . 84
5.3 Results . . . 87
5.3.1 Bayesian Networks . . . 87
5.3.2 Conditional Probability Queries . . . 90
5.4 Discussion . . . 96
5.5 Conclusion . . . 99

6 Predicting Physics Course Grades Using Bayesian Networks 100
6.1 Introduction . . . 101
6.1.1 Research Questions . . . 102
6.1.2 Bayesian Networks and Grade Prediction . . . 103
6.2 Methods . . . 105

6.2.1 Sample . . . 105
6.2.2 Identifying Conditional Probabilities . . . 107
6.2.3 Predicting Course Outcomes . . . 110
6.3 Results . . . 114
6.3.1 Identifying Conditional Probabilities . . . 114
6.3.2 Predicting Course Outcome . . . 119
6.4 Discussion . . . 123
6.5 Recommendations . . . 126
6.6 Conclusion . . . 129

7 Identifying Curricular Patterns Using Curricular Analytics 130
7.1 Introduction . . . 131
7.1.1 Research Questions . . . 132
7.1.2 Results of prior research . . . 133
7.1.3 Curricular Analytics . . . 133
7.2 Methods . . . 137
7.2.1 Sample . . . 137
7.2.2 Curricular Analytics . . . 138
7.3 Results . . . 143
7.3.1 Curricular analytics across multiple institutions . . . 143
7.3.2 The role of math readiness . . . 146
7.3.3 The effect of degree tracks . . . 149
7.4 Discussion . . . 151
7.4.1 Research Questions . . . 151
7.4.2 Other Observations . . . 156
7.5 Simplifying Curriculum by Making Prerequisite Adjustments . . . 157
7.5.1 Curriculum A . . . 158
7.5.2 Curriculum B . . . 158
7.6 Implications . . . 162
7.7 Limitations . . . 164
7.8 Conclusions . . . 165

8 Exploring Student Knowledge Structures in the BEMA as measured by MIRT 167
8.1 The Brief Electricity and Magnetism Assessment . . . 168
8.1.1 Research Questions . . . 169
8.2 Item Response Theory . . . 170
8.2.1 Prior constrained MIRT studies . . . 170
8.3 Prior Studies of the BEMA . . . 173
8.3.1 Studies comparing the BEMA and the CSEM . . . 174
8.4 The Structure of Knowledge . . . 175
8.5 Methods . . . 176
8.5.1 Sample . . . 176
8.5.2 Item Response Theory . . . 176
8.5.3 Model Fit Statistics . . . 179

8.6 Results . . . 182
8.6.1 Exploratory Analyses . . . 182
8.6.2 Confirmatory Analyses . . . 184
8.6.3 Topical Model . . . 195
8.6.4 Principle Model . . . 198
8.7 Discussion . . . 200
8.7.1 Research Questions . . . 200
8.7.2 Synthesis . . . 210
8.7.3 Future work . . . 212
8.8 Limitations . . . 213
8.9 Conclusions . . . 213
8.10 Acknowledgement . . . 214

9 Conclusions and Future Work **215**

Bibliography **220**

List of Tables

3.1 Confusion matrix. 33

4.1 Descriptive statistics applying a variety of filters for Institution 1. Filters are abbreviated: HS (high school) for students with HSGPA and ACT or SAT scores, P1 (Physics first) for students whose first declared major is physics, FTF (First-Time Freshman) students admitted as first-time freshmen, Fall First, students whose first semester was the fall semester. Different windows were also applied to investigate persistence and graduation. Grad (Graduation) removes the last six years of records, 1Year (One year) removes the last year of records, 2Year (Two year) the last two years, and 3Year (Three year) the last three years. Columns are abbreviated: ACTM% (ACT or SAT mathematics %), ACTV% (ACT or SAT verbal %), HSGPA (high school GPA), CGPA (college GPA), Grad Phys % (percentage of student graduating with a physics degree), Grad Other % (percentage of student graduating with a degree other than physics), Not Grad % (percentage of students who do not graduate with any degree), Surv Soph % (percentage of students enrolled as physics majors in their sophomore year), and Surv Junior % (percentage of students enrolled as physics majors in their junior year). Note, Grad Phys %, Grad Other %, and Not Grad % should add to one; for rows in which they do not, it is a result of the cumulative rounding of the numbers. 42

4.2 Institution 1 major election sequences. 45

4.3 Logistic regression. All regressions are significant improvements over the null model ($p < 0.001$). β is the normalized regression coefficient, SE is its standard error, z is the z-score of the coefficient, p the probability a value larger than z occurred by chance, and e^{β} is the odds ratio. 55

4.4 Logistic regression including first semester GPA. All regressions are significant improvements over the null model ($p < 0.001$). β is the normalized regression coefficient, SE is its standard error, z is the z-score of the coefficient, p the probability a value larger than z occurred by chance, and e^{β} is the odds ratio. 57

5.1 Descriptive statistics for data from Institution 1 after applying filters. Filters are abbreviated: HS (high school) for students with HSGPA records, P1 (Physics first) for students whose first declared major was physics, P112 (PHYS 112) for students who enrolled in PHYS 112, and P314 (PHYS 314) for students who enrolled in PHYS 314. Different windows were used to ensure that the samples only included students who could have met a particular milestone: 2year (Two year) removes the last two years of records, 3year (Three year) removes the last three years of records, Grad (Graduation) removes the last six years of records. HSGPA is the average High school GPA of the sample, and Math Ready % reports the percentage of students ready to take Calc 1 or higher upon enrollment. The last three columns report the percentage of students who met one of the three milestones. 83

6.1 List of courses used as variables in the analyses, as well as the college GPA variables. 106
6.2 The sample sizes of each dataset used to predict the target variable, as well as the dependent variables and the AB% of the dataset. The dependent variable numbers refer to Table 6.1. 120
6.3 Results of course predictions, averaged over 100 iterations. The decision threshold is the threshold for the probability that a student will "Struggle". The Blacklist column indicates whether the blacklist was used in learning the network structures. 120

7.1 Summary of the structural complexities of each tier. The table presents the mean, standard deviation (SD), standard error (SE), and 95% confidence interval. 144
7.2 Comparison between the first and last quartiles (ignoring tier placement) of the institutions included in the study 152
7.3 Two example curricular structures with differing complexity. 159

8.1 MIRT fit statistics for an Exploratory Factor Analysis of the BEMA. 184
8.2 Factor structure for the five-factor model. Only loadings greater than 0.3 are shown. The factors are labeled FC1 to FC5. 185
8.3 Theoretical model tested by the BEMA. An × indicates that the principle is used in the CSEM. 186
8.4 A continuation of Table 8.3 187
8.5 Model transformation table. Each entry presents the result of modifying a prior model (the original model) with one of the planned transformations to produce a modified model (the transformed model). These two models are compared and the model with superior fit statistics identified (the superior model). 193
8.6 Subscale scores for each topic. The mean ± the standard deviation (SD) are shown. The mean calculates the average fraction of item in the subscale answered correctly by the students 195

8.7 Best-fitting principle and topical MIRT models. The first column shows the item number (#). Not all items of the BEMA were modelled. The discrimination for principle k on item j, a_{jk}, is given by the number in parentheses following the principle label. The overall discrimination of item j on a knowledge of electromagnetism is given by a_{j0}. The difficulty of each item is related to d_j; items with larger positive d_j are easier, items with more negative d_j, harder. The discrimination of the item on the subtopics of the topical model is given by a^s_{jk}. 196
8.8 Comparison of BEMA and CSEM. DF, L, R, C, and LM represent principles in each instrument. The number in parenthesis is the number of the principles also in the other instrument. The Items column refers to the number of items in the instrument grouped into the electricity and magnetism subtopics; Mechanics and Superposition are not subtopics specific to electricity and magnetism, so their Items columns are 0. 206
8.9 Comparison of conceptual instruments. DF, L, R, F, C, LM, and RS represent principles in each instrument. Independent is abbreviated "ind" and principle "pcpl" when needed for spacing. 208

List of Figures

1.1 Results of Gain vs Pretest score in [1]. Lines represent normalized gain thresholds, with steeper lines representing greater gains. Shaded markers indicate traditional teaching methods, while empty markers represent reformed instruction or active learning strategies. 5

4.1 Sankey plot showing major changing and graduation patterns for students who elect a physics major at any point in their undergraduate career. Each group of two bars represents an academic year; fall semesters are odd numbers, spring semesters even. 45
4.2 Fraction departed or graduated for students entering the university declared as physics majors. 49
4.3 Hazard functions. The graduation hazard is plotted on a different scale shown by the right vertical axis. For Institution 1, each semester plotted has at least 50 students enrolled as physics majors. 50
4.4 Hazard function for required physics and math courses at Institution 1. The hazard in this case is calculated as the number of students who departed the program after taking the course over the number of students who took the course. The axis for leaving the program through graduation is on the right. The abbreviations in the figure are for various subjects in physics: Classical Mechanics 1 & 2 (CM1 & 2), Electricity and Magnetism 1 & 2 (EM1 & 2), Quantum Mechanics 1 (QM1), and Statistical Mechanics (SM). 51
4.5 Decision tree for persisting in physics to the sophomore year 60
4.6 Decision tree for persisting in physics to the junior year 61
4.7 Decision tree for persisting in physics to graduation 62
4.8 Traversing the major from entry to Modern Physics for students at Institution 1 who elect a physics major in their first semester. The figure uses the abbreviations <Calc for students whose first mathematics class is less advanced than Calculus 1, Calc for students whose first mathematics class is is Calculus 1, and >Calc for students whose first mathematics class is more advanced than Calculus 1. 65

5.1 Sample network with variables A, B, C, D, E, & F. 79
5.2 Bayesian networks for the milestones of enrolling in PHYS 112, enrolling in PHYS 314, and graduating physics. The caption of each network indicates the sample used to build the network. Only pre-college academic factors are included as variables. 88

5.3 Bayesian networks for the milestones of enrolling in PHYS 314 and graduating in physics. The caption of each network indicates the sample used to build the network. College physics course grades were included as variables in these networks, as shown by the P112 nodes and P314 nodes. 90
5.4 CPQ results for each milestone variable and each of its pre-college independent variables. The probabilities shown are the probabilities of a "1" outcome (i.e. reaching the milestone). Probabilities queried from the networks in Fig. 5.2. 91
5.5 Conditional probability tables for each milestone variable and their parent variables. Probabilities of reaching or not reaching the milestone are given for each possible combination of the parent variables. Probabilities queried from the networks in Fig. 5.2. 93
5.6 Observations per combination of MathEntry and HSGPA for the CPTs in Fig. 5.5. 94
5.7 CPQ results for milestone variables TakeP314 and EndPhys, including results for pre-college variables and college physics course grades. The probabilities shown are the probabilities of a "1" outcome (i.e. reaching the milestone). Probabilities queried from the networks in Fig. 5.3. 95

6.1 Bayesian networks for PHYS.112, PHYS.314, PHYS.331, and PHYS.333 . . 115
6.2 Bayesian networks for PHYS.341, PHYS.451, and PHYS.461 116
6.3 Probabilities of receiving an AB grade in a target course based on a grade received in a prior course. 117
6.4 Probabilities of receiving an AB grade in a target course based on a grade received in a prior course. CGPA has levels A, B, C, D, and F and the course variables have levels A, B, C, DFW. 118
6.5 Variable importance based on mean decrease of balanced accuracy by dependent variables. The error bars represent the standard error of the difference between mean balanced accuracies. 121
6.6 Variable importance based on mean decrease of balanced accuracy by dependent variables. The error bars represent the standard error of the difference between mean balanced accuracies. 122

7.1 Example curriculum graph for a mid-tier institution. General Physics 1 is shaded red and the courses it blocks are shaded gray. The count of the gray courses is the blocking factor of General Physics 1. The delay factor of General Physics 1 can be found by counting the number of courses in its longest path, in this case 6. The university divides general education requirements into seven categories labeled F1 to F7 in the figure. 141
7.2 Distribution of curricular complexity for physics programs with different rankings. 145
7.3 The complexity of the graduate-intending degree track plotted against the first mathematics class in which the students enrolls. 149
7.4 The complexities of various degree tracks versus the first mathematics class the student takes in college. 149

7.5 Curriculum A, with 20 required physics and math courses, and a structural complexity 290. 160
7.6 Curriculum B, the adjusted curriculum, with 20 required physics and math courses, and a structural complexity of 222. 161

8.1 Probability of selecting the correct response, $\pi(\theta)$, versus ability θ using $d = 0$ and $a = 1$. The dashed line represents the slope at $\theta = 0$ and has slope $a/4 = 0.25$. 178
8.2 Correlation matrix. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations. Thicker lines represent larger correlations. . 182
8.3 Partial correlation matrix. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations. Thicker lines represent larger correlations. 183

Chapter 1

Introduction to Physics Education Research

1.1 Introduction

The 2012 President's Council of Advisors on Science and Technology emphasized the need to improve STEM student retention to avoid a candidate shortfall of 1 million STEM jobs [2]. At the time, they estimated that less than 40% of students who enroll in STEM degrees complete those degrees, and the completion rate is even more concerning for students who are under-represented minorities (URM). A decade later we still see the need to improve STEM student graduation rates to fill jobs in STEM. This has led to increased focus in Discipline Based Education Research (DBER) to improve retention and graduation rates of students seeking STEM degrees. Similar concerns in the 1960's and 1970's, and perhaps added pressure from the Cold War space race, prompted government funding into the new field of Physics Education Research (PER) in hope of increasing the number of students seeking careers in the space industry and other physics-related fields [3]. This chapter will present a brief history of PER, starting with its origins in conceptual understanding, and moving through the development of conceptual inventories and other research-based curricular materials and instructional strategies.

The research contained in this dissertation will introduce and demonstrate methods that can be used to improve student retention to degree completion. While the work here focuses on the domain of physics, it should also be applicable to other STEM fields. It is subdivided into four main parts: Part 1 explores patterns of physics student retention at a public R1 institution while also investigating the attrition points and critical courses in the physics program; Part 2 uses Bayesian network methodologies to predict physics course grades, Part 3 int;oduces a new analytic technique to PER, Curricular Analytics, that will

be used to quantify the complexity of physics academic programs; and Part 4 investigates the knowledge structure of a commonly used conceptual instrument, the Brief Electricity and Magnetism Assessment (BEMA) and further supports the need to improve instruments that measure modern physics students understanding so that physics educators can better serve undergraduate students in physics classrooms.

1.2 Conceptual Understanding

In the 1970's, physics instructors began to recognize a problem in physics education; many of the misconceptions that students had about physics before taking a physics course were still present after successful completion of a physics course [4–6]. These misconceptions, in theory, should have been remedied by completing a physics course, where students should have connected the physical laws and principles with real life experiences. This led to an examination of educational practices; instruction was modified to better serve students, helping them overcome their misconceptions. Prior to these studies, qualitative understanding had not been emphasized, but rather mathematical logic and reasoning were the focus of instruction. This focus on qualitative or conceptual understanding led to the development of several conceptual inventories that measure conceptual knowledge in introductory physics courses.

1.2.1 Conceptual Inventories

One of the first and the most widely used conceptual inventory was the Force Concept Inventory (FCI), developed Hestenes *et al.* to measure conceptual understanding of forces and kinematics in introductory classical mechanics courses [7]. Other conceptual instruments

have been developed in many different areas of physics, but the most popular are the FCI and the Force Motion Conceptual Evaluation (FMCE)[8] for classical mechanics and the Conceptual Survey of Electricity and Magnetism (CSEM) [9] and the Brief Electricity and Magnetism Assessment (BEMA) [10] for introductory classes in electricity and magnetism. Each of these instruments have undergone various forms of reliability and validity testing to ensure they accurately measure conceptual understanding in their specified domain; recent studies employing Item Response Theory (IRT) have shown that these instruments are less accurate than initially thought, and the concepts taught in introductory physics courses are not completely reflected in the conceptual coverage of the instruments [11–14]. These studies have proposed that the instruments be updated or new instruments be written to better serve physics instructors and students. One of these studies is presented in Chapter 8 of this dissertation.

1.2.2 Normalized Gain

These conceptual instruments are often employed as a pretest before the respective introductory physics course and then re-administered as a post-test at the completion of the course. Often they are used as an evaluation tool to gauge the effectiveness of an instructor at improving students' conceptual understanding. This allows the study of the types of instruction that most improve conceptual understanding. Hake, using FCI scores from 62 different courses at several institutions, compared instruction types by examining how much the FCI score increased from pretest to post-test [1]. To compare different student populations he used:

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100 - \langle S_i \rangle} \tag{1.1}$$

which is referred to as the normalized gain or the Hake gain, where $\langle S_i \rangle$ is the class pretest average and $\langle S_f \rangle$ is the class post-test average, both on a scale from 0 to 100. This normalized gain scales the pretest to post-test gain by the maximum possible gain. Hake claimed that this allowed comparison across institutions.

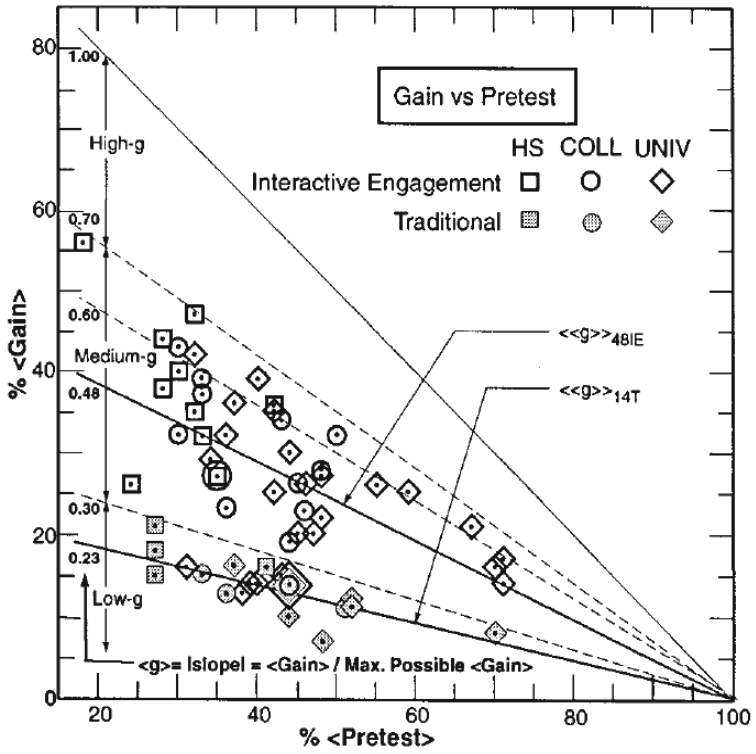


Figure 1.1: Results of Gain vs Pretest score in [1]. Lines represent normalized gain thresholds, with steeper lines representing greater gains. Shaded markers indicate traditional teaching methods, while empty markers represent reformed instruction or active learning strategies.

Hake showed that instructors who used active learning strategies, or engaged learning strategies, had greater gain scores at the end of instruction than instructors who used traditional lecture methods. This is shown in Fig 1.1 where each marker represents the FCI Hake gain score at one of the 62 courses involved in the study. This study led to increased research

into active learning strategies and their effectiveness, and was instrumental in supporting the broad adoption of these strategies.

1.3 Research-Based Instructional Strategies

The research in active learning strategies, or research into improving instruction and curriculum, has led to the development of many research-based instructional strategies (RBIS). Docktor and Mestre, in their synthesis of PER, divide RBIS into 5 groups: lecture-based methods, recitation or discussion methods, laboratory methods, structural changes to classroom environment, and general instructional strategies and materials [15].

1.3.1 Lecture-based Strategies

Lecture-based RBIS center around the goal of improving student interactions with their peers and their instructors in a lecture setting. One of the most common forms of lecture-based RBIS is the use of polling technology, often in the form of "clickers". One of the earliest methods of using "clickers" was Peer Instruction introduced by Mazur [16]. In Peer Instruction or other forms of polling, the instructor presents a conceptual or qualitative multiple-choice question, and students discuss with their nearby peers, and then select an answer. A class-wide discussion takes place, and together the class arrives at the correct solution. This method has been shown to be effective in increasing normalized gains in courses when compared to traditionally taught courses, as well as leading to decreased course attrition [17]. Interactive Lecture Demonstrations (ILDs) is another lecture-based RBIS , where a physical demonstration or experiment is presented to the students. The students, in discussion with their peers, make a prediction of what will occur during the demonstration,

observe the demonstration, and then compare their predictions to their observations [18].

1.3.2 Recitation-based Strategies

Recitation-based RBIS are designed to make the recitation setting more active for the students, with the intention of developing conceptual understanding of physics. *The Tutorials in Introductory Physics* (TIP) [19] was developed at the University of Washington to replace the traditional format of recitations where teaching assistants review homework problems on the board. A recitation using TIP consisys of pretests, worksheets, and homework assignments. Students work in groups of 3-4 and are taken through the process of understanding and thinking critically about the physics under consideration, while also confronting their misconceptions. Some variations to the TIP curriculum have been published by researchers at the University of Maryland: the Activity-Based Tutorials (ABT) and the Open-Source Tutorials (OST) [20, 21]. Each of these tutorial programs showed improved student understanding and course outcomes, as well as improved normalized gains compared to traditional recitation methods [22–24]. Another recitation-based RBIS that was shown to improve problem-solving capabilities is cooperative learning [25], a strategy for collaborative group problem solving. Students are split into groups of 2 or 3 and collaboratively work on context-rich problems.

1.3.3 Laboratory-based Strategies

Traditional labs have often been described as "cook book" or "cookie-cutter" because they require students to follow a step-by-step procedure with little thought required [26]. Attempts to rectify these problems included the use of various technological tools such as sonic

rangers or video-analysis software to get real-time data of kinematic motion and to generate graphical outputs. Students in these types of labs were shown to have greater understanding of graphs and kinematic concepts [27, 28]. Another strategy is to engage students in the process of science during their lab; for example, the Investigative Science Learning Environment (ISLE) labs remove the pre-built nature of traditional labs and students are required to make their own hypothesis based on a new phenomenon and test them with their own experiments [29]. Students in ISLE labs showed improved scientist-like thinking, as well as improved skills associated with scientists, such as data analysis and experiment design. ISLE labs pose a question of general interest to physics instructors everywhere, which is "What is the purpose of introductory physics laboratory courses? Should students in these labs be learning scientific skills in a physics setting, or is the purpose strictly for students to learn physics concepts?" The overall effectiveness of introductory labs and lab-based instruction has been called into question [30], and in a multi-institution and multi-course study there was no difference on final exam scores between students who enrolled in a lab-course and those who did not, further calling into question the focus of interventions to improve laboratory-based courses.

1.3.4 Classroom Environment-Based Strategies

Many studies have analyzed the effect of changing the classroom environment in ways such as rearranging seating, including technology, and combining laboratory activities with lecture activities in a workshop-like or studio setting. One of these RBIS is Student-Centered Active Learning Environment for Undergraduate Programs (SCALE-UP) [31]. In SCALE-UP classrooms, students sit at round tables and are separated into groups at each table. Each

group has access to a laptop and whiteboards to work collaboratively on hands-on activities and problems that occur in tandem with lecture. An extension of the SCALE-UP RBIS is the Technology-Enabled Active Learning project (TEAL) [32]. TEAL uses the same classroom design as SCALE-UP, where laboratory activities and lecture activities are combined into a single experience, and expands upon it by implementing technology-enhanced visualizations and activities. Both methods have been shown to improve conceptual gains and student completion rates [31, 32]. Students were slightly more favorable toward studio-style classes than traditional lecture classes.

1.3.5 General Instructional Materials

Several textbooks have been written based on results from PER, such as *Understanding Physics* [33] and *Physics for Scientists and Engineers: A Strategic Approach* [34] and *Six Ideas that Shaped Physics* [35]. Another PER based curriculum is *Matter and Interactions* [36], which is intended for introductory calculus-based physics courses. Other materials include simulation resources such as the University of Colorado's PhET project, which provides a range of simulations of phenomena in the sciences, as well as resources for activities that accompany the simulations [37].

Conclusion

At its core, PER is focused on improving the success of students in physics classes and physics programs. The past research discussed in this chapter describes much of the work that has been done to improve student learning and success in physics classrooms. The inclusion of these RBIS in physics classes will improve the success rate of students in those

classes. At the core of student success in physics programs is the success of students in the required classes of that program. The following chapter gives a broad overview of research specific to college student retention, in preparation for the research presented in Chapters 4-7 of this dissertation.

Chapter 2

Student Retention and Educational Data Mining

The research presented in this dissertation is focused on introducing several different types of analysis that physics and other STEM departments can use to investigate student retention and to inform decision making regarding curricular changes that improve student success via degree completion. This chapter presents a literature review of the work that has been done to recognize patterns in student retention and effective practices to improve student retention, while also giving a brief overview of the use of educational data mining in answering questions regarding student retention.

2.1 Retention

While little research into physics major persistence has been performed within PER, substantial research has investigated general college persistence and success as well as persistence in science, technology, engineering, and mathematics (STEM) majors. Within PER, a substantial research strand has investigated factors influencing student success in physics classes, a key component of college retention. The work contained herein focuses on quantitative factors that affect physics student retention. As such this review focuses on studies that examine quantitative factors in retention. There are many studies that examine qualitative factors in STEM and general student retention [38–40], as well as the retention of several demographic groups [41–44]. This qualitative body of research lends greater context to the factors that ultimately cause a student to leave the sciences.

2.1.1 Physics retention

Some studies have explored the issue of retention in physics including retention of majors to physics degrees, retention within the introductory sequence, and intention to

persist in physics. Aiken *et al.* used a random forest machine learning model to examine the factors most important in predicting whether a student would earn a physics degree [45]. They found that taking Modern Physics and taking an engineering class were the variables most important in the prediction.

Zwolak *et al.* examined the retention of students in a physics course sequence, which included other scientists and engineers. Zwolak *et al.* used network analysis to determine students' social and academic integration which was used to predict if students who enrolled in the first course of an introductory physics course sequence would persist to the second course in the sequence [46]. They found that by using a student's centrality measures in the integration network, they could predict a student's persistence in the sequence at a rate of seventy-five percent. This is similar to work done by Forsman *et al.* who used complexity science in analyzing social and academic networks of students in physics courses to explain student retention [47].

A largely qualitative study by Stiles-Clark and MacLeod surveyed students after the second course of a two-course calculus-based introductory physics sequence and asked about factors that influenced the decision to continue in the physics program or a different program at the university. They found that the primary reasons for persistence were the student's interest in the subject matter, the quality of their physics instructors, and their perceived career opportunities with a physics degree [48]. The researchers noted the need for physics faculty to engage students in research-based classroom and lecture techniques, as well, as the need to combat misconceptions about career opportunities for physics degrees.

2.1.2 General college retention

College retention and college persistence are major research strands in general education research. High school academic preparation is an important predictor of college success. Composite SAT scores are highly correlated with GPA in the first year of college [49]. Benchmarks for ACT composite scores have been created indicating the score required for 50% chance of earning at least a B in introductory college classes [50]. High school GPA is more variable due to the variety of high school curricula [51] but is still a strong predictor of first year GPA [52, 53] and overall college GPA [54]. One educational data mining study found that factors associated with the socioeconomic status and first generation status were highly predictive of retention after a student's third year as was a lack of academic preparedness based on ACT and COMPASS scores [55]. The COMPASS tests are administered by ACT Inc.; COMPASS scores are designed to help place students in the appropriate college classes.

Research into college student retention represents a major strand in general education research. A book with a foreword by Tinto [56] reviews the history of the field including differing models of student retention, economic considerations of student retention, retention in less traditional colleges such as community colleges and online colleges, as well as suggested actions to improve student retention. Although several models of student retention have been postulated, the most widely applied model was developed by Tinto [57, 58]. Tinto proposed that a student's persistence depends on their skills, attributes, intentions, commitment, and interactions with students and faculty within the college. He claimed the most important factor in student retention was the student's experiences in the college, and as a student became more integrated into the academic and social communities at the college the more

likely they were to persevere until graduation. Social integration refers to student-to-student interactions and involvement in extra-curricular activities available at the college. Academic integration is described as the congruence of a student's abilities, skills, and interests with the academic demands of the institution and also interactions between the student and faculty and staff. In 2012, Tinto introduced a framework for institutional actions to improve student retention [59]. His framework focused on improving teaching methods and classroom interventions as this is the primary interaction between students and faculty and thus the primary way they can become integrated into the college's academic community. While improving retention is often an institutional priority, a study by Henderson *et al.* [60] showed that among physics faculty, only 48% use methods that have been empirically proven to improve student learning, and only 23% used them at a high level.

2.1.3 STEM retention

The demand for employees having at least a bachelor's degree in a STEM discipline continues to grow [61]. Despite the critical need, only 40% of STEM students graduate with a STEM degree [2]. In a 2014, the U.S. Department of Education reported wide variation in the attrition rates (defined as leaving the university or the degree) of different STEM disciplines with an average rate of 48%. Attrition was highest for computer/information science majors (59%) and lowest for mathematics majors (38%) [62]. This attrition rate was lower than the attrition rate of students in the humanities or education (56-62%) and approximately equal to the rate for students in business and social/behavioral science [62].

Many studies have investigated STEM degree retention and methods to improve retention [63, 62, 64–69]. In general, measures of prior high school preparation (high school GPA

and ACT/SAT scores) as well as college performance metrics such as credit completion and college GPA were important factors in predicting student retention. Other factors that have been found to be important include relationships between faculty and students [70, 71], the use of learning communities [72], the implementation of a career planning seminar or career planning course [73, 74], a scientific thought and methods course [75], and for engineering students their grades in introductory physics courses [76]. A study using self-reported survey data [77] found that an institution's academic environment was important for students deciding to stay in STEM: features such as smaller class sizes, more integration of undergraduate student research, faculty teaching skills and whether or not students were engaged in active learning strategies were important. A review article by Sithole *et al.* synthesizes many reforms or changes that have been suggested to improve student retention such as improved academic advising, blending courses, peer mentoring, instruction in time management and study habits, and improving high school STEM curriculum and instruction [78].

2.1.4 Physics course success

Many PER studies have examined factors which influence student success in physics courses (generally introductory courses) using metrics such as final exam grade, course grade, and conceptual post-test scores. A certain level of success in physics courses is typically required for persistence in the major. One would also hypothesize that students who are more successful in their introductory physics courses are more likely to persist in the physics major. Much of this research has examined either instructional methods to increase success or remove conceptual barriers (misconceptions) that prevent success. Meltzer and Thornton provide an extensive review of research into interactive instructional methods and the efficacy

of these methods [79]. Research into student misconceptions spans the history of PER [80–83]. In 2014, the National Academy of Sciences published a synthesis of results from many disciplines showing interactive instruction improved conceptual performance as well as course outcomes [84]. A further meta-analysis demonstrated the efficacy of these methods at the college and pre-college level [85].

Recent studies have examined how general high school preparation metrics (ACT and SAT scores) and prior preparation in physics measured by conceptual pretest scores affect course outcome measures including final exam grades, overall course grades, and conceptual post-test scores [86–88]. These studies show that both general high school preparation and specific preparation in physics are important in predicting student outcomes; they also show that different factors are of varying importance for different demographic groups. Studies have also investigated the details of high school physics preparation as well as non-cognitive variables such as parental support as predictors of success in college physics classes [89].

Success in calculus-based introductory physics courses is also key for engineering and other science majors, who generally make up the majority of the students in an introductory physics class. A recent study by Wingate *et al.* [76] found that success in introductory physics courses was predictive of success in later engineering courses and persistence to an engineering degree. Most students who received a high grade in the introductory physics sequence continued to achieve high grades through the rest of the engineering coursework, while those who received a lower grade continued to struggle through their remaining classes.

2.2 Educational Data Mining

Educational data mining (EDM) involves the use of statistical, traditional, and machine learning data mining techniques to interpret and analyze educational data. With the advent of university learning management systems and increases in computing power, a very substantial branch of education research has attempted to use these large data systems and emerging computer technologies to predict both in-class success and retention to graduation. These techniques are called educational data mining (EDM) or learning analytics. Multiple reviews have summarized the efficacy of the numerous algorithms used by EDM to predict both in-class and overall student performance [90–97]. The application of data mining to the university retention problem began in the early nineties; Nandeshwar *et al.* provides a review of this work [55]. They report that college performance, high school GPA, ACT scores, and some socio-family factors affect student retention.

These techniques have been used in multiple studies to predict student first-year retention and persistence through graduation for engineering students [98–102, 73, 103]. Engineering students form the majority of the students in the calculus-based introductory physics classes taken by physics majors. Machine learning has recently be applied in PER to understand student performance in physics classes [104, 105]. The work in Chapters 4-6 of this dissertation uses EDM techniques in the analyses presented.

Conclusion

It is a responsibility of physics departments to make their curriculum, or program of study, effective in the retention of, instruction of, and preparation of physics students,

so those students can pursue meaningful careers in education, industry, or academia. A failure in any of the areas of retention, instruction, or preparation should be addressed by physics departments and changes should be made to improve student success. The research presented in this thesis is intended to inform university physics departments on methods to analyze the picture of retention in their department so they have better information to make decisions regarding changes to their program's course structure, instruction, and advising.

Chapter 3

Statistical Methods

Statistics is the collection, organization, analysis, interpretation, and presentation of data. Generally it is divided into two categories: descriptive statistics and inferential statistics. This chapter will introduce several statistical methods or techniques that were used in the research presented in this manuscript. Additional methods will be introduced in this manuscript as needed.

3.1 Descriptive Statistics

Generally, the first step in quantitative analysis of data is an exploration of the data. This often includes visual exploration via scatter plots or bar charts and the calculation of descriptive statistics for relevant variables. The most important statistics are those that measure central tendency and those that describe the variability of the data.

3.1.1 Measures of Central Tendency

Central tendency measures include the mean, median, and the mode. These measures give an estimate of a "typical" value for a certain variable. In the research that follows hereafter, only the mean and median of a dataset are used. The mean, or average, of a sample is defined as

$$M = \frac{\sum_i x_i}{n} \tag{3.1}$$

where the value of a specific variable is summed for all entries in the data, and divided by the total number of data points in the sample. The sample mean is an estimate of the population mean, defined as

$$\mu = \frac{\sum_i X_i}{N} \tag{3.2}$$

where the value of a specific variable is summed for all members in a population, and divided by the size of the population [106]. The median is defined as the middlemost value of a variable when the data points are listed in rank order of the variable. This measure splits the data in half, as 50% of the data points have a value less than or equal to the median, and 50% of the data points have a value greater than or equal to the median.

3.1.2 Variability

Measures of variability describe how the data is spread about the mean. The sample variance illustrates this, though the standard deviation - the square root of the variance - is more useful as it is in the same units as the data and the mean. The standard deviation of a sample is defined as

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2} \tag{3.3}$$

where the denominator of $n-1$ is the degrees of freedom, and allows the sample standard deviation to serve as an un-biased estimator of the population standard deviation, which would have a denominator of N, the size of the population [106].

When comparing means of different samples in a population, the standard error (SE) is a more useful measure of variability than the standard deviation. It describes variability of the sample mean about the population mean. For a normally distributed sample

$$SE = \frac{SD}{\sqrt{n}} \tag{3.4}$$

where n is the sample size [106]. A confidence interval (CI) gives an estimated range for an unknown parameter, often a population mean. The confidence level at which a CI is computed affects the width of the CI. A confidence level of 95% is the most common, though other levels such as 90% and 99% are common as well, with higher confidence levels typically giving a wider CI. Samples with a smaller standard error will have a narrower CI. A 95% CI can be calculated as

$$CI = M \pm 1.96(SE) \tag{3.5}$$

where 1.96 is the z-score related to a 95% confidence level (a z-score measures how far an observation is from the mean in terms of the standard deviation of the sample). A 95% CI is sometimes described as a range of values that the unknown parameter lies within at a 95% probability. This is incorrect, but rather the CI is a range of values that are not significantly different from the estimated unknown parameter, at a level of significance appropriate to the confidence level (a 95% confidence level would have a level of significance of 0.05). Significance levels are defined in the following section.

3.2 Inferential Statistics

Inferential statistics are methods that allows the researcher to test assumptions about the data, such as testing how likely an observed difference in mean scores happened by chance.

3.2.1 Hypothesis Testing

Hypothesis testing, or null hypothesis significance testing, is one of the most common forms of inferential statistics. Generally it is used to compare the means of two samples or the mean of a sample to the population mean. To do this one states a null hypothesis H_0 (e.g. the means of the samples are not significantly different, $H_0 : M_1 = M_2$). One then selects a mutually exclusive assumption, the alternate hypothesis H_1. This hypothesis could be one-sided (e.g. the mean of sample 1 is greater than the mean of sample 2) or two-sided (e.g. the mean of sample 1 is different than the mean of sample 2, meaning it could be either greater than or less than). Testing the hypothesis consists of assuming the null hypothesis is true and then calculating a test statistic. The distribution of the test statistic is known; the probability p that the calculated test statistic value occurred by chance is then computed. The p-value is compared to the chosen significance threshold, α, and results where $p < \alpha$ are considered to be significant. In these cases the null hypothesis is rejected, and the alternate hypothesis is accepted. There are several common test statistics such as the t-score, the z-score, and the F-score.

ANOVA

Analysis of variance (ANOVA) is a hypothesis testing method for comparing the means of two or more groups within a sample. ANOVA uses the F test, which calculates the ratio of the explained variance to the unexplained variance to determine if the group means are significantly different from each other. In this manuscript, one-way ANOVA testing is used, which determines if group means are different, where the groups are defined based on a

specific factor. In this case, the null hypothesis takes the form $H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ where k is the number of groups. The alternate hypothesis H_1 simply states that all or some of the means of the groups are not equal. Like general hypothesis testing, the F-statistic is compared to a critical value, determined by α and the degrees of freedom of the groups, and if the F-value is greater than or equal to the critical value the null hypothesis is rejected.

3.2.2 Effect Size

While hypothesis testing estimates whether or not a difference is significant, it cannot be used to determine whether a difference is practically meaningful, nor can it be used to determine the functional size of the difference. Cohen introduced the "effect size" which provides a measure of the size of the difference in two random variables [107]. Effect sizes classify differences in means as small, medium, or large effects. For differences in means, the most common effect size is Cohen's d, which is defined as

$$d = \frac{M_1 - M_2}{\sigma} \tag{3.6}$$

where the numerator is the difference in the means and the denominator is the pooled standard deviation. The criteria for the effect size of d is that a value of 0.2 is considered a small effect, 0.5 is considered a medium effect, and 0.8 is considered a large effect.

The uncertainty of the difference between means can be calculated with the standard error for the difference between means, defined as

$$SE_{M_1 - M_2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}. \tag{3.7}$$

3.2.3 Error

Two types of error can occur during hypothesis testing; Type I and Type II errors. A Type I error occurs when the null hypothesis is rejected when it should not be, when there is actually no difference in the means. This is defined as a false positive and occurs due to random fluctuations in the data being interpreted as an actual effect. When using a significance threshold of $\alpha = 0.05$, this error should occur once in 20 statistical tests. The most common correction for this type of error is the Bonferroni correction, which adjusts the significance threshold based on the number of statistical tests performed. Type II errors are defined as false negatives, where the null hypothesis is accepted when it is actually false. A common source of Type II errors is insufficient sample size, which leads to a lack of statistical power. To avoid this type of error, one can perform a power analysis to determine whether the sample size is sufficient to reliably detect the effect and significance.

3.2.4 Beyond Significance Testing

The correct use and interpretation of statistical methods is central to the effectiveness of PER. As such the PER community should be up to date in the latest advancements and changes in statistical research. In the last decade the topic of hypothesis testing and reporting significance determined by p values has been called into question, to the point that some journals discourage the use of null hypothesis significance testing (NHST). One journal has even banned the use of p values in its publications [108]. One of the biggest arguments for the elimination of p-value reporting is that it is so poorly understood and often incorrectly interpreted so that many faulty conclusions are drawn from valid statistical

work [109]. Greenland *et al.* [108] summarizes many of the misinterpretations that plague p-values, as well as misinterpretations of confidence intervals and power testing. Other issues arise with the use of p-values, such as dichotomous thinking that something is significant or not significant based on a p-value being above or below a fairly arbitrary value of 0.05. This dichotomous thinking has led to misrepresentation of statistical analyses, where only studies that find significance are reported while those that do not find significant results are not reported [109]. An effect that is found to be significant in several studies may, in fact, be insignificant in many other studies, but because the insignificant findings were not reported, the public receives a skewed or misleading interpretation of findings. Cumming [110] sets forth a program for nearly eliminating the reporting of p-values and significance testing and suggests studies focus on effect sizes and estimation. Other studies [111] suggest the use of a Bayesian approach that focuses on posterior distributions as opposed to the frequentist NHST approach. These changes in the use of statistics for research purposes are not widespread in PER.

3.2.5 Boot-strapping

Boot-strapping is a re-sampling technique that removes the need to assume that a distribution is normal in hypothesis testing. Boot-strapping creates many sub-samples from a sample (with replacement). The desired test statistic can be be calculated for every sub-sample, which produces a distribution of the test statistic, which will follow the normal distribution by the central limit theorem.

3.3 Regression Analysis

Regression analysis is a category of inferential analysis that quantifies how one variable will change with respect to another variable or variables. Perhaps the two most common types are linear regression and logistic regression.

3.3.1 Linear Regression

Linear regression is used to model the variation of a continuous dependent variable with a linear combination of independent variables, which can be continuous, dichotomous, or categorical variables. An example of a multivariate linear regression model is shown in Eqn. 3.8

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{3.8}$$

where y is the dependent variable, x_i refers to the various independent variables, and β_i refers to the regression coefficient of variable x_i, β_0 is the intercept and ϵ represents the error of the regression equation. The variance not explained by the predictors is the mean square ϵ. Linear regression minimizes the error ϵ in the regression equation by optimally finding the regression coefficients β_i. This essentially maximizes the explained variability in the distribution of the continuous dependent variable, minimizing the unexplained variability.

3.3.2 Logistic Regression

Logistic regression models how the probability distribution of one of the levels of a dichotomous variable depends upon the independent variables. These models are generally

more difficult to interpret than linear regression models, and are explained in more detail in Chapter 4.

3.4 Factor Analysis

Factor analysis, introduced by Spearman [112], uses a smaller set of unobserved or unobservable variables to explain the variance in the observed variable; the unobserved variables are called latent variables. Often these observed variables are item scores on an assessment instrument, such as a conceptual instrument, and factor analysis describes the internal structure of the instrument. These unobservable variables are referred to as latent variables, and are the factors extracted by the analysis. They represent the constructs measured by the instrument, e.g. Newton's second law in the FCI. There are two types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

3.4.1 Exploratory Factor Analysis

In EFA, a set of linear relationships between the items (observed variables) is proposed where the variables y_{ji} represent the score on item j by participant i. A set of latent traits (factors), x_{ik}, explains the variation in y. The latent trait x_{ik} is the trait of participant i associated with factor k. The factor loadings f_{jk} relate the latent traits measured by an item j to the factor k based on the observed data. These linear relationships are shown in Eqn. 3.9

$$
\begin{aligned}
y_{1i} &= f_{11}x_{1i} + f_{12}x_{2i} + f_{13}x_{3i} + u_{1i}. \\
y_{2i} &= f_{21}x_{1i} + f_{22}x_{2i} + f_{23}x_{3i} + u_{2i}. \\
&\dots \\
y_{ni} &= f_{n1}x_{1i} + f_{n2}x_{2i} + f_{n3}x_{3i} + u_{3i}.
\end{aligned}
\tag{3.9}
$$

where u_{1i} is the residual error for student i on item 1. In EFA, items are allowed to load onto any factor, and are not constrained by any input from the researcher. Generally EFA creates a set of models where each model extracts one more factor than the previous model, and these models are compared based on a set of model fit statistics, and the best fitting model is retained. One of the goals of EFA is to maximize the variance explained with the fewest number of factors.

The results of EFA form a set of coordinate vectors in the K-dimensional space defined by the K factors. This coordinate system can be arbitrarily rotated to be easier to interpret. Many factor rotations exist. A common rotation is varimax rotation which seeks factor loadings with a few large values and as many zeros as possible and leads to orthogonal factors. In EFA, other rotations allow factors to be correlated (this is theoretically reasonable in many cases). The goal of rotation is to find the simplest structure of the correlation matrix that gives easily interpretable results and retains all the pertinent correlations [113]. The simplest structure possible is one where each item loads only on one factor, and there are no interfactor correlations. Rotation methods should be carefully chosen, as many researchers oversimplify the structure through rotation and lose valuable interfactor correlations. For a

much more in depth review of rotations in EFA and how to select a method see Sass and Schmitt [113].

3.4.2 Confirmatory Factor Analysis

In CFA, a model is chosen, based on theoretical considerations, a priori to the analysis and is compared to the observed data. Adjustments are made to the model to try to improve model fit to the observed data. Generally the model consists of a set of constructs (factors) that the instrument purports to measure and each item is assigned to load onto a subset of factors. The model is adjusted by adding or removing items from loading onto factors. To ensure robsustness, models are often compared using a set of model fit statistics. Common statistics or fit indices include the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). These statistics are explained in greater detail in Chapter 8.

3.5 Machine Learning

Machine learning algorithms build models from sample data, often called training data, and then make decisions or predictions based on what they "learned" from the data. Often in PER, machine learning algorithms are used for classification tasks or predicting some type of outcome, though it has also been applied to intelligent tutoring systems and automated grading of assignments [114]. When machine learning is applied to educational systems, it is classified as part of the broader field of educational data mining. There are several approaches to machine learning that are used depending on the task at hand. Here only supervised learning will be discussed.

3.5.1 Supervised Learning

In supervised learning, the training data contains both the input information and the output associated with specific inputs. Perhaps the most common types of supervised learning algorithms are classification algorithms.

Prediction and Classification

Prediction algorithms attempt to learn characteristics from a set of data, and then predict the target variable of those data points based on their characteristics. The algorithm processes the training set or input, recognizing patterns between the independent variables and the known dependent variables. Once training is complete, the model can be used to process data where there are no values for the dependent or target variable. The model "predicts" or assigns a value to the dependent variable. Typically, if the dependent variable is categorical this is process is referred to as classification. The main goal of prediction and classification models is to maximize some type of predictive accuracy (different types of accuracy are discussed in the following section). A commonly used method for improving predictive accuracy is to use a group of theoretically independent models instead of a single model. Each individual model "votes" on the prediction and the majority decides the final prediction. These types of predictors or classifiers are referred to as ensemble predictors.

Model Validation and Evaluation

To ensure that models are learning effectively from the data, some type or several types of model validation should be employed. These methods generally are used to measure and improve model prediction accuracy. The holdout method splits the data into a training set

and a test set. The model learns from the training data, which includes the labels or values for the dependent variable. The model then is "tested" on the test dataset, where the values of the dependent variable are hidden from the model, and the model predicts or classifies the dependent variable. The model prediction can then be compared to the true observed values, and accuracy can be assessed.

Results of a classification algorithm are summarized in a confusion matrix, as displayed in Table 3.1.

	Actual Negative	Actual Positive
Predicted Negative	True Negative (TN)	False Negative (FN)
Predicted Positive	False Positive (FP)	True Positive (TP)

Table 3.1: Confusion matrix.

For a dichotomous classification, the algorithm predicts the observation to be "positive" or "negative", which are assigned to the dichotomous outcome of the target variable by the researcher. The elements of a confusion matrix are used to calculate most performance statistics or metrics. The most straight forward metric is the overall classification accuracy, which is the fraction of correct predictions and shown in Eq. 3.10:

$$accuraccy = \frac{TN + TP}{N_{test}}, \tag{3.10}$$

where $N_{test} = TP + TN + FP + FN$ is the total size of the test dataset. The true positive rate (TPR) or "sensitivity" (Eq. 3.11) is the fraction of positive observations that were correctly classified, and characterizes the accuracy of the model in predicting the positive class. Its converse is the true negative rate (TNR) or "specificity" (Eq. 3.12), which is the fraction of negative observations that were correctly classified.

$$\beta_1 = \frac{TP}{TP + FN}. \tag{3.11}$$

$$\beta_2 = \frac{TN}{TN + FP}. \tag{3.12}$$

In Eqns. 3.11 and 3.12, β_1 is the sensitivity and β_2 is the specificity.

Another common metric is the balanced accuracy, which is the arithmetic mean of the sensitivity and the specificity. It is a good indicator of how well the model predicts both the positive and negative class, and is particularly of importance if the dataset is imbalanced or heavily favors one of the two dichotomous classes. Balanced accuracy can range from 0 to 1 (or 0% to 100%); a balanced accuracy of 0 indicates that there were no correct predictions in the model, and a balanced accuracy of 1 indicates a model that predicted each observation correctly. For a model that predicts every observation to be the majority class, the balanced accuracy would be 0.5, and the overall accuracy would be equal to the ratio of the frequency of the majority class to the sample size of the test set. The balanced accuracy $\mathcal{B}$ is shown in Eq. 3.13.

$$\mathcal{B} = \frac{\beta_1 + \beta_2}{2}. \tag{3.13}$$

Cross-validation is a resampling method that trains a model in different iterations based on different splits of the data. A common form is K-folds cross-validation which randomly partitions the data into K subsets, and then the model is trained K times where each training uses one of the subsets as the test set and the other $K-1$ subsets are used as the training set. Cross-validation gives an estimate of the accuracy of a predictive model. Other methods are

also common, and often methods are combined. Different sampling techniques can also be used to improve model stability, such as bootstrapping, which is discussed in Section 3.2.5

Conclusion

The methods in this chapter are common quantitative tools used in PER, and are used throughout this manuscript. Other methods used in the research presented herein, such as survival analysis, Bayesian networks, decision trees, curricular analytics, and multi-dimensional item response theory, are specific to particular analyses and will be discussed in the chapters in which they were used in the analysis.

Chapter 4

Exploring the Retention of Physics Students

*

*Parts of this chapter were published in "Stewart, J., Hansen, J., & Burkholder, E. (2022). *Visualizing and predicting the path to an undergraduate physics degree at two different institutions.* Physical Review Physics Education Research., **18(2), 020117.**"

4.1 Introduction

Since its inception, Physics Education Research (PER) has investigated issues of critical importance to university physics departments and to the physics community in general. Much of this research has explored issues specific to the teaching and learning of physics [15]. A second more recent strand has explored another central issue, the promotion of diversity, equity, and inclusion in physics programs and physics classes [115, 116]. A third issue of central and sometimes existential importance to physics departments is the retention of physics majors to degree. While the American Institute of Physics maintains detailed data on the number of physics graduates [117] as well as junior and senior undergraduate physics enrollment, little is known about how many students enter physics programs and fail to complete the degree. For many programs, because of the relation between the number of physics majors and university economic support for the department, the retention and recruitment of physics majors represents one of the most important departmental priorities. For some programs, because of state laws closing smaller academic units, retention of majors is a matter of survival [118].

4.1.1 Research Questions

This work explores physics major retention at one institution with a student body with an average level of high school academic preparation. This work investigates factors influencing students departing physics programs through two modes: leaving college entirely and changing to a different major while staying in college.

RQ1: At which point in their undergraduate physics career are students most at risk of

leaving the physics major? How does this differ by modes of leaving the major?

RQ2: What pre-college academic factors influence a student's risk of leaving the major through each mode? How does this change if first semester GPA is added as an independent variable?

This work focuses on pre-college academic factors because these factors largely control the students progression through the first year of college, which will be shown to be key to retaining physics majors. These factors determine the first mathematics classes in which a student enrolls which largely sets the progression of future courses the student must take. Pre-college factors such as ACT scores also form the primary data available to physics programs to inform the adjustment of course structures and the placement of students in those structures to allow more students to succeed.

This work also introduces a number of methods to visualize physics retention which may be useful for physics departments to understand and improve the retention of majors.

4.1.2 Results of prior research

Student retention is a topic of great importance to institutions of higher learning, and has been discussed frequently and extensively in academic publications. Little research has explored physics student retention. The aim of this study is to begin the quantitative analysis of physics student retention within PER. For a brief synopsis of retention research in physics and higher learning, see Chapter 2.

4.2 Methods

4.2.1 Sample

This study investigates retention using samples drawn from a single institution which will be referred to as Institution 1 throughout this chapter. Institution 1 is a large land-grant research university in the eastern United States with total undergraduate enrollment in fall 2020 of 20,500 students. The overall demographics of the undergraduate population were 82% White, 4% Black or African American, 4% Hispanic/Latino, 4% non-resident alien, 4% two or more races, with other groups 2% or less. The ACT composite scores range was 21 to 27 for the 25th percentile to the 75th percentile of students scores [119]. This range of ACT composite scores represents a range of ACT percentile scores of 21 (59%) to 27 (85%). Thirty-one percent of undergraduate students were eligible to receive Pell grants. Pell grants are only given to students of lower socioeconomic status (SES) and are a common measure of the fraction of low SES students.

The dataset included all students who elected a physics major at any point in their undergraduate career from the spring 2001 semester to the fall 2019 semester. The university undergraduate population grew during this time from 16,000 in 2001. The university became more diverse over the time period; White students formed 90% of the undergraduate population in 2001. The ACT score range increased slightly over this period. The details of the filtering of this raw dataset to the analysis dataset are given in Sec. 4.3.1 to show the reader some of the complexities of working with institutional data.

This work discusses four classes commonly taken by physics majors. Calculus 1 is the first semester calculus course introducing integration and differentiation. Physics 1 is the in-

troductory calculus-based mechanics class taken by physical scientists and engineers. Physics 2 is the introductory calculus-based electricity and magnetism course. Modern Physics is taken primarily by physics majors and covers multiple topics including relativity, quantum mechanics, and statistical mechanics. Physics 1 and 2 are presented in the large lecture format with a required co-requisite laboratory session.

4.2.2 Variables

This work uses a set of variables drawn from institutional records. This study used high school GPA (HSGPA), ACT/SAT mathematics percentile score (ACTM), ACT/SAT verbal percentile scores (ACTV), a variable indicating the number of transfer courses for which a student had credit (TranCount), a variable indicating the number of Advanced Placement (AP) courses for which a student had credit (APCount), dichotomous variables indicating whether a student had credit for any AP physics or math courses (APMath, APPhys), and a dichotomous variable MathReady. MathReady was one if the student enrolled in Calculus 1 or a more advanced mathematics class his or her first semester of college, zero otherwise. APPhys and APMath were one if the student had AP credit for any physics or math class regardless of whether the class was required for the physics major, and were zero otherwise. Taking Calculus 1 the first semester of college was required by the four-year physics degree plans. Later, the analysis was repeated with the inclusion of one college-level variable; cumulative GPA after a student's first semester (CGPA).

4.2.3 Statistical and Graphical Methods

This work presents a number of graphical representations of retention and statistical methods to characterize retention. Each will be introduced as it is used. All analyses were performed with the "R" software system [120].

Sankey Plots : Sankey plots give an overall visual picture of retention in physics, drawing retention patterns as flows through a series of semesters.The Sankey plots were drawn with the "ggalluvial" package [121] in "R".

Survival Analysis : Survival analysis was used to calculate a student's risk of leaving the physics major each semester.

Logistic Regression : Logistic regression was used to predict the probability of several outcomes including graduation, one-year persistence, and persistence from Calculus 1 to Modern Physics.

Decision Trees : Decision trees were used to characterize the variables that are the most important in predicting whether a student will persist as a physics major to their sophomore and junior years.

4.3 Results

4.3.1 Descriptive Analysis

This section presents basic descriptive statistics for the various datasets used in the study. To study retention, one must restrict the temporal range of the data to allow time for persistence or graduation. Different time windows were applied for different outcomes

#	Filter	N	Math Ready %	ACTC %	ACTM %	ACTV %	HSGPA	CGPA	Grad Phys %	Grad Other %	Not Grad %	Surv Soph %	Surv Junior %
Institution 1 - Complete Dataset													
1.1	None	586	63					2.99					
1.2	Grad	411	68					3.01	38	29	32		
1.3	Grad, HS	352	68		80	77	3.58	3.00	37	31	32		
Institution 1 - Admit Code Dataset													
1.4	None	463	63					3.00					
1.5	Grad	314	68					3.00	36	30	35		
1.6	Grad, HS	296	69		80	77	3.59	2.99	36	30	34		
1.7	Grad, P1	198	68		76	74	3.51	2.91	31	31	38		
1.8	Grad, HS, P1	187	69		81	78	3.60	2.90	31	32	37		
1.9	1Year, HS, P1	247	66		79	78	3.63	2.92				64	
1.10	2Year, HS, P1	231	67		79	78	3.62	2.91				64	46
1.11	3Year, P1	227	66		75	74	3.53	2.93				64	46
1.12	Grad, P1, First Fall, FTF	143	68					2.94	34	28	38	64	43

Table 4.1: Descriptive statistics applying a variety of filters for Institution 1. Filters are abbreviated: HS (high school) for students with HSGPA and ACT or SAT scores, P1 (Physics first) for students whose first declared major is physics, FTF (First-Time Freshman) students admitted as first-time freshmen, Fall First, students whose first semester was the fall semester. Different windows were also applied to investigate persistence and graduation. Grad (Graduation) removes the last six years of records, 1Year (One year) removes the last year of records, 2Year (Two year) the last two years, and 3Year (Three year) the last three years. Columns are abbreviated: ACTM% (ACT or SAT mathematics %), ACTV% (ACT or SAT verbal %), HSGPA (high school GPA), CGPA (college GPA), Grad Phys % (percentage of student graduating with a physics degree), Grad Other % (percentage of student graduating with a degree other than physics), Not Grad % (percentage of students who do not graduate with any degree), Surv Soph % (percentage of students enrolled as physics majors in their sophomore year), and Surv Junior % (percentage of students enrolled as physics majors in their junior year). Note, Grad Phys %, Grad Other %, and Not Grad % should add to one; for rows in which they do not, it is a result of the cumulative rounding of the numbers.

(i.e. graduation or first-year retention) generating datasets with different overall averages. Further, not all variables were available for all students; restricting to complete records may change the overall average of some variables. The general descriptive statistics are shown in Table 4.1.

One goal of this work is to inform readers interested in replicating this analysis about some of the complexities they may encounter in working with institutional data. The dataset studied included all students who elected a physics major at any time during their undergraduate career from the spring 2001 semester to the fall 2019 semester and course taking data for the same time period, a total of $N = 659$ students. For students early in the dataset, additional course records were obtained to ensure a complete academic record was

available for all students. Of these, 30 students elected the physics major prior to attending the university but were never enrolled as physics majors for a semester in which they took classes; 23 students never took a class in a semester where they were enrolled as a physics major. These students were removed leaving 606 students. An additional 20 students elected a physics major only after completing a degree in another discipline and did not complete the physics major. These students were also removed, leaving 586 students. Descriptive statistics for this set of students are included in the Complete Dataset section of Table 4.1 (Dataset 1.1).

Students were admitted to the university under 11 different admission codes (Admit Codes). The largest group was First-Time Freshman (FTF), 356 students, followed by students readmitted to the university, 76 students, and transfer students, 70 students. Students with admit codes suggesting they might have academic trajectories distinct from other students were removed to form the Admit Code Dataset in Table 4.1. Students without an Admit Code ($N = 7$) were removed as well as visiting students ($N = 5$), transfer students ($N = 70$), non-degree students ($N = 13$), and second degree students ($N = 18$). This resulted in a dataset with 463 records (Dataset 1.4, Table 4.1). Transfer students would be a fascinating cohort to study, but there were not enough of them in the dataset for statistical analysis.

High school academic control variables, HSGPA, ACTM, and ACTV, were not available for all students. Descriptive statistics for students for which these variables were available are shown in the HS rows of Table 4.1. To investigate graduation or persistence to either sophomore year (1-year persistence), junior year (2-year persistence), or Modern Physics (3-year persistence), the latest records must be removed so all students have the same time

to either graduate or persist; the data must be windowed. On sequence students should take Modern physics in the spring sophomore semester; however, Modern is only offered once per year, and therefore off sequence students must often wait until their junior year to take the class. Removing these records changed the overall statistics of the sample little as shown in Table 4.1. A six-year window was used to investigate graduation. With this window applied, the percentage of students graduating with a physics degree (Grad Phys %), graduating with a degree in another discipline (Grad Other %), and not graduating (Not Grad %) was calculated. Each of these outcomes is approximately equally likely in both the Complete Dataset and the Admit Code dataset. One-year and two-year persistence was studied by windowing the data to remove the final one year or two years of records (the codes 1Year and 2Year in Table 4.1). For the one-year, two-year, three-year, and graduation window, the fraction of students surviving to sophomore year as physics majors was calculated (Surv. Soph. %). For the two-year, three-year, and graduation window, the fraction of students surviving to junior year as physics majors was calculated (Surv. Junior %).

4.3.2 Visualizing Retention

College retention is intrinsically a time-dependent process. One method of visualizing the transitions students make between majors and into college outcomes is a Sankey plot. The Sankey plot using the admit code filtered datasets with a graduation window (Table 4.1, Dataset 1.5) are shown in Fig. 4.1. Students' active majors are classified as physics, engineering, other STEM, and non-STEM. Students' outcomes are classified as leaving college, graduate physics, and graduate other. The height of the bar in the Sankey plot represents the number of students in each category each semester. Semesters are numbered from 1

(fall freshman) to 12 (spring year 6); summer semesters have been suppressed. Two vertical bars represent an academic year. Curves are drawn showing transitions between semesters; the color of the curve shows the classification in the later semester; the width of the curve represents the number of students making the transition.

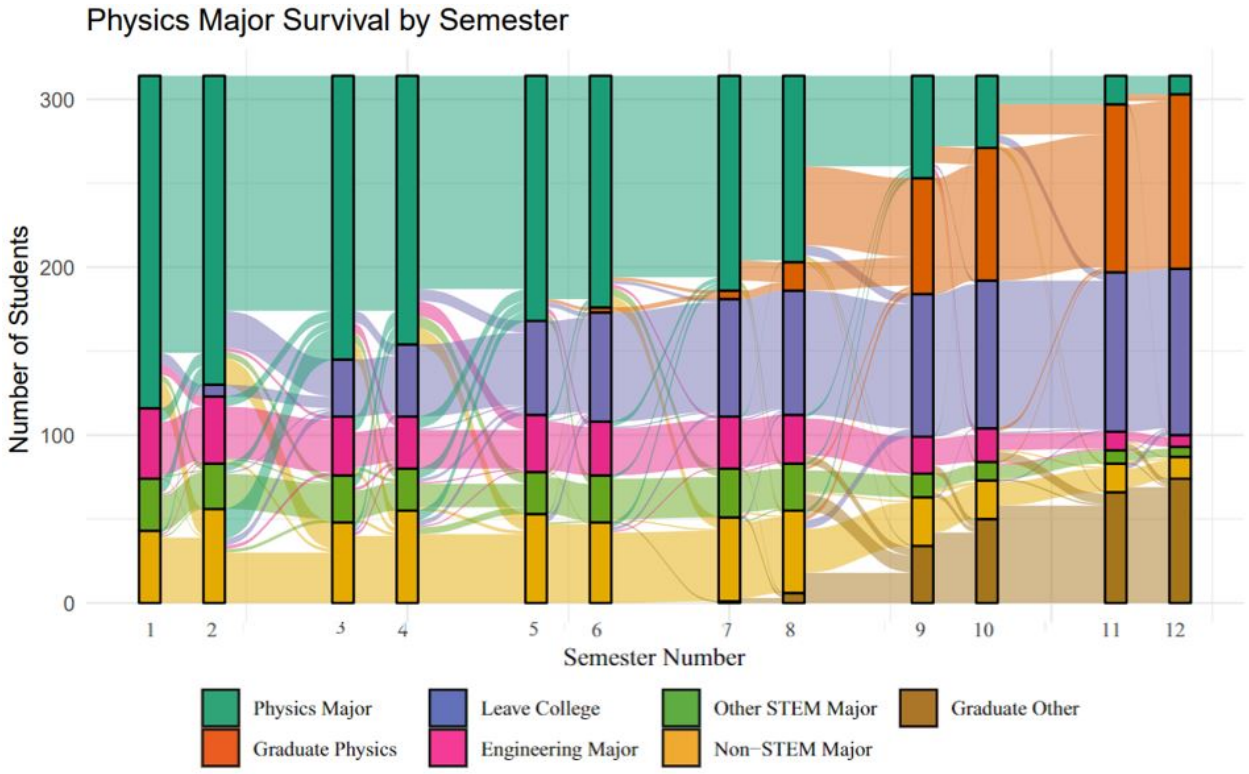


Figure 4.1: Sankey plot showing major changing and graduation patterns for students who elect a physics major at any point in their undergraduate career. Each group of two bars represents an academic year; fall semesters are odd numbers, spring semesters even.

Sequence	Grad Phys %	Grad Other %	Not Grad %
Physics	54 (54%)	0 (0%)	46 (46%)
Other - Physics	50 (76%)	1 (2%)	15 (23%)
Other - Physics - Other	0 (0%)	31 (62%)	19 (38%)
Physics - Other	0 (0%)	61 (68%)	29 (32%)
Physics - Other - Physics	8 (100%)	0 (0%)	0 (0%)

Table 4.2: Institution 1 major election sequences.

Table 4.2 summarizes the patterns observed in Fig. 4.1. These use the same dataset

which was used to construct the Sankey plot. Only 100 of the 314 students are physics majors for their entire undergraduate career; these students graduate with a physics degree 54% of the time. Unfortunately, 46% of these students do not earn a college degree. This college graduation rate is lower than that of the 90 students who start in physics and leave the major for another degree; these students earn college degrees 68% of the time. A substantial group of students, $N = 66$, begin college in other majors and switch to physics; these students graduate with physics degrees 76% of the time and graduate college 77% of the time. One student in the "Other-Physics" pathway earned a degree in another discipline, but not physics. This student was a physics major until the end of their undergraduate career, but applied to graduate with a different major once classes were over.

4.3.3 Survival Analysis

The time dependent nature of college retention and retention to major can be thought of as the process of surviving to graduation. As such, survival analysis, a statistical analysis method originally developed to model the survival of patients with life threatening diseases, represents a promising method to model the process of successfully graduating with the physics major.

Normally, survival analysis attempts to make predictions about a continuous random variable T which represents the time a state-changing event happens (such as dying or quitting school). The variable has probability density $f(t)$ and cumulative distribution function $F(t) = \int_{-\infty}^{t} f(t)dt = P(T < t)$; $F(t)$ is the probability the event has already happened. The survival function $S(t) = 1 - F(t) = \int_{t}^{\infty} f(t)dt$ is the probability the event happens after t or the probability you have survived to t.

The hazard function $\lambda(t)$ is the probability the event happens in the range $[t, t+\Delta t]$ given the event has not already happened at t, the rate the event is happening at time t as shown in Eqn. 4.1.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{4.1}$$

Survival through college to earn a physics degree is an intrinsically discrete process because information on changing majors and leaving college only exists at the semester level. For the discrete case, Eqn. 4.1 simplifies dramatically. For example, the leaving college hazard in semester j, λ_j^{LC}, is the ratio of the students enrolled in semester j who have left college by semester $j+1$, $\Delta N_{j,j+1}^{LC}$, to students enrolled in semester j, N_j, as shown in Eqn. 4.2.

$$\lambda_j^{LC} = \frac{\Delta N_{j,j+1}^{LC}}{N_j} \tag{4.2}$$

A similar definition can be given for the changing major hazard, λ_j^{CM}. The graduation hazard is the fraction of students enrolled in semester j who graduate that semester, N_j^G; $\lambda_j^G = N_j^G / N_j$

For the survival analysis, the data were filtered to a set of maximally homogeneous students after applying a graduation window. The admit code dataset was restricted to include only students who began in the fall semester, who were admitted as first-time freshmen, and who elected physics as their first college major (Table 4.1, Dataset 1.12, $N = 143$). This strong filter was necessary because students who enter in a semester other than the fall have less time until the critical first summer semester. Students who are not initially physics majors may have different course trajectories and require more time to graduate. For this analysis, three modes of leaving the physics major were considered: changing to another

major while staying in college (Change Major), leaving college without earning a degree (Leave College), and graduating with a physics degree (Graduate Physics). The fraction of students in this dataset that leave physics through each mode is shown in Fig. 4.2. The figure shows that approximately twice as many students starting with a physics major leave physics by changing to a different major than those who leave physics by leaving college. The fraction of students leaving college is not directly comparable to the Not Grad % in Table 4.1 because the plot shows the fraction who leave college while still enrolled as physics majors. Note, these results are somewhat different than those shown in the Sankey plot. These differences are a result of the different datasets used. The students used in the survival analysis are students who have the general academic trajectory (first-time freshmen entering in fall semester) around which the undergraduate physics program was designed and are a particularly interesting subpopulation.

The hazard function for all three modes of leaving physics is shown in Fig. 4.3. Note, the graduation hazard (rate) is plotted against the right axis. There is a strong peak in the leaving college hazard at Semester 2. This hazard is understandable; students not thriving at college return home after their freshman year and do not return. There is a peak in the change major hazard at Semester 3, the fall sophomore semester. This likely results from students returning from the summer between freshman and sophomore years and changing their major upon their return. All semesters plotted in the hazard plot enroll at least 50 students.

Also of interest to many academic departments are the courses that may lead to a student leaving the major. An approach that could be used to identify these courses is a form of the hazard function. The per course hazard, λ_i^{LC}, is defined as the hazard of a

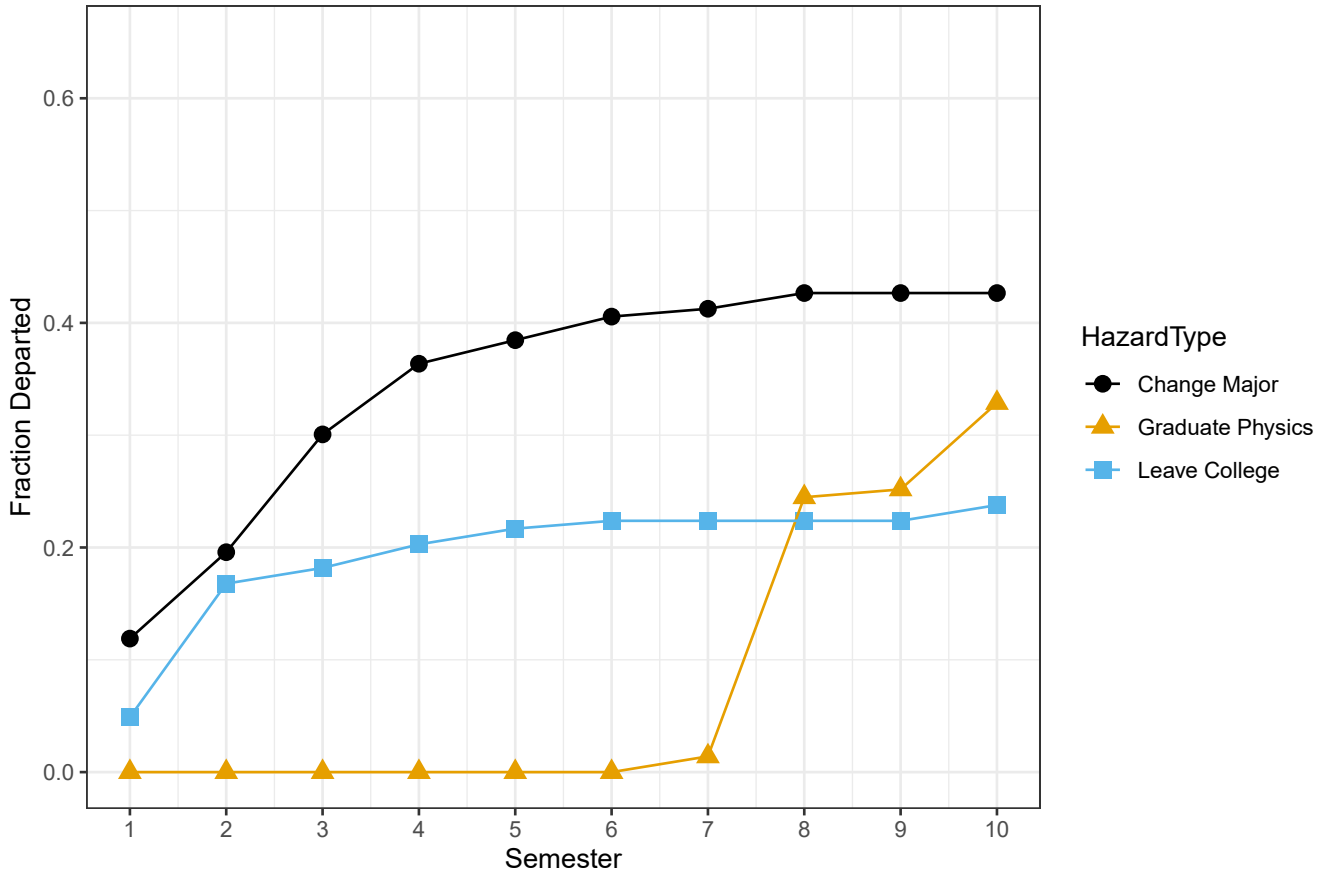


Figure 4.2: Fraction departed or graduated for students entering the university declared as physics majors.

student leaving college immediately after taking a particular course i as shown in Eq. 4.3.

$$\lambda_i^{LC} = \frac{\Delta N_{i,i+1}^{LC}}{N_i} \tag{4.3}$$

which is the ratio of the students who enrolled in course i who left college directly after completing the course (i.e., before taken a subsequent course $i+1$), $\Delta N_{i,i+1}^{LC}$, to students who enrolled in course i as a physics major, N_i. Similarly the hazard of changing majors after a particular course, λ_i^{CM}, is the ratio of students who enrolled in course i who changed majors before enrolling in a subsequent required course to students who enrolled in course i as a physics major. The graduation hazard of a course is the ratio of students who graduated the

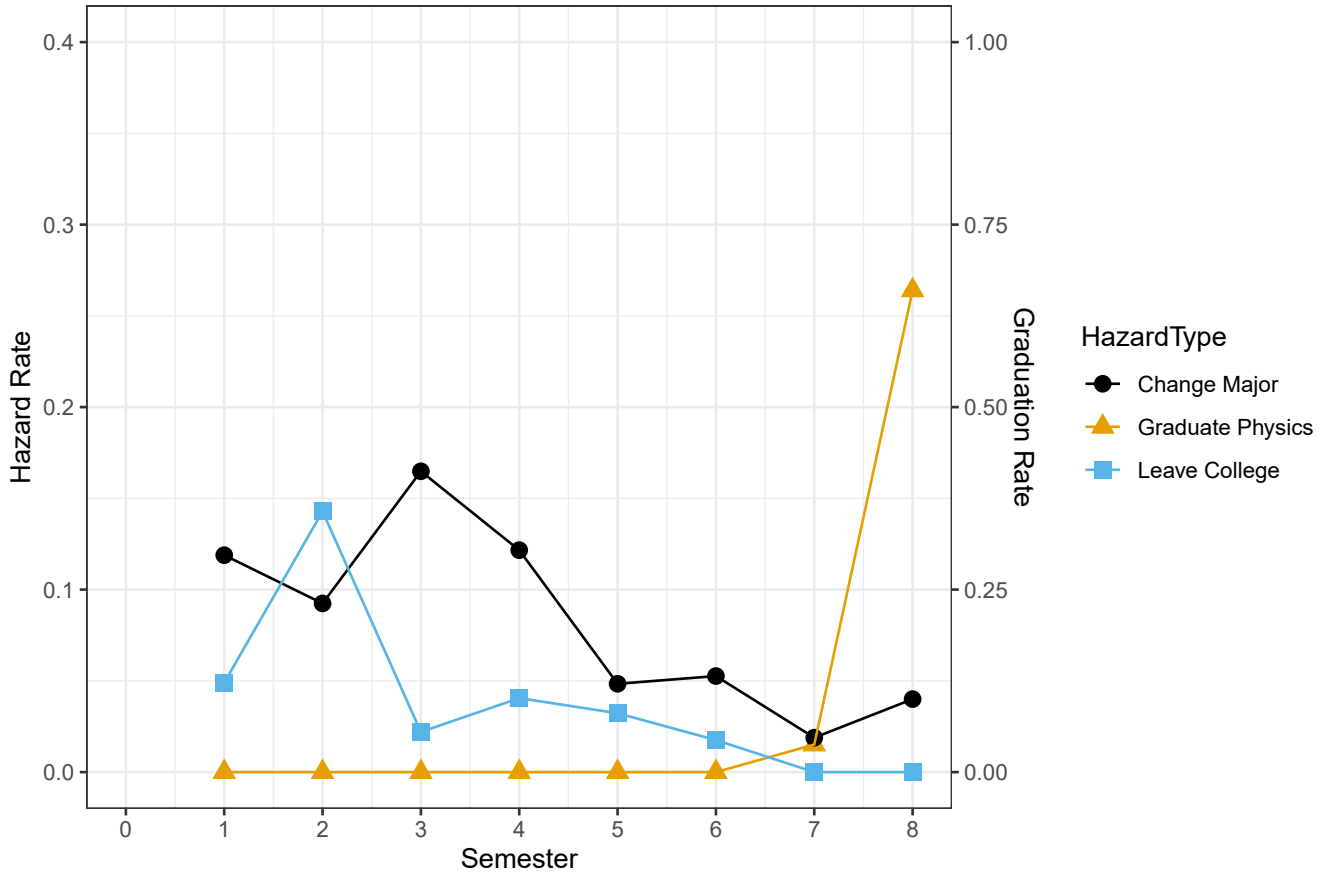


Figure 4.3: Hazard functions. The graduation hazard is plotted on a different scale shown by the right vertical axis. For Institution 1, each semester plotted has at least 50 students enrolled as physics majors.

semester they took course i to students who enrolled in course i. These 'course hazards' are shown in Fig. 4.4. The analysis for these hazard functions was done using the same dataset used for the semester hazard functions (Table 4.1, Dataset 1.12, $N = 143$). Twenty of the students contained in the dataset left the physics major, either through changing majors or leaving college, before they took a physics or math course. These students are not reflected in Fig. 4.4.

The greatest hazard for students leaving college occurs in preparatory math courses such as college algebra, trigonometry, and the stretch calculus course. The stretch calculus course is a course designed for students not yet ready to take Calculus 1, which stretches

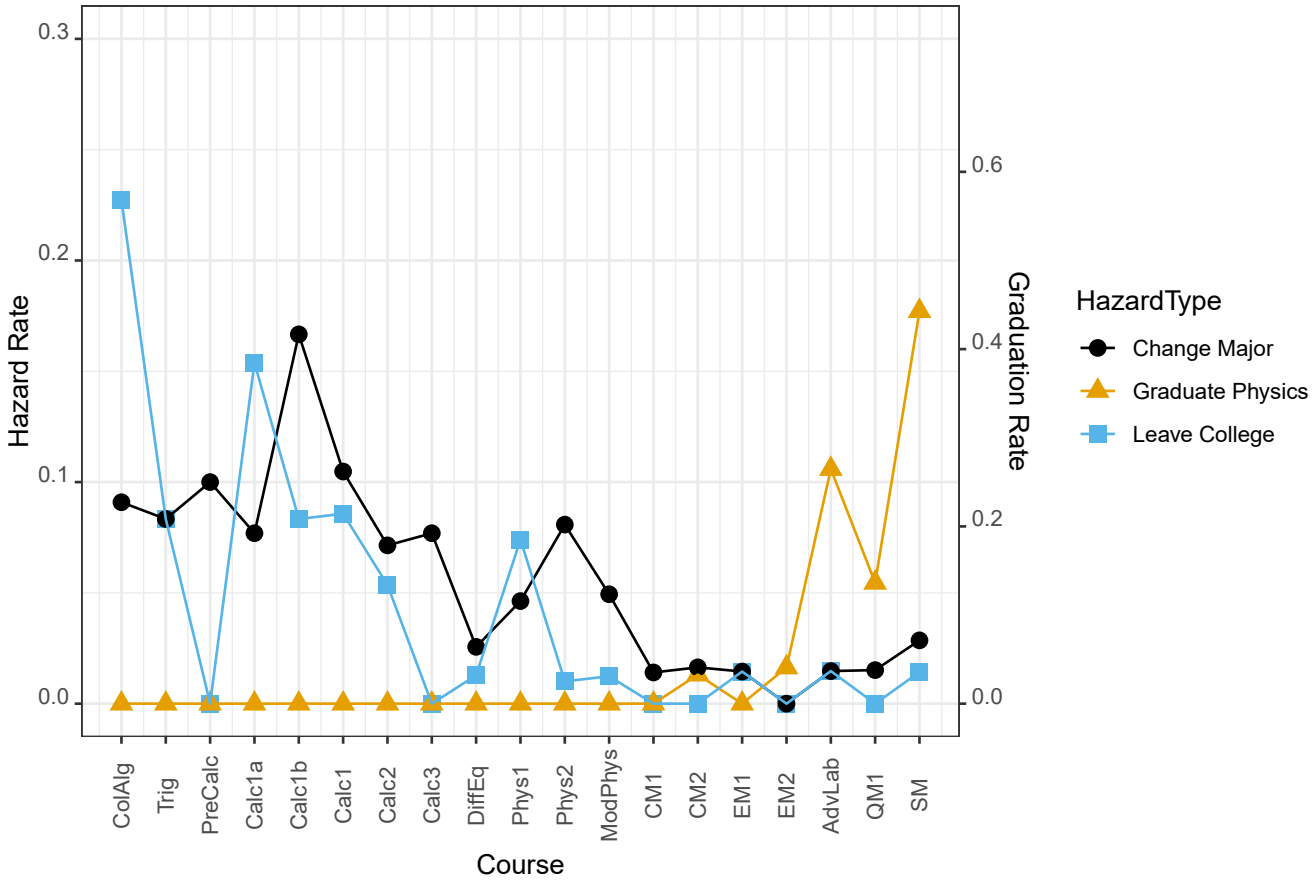


Figure 4.4: Hazard function for required physics and math courses at Institution 1. The hazard in this case is calculated as the number of students who departed the program after taking the course over the number of students who took the course. The axis for leaving the program through graduation is on the right. The abbreviations in the figure are for various subjects in physics: Classical Mechanics 1 & 2 (CM1 & 2), Electricity and Magnetism 1 & 2 (EM1 & 2), Quantum Mechanics 1 (QM1), and Statistical Mechanics (SM).

the content of Calculus 1 over two semesters, Calc1a and Calc1b in Fig. 4.4, and includes pre-calculus content. Students enroll in these courses because they are not ready to enroll in Calculus 1, and as such experience a more difficult path to the completion of the physics degree due to the increased number of required math courses. The leaving college hazard also spikes after students take Physics 1, and then settles down near zero for upper level physics courses. The hazard for changing majors in also greatest for the preparatory math classes and spikes after the second semester of the stretch calculus course. The pre-requisite

math courses for the physics major (Calculus 1, Calculus 2, and Calculus 3) also have a relatively high hazard for changing majors, as does Physics 2, after which the hazard drops near zero.

4.3.4 Logistic regression

Logistic regression allows the modeling of how factors affect a dichotomous dependent variable. Logistic regression predicts the probability of the high level of the dichotomous variable ($Y = 1$); the variable Y is coded so the low level is zero and the high level is one. The probability that $Y = 1$ is observed for student i is modeled by the probability function $P_i(Y = 1)$. The odds of the $Y = 1$ outcome for student i is then calculated as $\text{odds}_i = P_i(Y = 1)/(1 - P_i(Y = 1))$, the ratio of probability of $Y = 1$ being observed to the probability of $Y = 0$ being observed. The range of the odds is from 0 to ∞. To project this quantity into an unbounded range, the log-odds is calculated as $\text{log-odds}_i = \ln(\text{odds}_i)$. The log-odds is then predicted with a set of independent variables very much as a continuous dependent variable would be in linear regression (but with differing underlying statistical assumptions). For example, Eqn. 4.4 predicts the log-odds using two independent variables X_1 and X_2. To do this, an intercept β_0 and two slopes β_1 and β_2 are estimated.

$$\text{log-odds} = \ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{4.4}$$

The intercept predicts the log-odds when X_1 and X_2 are both zero. The slope β_1 is the change in log-odds for a one unit increase is X_1. Log-odds, however, is a fairly difficult quantity to interpret qualitatively. It is much more intuitive to discuss changes in the odds.

To calculate the odds, both sides of Eqn. 4.4 are exponentiated yielding Eqn. 4.5.

$$\text{odds} = \frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0} \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \tag{4.5}$$

As such, e^{β_0} is the base odds when $X_i = 0$ and e^{β_1} multiplies this base odds when $X_1 = 1$.

Logistic regression was used to explore factors influencing persistence to the sophomore year, the junior year, and to graduation. For this analysis, the admit code dataset was filtered to retain only students electing physics as their first college major for whom high-school-level data were available; the data were then windowed for each outcome variable.

This produced the three datasets shown in Table 4.1: 1-year persistence, Dataset 1.9, $N = 247$; 2-year persistence, Dataset 1.10, $N = 231$; graduation, Dataset 1.8, $N = 187$). Table 4.3 presents the logistic regression results for several outcome variables: leaving college by the sophomore year, leaving college by the junior year, leaving the physics major but staying in college though the sophomore year, leaving the physics major but staying in college through the junior year, and graduating with a physics degree. These models were initially fit using HSGPA, MathReady, ACTM, ACTV, TranCount, APCount, APMath, and APPhys as independent variables. The models were fit again using the same independent variables with the addition of the college-level independent variable CGPA. The full regression equations are shown in Eq. 4.6 and Eq. 4.7.

$$\begin{aligned}\text{log-odds}(Outcome) =& \beta_0 + \beta_1 \cdot HSGPA + \beta_2 \cdot ACTM + \\ & \beta_3 \cdot ACTV + \beta_4 \cdot MathReady + \beta_5 \cdot TranCount + \\ & \beta_6 \cdot APCount + \beta_7 \cdot APMath + \beta_8 \cdot APPhys\end{aligned} \tag{4.6}$$

$$\begin{aligned}\text{log-odds}(Outcome) =& \beta_0 + \beta_1 \cdot HSGPA + \beta_2 \cdot ACTM + \\ & \beta_3 \cdot ACTV + \beta_4 \cdot MathReady + \beta_5 \cdot TranCount + \\ & \beta_6 \cdot APCount + \beta_7 \cdot APMath + \beta_8 \cdot APPhys + \\ & \beta_9 \cdot CGPA\end{aligned} \tag{4.7}$$

where β_0 is the intercept, β_i are the slopes, and *Outcome* is one of: graduation in physics, leaving college by sophomore year, leaving college by junior year, leaving physics while staying in college by sophomore year, and leaving physics while staying in college by junior year.

For all models, the model using the independent variables was a statistically significant improvement over the null model. For logistic regression, the null model is the model including only the intercept term. Once the full model shown in Eqn. 4.6 was fit, it was examined for statistically insignificant independent variables. A variable i is determined to be statistically insignificant if its slope β_i is not significantly different from 0. Changing the value of an independent variable with a slope that is not significantly different from zero would have no significant effect on the log-odds of the dependent variable, indicating that the

independent variable does not give any useful probabilistic information about the outcome of the dependent variable. These variables were removed producing a more parsimonious model. An ANOVA test showed the model removing insignificant independent variables was not significantly less well fitting than the full model in all cases. This model is shown in Table 4.3. For the majority of models only one variable was retained; however, models predicting graduating with a physics degree and passing Physics 2 or Modern Physics as a physics major retained more than 1 variable.

Variable	β	SE	z	p	e^{β}
Leave College by Sophomore Year ($N = 247$)					
(Intercept)	-2.12	0.22	-9.62	0.0000	0.12
HSGPA	-0.69	0.19	-3.65	0.0003	0.50
Leave Physics Stay in College by Sophomore Year ($N = 247$)					
(Intercept)	-0.39	0.22	-1.73	0.0827	0.68
MathReady	-1.32	0.31	-4.25	0.0000	0.27
Leave College by Junior Year ($N = 231$)					
(Intercept)	-1.83	0.20	-8.94	0.0000	0.16
HSGPA	-0.72	0.18	-3.93	0.0000	0.49
Leave Physics Stay in College by Junior Year ($N = 231$)					
(Intercept)	0.34	0.23	1.47	0.1404	1.41
MathReady	-1.29	0.29	-4.40	0.0000	0.28
Graduate Physics ($N = 187$)					
(Intercept)	-1.64	0.40	-4.14	0.0000	0.19
HSGPA	0.91	0.23	4.02	0.0000	2.49
MathReady	0.88	0.45	1.96	0.0504	2.41
Enroll Calculus 1 - Pass Physics 2 as Major ($N = 132$)					
(Intercept)	0.25	0.21	1.19	0.2323	1.29
APCount	0.72	0.28	2.53	0.0114	2.05
TranCount	0.48	0.23	2.10	0.0358	1.61
HSGPA	0.89	0.21	3.58	0.0003	2.44
Enroll Calculus 1 - Pass Modern as Major ($N = 132$)					
(Intercept)	-0.65	0.22	-2.94	0.0032	0.52
APCount	0.68	0.24	2.83	0.0047	1.98
HSGPA	.91	0.27	3.39	0.0007	2.48

Table 4.3: Logistic regression. All regressions are significant improvements over the null model ($p < 0.001$). β is the normalized regression coefficient, SE is its standard error, z is the z-score of the coefficient, p the probability a value larger than z occurred by chance, and e^{β} is the odds ratio.

Using only the pre-college independent variables, the results for persistence in physics were quite different than the results for persistence in college. Persistence in college while leaving the physics major was most strongly related to math readiness (being able to enroll in Calculus 1 the first semester of college). The base odds of leaving physics while staying in college (the odds, e_0^{β}, of the intercept) was reduced by a factor of 0.27 for the sophomore year and 0.28 for the junior year for math ready students. As such, being math ready decreases the odds of leaving the major by $(1/0.28 - 1) \cdot 100\% = 260\%$. In other words, it reduces the odds of leaving the major by a factor of 2.6. The relation of math-readiness to leaving the physics major but remaining in college is very understandable; non-math-ready students have to take a sequence of mathematics classes, often a year and a half of mathematics classes, before ever enrolling in their first physics class. They also are very unlikely to complete their degree in four years. These factors make them very hard to retain and add financial pressures to the student to change to a less math intensive major. This is also reflected in Fig. 4.4, where the preparatory math courses (those before the traditional Calculus 1 course) have the highest hazard rates for leaving the major but staying in college.

The variables important in predicting whether a physics student would leave college by the sophomore or junior year were quite different; HSGPA was the most important variable. While high school classes and curricula are extremely variable, HSGPA provides a measure of how successful a student has been in the high school academic system. This success is an important indicator of whether the student will successfully navigate college. Both MathReady and HSGPA were important in predicting graduation with a physics degree (MathReady was $p = 0.0004$, below the 0.05 significant threshold). A student who graduates with a physics degree must avoid both leaving the major and leaving college, so it is reasonable that both

factors are involved. Both factors have similar odds ratios in predicting graduation; math readiness increased the odds of graduating with a physics degree by $(2.41-1)\cdot 100\% = 141\%$ and a one standard deviation increase in HSGPA increases the odds by 149%.

Variable	β	SE	z	p	e^β
Leave College by Sophomore Year ($N = 247$)					
(Intercept)	-2.28	0.24	-9.47	0.0000	0.10
CGPA	-0.94	0.18	-5.35	0.0000	0.39
Leave Physics Stay in College by Sophomore Year ($N = 247$)					
(Intercept)	-0.39	0.22	-1.73	0.0827	0.68
MathReady	-1.32	0.31	-4.25	0.0000	0.27
Leave College by Junior Year ($N = 231$)					
(Intercept)	-1.96	0.22	-8.88	0.0000	0.14
CGPA	-0.98	0.18	-5.57	0.0000	0.37
Leave Physics Stay in College by Junior Year ($N = 231$)					
(Intercept)	0.34	0.23	1.47	0.1404	1.41
MathReady	-1.29	0.29	-4.40	0.0000	0.28
Graduate Physics ($N = 187$)					
(Intercept)	-1.34	0.26	-5.21	0.0000	0.26
CGPA	1.79	0.37	4.81	0.0000	5.98
Enroll Calculus 1 - Pass Physics 2 as Major ($N = 132$)					
(Intercept)	0.09	0.25	0.36	0.7190	1.09
APCount	0.59	0.29	2.06	0.0394	1.80
TranCount	0.66	0.28	2.32	0.0203	1.94
CGPA	1.62	0.39	4.18	0.0000	5.08
Enroll Calculus 1 - Pass Modern as Major ($N = 132$)					
(Intercept)	-1.07	0.31	-3.43	0.0006	0.34
APCount	0.57	0.25	2.32	0.0203	1.77
CGPA	1.95	0.50	3.87	0.0001	7.04

Table 4.4: Logistic regression including first semester GPA. All regressions are significant improvements over the null model ($p < 0.001$). β is the normalized regression coefficient, SE is its standard error, z is the z-score of the coefficient, p the probability a value larger than z occurred by chance, and e^β is the odds ratio.

This picture is changed, however, by including the college-level independent variable CGPA. As shown in Table 4.4, the significant variable for staying in college but leaving physics is still MathReady, but the significant variable for leaving college is now CGPA for leaving college by the sophomore year and the junior year. CGPA is significant in graduating with a physics degree, but MathReady is not. It is not surprising that a student's college GPA

would be more important than their high school GPA in determining their persistence; college GPA indicates how successfully a student traverses the university or college academic system, and would be a more accurate measure of success than high school GPA. Interestingly, it is still a student's math readiness that is predictive of whether they will leave the physics major by their sophomore and junior year, indicating that calculus ready students are better equipped to navigate the physics major through the first two years of college.

4.3.5 Decision Trees

Decisions trees are a common machine learning algorithm that are used for describing data and classification tasks. A decision tree predicts the outcome of a target variable based on a model built from the input of independent variables. The algorithm takes the dataset or "root node" and splits it by each independent variable, and measures which variable splits the data into the "most" homogeneous subsets (each subset should be heavily weighted to one of the outcomes of the target variable). The criterion associated with a split is the threshold the tree uses to make a decision for the split; for example, whether a student has a CGPA greater than or equal to 3.5. Each subset is then split using the same method, and the process continues until the final subsets are perfectly homogeneous. This creates a tree of nodes, where each internal node represents a subset of a prior split and is characterized by the criterion that splits the subset in a way that maximizes homogeneity. Typically decision trees are "pruned back" so as to balance complexity with predictive power, and the terminal nodes or "leaves" are not always purely homogeneous. Decision trees are a good indicator of the relative variable importance for a model, as variables that appear closer to the root node are more important in predicting student outcomes. Decision trees are less susceptible

to multicollinearity when compared to other common PER statistical methods such as linear regression.

A decision tree was formed for the three outcomes of surviving as a physics major to the sophomore year, surviving as a physics major to the junior year, and graduating as a physics major. In these analyses, students who have a negative outcome ("NotSurvive" for surviving to sophomore and junior year and "NotPhysGrad" for graduating as a physics major) for the target variable could have left the program either by leaving college or changing majors. This differs slightly from the logistic regression analyses, where only one method of leaving the physics program was investigated at a time. CGPA was not included as an independent variable in constructing these decision trees, with the intent of identifying the variables which are useful in predicting when students may struggle in the physics program before those students begin classes.

Fig. 4.5 shows the decision tree for predicting whether a student persists as a physics major to their sophomore year. Each node is labeled by the majority class of that subset, either "Survive" or "NotSurvive", and the percentage at the bottom of the node is the percentage of the original dataset the node represents (the root node shows 100%). The middle numbers show the distribution of the node for the target variable (surviving to the sophomore year in this case) and the text below each node indicates the variable and associated criterion for the subsequent split or "decision". Unsurprisingly, MathReady is the most important variable in determining whether a student will be retained in the physics program by their sophomore year, agreeing with the outcome in the logistic regression analysis of whether a student leaves physics but stays in college by their sophomore year. With the decision tree, we can see the difference being math ready makes; 75% of students who are

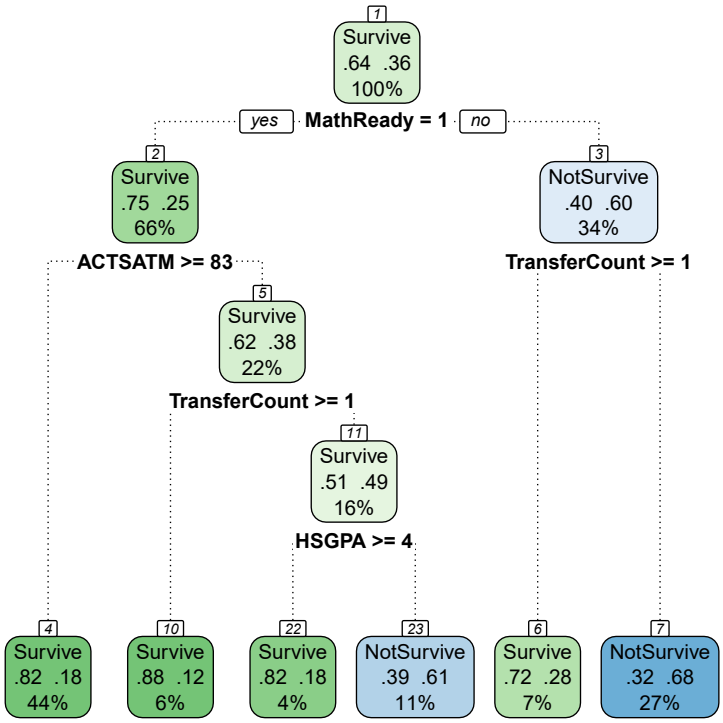


Figure 4.5: Decision tree for persisting in physics to the sophomore year

math ready are retained while only 40% of students who are not math ready are retained. Students who are not math ready but have taken some college transfer courses (generally as a dual-enrollment course in high school) are 40% more likely to be retained in physics by their sophomore year than students who did not have college transfer credit and were not math ready. Other important variables in "deciding" whether a student is retained by their sophomore year include ACTSATM and HSGPA.

Fig. 4.6 presents the decision tree of persistence to junior year, and overall a student starting in physics is more likely to leave physics by their junior year. MathReady is still the most important variable, with the same 40% increase between surviving and not surviving based on whether or not a student was ready for calculus in their first semester. In this

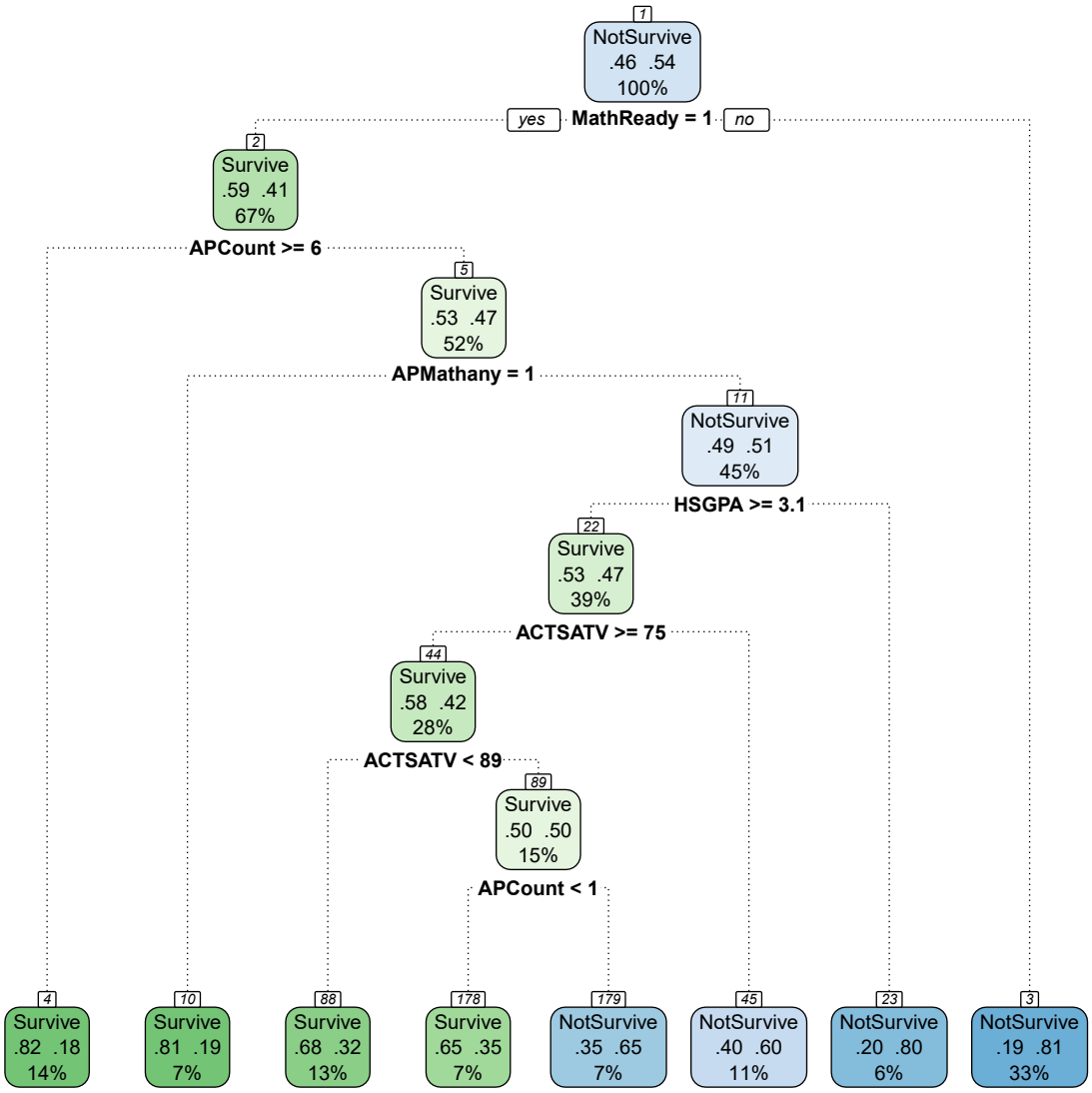


Figure 4.6: Decision tree for persisting in physics to the junior year

case, instead of the number of transfer courses a student has credit for being important, the number of AP classes they earned credit for is important, followed by whether they took an AP math course, HSGPA, and ACTSATV.

Fig. 4.7 shows the important variables for determining whether an incoming freshman will graduate as a physics major. HSGPA is the most important variable, and students with

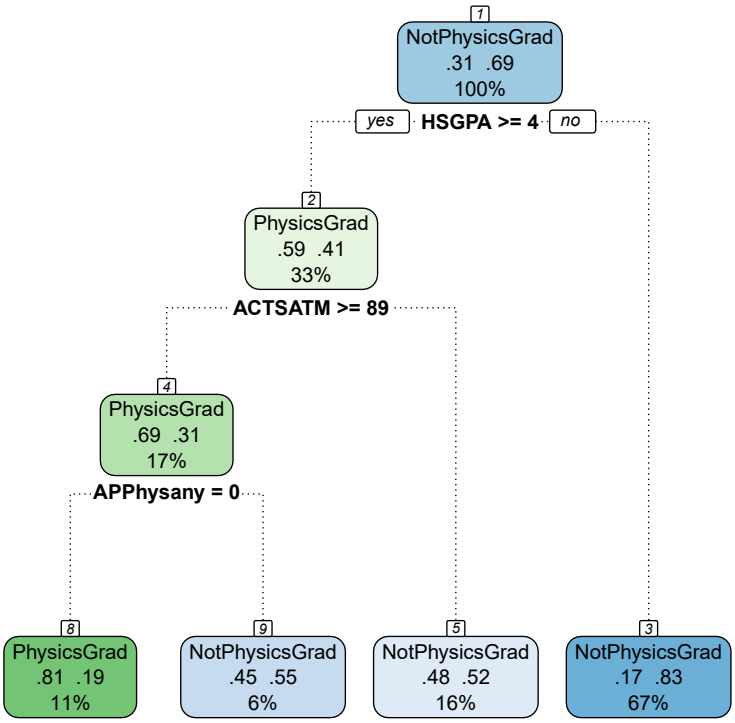


Figure 4.7: Decision tree for persisting in physics to graduation

a HSGPA greater than 4 are 42% more likely to graduate in physics that those with a lower HSGPA. Typically, if a student has a GPA higher than four it is indicative that their school district uses a bonus point system, where AP courses, dual-enrollment courses, and other college preparatory courses are worth more than the traditional four GPA points of a regular high school course. A HSGPA greater than four indicates the student has enrolled in and done well in these types of courses, whether they earned transfer credit and AP credit or not.

4.3.6 Traversing the course network

As a student persists in college they traverse a network of required courses. For a physics major at Institution 1, the key sequence of courses early in college is Calculus 1, Physics 1, Physics 2, then Modern Physics. The logistic regression analysis was repeated to explore the factors influencing whether a student who enrolls in Calculus 1 persists to either Physics 2 or Modern Physics.

HSGPA, APCount, and TranCount were the most important predictors of a student who enrolled in Calculus 1 passing Physics 2 as a major as shown in Table 4.3. The same is true for passing Modern Physics as a major, except TranCount is not retained in that model. A one standard deviation higher HSGPA increased the odds of staying a physics major through Modern by 150%. For the set of models that include CGPA, a similar result is found, except HSGPA is replaced by CGPA (see Fig. 4.4). In this case, a one standard deviation higher CGPA increased the odds of staying a physics major through modern by 600%.

Examining the progression of students through the network also provides additional insights. Figure 4.8 shows the progression of students who enter Institution 1 declared as physics majors through Modern Physics and to graduation. For this analysis, a 3-year window was applied to the admit code filtered dataset (Table 4.1, Dataset 1.11, $N = 227$). Students first enrolling in Modern Physics or a more advanced physics class were removed (8 students); students who never took a mathematics class were also removed (10 students) leaving 209 students for analysis. The figure uses the abbreviations "<Calc" for students whose first mathematics class is less advanced than Calculus 1, "Calc" for students whose

first mathematics class is is Calculus 1, and ">Calc" for students whose first mathematics class is more advanced than Calculus 1.

The figure starkly shows the importance of math readiness for this population. Of the 209 students, 41% first enroll in a mathematics class less advanced the Calculus 1; 59% of these students leave physics before enrolling in Physics 1. Of the 37% of the students who first enroll in Calculus 1; only 26% of these leave physics before enrolling in Physics 1. Students with AP or transfer credit for Calculus 1 first enroll in a mathematics class more advanced than Calculus 1; only 7% of these students fail to enroll in Physics 1. The advanced math entry students have a persistence advantage over other students through Modern Physics. Once either a non-math-ready or a Calculus 1 entering student enrolls in Physics 1, they persist to Physics 2 at about equal rates. This indicates that pre-college factors are most important in allowing students to persist to enroll in a physics class; once the student successfully enrolls in physics, pre-college factors become less important. From Physics 2, the non-math-ready student persists to Modern at a somewhat lower rate than the Calculus 1 entry student. Of the 209 initial physics majors, 19 of the 82 non-calculus-ready students enroll in Modern Physics as a physics major, 23%; 44 of 84 Calculus 1 entry students enroll in Modern Physics, 52%; 28 of the 43 advanced math entry students enroll in Modern Physics, 65%.

For the graduation probabilities after enrolling in Modern Physics in Fig. 4.8, a 6-year window was applied (Table 4.1, Dataset 1.7, $N = 198$). As before, students who first enroll in Modern or a more advanced physics class and students who never enroll in a mathematics class were removed leaving 181 students. Figure 4.8 presents the graduation probability of these students once they enroll in Modern Physics. The graduation rates for all math entry

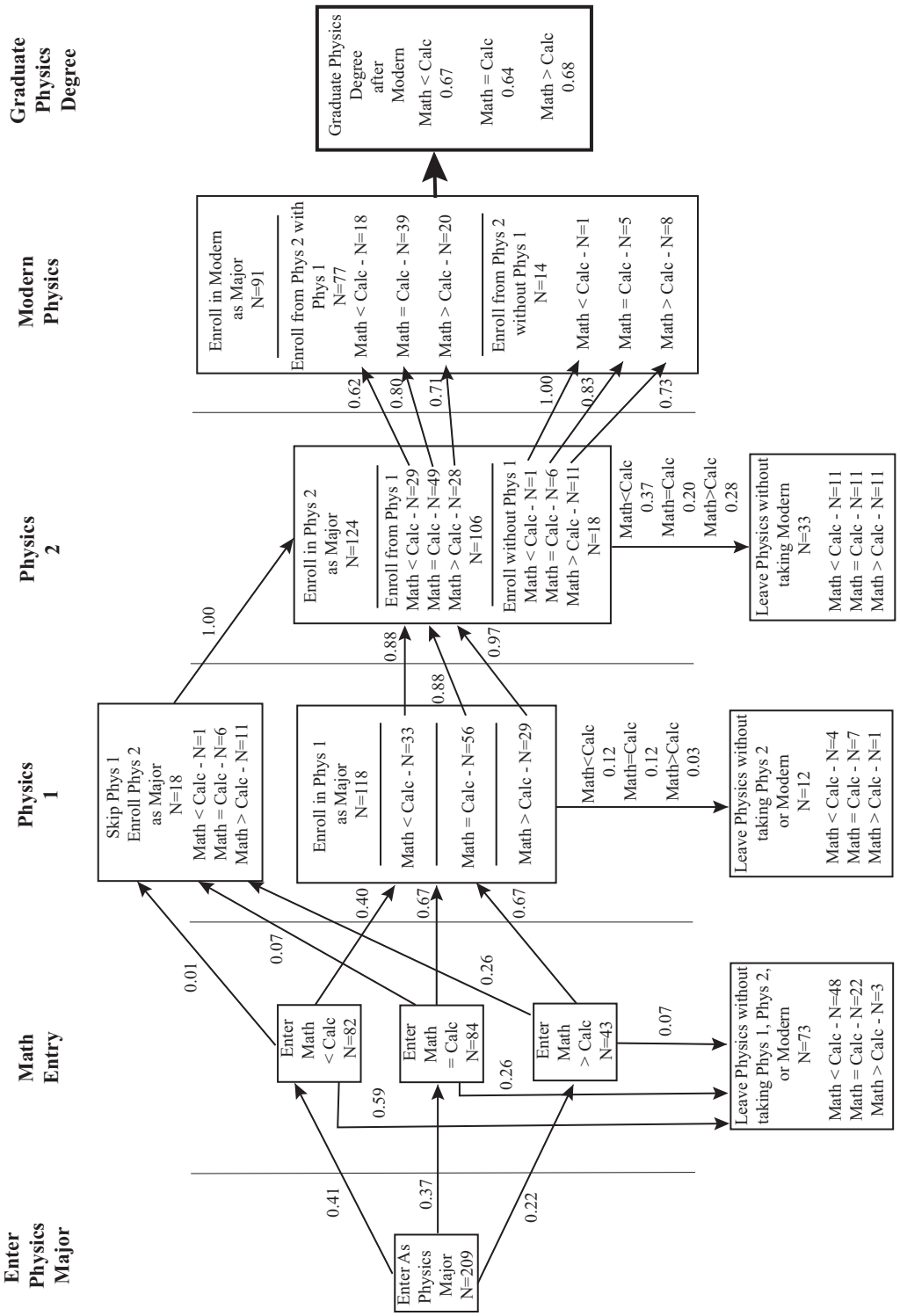


Figure 4.8: Traversing the major from entry to Modern Physics for students at Institution 1 who elect a physics major in their first semester. The figure uses the abbreviations <Calc for students whose first mathematics class is less advanced than Calculus 1, Calc for students whose first mathematics class is is Calculus 1, and >Calc for students whose first mathematics class is more advanced than Calculus 1.

points are approximately equal; all students who persist to Modern have an equal chance of graduating with a physics degree.

For the graduation filtered dataset, overall graduation probabilities in physics were calculated for each stage of the progression through the network. Of the 181 students who initially enrolled as physics majors, 31% graduated with a physics degree. Disaggregating by math readiness, of the 65 students not ready to take Calculus 1, 15% graduated; of the 76 students who initially enrolled in Calculus 1, 34% graduated with a physics degree; and of the 40 students who initially enrolled in a mathematics class more advanced than Calculus 1, 53% graduated with a physics degree. Of the 100 students who enrolled in Physics 1 as a physics major, 50% graduated with a physics degree (<Calc 1 42%, Calc 1 46%, >Calc 1 65%). Of the 107 students who enrolled in Physics 2 as a physics major, 53% graduated with a physics degree (<Calc 1 45%, Calc 1 53%, >Calc 1 58%). Of the 79 students who enrolled in Modern Physics as a physics major, 65% graduated with a physics degree (<Calc 1 67%, Calc 1 64%, >Calc 1 68%). As such, the additional advantage confirmed by a more enriched high school STEM experience was important in the early years of college, but ceased to be important once a student progressed to their advanced coursework. We note the 65% graduation rate for students who enroll in Modern Physics is much smaller than the department would like and this will be one target of retention efforts.

4.4 Discussion

This study sought to answer two research questions; they will be addressed below. The detailed results were discussed above; the following will synthesize the most important

points.

RQ1: At which point in their undergraduate physics career are students most at risk of leaving the physics major? How does this differ by modes of leaving the major? The risk (hazard) profiles for the two modes of leaving the physics major (leaving college or leaving the major while staying in college) were quite different as shown in Fig. 4.3. At Institution 1, there was a peak in the leaving college hazard in the spring freshman semester as students failed to return to campus for the fall sophomore semester. This hazard decreases dramatically after this point.

This hazard for leaving the major while staying in college peaked in the fall sophomore semester. Students made the major-changing decision when they returned to campus for their sophomore year. The hazard declined after this point, but did not reach zero until the fifth year. Students who do leave physics appear to enter both STEM and non-STEM majors at similar rates; non-STEM majors approximately equal STEM majors (including engineering) as alternate majors selected by physics students in Fig 4.1.

The course hazard function in Fig. 4.4 mirrors this. For on-sequence students, Physics 1 and Calculus 2 are typically taken together in the second semester and these courses have a similar hazard for leaving college, which mirrors the result of the spike in Fig. 4.3 for leaving college after semester 2. This holds for Physics 2 and Calculus 3, which are typically taken together in semester 3 and have a similar hazard for leaving physics but staying in college, and mirror the spike in leaving physics but staying in college for semester 3 in Fig. 4.3. Calculus 1 has the greatest hazard for leaving college and leaving the major out of all of the required courses for physics major. It appears that once a student has completed Calculus 2 and Physics 1 they are far less likely to leave college, and once a student completes Modern

Phyiscs and Differential Equations they are less likely to leave the major. Unfortunately, the hazard of leaving the physics major by either leaving college or changing majors is still non-zero for the upper-level physics courses, indicating that students are still facing challenges within the program after completing the introductory course sequence and math prerequisites. The hazards for the preparatory math courses (College Algebra, Trigonometry, Pre-Calculus, Calculus 1a, and Calculus 1b) are some of the highest in the figure, but there are far fewer students who enrolled in those courses as physics majors; between 20 and 10 students enrolled in College Algebra, Trigonometry, Pre-Calculus, Calculus 1a, and Calculus 1b, whereas there are between 120 and 50 students for the other courses in Fig. 4.4.

RQ2: What pre-college academic factors influence a student's risk of leaving the major through each mode? How does this change if first semester GPA is added as an independent variable? The factors influencing different outcomes, one year persistence, two year persistence, and graduation, differed between the different modes of leaving the major. These factors were explored using logistic regression as shown in Table 4.3. Leaving the major while staying in college was most strongly related to math-readiness. The odds that a math ready student would leave the physics major for another major were 260% lower than a non-math-ready student. Not being math ready increases time to degree and delays entry into physics classes, making retention difficult, and other majors with less restrictive mathematics requirements more attractive. Leaving college was more related to general high school preparation and success measured by HSGPA. Each standard deviation increase in HSGPA lowered the odds of leaving college by the junior year by 100%.

Table 4.4 explored the same outcomes as Table 4.3 except it included first-semester college GPA as an independent variable. The results were the same, except CGPA replaced

HSGPA as the significant variable in predicting if a student leaves college by their sophomore and junior year. This indicates that once a student has some college experience, their performance in college is more predictive than their high school preparation in determining whether they will stay in college. A standard deviation increase in CGPA lowers the odds of leaving college by the junior year by a factor of 1.7. CGPA also is the single significant variable in predicting whether a student graduates in physics, with each standard deviation increase of CGPA increasing the odds of graduation by 500%. For this dataset (Dataset 1.8 in Table 4.1) the standard deviation of first semester college GPA is 1.1, or roughly one letter grade.

The decision tree analysis presented in Sec. 4.3.5 examined the effect of the variables on a simplified outcome of leaving physics (either by leaving college or leaving the major) at a specified point. Students who enter college ready to take Calculus 1 are 35% more likely to persist to their sophomore year as a physics major. Of those students who are not math ready, if they have some transfer credit, they are 40% more likely to persist in physics to their sophomore year. The MathReady variable continued to be the most important variable in predicting a student's persistence in physics to the junior year, as students who were math ready were 40% more likely to persist in physics.

The progression through the major and the role of math readiness was further explored by examining the progression through the course network in Fig. 4.8. At this institution, 41% of students enrolled as physics majors their first semester were not ready to enroll in Calculus 1; 59% of these students left physics without ever enrolling in Physics 1. Only 15% of these students graduated with a physics degree. For students whose first mathematics class was Calculus 1, 34% graduated with a physics degree; for students who first enroll in

a mathematics class more advanced than Calculus 1, 53% graduated with a physics degree. This illustrates the importance of access to advanced high school course offering to success in physics. Some students underrepresented in physics may have limited access to these courses [122]. There were few differences in physics graduation rates for students who remained in the major long enough to enroll in Modern Physics. This is somewhat reflected in the variable importance as found in the decision tree analysis for graduating physics (Fig. 4.7). For graduating in physics, if a student had a HSGPA greater than or equal to four they were 42% more likely to graduate in physics; MathReady was not a significant variable. Once a students enrolls in Modern Physics, math readiness ceases to be important, and HSGPA becomes the most important characteristic of students who are retained and students who are not.

4.5 Implications

For Institution 1, the analysis suggests three points where retention efforts could be directed. Non-math-ready students succeed in the major at very low rates and often leave the major before taking Physics 1. Exploring methods to allow these students to begin taking physics while they catch up in mathematics might retain more to the major. This might involve allowing these students to take the algebra-based physics sequence and accepting these for the calculus-based Physics 1 and 2 with successful completion of Modern Physics and Calculus 1. There is a continuous slow attrition of majors after semester 4 (spring sophomore semester) when students are taking their advanced coursework. This suggests Institution 1 should examine the features of their advanced undergraduate program that

cause students to leave late in the program. Finally, the institution loses majors at the highest rate after the spring freshman semester to the leaving college hazard and after the fall sophomore semester to the changing major hazard (the changing major decision may have been made the semester before). This suggests substantial efforts be focused on retention in the first year of college. Efforts currently under discussion include a redesigned freshman seminar course focused on retention, a freshman research experience with a cohort building element, and an introductory laboratory section for physics majors taught by faculty.

4.6 Limitations and Future Work

This study was performed at one institution with a relatively small physics undergraduate program. This work should be replicated at other programs, both at larger programs and similar programs with different demographic composition, so as to map out the spectrum of physics retention. This work was unable to explore differences in retention of demographic groups underrepresented in physics; these differences should be explored in future studies.

4.7 Conclusions

This work examined the retention of physics majors through multiple points in their undergraduate career at one institution. At Institution 1, many students arrive on campus who are not ready to enroll in Calculus 1. There was a peak in the risk of leaving the physics major by leaving college in the spring freshmen semester. The changing major risk was highest in the fall sophomore semester. Math readiness emerged as the key factor predicting changing to major other than physics while staying in college; students who were math

ready were 260% more likely to be retained in physics up to their junior year. Math ready students are prepared to enroll in Calculus 1 or a more advance mathematics class their first semester of college. 41% of students electing a physics major their first semester were not math ready; only 15% of these graduated with a physics degree; 37% of incoming physics majors enrolled in Calculus 1 their first semester; 34% of these graduated with a physics degree. This analysis also suggested advanced high school college preparatory curriculum was important in physics student success; 22% of incoming physics majors had high school credit for Calculus 1 and enrolled in a more advanced class; 53% of these students graduated with a physics major.

Different factors were important in predicting leaving college and graduating. High school GPA was the most important factor in predicting retention to college and graduation with a physics degree; math readiness was the most important factor predicting leaving physics while staying in college.

Chapter 5

Examining the Conditional Probabilities of Physics Student Retention with Bayesian Networks

5.1 Introduction

In the study presented in the prior chapter, two critical points were identified in which the hazard of leaving the physics program of Institution 1 was greatest; at the end of a student's first year for leaving college, and at the start of a student's sophomore year for switching majors. Surviving in the physics program beyond these two points can be considered a "milestone" in a student's academic progress towards completion of the physics program. Enrolling in introductory Physics 2 (PHYS 112), the second course in the calculus-based introductory sequence at Institution 1, roughly co-incides with this milestone. PHYS 112 is usually taken in the fall of a student's sophomore year, though it is often taken in the spring of a student's sophomore year as well, depending on the student's math readiness. Students who enroll in PHYS 112 have survived past these two critical points of attrition. Another critical point in the progression of physics students was the enrollment of students in Modern Physics, or PHYS 314 at Institution 1. It was at this point that the effect of a student's pre-college math preparation diminished, as students who enrolled in PHYS 314 completed the degree at roughly the same rate regardless of math readiness. Enrolling in PHYS 314 could be considered another milestone in student progress.

The study presented in the previous chapter identified a kind of "hierarchy" to the pre-college academic factors that influence if and when a student is likely to leave the physics program at Institution 1, with HSGPA and a student's math readiness being the most influential. This study examines the probabilities of students reaching particular milestones in the physics curriculum based on their pre-college academic factors. To be able to determine the probability of retaining a student, or of the student reaching the milestones discussed above,

could be of great value to physics departments. Students with different pre-college characteristics likely have different probabilities of reaching specific milestones. Knowing which students will struggle to reach a particular milestone gives physics departments the ability to offer an intervention or change the structure of their program to retain more students. The implementation of an intervention that successfully improves students' probabilities of reaching these milestones has been reserved for future research.

5.1.1 Research Question

This study investigates critical points of progression in the physics curriculum at Institution 1. The critical points or milestones investigated are enrolling in PHYS 112, enrolling in PHYS 314, and graduating from the physics program.

RQ1: What is the probabilistic relationship between various points of progression in a physics curriculum and pre-college academic factors? How do these relationships change when physics course grades are added to the model?

The relationship between reaching these milestones and pre-college factors is investigated because the pre-college factors are available as soon as a student enrolls in the university. The addition of some physics course grades gives an indication of how the usefulness of the pre-college academic factors changes as a student progresses in the program.

This study also introduces Bayesian networks into PER. Bayesian networks encode the global probability distribution of a set of random variables, and are an useful tool in determining the conditional probabilities between variables.

5.1.2 Bayes' Theorem

One of two main theoretical underpinnings of Bayesian networks is Bayes' Theorem or Bayes' Rule developed by Rev. Thomas Bayes [123]; it is shown in Eqn. 5.1

$$P(A \mid B, c) = \frac{P(A \mid c) \times P(B \mid A, c)}{P(B \mid c)}. \tag{5.1}$$

The term $P(A \mid B, c)$ represents the probability of some observation A given some evidence B and background context c, and is known as the "posterior probability", $P(A \mid c)$ is the "prior probability" of event A with regard to the background context c, and $P(B \mid A, c)$ is the "likelihood" and returns the probability of the given evidence B on the assumption that A occurred and the context c is true. The denominator, often regarded as a normalizing factor, is the probability of the evidence given the context alone [123]. Each of the probabilities in Eqn. 5.1 is a conditional probability; a probability of one event occurring assuming that another event has already occurred. Often Bayes' theorem is simplified as

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(B)}. \tag{5.2}$$

This simplification comes from assuming that the background context c remains constant throughout the analysis.

For a set of random variables $\mathbf{X}$, the global probability distribution of $\mathbf{X}$ given the context c, $P(\mathbf{X} \mid c)$, gives the probabilities of all possible occurrences of all the variables $X_i \in \mathbf{X}$. This joint probability function is calculated using the chain rule, or probability product rule:

$$P(\mathbf{X} \mid c) = \prod_{i=1}^{N} P(X_i \mid X_1, \ldots, X_{i-1}, c). \tag{5.3}$$

Eqn. 5.3 is easily decomposed for small sets of random variables. Imagine a set $\mathbf{X}$ with variables A, B, and C. The global probability distribution can be decomposed to $P(\mathbf{X}) = P(A \mid B, C)P(B \mid C)P(C)$. Each $P(X_i \mid X_1, \ldots, X_{i-1}, c)$ in Eqn. 5.3 can be considered as the posterior probability in Eqn. 5.1 and can be calculated using Bayes' theorem. Bayesian networks simplify the global probability distribution of a set of random variables by identifying the conditional dependencies and independencies between variables. This identification of conditional dependencies and independencies allows the global probability distribution to be decomposed to a set of local probability distributions.

5.1.3 Bayesian Networks

Graph theory is the other main theoretical foundation of Bayesian networks. A graph G contains a set of nodes $\mathbf{V}$ and a set of arcs $\mathbf{A}$ which are identified by the two nodes the arc connects, e.g. $a_{ij} = (v_i, v_j)$, where $a_{ij} \in \mathbf{A}$ and $v_i, v_j \in \mathbf{V}$. For a given $\mathbf{V}$, G is uniquely defined by $\mathbf{A}$, with the assumption that there is no more than one arc between a pair of nodes in $\mathbf{V}$. For a Bayesian network, the arcs contained in $\mathbf{A}$ must be directed; each arc must point from one node to another. Bayesian networks are also acyclical; if one starts at any node v_i in the graph and moves along the directed arcs in the graph, it is impossible to return to node v_i. Because of these criteria, the structure of a Bayesian network is a directed acyclic graph, or DAG. A Bayesian network, B, is a combination of a DAG G and the global probability $P(\mathbf{X} \mid c)$ of a set of random variables. The set of nodes $\mathbf{V}$ in G must

have a one-to-one correspondence with the variables in $\mathbf{X}$. One of the benefits of Bayesian networks is they allow the decomposition of the global probability distribution to the set of local probability distributions Θ. Θ is referred to as the parameters of a Bayesian network B. These local probability distributions are the conditional probabilities of each random variable with respect to the other random variables in $\mathbf{X}$, and so for every X_i there is a local probability distribution such that $P(X_i \mid \mathbf{X}, c) \in \Theta$. In its simplest form, the Bayesian network $B = (G, \Theta)$.

The relationship between G and Θ illustrates another benefit of Bayesian networks. For each local probability distribution in Θ, the probability of variable X_i is determined based on outcomes in each of the remaining variables in $\mathbf{X}$. For sets of many random variables, this quickly becomes extremely difficult to calculate. The Markov property of Bayesian networks, which is a direct application of the probability chain rule, simplifies the conditional probabilities in Θ, so the conditional probability of variable X_i is only dependant on the set of variables that make up the "parents" of X_i [124, 125]. The parent nodes of a given variable X_i have an arc pointing to X_i, which is considered the "child" node. This simplifies Eqn. 5.3 to

$$P(\mathbf{X} \mid c) = \prod_{i=1}^{N} P(X_i \mid \Pi_{X_i}, c). \tag{5.4}$$

where Π_{X_i} is the set of parents of X_i, and each $P(X_i \mid \Pi_{X_i}, c)$ is a local probability distribution in Θ. A parent child relationship, or an arc, in the DAG G represents a conditional dependence, or a direct probabilistic relationship, in Θ. The direction of an arc indicates the direction of the conditional dependence. Given the correct assumptions, the direction of an arc can indicate causality. However, in this study and the work in the following chapter, no

causality assumptions are made, and an arc's direction is determined by the network score in learning the network structure (see Sec. 5.2.2).

If a parent-child relationship does not exist between two nodes in G, then the corresponding variables X_i, X_j have a nuanced relationship that is determined by the structure of the network. Fig. 5.1 shows an example of a simple Bayesian network with six random variables A, B, C, D, E, and F. Fig. 5.1 includes three key node structures. The first is a

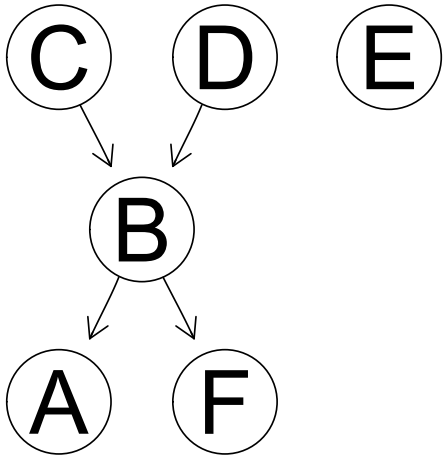


Figure 5.1: Sample network with variables A, B, C, D, E, & F.

sequential or serial structure: $C \rightarrow B \rightarrow F$. In this structure, F is considered conditionally independent from C given B. However F and C are not considered independent; if no information was known about B, knowing C would influence the probability of F, and vice versa. The joint probability of the three variables is $P(C, B, F) = P(F \mid B)P(B \mid C)P(C)$ [126]. The second structure is called a converging structure: $C \rightarrow B \leftarrow D$. In a converging structure, the parents are considered independent; knowing information about C does not affect the probability of D. However, the parents are not conditionally independent. Knowing information about B and D would affect the probabilistic outcome of C, and vice versa. The joint probability of C, D, B is $P(B, C, D) = P(B \mid C, D)P(C)P(D)$ [126]. The third

key structure is a diverging structure: $A \leftarrow B \rightarrow F$. In this structure, A and F are not independent; given some information about A, information can be inferred about B and then a probabilistic outcome of F can be determined. A and F are conditionally independent; knowing B affects the probabilistic outcome of A and F, but the probabilistic outcome of A is not affected by the probabilistic outcome of F, and vice versa. The joint probability distribution of the variables A, B, F is $P(A, B, F) = P(A \mid B)P(F \mid B)P(B)$ [126]. The variable E is not connected to any other variable and is considered independent from all variables, as such it is considered to be excluded from the network.

For a set of random variables and its associated Bayesian network B, Eqn. 5.3 and Eqn. 5.4 are exactly equivalent if and only if the Bayesian network has the correct set of parent-child relationships between the nodes in B; the arcs in B must correctly described the independence and dependence relationships between the variables in $\mathbf{X}$. The Bayesian network is considered to be the "true" network if this is the case. These independence and dependence relationships are not always known, and so the structure of the Bayesian network must be determined either through a learning algorithm or with expert knowledge. This action of finding the Bayesian network structure is referred to as "structure learning" and is considered to be the most important step in probabilistic modeling with Bayesian networks by some [127]. Structure learning is discussed in greater detail in Sec. 5.2.2.

5.1.4 Prior Studies of Bayesian Networks in Retention

While Bayesian networks have not been used in PER, Bayesian networks have been applied in many educational research fields. Bayesian networks have been used in the development of intelligent tutoring systems (ITS) [128, 129]. An ITS is software with which a

student interacts; as the student completes modules and assignments, the ITS learns the student's knowledge deficiencies and then assigns additional modules in areas the student needs to improve. Using Bayesian networks as part of an ITS is an extension of using Bayesian networks to assess student learning and performance [130, 131], another common research strand in educational research fields. Bayesian networks have been applied in student assessment research including using Bayesian networks to model students' test responses and identify common mistakes [132], whether they are using proper problem solving techniques and physical principles correctly [133], and to give personalized feedback to students on engineering design tasks [134].

There has been a substantial amount of work applying Bayesian networks to the problem of college student retention, with the majority of these studies focused on identifying the variables that have the greatest effect on student retention [100, 55, 135, 102, 136], and predicting student retention [135, 102, 136–139]. Different studies investigated the effects of different types of variables, such as pre-college academic factors, college academic factors, and demographic and socio-economic factors. McGovern *et al.* [100] investigated the factors related to retention among minority engineering students. They found that HSGPA was important in predicting retention, as well as student ethnicity. The amount of engineering related work experience a student had also positively affected retention. Nandeshwar *et al.* [55] examined important factors affecting retention of students at a mid-sized U.S. university; good high school performance metrics, such as ACT scores and HSGPA, positively affected retention. Other factors that positively impacted student retention were if the student lived on campus, and if the student or student's family had a higher income. In a studying analyzing the retention of computer science students, Lacave *et al.* [102] found that how many

courses a student had passed positively impacted their retention. Arcuria applied Bayesian networks to community college retention, analyzing the factors that affected retention for a student's first six terms. They found that students who received more need-based financial assistance, attempted fewer credits in the prior term, and enrolled in more daytime courses were more likely to be retained term-to-term. The studies that predicted student retention noted that models that included college-level academic factors were better predictors of student retention [135, 136, 102, 139]. These studies predicited general college retention with two exceptions; one study predicted computer science student retention [102] and another predicted engineering student retention [100].

5.2 Methods

This section discusses the methods and processes used to construct and query Bayesian networks. All of the networks were built using the bnlearn package [140] as implemented in the R software system.

5.2.1 Sample

The sample used in this study is the same sample that was used in the previous chapter as described in Sec. 4.2.1. Because the analysis presented in this chapter examines probabilistic relationships between academic factors and specific milestones in the completion of the physics degree, the analysis for each milestone uses a filtered subset of the original data, with each filter including only students who have been enrolled for an amount of time that reasonably allows them to have met the milestone. The applications of these filters to the original data and the subsequent descriptive statistics are shown in Table 5.1.

#	Filter	N	HSGPA	Math Ready %	Enroll P112 %	Enroll P314 %	Grad Physics %
1.1	None	586		63			
1.2	2year, HS, P1	274	3.6	64	58		
1.3	3year, HS, P1	267	3.6	66		42	
1.4	Grad, HS, P1	236	3.6	69			31
1.5	3year, HS, P1, P112	148	3.8	82		70	
1.6	Grad, HS, P314	171	3.7	84			67

Table 5.1: Descriptive statistics for data from Institution 1 after applying filters. Filters are abbreviated: HS (high school) for students with HSGPA records, P1 (Physics first) for students whose first declared major was physics, P112 (PHYS 112) for students who enrolled in PHYS 112, and P314 (PHYS 314) for students who enrolled in PHYS 314. Different windows were used to ensure that the samples only included students who could have met a particular milestone: 2year (Two year) removes the last two years of records, 3year (Three year) removes the last three years of records, Grad (Graduation) removes the last six years of records. HSGPA is the average High school GPA of the sample, and Math Ready % reports the percentage of students ready to take Calc 1 or higher upon enrollment. The last three columns report the percentage of students who met one of the three milestones.

The pre-college academic variables in this study are similar to those used in the prior chapter, except they have been adjusted to be categorical variables to permit the use of discrete Bayesian networks. ACTSATM and ACTSATV are three-level ordinal variables with categories “High”, “Mid”, and “Low”. The breaks for these categories were derived from the tertile breaks of the ACTSATM and ACTSATV continuous percentile scores; for ACTSATM the breaks are 89 and 74, for ACTSATV the breaks are 89.5 and 73. HSGPA is an ordinal variable with classes 2, 3, 4, and 5. These variable levels were discretized from the continuous HSGPA (cHSGPA) scores, with the following discretization bins: a cHSGPA score greater than 4 became HSGPA 5, cHSGPA between 3.5 and 4 became HSGPA 4, cHSGPA between 3 and 3.5 became HSGPA 3, and any cHSGPA score less than 3 became HSGPA 2. APPhys and APMath are dichotomous variables that indicate whether the student has any AP Math or Physics credits, with a 1 indicating that the student does have credit and a 0 indiciating no credit. MathEntry is a 3-level variable with levels “<Calc1” indicating

the student was not calculus ready, "Calc1" indicating the student's first math course was Calculus 1, and ">Calc1" indicating the student's first college math course was Calculus 2 or a more advanced math course. The outcome variables, or the variables that indicate whether a student reached a milestone in the program, are TakeP112, TakeP314, and EndPhys. Each of these are dichotomous variables, with a 1 indicating that the student reached the milestone of enrolling in PHYS.112, enrolling in PHYS.314, and graduating from the physics program, respectively.

A second analysis was performed for the probabilities of enrolling in PHYS 314 and graduating in physics. In this analysis, the course grades of PHYS 112 were included as a variable (P112) in the model for enrolling in PHYS 314, and the grades for PHYS 112 and PHYS 314 (P314) were included as variables in the model for graduating in physics. In the case of enrolling in PHYS 314, the data was filtered to include students who had enrolled in PHYS 112 and had started their college career as a physics major; in the case of graduating physics, the data was filtered to include students who had enrolled in PHYS 314. These variables had categories corresponding to course grades: A, B, C, D, and F.

5.2.2 Building Bayesian Networks

One function of a Bayesian network is to identify the joint probability distributions and conditional probability distributions of a set of random variables [124]. These conditional probabilities can give insight to how states of an independent variable affect outcomes in the dependent variable. A Bayesian network was constructed for the three milestones to identify the conditional probabilities between reaching the milestone and pre-college academic factors. Two more networks were built to include P112 and P314 as variables for enrolling in PHYS

314 and graduating physics.

The method of determining the conditional probabilities is referred to as a conditional probability query (CPQ). A CPQ investigates the posterior distributions of a learned Bayesian network B for a specific outcome of a variable in $\mathbf{X}$ based on a piece of evidence $\mathbf{E}$ [125]. There are two types of evidence that can be provided to a CPQ: hard evidence, which is a new observation of one or more random variables in $\mathbf{X}$, or soft evidence where the distribution of one or more variables is changed. This study uses CPQs with hard evidence, where the outcome X_i is whether a student reaches the milestone, and the hard evidence X_j is the value of a pre-college factor or a prior course grade. A CPQ returns the probability for all possible values of the target variable, as shown in Eqn. 5.5.

$$P(X_i \mid \mathbf{E}, B) = P(X_{i_1}, \ldots, X_{i_k} \mid X_{j_k}, G, \Theta) \tag{5.5}$$

where X_{i_k} represents the k^{th} level of the target variable X_i, and X_{j_k} represents the k^{th} level of variable X_j.

The structure of a Bayesian network can be learned with a structure learning algorithm or it can be manually defined. Structure learning algorithms fall into three categories: constraint-based algorithms, score-based algorithms, and hybrid algorithms. Constraint-based algorithms use various statistical tests to learn the conditional independence relationships (or "constraints") found in the data [127]. Score-based algorithms build many different DAG structures and measure the likelihood of the data given the proposed DAGs, selecting the DAG that maximizes the likelihood of the data [127] (i.e. the DAG with the structure that best fits the data [141]). Hybrid algorithms are a mixture of the score-based and

constraint-based algorithms, in that they conduct conditional independence tests to learn at least part of the conditional independence relationships in the data, and then try to maximize the goodness of fit based on the found constraints [125]. Manually defining a Bayesian network is often referred to as "expert elicitation", and consists of an expert determining the inclusion and direction of arcs between nodes [142]. In the bnlearn framework, each structure learning algorithm can be constrained by user input, allowing a network to be created that is both learned from the data and determined by an expert. The purpose of structure learning is to determine the true Bayesian network structure associated with the data, and in turn correctly defining the probabilistic effects between variables. It is often the case that the true structure can only be approximated for a specific dataset; in these cases, the joint probability distribution calculated in Eqn. 5.4 is an estimate or approximation of the true joint probability distribution calculated in Eqn. 5.3.

The networks in this study were built using expert elicitation and the hill-climbing algorithm in tandem. The hill-climbing algorithm is a score-based algorithm that maximizes the likelihood of the data given the proposed structure [127, 143]; it begins with an initial DAG that is typically empty (no arcs), and then adds, deletes, and reverses arcs in the DAG, retaining arcs that improve the likelihood. Once no arcs can be added, deleted, or reversed to improve the likelihood, the algorithm selects the remaining DAG as the network structure. The algorithm measures the effect of structure changes to the likelihood by calculating a network score. The hill-climbing algorithm can be set to maximize any type of network score; in this study the network score is based on the Bayesian Information Criterion (BIC). For more information on BIC, see Chapter 8. Expert elicitation was used to preserve relationships between pre-college factors and the milestones of the physics program that were found in

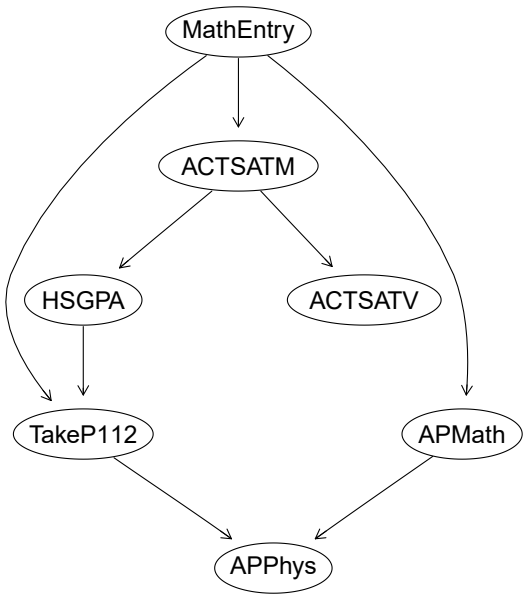
Chapter 4. To do this, a whitelist was constructed for each network. A whitelist is a list of arcs that is supplied to the hill-climbing algorithm that must be present in the final network structure. For the networks using only pre-college factors, the whitelist constrains the network to include an arc from MathReady to the milestone variable, and an arc from HSGPA to the milestone variable. These variables were found to be the most influential variables in determining retention in the prior study. For the networks including prior physics course grades, the whitelist consisted of an arc pointing from the prior courses to the target outcome, and arcs from MathReady and HSGPA pointing towards the prior course variable.

5.3 Results

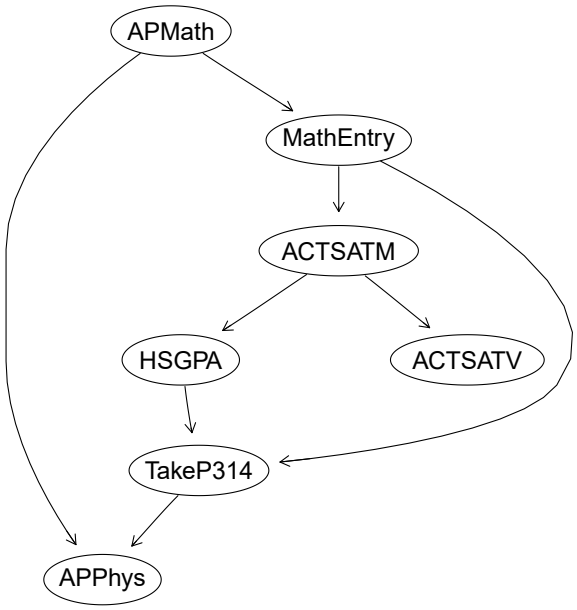
A Bayesian network was constructed for Samples 1.2-1.6 in Table 5.1. Each network was built using a combination of the hill-climbing algorithm and expert input, and then was queried to determine the conditional probabilities between the independent variables and the dependent milestone variables. The networks are discussed in the following section, and the results of the CPQ are in the section after that.

5.3.1 Bayesian Networks

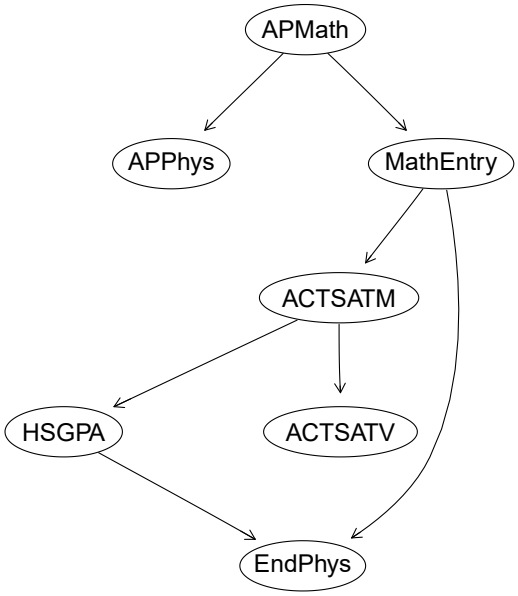
Each Bayesian network was built using the hill-climbing algorithm and a whitelist to constrain the structure. The networks using only the pre-college academic factors are shown in Fig. 5.2. These networks share many of the same arcs; this is unsurprising as the samples used to construct each network only differ by the window of time used to select them. Perhaps the most notable difference is the exclusion of the arc from the variable EndPhys



(a) 1.2, Bayesian network for enrolling in PHYS 112.



(b) 1.3, Bayesian network for enrolling in PHYS 314.



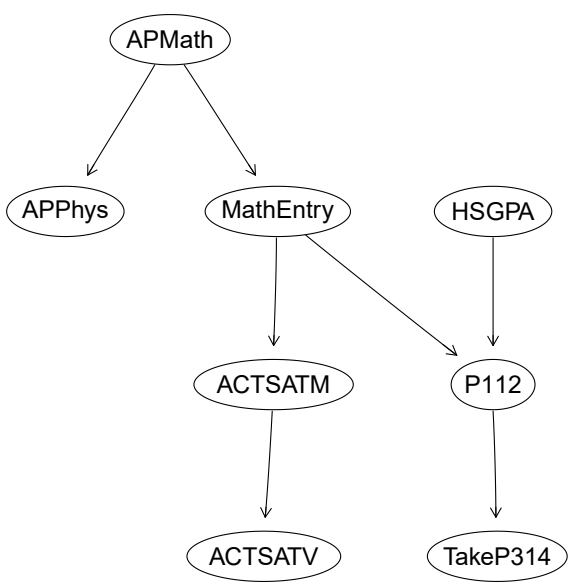
(c) 1.4, Bayesian network for graduating physics.

Figure 5.2: Bayesian networks for the milestones of enrolling in PHYS 112, enrolling in PHYS 314, and graduating physics. The caption of each network indicates the sample used to build the network. Only pre-college academic factors are included as variables.

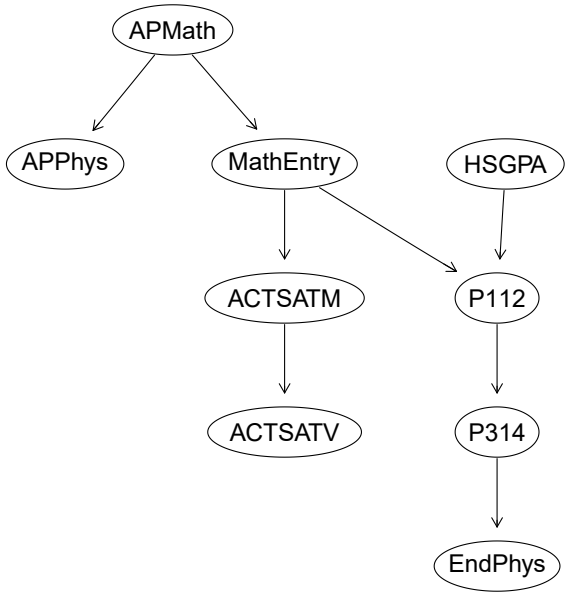
(Fig. 5.2c) to the variable APPhys. This arc is present in Figs. 5.2a and 5.2b, pointing from the milestone variables TakeP112 and TakeP314 to APPhys. This is indicative of a direct

probabilistic relationship between whether a student has AP Physics credit and whether they reach the milestones of enrolling in PHYS 112 and PHYS 314. This relationship either does not exist between whether a student graduates in physics and has AP Physics credit or the probabilistic relationship is conditioned on other variables in the network. The direction of this arc in Figs. 5.2a and 5.2b may seem counter-intuitive; AP Physics credit would be earned before a student tries to enroll in either PHYS 112 or PHYS 314, and so the direction of causation should be the reverse of the direction in the graphs. Similarly, the arc between MathEntry and ACTSATM is in the opposite direction of what intuition would dictate; ACT and SAT math scores are used to determine a student's first math course. The reversal of some arcs is due to the nature of the nature of the hill-climbing algorithm. The parameters Θ of a Bayesian network should match the conditional dependencies and independencies found in the global distribution $P(\mathbf{X} \mid c)$. When learning the structure of a Bayesian network from a set of random variables, the hill-climbing algorithm builds a model that has the best fit to the conditional dependencies in $P(\mathbf{X} \mid c)$. It does this by maximizing network score, and so improvement of network score is the determining factor in the direction of any arc in the network, and a sense of causality or chronology is ignored by the algorithm.

The networks for reaching the milestones of enrolling in PHYS 314 and graduating physics that include college physics course grades are shown in Fig. 5.3. These networks are nearly identical, with the network in Fig. 5.3b including the milestone variable EndPhys, which is a child node of P314, which replaces TakeP314 in Fig. 5.3a (P314 represents the earned grade in PHYS 314, TakeP314 indicates whether the student enrolled in PHYS 314). The whitelists used in constructing these networks were more constraining than those used in the networks in Fig. 5.2; they consisted of constraining MathEntry and HSGPA to be



(a) 1.5, Bayesian network for enrolling in PHYS 314.



(b) 1.6, Bayesian network for graduating physics.

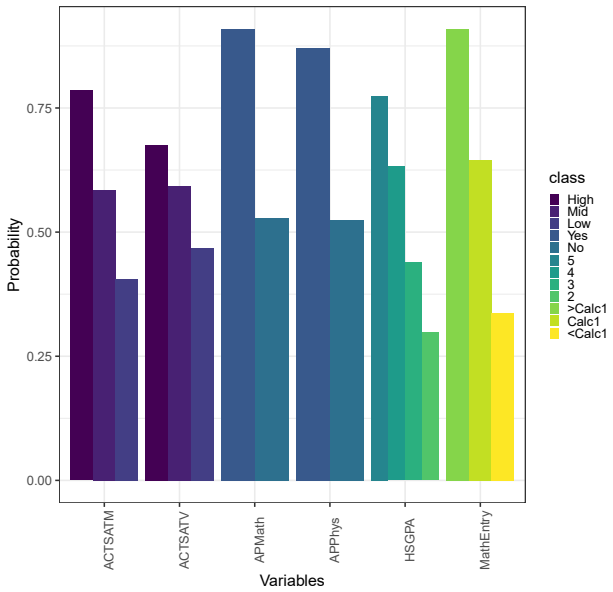
Figure 5.3: Bayesian networks for the milestones of enrolling in PHYS 314 and graduating in physics. The caption of each network indicates the sample used to build the network. College physics course grades were included as variables in these networks, as shown by the P112 nodes and P314 nodes.

parent nodes of P112, constraining P112 to be a parent node of TakeP314 in Fig. 5.3a and P314 in Fig. 5.3b, and constraining P314 to be a parent node of EndPhys in Fig. 5.3b.

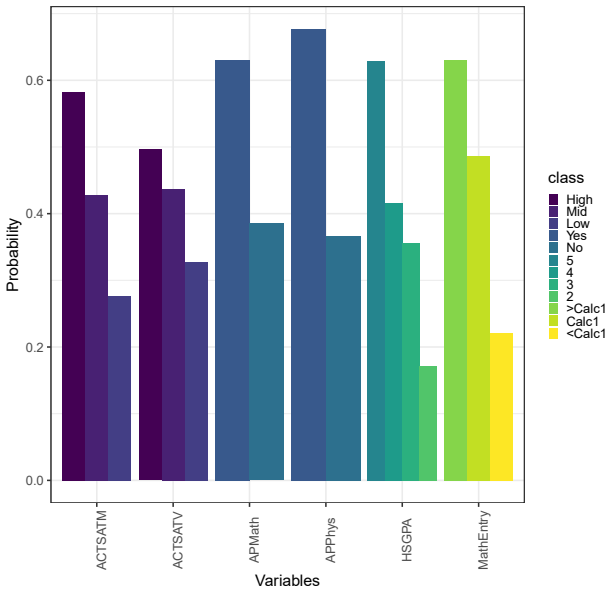
5.3.2 Conditional Probability Queries

The probabilities of successfully reaching a milestone were determined using CPQs. The outcome variable of the CPQs was the specified milestone variable, and for each milestone variable each level of each variable was used as evidence. The results of these CPQs are shown in Fig. 5.4, where the probability of reaching the milestone for each variable is shown.

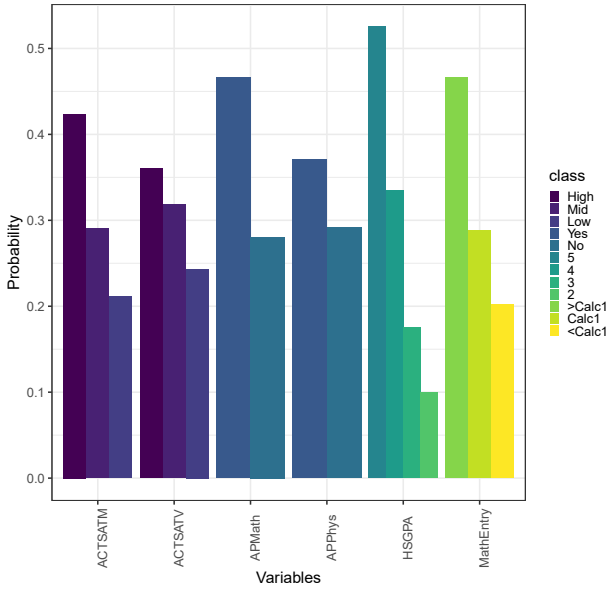
For each milestone, as the evidence progresses down through a variable's levels, the probability of reaching the milestone decreases as expected. Students with a "High" ACT math score would be more likely to progress to enrolling in PHYS 314 than a student with



(a) Conditional probabilities for TakeP112.



(b) Conditional probabilities for TakeP314.



(c) Conditional probabilities for EndPhys.

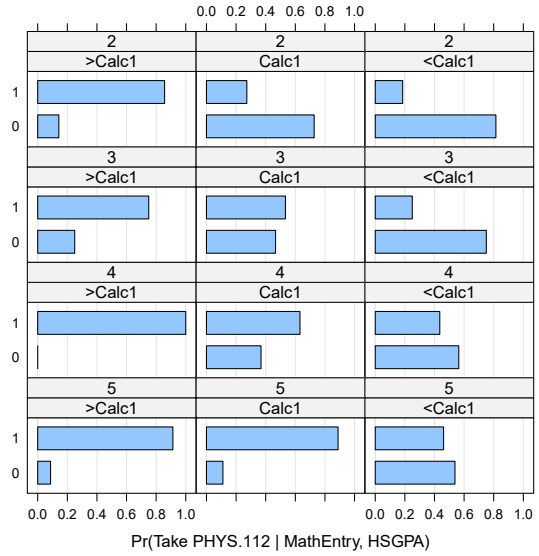
Figure 5.4: CPQ results for each milestone variable and each of its pre-college independent variables. The probabilities shown are the probabilities of a "1" outcome (i.e. reaching the milestone). Probabilities queried from the networks in Fig. 5.2.

a "Low" ACT math score. For the milestone TakeP112, a MathEntry value of >Calc1 and having AP Math credit return the highest probability for enrolling in PHYS 112 (Fig. 5.4a).

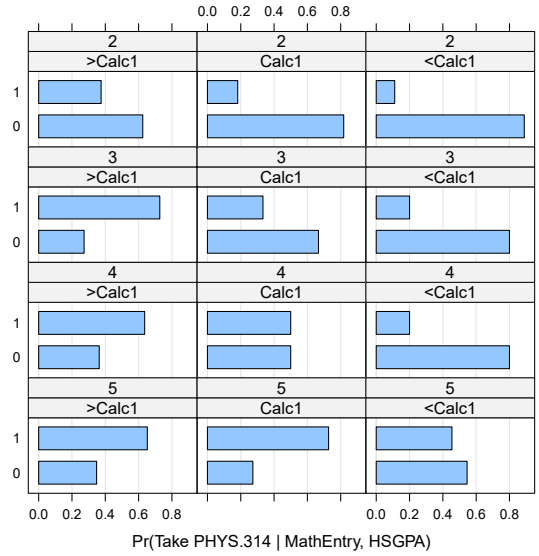
For TakeP314, having credit in an AP Physics course returns the highest probability for reaching PHYS 314, followed closely by a HSGPA of 5 and a MathEntry of >Calc1 (Fig. 5.4b). Only 31% of students who begin in physics as freshmen graduate physics (Table 5.1). As such, for nearly every possible value of the pre-college factors a student is more likely to leave physics than complete the program (Fig. 5.4c). Only HSGPA of 5 returns a greater probability of graduating physics than not for the milestone variable EndPhys.

If the outcome variable has more than one parent node, a simple CPQ using only one parent variable as evidence can fail to capture the intricacies of the interaction of two or more parent nodes and the child node. To capture these relationships, all possible combinations of the parent variables were used as evidence in a CPQ, and a conditional probability table (CPT) was formed to show these relationships. The milestone variables in the networks in Fig. 5.2 all have the same two parent nodes of HSGPA and MathEntry. CPTs for the networks for each milestone variable are shown in Fig. 5.5.

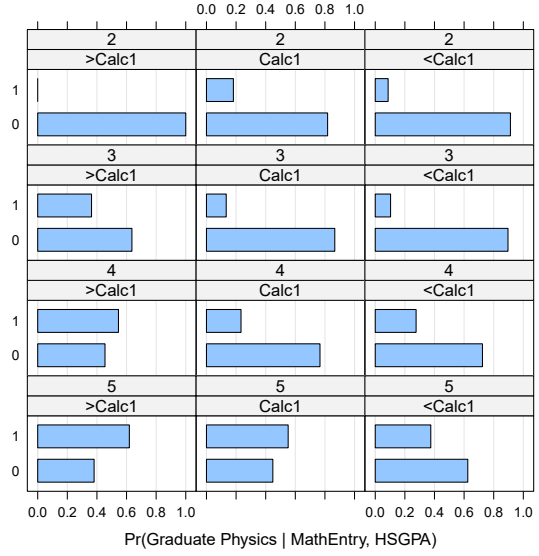
A CPT shows how the probability of the outcome changes with changes of level in one variable while the other variable is held constant. For example, the right most column in Fig. 5.5a shows the probabilities of the various levels of HSGPA when MathEntry is held constant at <Calc1. Moving down the column changes the HSGPA from 2 to 5. In this case, the probability of enrolling in PHYS 112 increases from 20% with a HSGPA of 2 to 45% with a HSGPA of 5. Similarly, the second row of Fig. 5.5a shows the changing probabilities for levels of MathEntry when HSGPA is held at 3. As MathEntry decreases from >Calc1 to <Calc1, the probabilty of enrolling in PHYS 112 decreases from 75% to 25% for a HSGPA of 3. Generally, each CPT in Fig. 5.5 shows an increasing probability of reaching the milestone for increasing HSGPA and MathEntry (i.e. moving down the table for increasing HSGPA



(a) Conditional probability table for TakeP112 and its parent nodes HSGPA and MathEntry



(b) Conditional probability table for TakeP314 and its parent nodes HSGPA and MathEntry.



(c) Conditional probability table for EndPhys and its parent nodes HSGPA and MathEntry

Figure 5.5: Conditional probability tables for each milestone variable and their parent variables. Probabilities of reaching or not reaching the milestone are given for each possible combination of the parent variables. Probabilities queried from the networks in Fig. 5.2.

and from right to left for increasing MathEntry). Each possible combination of HSGPA and MathEntry is filled with some fraction of students from the sample that match the criteria.

The number of observations per possible combination is shown in Fig. 5.6.

TakeP112		MathEntry		
		>Calc1	Calc1	<Calc1
HSGPA	2	7	11	27
	3	8	15	36
	4	21	38	39
	5	23	36	13

(a) Observations per combination of MathEntry and HSGPA for PHYS 112

TakeP314		MathEntry		
		>Calc1	Calc1	<Calc1
HSGPA	2	8	11	27
	3	11	15	35
	4	22	36	35
	5	23	33	11

(b) Observations per combination of MathEntry and HSGPA for PHYS 314

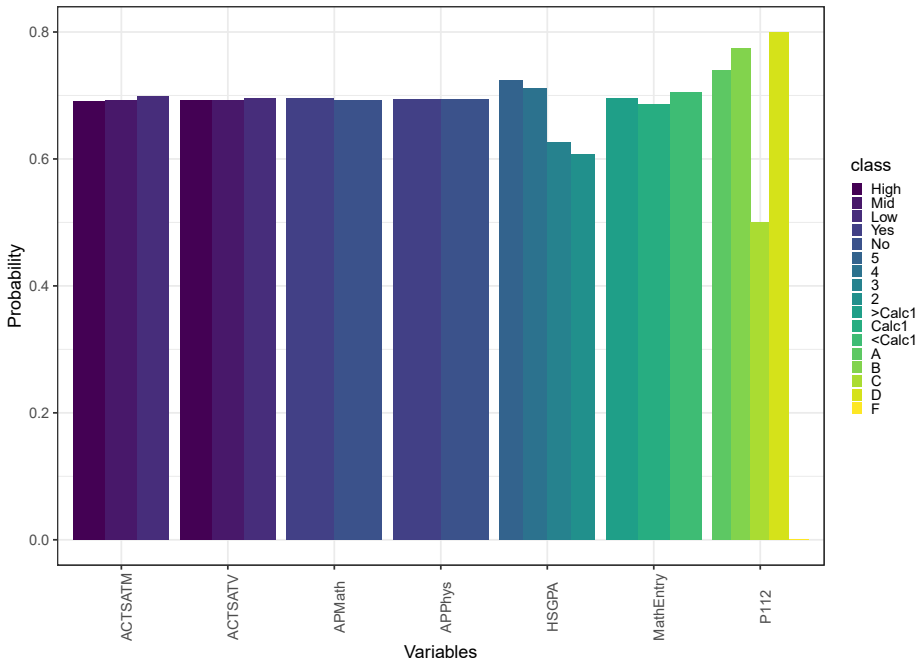
EndPhys		MathEntry		
		>Calc1	Calc1	<Calc1
HSGPA	2	8	11	23
	3	11	15	29
	4	22	30	29
	5	21	29	8

(c) Observations per combination of MathEntry and HSGPA for EndPhys.

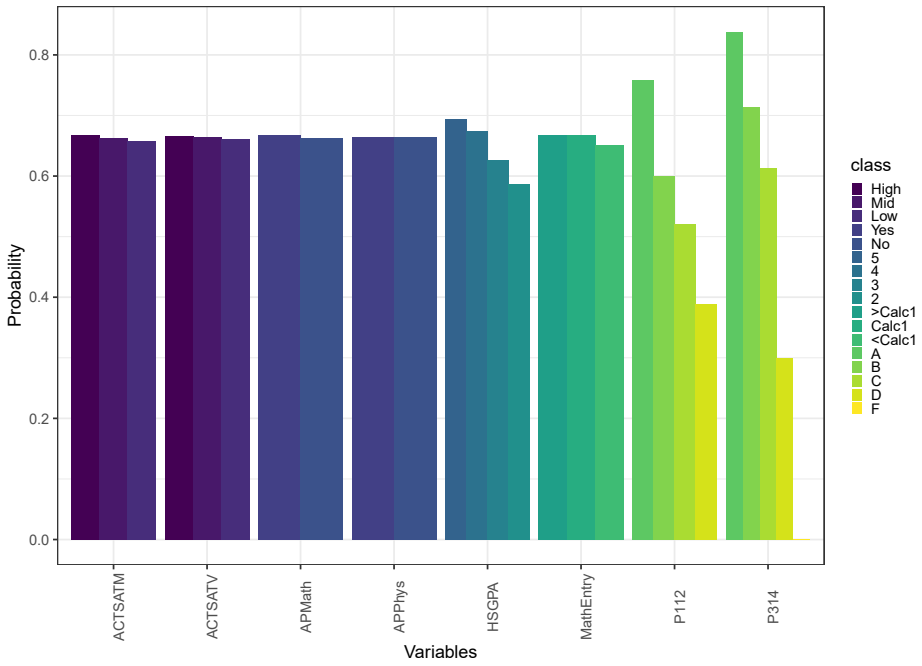
Figure 5.6: Observations per combination of MathEntry and HSGPA for the CPTs in Fig. 5.5.

The networks shown in Fig. 5.3 that include college physics course grades were also queried to determine conditional probabilities. These results are shown in Fig. 5.7. Because the outcome variables of TakeP314 and EndPhys have only one parent node each (P112 and P314 respectively), a CPT is not reported.

The conditional probabilities for the pre-college factors in Fig. 5.7 are markedly different than those in Fig. 5.4. The different levels of pre-college factors in Fig. 5.7 show very little variation in probability of reaching the milestone. This indicates that for students who have reached an early milestone in the physics program (e.g. enrolling in PHYS 112), their probability of reaching a later milestone such as graduating physics or enrolling in PHYS 314 is mainly affected by their college performance, and any effect that a pre-college factor may have on reaching a particular milestone is explained by their college performance. Also of note is the much greater percentage of students that reach the milestones of enrolling in



(a) Conditional probabilities for TakeP314.



(b) Conditional probabilities for EndPhys.

Figure 5.7: CPQ results for milestone variables TakeP314 and EndPhys, including results for pre-college variables and college physics course grades. The probabilities shown are the probabilities of a "1" outcome (i.e. reaching the milestone). Probabilities queried from the networks in Fig. 5.3.

PHYS 314 and graduating in physics in samples 1.5 and 1.6, which were used to build the networks that produced these probabilities. Sample 1.5 is filtered to only include students who enrolled in PHYS 112. This filter increased the probability of enrolling in PHYS 314 from 42% in Sample 1.3 to 70%, a 170% increase. Similarly, Sample 1.6 is filtered to only include students who enrolled in PHYS 314. This filter increased the percentage of students who graduate in physics from 31% in Sample 1.4 to 67%, a 220% increase.

5.4 Discussion

This study sought to answer a single research question. The result and its implications are discussed below.

RQ1: What is the probabilistic relationship between various points of progression in a physics curriculum and pre-college academic factors? How do these relationships change when select physics course grades are added to the model? Fig. 5.4 shows the specific probabilistic relationships between pre-college factors reaching the milestones of PHYS 112, PHYS 314, and graduating physics. For every pre-college factor, the probability of reaching any milestone decreases as the level of the pre-college factor decreases. However, for every pre-college factor, regardless of the level of the factor, students are more likely to not reach the milestone of graduating in physics, with the exception of students who have a HSGPA of 5. When the college physics courses were included in the models, the relationship between pre-college factors and the milestones of enrolling in PHYS 314 and graduating in physics changed dramatically (see Fig. 5.7). These models were built from samples that only included students who had already met the milestone of enrolling in PHYS 112 (Fig. 5.7a) and the

milestone of enrolling in PHYS 314 (Fig. 5.7b). Essentially, the probability of reaching a milestone is the same for all levels of the pre-college factors, with the exception of HSGPA, where there is a slight decrease in probability of reaching the milestones when the HSGPA decreases from 5 to 2. For these models, a strong probabilistic relationship exists between the milestones and the college physics course grades; as the grade in PHYS 112 and PHYS 314 decreases, so does the probability of graduating physics. For the probability of enrolling in PHYS 314 (Fig. 5.7a), there is a strong drop when moving from a B in PHYS 112 to a C in PHYS 112, then the probability suddenly rises with a D in PHYS 112. This inconsistency is likely due to a lack of statistical power. Only five students in Sample 1.5 received a D in PHYS 112; against all odds, four of these students enrolled in PHYS 314.

This strong dependence on prior course grades such as PHYS 112 may indicate that the physics department at Institution 1 should focus on helping incoming students in their first two to three semesters at the university, especially those who are likely to not enroll in PHYS 112 or likely to struggle in PHYS 112. This is because once the first milestone of enrolling in PHYS 112 was met, their college course performance is more probabilistically indicative of their future outcomes. Fig. 5.4a indicates that pre-college factors are strongly related with reaching the first milestone of enrolling in PHYS 112, and Figs. 5.7a & 5.7b imply that it is the grades in prior college physics courses that are strongly related with reaching the milestones of enrolling in PHYS 314 and graduating physics. Understanding the relationship of these pre-college factors and the probability of reaching an early milestone in a physics curriculum (like the milestone of PHYS 112 at Institution 1) can help physics departments to implement the necessary interventions in their program to increase the fraction of students who reach these early milestones and in turn, increase the number

of students who successfully reach subsequent milestones.

The strong dependence of reaching the first milestone of PHYS 112 and pre-college academic factors is somewhat troublesome. The three variable states that had the greatest probability of enrolling in PHYS 112 were having credit in any AP Math course, having credit in any AP Physics course, and having a MathEntry point of >Calc1. Having a MathEntry point of >Calc1 typically indicates that a student enrolled in an AP Calculus course in high school or enrolled in a dual-enrollment course in high school. Often, these college-level high school courses are considered bonus point courses in the school districts in which they are offered, and a student can receive up to five GPA points for a successful completion of the course, as opposed to the typical four GPA points. For a student to have a HSGPA of 5, they need to have attended a school district that offers AP or dual-enrollment courses. For graduating physics, only students who had a HSGPA of 5 were likely to reach that milestone; any student without a HSGPA of 5, regardless of their other pre-college factors, were more likely to leave the physics program. AP courses and dual-enrollment courses are not options for many students at under-resourced school districts. Persons from traditionally marginalized communities in STEM fields disproportionately attend these under-resourced school districts [122]. If this trend found at Institution 1 is universal, then these students are at a strong disadvantage of reaching early milestones in their physics programs. Assisting students who are less likely to reach early milestones becomes an issue of equity, and physics departments have the responsibility to improve their programs to be more equitable and better serve all students.

5.5 Conclusion

The probabilistic relationships between reaching specific milestones in the physics curriculum at Institution 1 and pre-college academic factors were explored by querying a Bayesian network. The milestones investigated were reaching the courses PHYS 112 and PHYS 314, and graduating from the physics program. Reaching PHYS 112 had a strong dependence with college preparation; students with some AP Physics and Math credit were nearly 40% more likely to reach PHYS 112 than those without credit, and students who enrolled in a higher math course than Calculus 1 their first semester were nearly 50% more likely to reach Physics 2. Reaching PHYS 314 had a strong probabilistic relationship between high HSGPA scores and math readiness. Graduating from the physics program had a strong probabilistic relationship with high HSGPA scores. When grades from a prior college physics course were added to the models, these factors had stronger probabilistic relationships with reaching later milestones than the pre-college factors.

Chapter 6

Predicting Physics Course Grades Using Bayesian Networks

6.1 Introduction

The preceding chapter introduced Bayesian networks as a method to calculate the probabilities of students reaching a particular milestone in the physics curriculum at Institution 1. The factors that had the greatest influence on reaching a milestone were students' grades in a prior physics course. This relationship between grades and reaching a milestone in the program is not surprising. Receiving a passing grade in a required course allows a student to continue to progress in the program, while a failing grade requires the student to re-take the course before they can progress. While Chapter 5 only looked at two courses, each required course in a physics curriculum can be considered a milestone, and reaching and passing each course is a necessary step in successfully completing a physics program. This study extends the analysis in the preceding chapter by finding the relationships between the grades in required physics courses and their pre-requisite course grades. Specifically, this chapter uses Bayesian networks to determine the conditional probabilities of a student being successful in a required physics course based on their grades in prior physics courses.

Determining the probability of student outcomes in a particular course has direct student advising applications. Effective student advising is a core responsibility of physics programs. Advising has been shown to increase rates of student persistence to graduation [144]. Quality advising has been cited as the second most important responsibility of academic programs, with quality instruction as the most important responsibility [144, 145]. Quality advising should not only instruct students on which courses they must complete to qualify for graduation, but also when courses are offered, when and in what sequence to take courses, and what course combinations are beneficial or detrimental. Knowing the condi-

tional probability of a successful outcome in each required physics course based on grades in prior courses could be an extremely useful tool for undergraduate physics advisors.

6.1.1 Research Questions

This study applies Bayesian networks to determine probabilistic relationships between outcomes in courses required in the physics curriculum at Institution 1. Bayesian networks are also used to predict student outcomes. These probabilistic dependencies and predictions are used to determine ways they could be applied by a physics department to improve its physics curriculum, with the hypothesis that an improved curriculum causes improved retention.

RQ1: What are the probabilistic dependencies between upper-level physics courses and their prerequisites?

RQ2: How accurate are Bayesian networks in predicting outcomes in upper-level physics courses? Which prior course is the most important predictor of the target course?

This study uses prior required physics and math course grades. The study in the preceding chapter showed that these grades had strong probabilistic relationships with reaching milestones in the curriculum, such as enrolling in a modern physics course and graduating the program. This study looks at seven of the required physics courses at Institution 1 and treats them similarly to the milestones discussed in the previous chapter. These courses are referred to often as "target courses" in this chapter.

6.1.2 Bayesian Networks and Grade Prediction

The study of student performance using educational data mining (EDM) and machine learning (ML) methods is an increasingly popular research strand in many educational research subfields, such as PER [104, 105, 55]. As discussed in the prior chapter, the use of Bayesian networks to study retention and student academic performance is not uncommon. Several studies have used Bayesian networks to predict student course grades [137, 101, 146, 147]. One of these [101] used an expert elicited Bayesian network to predict student grades in three core courses of the engineering program at a university in the midwestern U.S. The courses they predicted (Physics 2, Calculus 2, and Intro to Computer Programming) were considered "gateway courses" in the engineering program; each was a required course in the engineering curriculum, and many students leave the engineering program after performing poorly in any of these courses. They compared the expert elicited Bayesian network with other common prediction methods such as random forests, decision trees, K-nearest neighbors, and others. They found that their expert elicited network outperformed all other predictive models, predicting Physics 2 outcomes with an accuracy of 70%, Calculus 2 outcomes with an accuracy of 73%, and Intro to Computer Programming with an accuracy of 36%. They used prior course grades and pre-college academic and demographic factors as independent variables in their predictions. Another study [137] used Bayesian networks to predict students' 3rd year overall academic performance at a university in London, United Kingdom. They used a mixture of data sources as independent variables in their prediction including pre-college demographic and academic information, final grades for all first and second year courses, and online and in-person engagement information. Their

data were highly imbalanced; there were far fewer students who were at a high risk of poor performance. They showed that using bootstrap aggregation (bagging) improved prediction accuracy of at-risk students by 15-20%.

Two studies used Bayesian networks specifically to create an advising tool for computer science [147, 146]. The first of these created a Bayesian network based on the pre-requisite structure of courses in the computer science program at a university in the eastern U.S., with some adjustment from experts (i.e. faculty members). This network was not built with student data or applied to the prediction of real students, rather the researchers created several different simulated students with different characteristics describing their mathematical and programming abilities, and used the network to predict the simulated students' outcomes. They compared the network's predicted outcomes with the outcomes that various undergraduate advisors predicted based on the simulated student information. They found that the network predictions agreed with the advisors predictions in most cases. The other study [146] was performed at a liberal arts college in the central U.S., and predicted student grades in all of the required courses of the computer science curriculum. The Bayesian network structure was built using the pre-requisite structure of the curriculum; arcs in the network corresponded to pre-requisite relationships between courses (e.g. there would be an arc pointing from Calculus 1 to Calculus 2). The network was used to predict each required course in the curriculum, with varying levels of success. The prediction accuracy of each course was better than the baseline accuracy of guessing the majority class for every prediction; however, in some cases the prediction accuracy was still less than 40%, and the highest prediction accuracy was 87% in a senior level computer science course. The work presented in this chapter is similar to these two studies, though it is the first application of Bayesian

networks to predict physics course grades.

6.2 Methods

6.2.1 Sample

The sample used in this study is the same sample that was used in the analysis in Chapter 4 as described in Sec. 4.2.1. The variables used in this study consist of the grades in some required physics and math courses for physics majors at Institution 1, as well as first or second-semester college GPA. All variables were ordinal categorical variables, where the possible categories of a variable are the possible outcomes of the course, or in other words the grade earned in the course. These variables are shown in Table 6.1. Some of the student records in the sample had missing data in some of the introductory physics and math courses (they had no recorded grade for the course.) This happened when some students received college credit for AP courses or dual-enrollment courses. These data were considered to be missing at random (MAR), because the missingness of the data does not affect the value the data would take if it was not missing. It is not missing completely at random (MCAR) because the reason the data were missing can be explained by the data (the students have AP or dual-enrolment credit), and it is not missing not at random (MNAR) because the fact that the data is missing is not explained by the values of the missing data. Because it is MAR, we can use multiple imputation methods to impute these missing values.

Reference No.	Variable Name	Canonical Course Name
1	MATH.155	Calculus 1
2	MATH.156	Calculus 2
3	MATH.251	Calculus 3
4	MATH.261	Differential Equations
5	PHYS.111	Intro Physics 1
6	PHYS.112	Intro Physics 2
7	PHYS.314	Modern Physics
8	PHYS.331	Classical Mechanics
9	PHYS.333	Electricity and Magnetism
10	PHYS.341	Advanced Lab
11	PHYS.451	Quantum Mechanics
12	PHYS.461	Statistical Mechanics
13	CGPA	Second-semester college GPA
14	CGPA1	First-semester college GPA

Table 6.1: List of courses used as variables in the analyses, as well as the college GPA variables.

Multiple Imputation

Multiple imputation follows a straightforward process. First the missing data are imputed multiple times to create M full datasets. Second, the analysis is performed on each dataset, resulting in M results. Lastly, the results are pooled following Rubin's rules [148]. In a recent article, it was shown that one can also average the results of the M imputations, and use the averaged full dataset for the analysis [149]. The imputation method used in this study is the structural expectation-maximization (SEM) algortihm as implemented by the bnlearn package [150]. The algorithm has three steps: the algorithm builds a Bayesian network and fits it to the dataset with missing values, then the missing values are imputed using Bayes' theorem and the parameters learned in the network, and lastly the algorithm maximizes a specified network score as it learns a Bayesian network from the completed dataset with a structure-learning algorithm. SEM was used to perform multiple imputations to handle the missing data and build models that better represented the conditional probabilities between

outcomes in the required courses for physics majors at Institution 1.

6.2.2 Identifying Conditional Probabilities

A Bayesian network was constructed for seven required physics courses at Institution 1. These Bayesian networks were built to identify the conditional probabilities between outcomes in the target course and grades in prior physics and math courses. The outcomes of the target variable were classified as "Succeed" or "Struggle", where a student who "Succeed[s]" is one who earned a grade of A or B, and a student who "Struggle[d]" is one who received a grade of C, D, F, or withdrew from the course (W). The other courses (independent variables) used to build the networks had variable levels that corresponded to the grade received in the course: A, B, C, and DFW, which indicates any failing grade or a course withdrawal. Second-semester college GPA (or first-semester college GPA in the case of PHYS.112) was also included as a variable and had levels corresponding to the letter grade associated with the grade point average (A, B, C, D, F). For each target course, the conditional probabilities of the target course outcome were calculated only for the courses that are typically taken prior to the target course. A separate network was built for each target course; in each case, the data were filtered to include only students who had enrolled in that course.

To determine probabilistic relationships between prior course grades and target course outcomes, conditional probability queries (CPQ's) were conducted, as described in Sec. 5.2.2. In this case, the outcome X_{i_k} is the k^{th} outcome (Struggle, Succeed) of the target course i, and the hard evidence X_{j_k} is a specific grade k (A, B, C, DFW) in a prior course j. The CPQ in this case takes the form

$$P(\mathbf{X} \mid \mathbf{E}, B) = P(X_{i_1}, \cdots, X_{i_k} \mid X_{j_k}, G, \Theta). \tag{6.1}$$

As in Sec. 5.2.2, G encodes the characteristics of the directed acyclic graph (DAG) associated with the Bayesian network B, and Θ represents the local probability distributions of B.

A network and its posterior probabilities were learned from the available data, and those probabilities were then used to calculate values for the missing data using Bayes' theorem. This was done 100 times; 100 networks were learned from the data and each imputed the missing data creating a set of 100 imputations. The mode of the 100 imputations was taken to create a dataset that had multiply imputed data. This multiply imputed dataset was then used to build the networks and find the conditional probabilities of the target courses as discussed in the following section.

Structure Building

The networks constructed to determine conditional probabilities of prior grades were built using the hill-climbing algorithm in conjunction with expert elicitation. One shortfall of the hill-climbing algorithm is that it can fall into a local maximum, and will fail to identify the overall best fitting structure [125]. This can be avoided by introducing random restarts to the algorithm, which causes it to "jump away from" the local maximum. The DAG that is "jumped to" is a perturbation of the local maximum DAG; some of the arcs in the local maximum DAG have been added, deleted, or reversed at random.

Expert elicitation was used to constrain the hill-climbing algorithm in the form of a blacklist. A blacklist is a list of directional arcs that are not allowed to be present in the DAG;

for example, if an arc from variable A that is directed to variable B is blacklisted, then it will not be present in the network structure after the hill-climbing algorithm is complete, though its reverse arc (from B to A) could be present. The blacklist used consisted of arcs that would violate the prerequisite relationships between courses in the physics program (e.g. Calculus 2 could not have an arc pointing towards Calculus 1). We found that using a blacklist instead of a whitelist (a list of arcs that must be included in the DAG) allowed the hill-climbing algorithm more freedom in determining the probabilistic relationships between the courses, allowing the identification of relationships between classes that were not expected or reflected in the prerequisite structure of the courses (the prerequisite structure of Institution 1 is shown in Fig. 7.1 in Chapter 7).

In learning the networks for the seven target courses using hill-climbing with random restarts and expert elicitation, the DAG structure that was determined to be the structure with the maximum network score was not always the same. To account for this variability, model-averaging was employed to find the final structure. Model-averaging is a method that combines a set of DAGs built with a set of random variables and "averages" them by counting the number of times an arc appears in the set of DAGs and retaining arcs that appear more times than a specified threshold. Arc directions are determined in a similar manner; the direction of an arc that appears most often in the set of DAGs is the direction that arc takes in the averaged DAG. This occasionally may lead to arcs that introduce cycles to the graph; to avoid this, the least-occurring arc whose deletion would remedy the cycle was removed from the graph, resulting in a valid averaged DAG. In this analysis, 10000 DAGs were learned for each target course and were averaged, and only arcs that occurred at least 1000 times were retained.

6.2.3 Predicting Course Outcomes

A Bayesian network can be used to predict outcomes of new observations. When the outcome to be predicted is categorical, this is often referred to as classification. This prediction is made by using Eqns. 5.1 and 6.1, and the model selects the outcome with the greatest probability, based on the new evidence and the prior distributions in the Bayesian network. Traditional Bayesian network classifiers include Naive Bayes classifiers and tree-augmented naive Bayes classifiers. In a Naive Bayes classifier, the target variable has an arc that points to each of the independent variables, and no other arcs are present in the DAG [151]. Tree-augmented Naive Bayes are similar, but there are additional arcs between independent variables that have strong probabilistic relationships. Both of these sacrifice the interpretability of a traditional Bayesian network for one that fits the data well for predictive accuracy.

In this study, an ensemble of traditional Bayesian networks (networks that were not Naive Bayes or tree-augmented Naive Bayes) was used to perform predictions for the seven target courses. Traditional Bayesian networks were selected instead of typical Bayesian classifiers after some preliminary predictions were performed, where the naive Bayes and tree-augmented naive Bayes models were outperformed by traditional networks. The dataset was evenly split into a training set and a test set, both of whose distribution of the target variable was the same as the full dataset. The variables used to predict each target course outcome consisted of the courses that are typically taken prior to the target course, as well as first or second-semester college GPA, depending on the target course. Students were predicted to either "Succeed" or "Struggle" in the course, which had the same definitions

as the similarly named outcomes analyzed in Sec. 6.2.2. When predicting each course, the dataset was filtered to only include students who enrolled in the target course. The process for selecting these models is outlined in the following section.

Model Selection

When building an ensemble of Bayesian networks, ensuring that each model in the ensemble is independent is difficult, as each network is learned from the same dataset and Bayes' theorem will create similar network parameters. To avoid this, the models built in the ensemble were learned through cross-validation, which allowed each model to be learned from a resampled subset of the training set. A 10-fold cross-validation was used, and the Bayesian network learned by each fold of the cross-validation was retained as a model for the ensemble. As such, the ensemble predictor for each target course had 10 Bayesian networks as part of the ensemble.

To build the networks for the ensembles, SEM with the hill-climbing algorithm was used as the learning algorithm for each fold of the cross-validated training set. Each network in the ensemble was learned with the available data in the fold, and as a by product the missing information in the fold was imputed. The test set was input to each model in the ensemble; each model imputed any missing values for independent variables in the test set using the posterior probabilities in the model learned from the training data and the evidence from the test set. Each model then predicted the dependent or target variable, resulting in 10 sets of predictions. The mode of the ten sets of predictions was then taken to create the final predictions for the target variable. The purpose of performing the imputations was to make the parameters of the Bayesian networks more robust, leading to better predictive

performance. In preliminary analysis, models that used this method of multiple imputation performed 4-5% better than models built without multiple imputation.

Typically in prediction or classification, the dataset consists of the dependent variable and the set of independent variables used to predict the dependent variable. However, the intent of this study in predicting an outcome in a course is not simply to predict that outcome, but rather to predict that outcome in the context of the entire physics course network, to better serve advisors in physics departments. The training set used in building the models for the ensemble predictor contains all available information about required physics and math courses in the physics major at Institution 1. The test set is filtered to only include the independent variables used in predicting the dependent variable; these independent variables are the courses that are typically taken before the target course and the first or second-semester CGPA. The missing courses in the test set (these would be upper level physics courses that are taken after the target course, referred to as "post-course variables") are treated as missing data and are not included in the evidence E in Eqn. 6.1 used to predict the target course outcome. By including the post-course variables in the learning and building of the Bayesian networks, the predictive models are able to use the probabilities associated with those variables in making predictions on the target course. In preliminary analysis, including these post-course variables in the model building phase improved predictive performance by up to 10%.

The blacklist used in building the models in Sec. 6.2.2 was implemented in creating the models for some of the target courses. Its use was determined by whether it improved predictive performance of the model. A predictor does not necessarily need to be intuitive and its goal is to predict as best as it possibly can, so constraining the model to maintain

some form of chronological order of courses was not deemed necessary.

The nature of educational data, especially course outcome data, is often unbalanced. This is problematic in creating predictive models; because the models are trained on mostly the majority class, they tend to over fit the model to predict that most observations will fall into the majority class. As such the accuracy of a model (Eqn. 3.10) may be very high, but the model is simply guessing that every observation is in the majority class. The balanced accuracy, the average of sensitivity and specificity and $\mathcal{B}$ in Eqn. 3.13, of such a model would be 50%; the model only predicts the majority class well. For models performing predictions on unbalanced data, $\mathcal{B}$ is a better metric of model performance, as it contains information as to how well the model predicts both classes, not just the majority class. This is particularly important in the models constructed in this study, as the vast majority of students in each of the target courses received an A or B grade. If the models were built with only overall accuracy as the primary performance metric, then the failing and struggling students would mostly be ignored. It is precisely the struggling students that we want to identify so some type of intervention can be used to assist them. By using $\mathcal{B}$ as the primary performance metric, the model that is selected is the model that has the highest $\mathcal{B}$, and as such is that model that predicts both the majority and minority class the best.

To improve the $\mathcal{B}$ of a model, decision threshold tuning was used. In this study, the Bayesian network predictor receives an observation of independent variables and then predicts the dependent variable. The prediction is determined by the probabilities of the possible outcomes based on the evidence contained in the observation; if the probability of a student receiving a C grade or lower is greater than 50%, that student is predicted to be a student that will "Struggle". The 50% threshold is the default decision threshold of

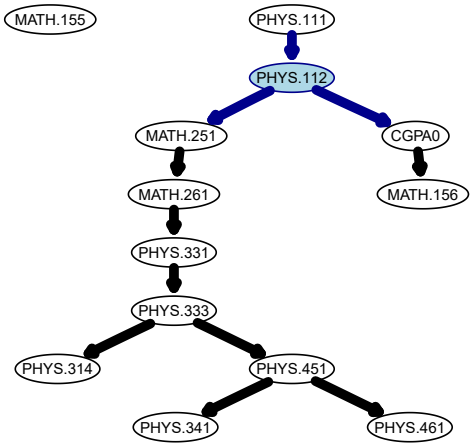
the model. This threshold can be tuned (changed) to a different value resulting in different predicted outcomes. For example, if the threshold was tuned to 25%, a student would have to have a probability of more than 25% of getting an C or lower grade to be determined as struggling; if their probability of C or lower was less than 25%, they would be predicted to succeed. By tuning the decision threshold, the model can be adjusted to predict both the majority and minority classes equally well, resulting in a maximized $\mathcal{B}$. The decision thresholds of the ensemble Bayesian network predictors were lowered in 5% increments. The resulting predictions were compared, and the model with the best $\mathcal{B}$ was selected. Typically the model with the highest $\mathcal{B}$ had a β_1 (sensitivity, Eqn. 3.11) equal to its β_2 (specificity, Eqn. 3.12), or nearly so.

6.3 Results

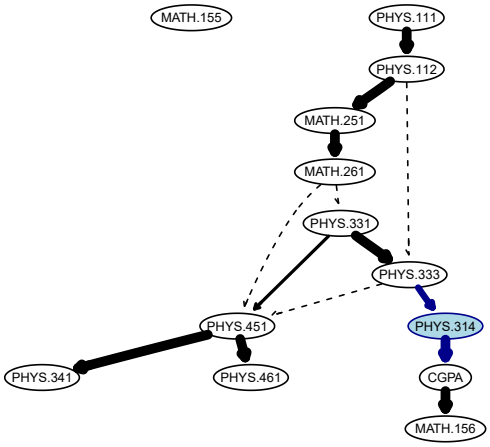
6.3.1 Identifying Conditional Probabilities

A Bayesian network was constructed for the following courses: PHYS.112 (Introductory Physics 2), PHYS.314 (Modern Physics), PHYS.331 (Classical Mechanics), PHYS.333 (Electricity and Magnetism), PHYS.341 (Advanced Lab), PHYS.451 (Quantum Mechanics), and PHYS.461 (Statistical Mechanics). A visualization of the averaged network for each target course was constructed. These are shown in Figs. 6.1 and 6.2. The target course is highlighted in blue, as well as the incoming and outgoing arcs to the target course. The linewidth of the arc is representative of the "strength" of the arc, or the frequency of the arc in the 10000 averaged networks used to construct the network; dashed lines represent the weakest arcs. Each network was built with only the records of the students who enrolled

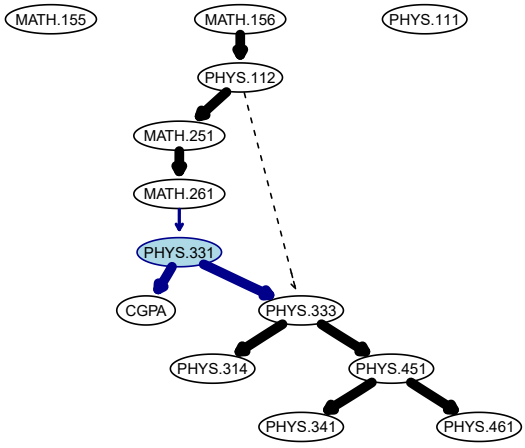
in the target course, and the target course was re-categorized from having classes A, B, C, DFW to the dichotomous classification of AB and CDFW. The differing network structures are due to the varying sample size used for each network and the re-categorization of each target course from a 4-level ordinal variable to a dichotomous variable. This focused the network on the relationships between the succeed or struggle outcome in the target course and specific grades in other courses. Nodes that have no incoming or outgoing arcs in the network are considered "excluded" from the network.



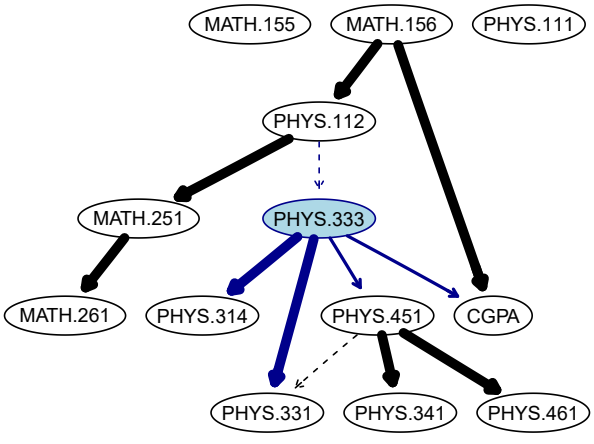
(a) PHYS.112



(b) PHYS.314

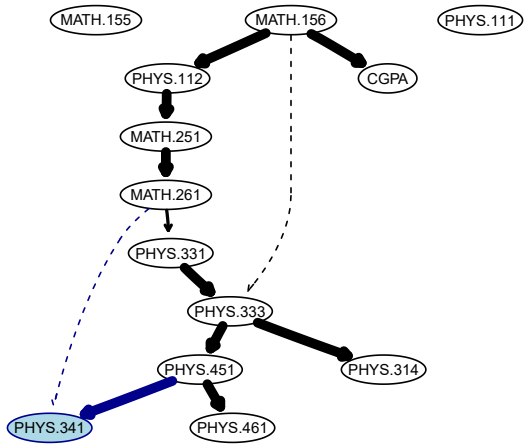


(c) PHYS.331

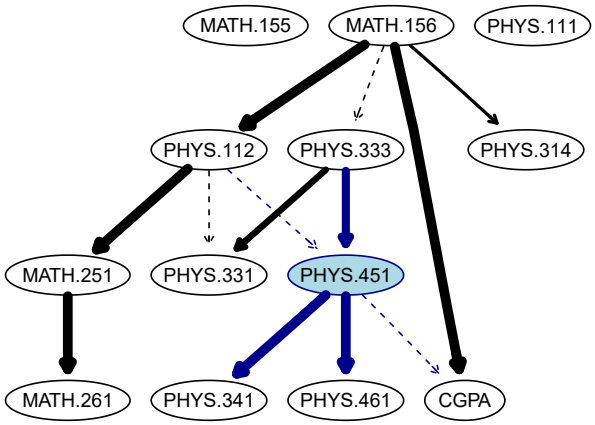


(d) PHYS.333

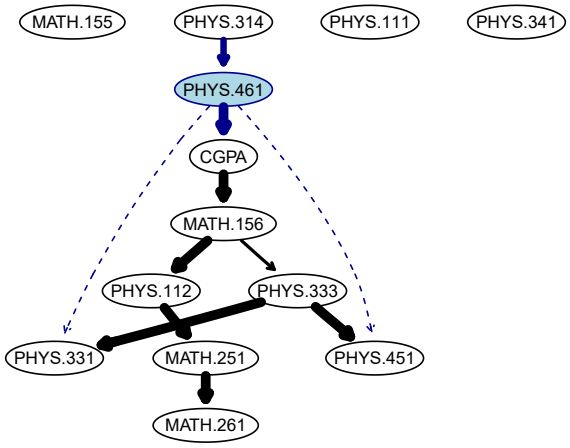
Figure 6.1: Bayesian networks for PHYS.112, PHYS.314, PHYS.331, and PHYS.333



(a) PHYS.341



(b) PHYS.451

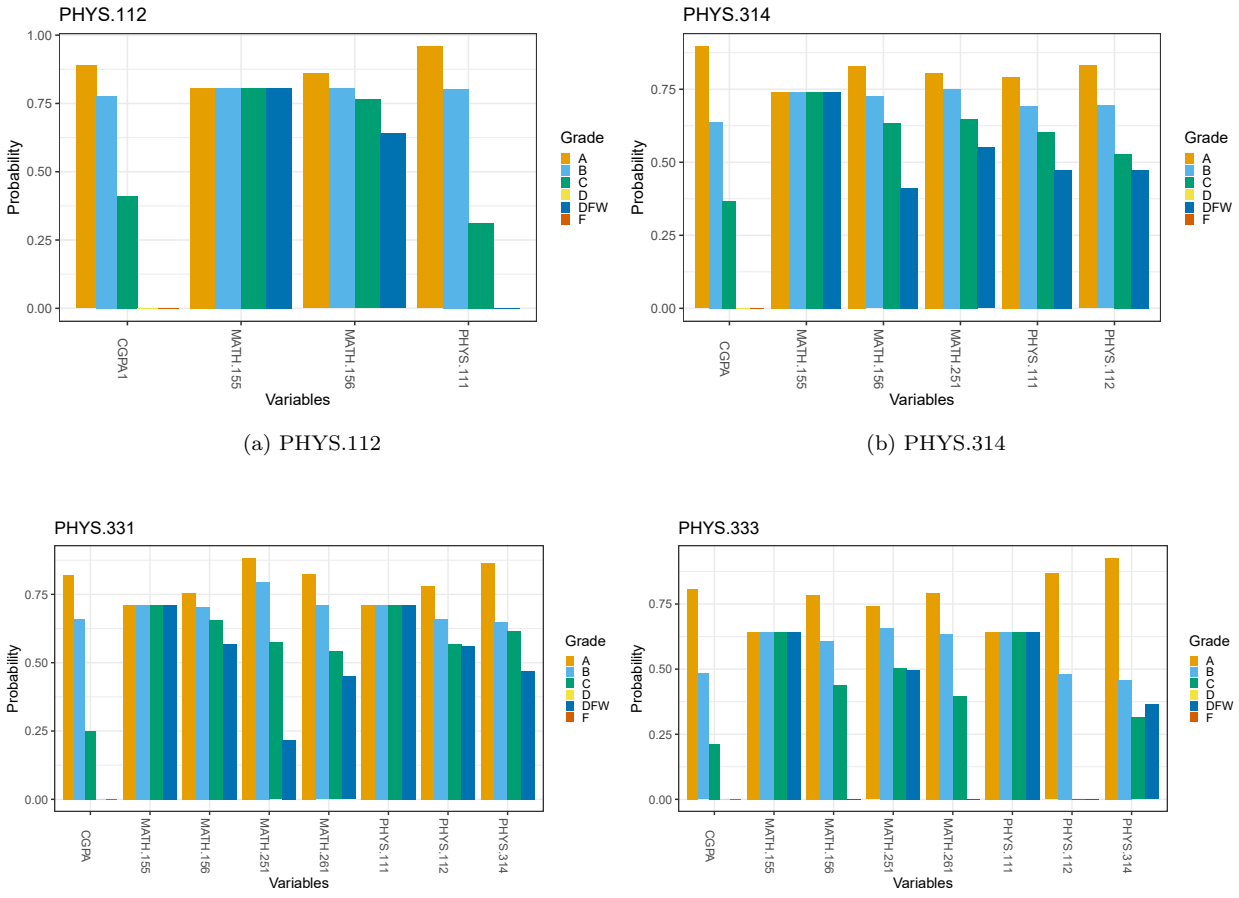


(c) PHYS.461

Figure 6.2: Bayesian networks for PHYS.341, PHYS.451, and PHYS.461

The networks were queried to determine the probability of an AB outcome in the target course based on an outcome in a prior course. The results of these CPQ's are shown for each target course in Figs. 6.3 and 6.4.

Each prior course provided a set of 4 probabilities based on the possible grades in that course. Possible grades in prior courses are A, B C, and DFW. If the set of probabilities for a prior course is nearly homogeneous (the same or nearly the same probability for an AB outcome in the target course regardless of grade in the prior course), it indicates that the



(a) PHYS.112

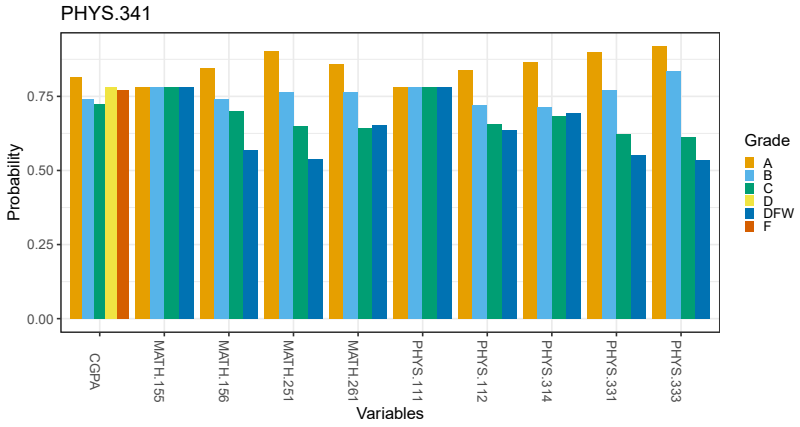
(b) PHYS.314

(c) PHYS.331

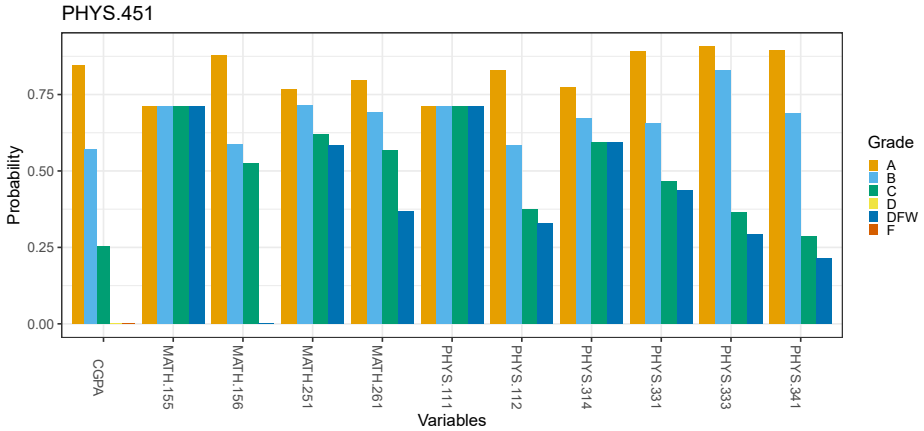
(d) PHYS.333

Figure 6.3: Probabilities of receiving an AB grade in a target course based on a grade received in a prior course.

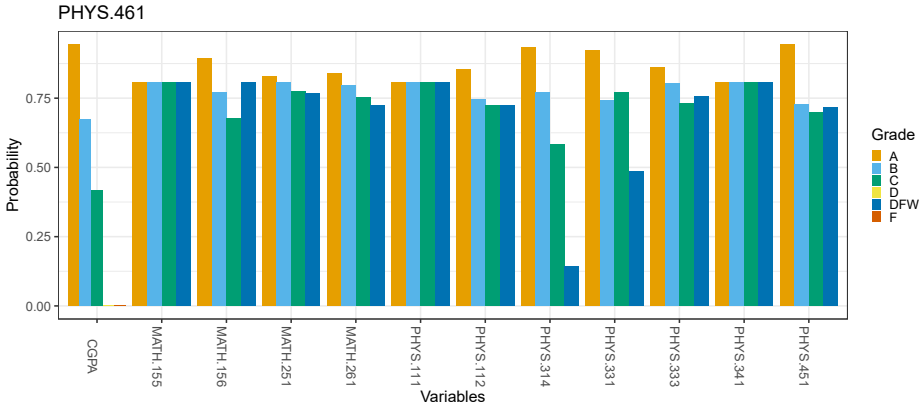
grade in the prior course has little or no correlation to the outcome in the target course. If there is significant variation in the set of probabilities for a prior course, it indicates that there is a relatively strong correlation between the grade received in the prior course and the the outcome in the target course. For each target course, the probability of a successful outcome in the target course is the same for each possible grade in MATH.155. In this case, the probability of a successful outcome in the target course is equal to the distribution of the target course, or in other words it is equal to the ratio of students who received a successful outcome in the target course to students who enrolled in the course. This is not surprising,



(a) PHYS.341



(b) PHYS.451



(c) PHYS.461

Figure 6.4: Probabilities of receiving an AB grade in a target course based on a grade received in a prior course. CGPA has levels A, B, C, D, and F and the course variables have levels A, B, C, DFW.

as MATH.155 is not connected by an arc to another node in any of the networks in Figs. 6.1 and 6.2, indicating that there is not a strong probabilistic relationship between its grades and outcomes in future courses. This is likely because MATH.155 is typically the first required course in which physics majors enroll if they are math ready. As majors progress through the program and take courses that are nearer to the target course, those course grades are more likely to have a strong probabilistic relationship with outcomes in the target course, and any relationship between the target course and MATH.155 is explained by the intermediary courses. Nearly every target course had several prior courses that had a relatively strong correlation between target course outcome and prior course grade. This was not true for PHYS.461, the Statistical Mechanics course. CGPA and PHYS.314 (Modern Physics) have the most variation across their grades for an AB outcome in PHYS.461, followed closely by PHYS.331 (Classical Mechanics). The other prior courses show very little variation in the posterior probability of a successful outcome in PHYS.461 across their possible grades (the range of probabilities is less than 20%). This may indicate that PHYS.461 does not fit well into the overall course network and should be examined for improvement.

6.3.2 Predicting Course Outcome

An ensemble predictor was built for each of the seven target courses. Each ensemble model consisted of 10 cross-validated Bayesian network predictors, and final predictions were based on the mode of the 10 Bayesian network predictions. The predictions for each course were performed 100 times; the results are averaged in Table 6.3. Table 6.2 shows the sample size used in each prediction, as well as the dependent variables used in the model. The dependent variable numbers refer to their numbers in Table 6.1.

Target Course	Dependent Variables	Sample Size	AB%
PHYS.112	1,2,5,14	318	79.2
PHYS.314	1-3,5,6,13	248	72.6
PHYS.331	1-7,13	241	71.4
PHYS.333	1-7,13	236	64.4
PHYS.341	1-9,13	199	76.9
PHYS.451	1-10,13	184	72.3
PHYS.461	1-11,13	175	81.1

Table 6.2: The sample sizes of each dataset used to predict the target variable, as well as the dependent variables and the AB% of the dataset. The dependent variable numbers refer to Table 6.1.

Target Course	Avg. % Accuracy	Avg. % Balanced Accuracy	95% C.I. for Balanced Accuracy	Decision Threshold	Blacklist
PHYS.112	76.9	76.8	75.9-77.7	0.15	No
PHYS.314	69.1	72.0	71.0-73.1	0.20	Yes
PHYS.331	73.9	74.0	73.2-74.9	0.25	Yes
PHYS.333	82.6	82.3	81.7-82.9	0.30	No
PHYS.341	73.6	72.9	72.0-73.8	0.20	No
PHYS.451	81.6	81.1	80.3-82.0	0.30	Yes
PHYS.461	73.8	72.6	71.4-73.7	0.20	Yes

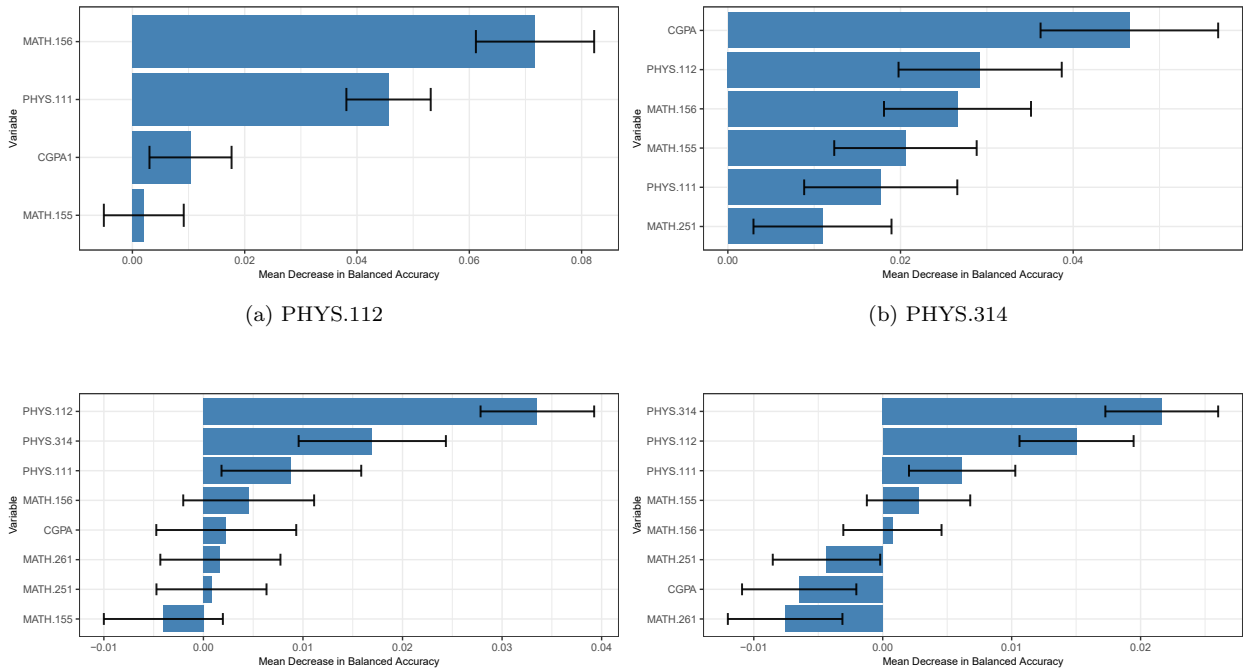
Table 6.3: Results of course predictions, averaged over 100 iterations. The decision threshold is the threshold for the probability that a student will "Struggle". The Blacklist column indicates whether the blacklist was used in learning the network structures.

PHYS.333 (Electricity and Magnetism) had the best performance based on the balanced accuracy of its model. It is followed closely by PHYS.451 (Quantum Mechanics). These are followed by PHYS.112 and PHYS.331, and then PHYS.341, PHYS.461, and PHYS.314. The last three have a balanced accuracy within one percent of each other. All models have a balanced accuracy greater than 70%, indicating a model that performs better than guessing the majority class for every observation (such a model would have a balanced accuracy of 50%). The decision thresholds for the models range from 0.15 to 0.3; the lower decision thresholds were used in the models whose dataset was more unbalanced (i.e. there were more AB students than CDFW students). The decision threshold for each

model was adjusted to optimize the prediction of the CDFW outcomes. A model with an un-adjusted decision threshold would err in the direction of predicting AB outcomes more accurately than CDFW outcomes.

Variable Importance

Once the optimal decision threshold for the balanced accuracy was selected, each model was checked to see the effect each independent variable had on the balanced accuracy. Each independent variable was removed from the model, and the balanced accuracy of the model without that independent variable was determined. Those results are shown in Figs. 6.5 and 6.6.



(a) PHYS.112

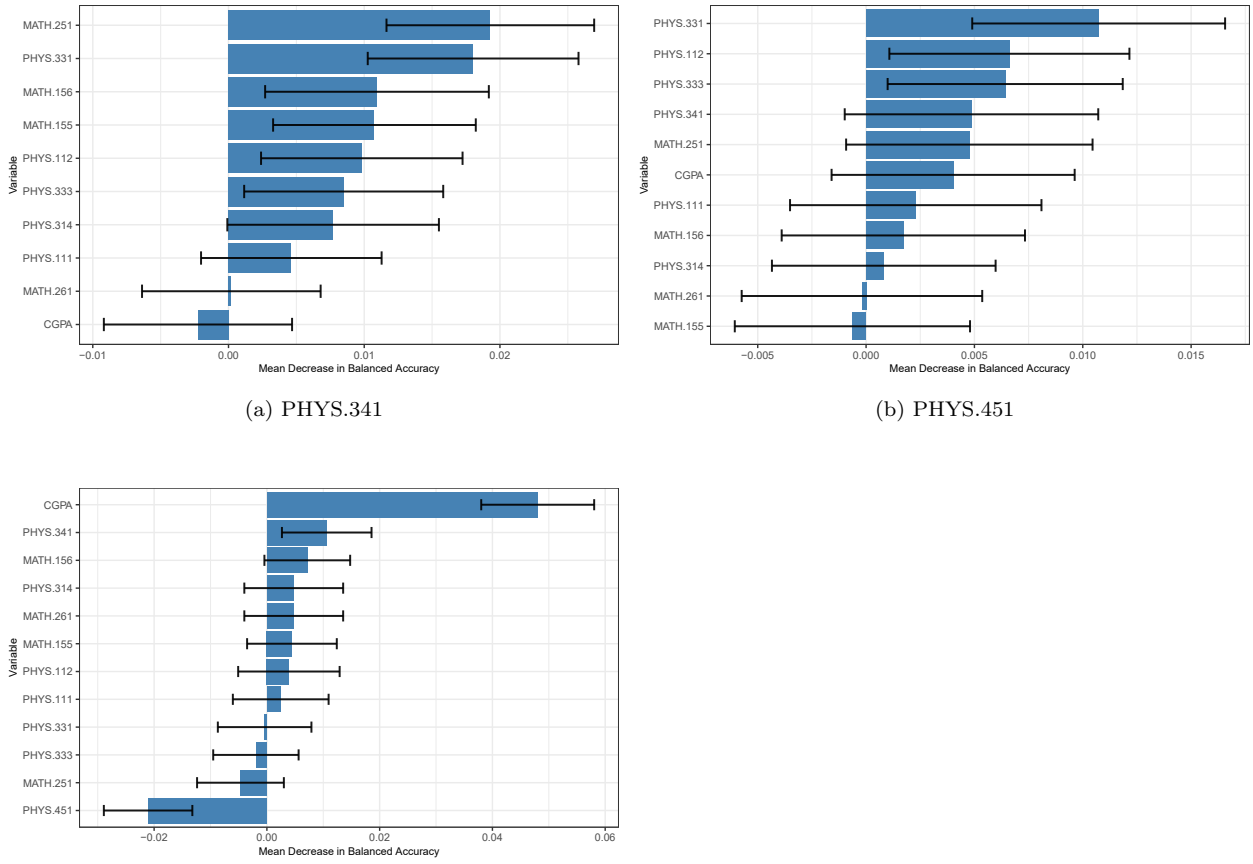
(b) PHYS.314

(c) PHYS.331

(d) PHYS.333

Figure 6.5: Variable importance based on mean decrease of balanced accuracy by dependent variables. The error bars represent the standard error of the difference between mean balanced accuracies.

The only variable to appear more than once as the most important variable in the prediction model was CGPA, which was the most important variable for predicting PHYS.314



(a) PHYS.341

(b) PHYS.451

(c) PHYS.461

Figure 6.6: Variable importance based on mean decrease of balanced accuracy by dependent variables. The error bars represent the standard error of the difference between mean balanced accuracies.

and PHYS.461. Interestingly, PHYS.314 and PHYS.461 were two of the three courses that were the most difficult to predict. Some models' balanced accuracy changed significantly with the removal of variables, such as the model for PHYS.112 where the removal of MATH.156 resulting in a mean decrease of nearly 7.5%. Conversely, the model for PHYS.451 only showed a mean decrease of about 1% for the removal of its most important variable PHYS.331. Both PHYS.112 and PHYS.341 had a math course as the most important variable, though for PHYS.341 the second most important variable was PHYS.331, which was within a tenth of a percent of MATH.251.

6.4 Discussion

This study sought to answer two research questions; they will be addressed below. The detailed results were discussed above; the following will synthesize the most important points.

RQ1: What are the probabilistic dependencies between upper-level physics courses and their prerequisites? The probabilities are shown in Figs. 6.3 and 6.4. As expected, as the grade received in a prior course decreases, the probability of an AB outcome in the target course also decreases. This is clearly shown by the probabilities of an AB outcome in PHYS.451 (Fig. 6.4b). Each of the courses that were not excluded from the network (an excluded course has no arcs) showed this trend, as can be seen by the clear and obvious left to right downward trend of the probability columns in the sets of probabilities for the prior courses. The larger the difference in probability between grades in a specific course results in a steeper set of columns, and indicates a course whose grade has a strong effect on the outcomes of the target course, in this case PHYS.451. This is also illustrated in the probabilities for PHYS.112, PHYS.314, PHYS.331, and PHYS.333, although the trend is less strong for PHYS.314 than the others. This is not the case for PHYS.341 and PHYS.461 (Figs. 6.4a and 6.4c). While the prior courses that are not excluded from the network do show decreasing probability for decreasing prior grade, the difference in probabilities for decreasing prior grade is much smaller and the set of columns are less steep for these target courses than the other target courses in the analysis. For example, a student who received a DFW grade in MATH.261 (Differential Equations) has nearly the same probability of an AB outcome in PHYS.461 as a student who received an A grade in MATH.261. The

only variables that seem to have a strong effect on the probable outcome in PHYS.461 are PHYS.314 and CGPA. This lack of strong dependence on prior course grades on the outcomes in PHYS.461 and PHYS.341 may indicate that these courses are not well integrated in the overall course network; they are not building upon the skills and knowledge gained in prior courses, or they are graded inconsistently with respect to other physics courses. Both of these courses are typically taken in a student's last three semesters in the physics program at Institution 1. Students in these courses have made it through the majority of their coursework and have been retained in the major regardless of prior course performance. If the weak dependence on prior grades was due to the course being a senior level course, the same trend should appear in all senior level courses. PHYS.451 also is typically taken during a students final three semesters, but PHYS.451 does not show the same weak dependence on prior grades. PHYS.314 is not a senior-level course, yet it shows trends similar to those of PHYS.461 and PHYS.341. In Fig. 6.3b, for PHYS.314 we see that each non-excluded prior course shows the trend of decreasing AB probability for decreasing prior course grades, but the differences in probabilities are relatively small compared to Figs. 6.3a, 6.3c, and 6.3d. For each prior course for PHYS.314, a DFW grade results in a probability of nearly 50% for an AB outcome; students who fail a prior course are just as likely to succeed in PHYS.314 as they are to struggle in PHYS.314. Similar to PHYS.461, only CGPA shows a strong effect on the probable outcome of PHYS.314. It also appears to not be well integrated in the overall course network.

The visualizations of the seven networks also give insight into the overall course network and how courses fit together within it. In nearly all of the networks, if the weakest three arcs were removed, the networks would break down into two smaller networks; one with

the introductory physics courses and required math courses, and the other with the upper-level physics courses (PHYS.314 (Modern Physics) appears in both of these "subnetworks", depending on the target course). The exception to this is the network for PHYS.112 (Fig. 6.1a). The network for PHYS.461 (Fig. 6.2c) may seem to be an exception, but removing the three weakest arcs creates two networks, one with only upper-level courses and one with the introductory courses, math courses, and PHYS.461. This implies that the upper level courses have stronger probabilistic relationships among themselves than they have with the introductory physics and math courses. The same is true for the introductory courses; they have stronger probabilistic relationships among themselves than they do with the upper level courses. This is not altogether unexpected; however, a stronger probabilistic relationship between the two groups would be preferred. At Institution 1, the introductory physics courses are not specific to physics majors; other STEM majors enroll in and make up the majority of students in these classes. This is also true for all of the required math courses for the physics major. The majority of students in these classes are enrolled in the various programs of engineering. This weaker connection between the upper and introductory courses may simply be that grades in courses specific to physics majors are much more probabilistically dependent on grades in other physics-specific courses. The data used in these courses do not include the other STEM majors that enroll in introductory physics and math courses; only physics majors are in the data.

RQ2: How accurate are Bayesian network predictors in predicting outcomes in upper-level physics courses? Which prior course is the most important predictor of the target course? Because of the unbalanced nature of the data, balanced accuracy, $\mathcal{B}$, was used as the primary performance metric for the prediction models. Each model performed at least

22% better than the baseline of 50% (a model that predicts the majority class for each observation). This is indicative of a generally good performance by each of the models. The range of $\mathcal{B}$ for the seven courses was 10.3%. The most predictable course was PHYS.333, and the least predictable course was PHYS.314, though PHYS.341 and PHYS.461 were nearly as equally un-predictable. The lower predictability of these courses indicates the outcomes in these courses do not reflect the grades received in prior courses. These are the same courses that were discussed in the prior research question as not fitting into the course network well. It is not inherently bad that a course outcome does not reflect prior grades, though it may be a cause for concern. If grades are assumed to be a measure of knowledge and skill mastery, then a course that is unpredictable is not building upon the knowledge and skills mastered in a prior course, or the knowledge and skills mastered in the prior courses are not sufficient for the content of the upper level course.

The most predictive variable for each target course was determined by finding the mean difference in $\mathcal{B}$ when the variable was removed from the dependent variables. These results may be useful to physics departments trying to implement interventions to help students progress more smoothly through the physics program, and the application will be discussed in the next section.

6.5 Recommendations

The ability to accurately predict student outcomes in a specific course could be a very informative tool for academic advisors and instructors in physics departments. If an advisor was concerned with how difficult an upcoming semester might be for a student, they could

input the student's prior course grades to a Bayesian network built from the records of prior students and get a probability of the outcomes in the various courses the student wishes to enroll in. Based on these probabilities, the advisor then can suggest to the student to alter their course load or delay a course to another semester if it is probable that the student will struggle with one or more of their upcoming courses. As students continue to enroll in physics courses and progress through the program, the Bayesian network can be updated with their course outcomes so the conditional probabilities of the network are constantly adjusted to better fit the data. A predictive model, similar to those built in this study, could also be built for the purpose of identifying students who may struggle in a course.

Often, there are certain required courses that are almost always taken concurrently in an academic program, even though there is no co-requisite structure between the courses. For example, physics students at Institution 1 typically enroll in PHYS.112 and MATH.251 in the same semester. Some of these course combinations have a high level of difficulty, which may be detrimental to certain students. If the probability of failing one or both courses when taken in the same semester is high, it may be advantageous for the student to take the courses separately. Although it was not explored in this study, Bayesian networks could be used to identify students that may struggle in certain course combinations, and advisors could use a Bayesian network or Bayesian predictor to find the probability of students' success in the course combination. The student could then be advised to delay taking one of the two courses.

These tools can also be useful to a department trying to reform their curriculum or improve their course outcomes. As discussed in prior sections, three of the courses that were analyzed appear to fit poorly in the overall network; PHYS.314, PHYS.341, and PHYS.461

were the least predictable and their outcomes did not have strong probabilistic dependencies with prior courses. A department doing a similar analysis could find courses that also are less predictable or similarly have weaker probabilistic relationships with prior courses. Courses such as these should be assessed, and perhaps changes need to be made to their content, instruction, or their place in the course network (i.e. change their prerequisite courses). The relationship between courses can also be taken into account by departments as they make decisions that will affect student outcomes. At Institution 1, PHYS.333 is the most predictable course. It also has the lowest AB%, meaning that it is the course that students struggle in the most. Although it is the most predictable, it clearly is a course in which Institution 1 would like to improve student performance. The most predictive course for PHYS.333 is PHYS.314 (Fig. 6.5d). This is reflected in the CPQ results for PHYS.333, where an A in PHYS.314 gives a high probability of an AB outcome in PHYS.333, but any other grade in PHYS.314 gives a low probability of an AB outcome in PHYS.333 (6.3d). Academic advisors and instructors could reach out to students who did not receive an A in PHYS.314, and encourage them to make use of resources in the department such as attending office hours and tutoring sessions, or encourage them to employ self-regulated learning techniques or metacognitive techniques. Due to the strong connection between PHYS.314 and PHYS.333, the physics department at Institution 1 could also adjust some of the content in PHYS.314 to better prepare students for the content they will see in PHYS.333.

6.6 Conclusion

Bayesian networks were used to analyze conditional probabilities between outcomes in seven physics courses and grades in their prior courses at Institution 1. Higher grades in previous courses resulted in higher probabilities of a successful outcome in the target courses. This analysis identified three courses, PHYS.314, PHYS.341, PHYS.461, which were less well predicted by their prior courses. An ensemble of Bayesian network predictors was used to predict outcomes in each target course. Balanced accuracy $\mathcal{B}$ was used at the metric of interest because of the unbalanced nature of the data. The predictions of each course resulted in a $\mathcal{B}$ of greater than 70%, with predictions for PHYS.333 having the greatest $\mathcal{B}$ of 82.3%. PHYS.314, PHYS.341, and PHYS.461 had the lowest $\mathcal{B}$. These courses may need to be examined to see if improvement can be made in their structure or instruction to better fit within the physics curriculum at Institution 1.

Chapter 7

Identifying Curricular Patterns Using Curricular Analytics

*

*The work in this chapter was submitted for peer review and publication in *Physical Review: Physics Education Research*. After lengthy review it was rejected, and is currently being revised for re-submission.

7.1 Introduction

Chapters 4 and 5 analyzed the pre-college factors that influence the retention of physics students. Chapter 6 analyzed the relationship between physics course outcomes and grades received in prior courses. Each of these chapters investigated which factors influence the progression of students through a physics program. The progression of students through academic programs is central to understanding student retention. Curricular Analytics (CA) [152] is a quantitative method developed to explore the pathways students traverse as they complete academic programs. CA is primarily a method to analyze the structure of a program's curriculum in order to quantify a program's complexity. The central hypothesis of CA is that, as a program's complexity is decreased, the student completion rate of the program will increase. As such, CA is a method that can inform the restructuring of program requirements and curriculum to improve student retention.

Physics curricula, the required courses and prerequisite relations in a physics degree, are superficially independent of issues of diversity and inclusion; however, this study will show that the complexity of the curriculum changes with the math readiness of the student. For most institutions, the four-year degree plans of physical science and engineering students assume a student is ready to enroll in Calculus 1 their first semester; these students are considered "math ready". A student's initial mathematics class is generally determined by their standardized test scores (ACT or SAT), often supplemented by a mathematics placement test. Several recent studies have shown that the prior preparation measured by standardized test score or conceptual physics pretest score of introductory physics students differs by demographic group [86–88]. This difference mediates the outcomes of students in

physics classes measured by course grades, final exam scores, or conceptual physics post-test scores. As such, students without access to advanced high school course offering may experience curricula with higher complexity than students with more enriched high school backgrounds. Often students from historically marginalized communities have less access to advanced high school coursework than other students [122]. Additional factors beyond academic preparation such as parental support can also influence success in college physics [89]. Equity is dependent on an institution identifying ways in which it can support timely graduation of STEM students who have been underserved in high school.

7.1.1 Research Questions

This work employs CA to investigate the program complexity of undergraduate physics programs. It explores curricular complexity across many institutions throughout the United States (US) and investigates the role that math readiness and chosen degree track has on complexity. A degree track is an area of academic focus which can be selected as part of the physics major such as a biophysics focus. The degree track generally modifies the requirements for the degree somewhat. This study seeks to answer three research questions:

RQ1: Is there a correlation between program ranking and program curricular complexity across physics programs in the US?

RQ2: How does a student's math readiness affect the complexity of their physics curriculum?

RQ3: How do different physics degree tracks alter the curricular complexity? Is the effect of math readiness different in some tracks than other tracks?

In this study, we only look at a program's curriculum, the required classes and their

prerequisite structure. For each student, the curriculum must be converted to an 8-semester degree plan which takes into account when courses are offered and the student's college preparation; this may modify complexity. The effect of converting the curriculum into a degree plan will be investigated in future works.

7.1.2 Results of prior research

Curricular Analytics represents a new research strand within PER studying the structure of physics curricula. This study represents the beginning of the strand; future work will examine how those curricula fit into academic semesters to become degree plans and how different degree plans predict student success. The purpose of such a research program is to understand the features of physics programmatic decisions such as the courses required, the prerequisites of those courses, and how often the courses are offered on the ultimate success of physics students measured by the rate of obtaining physics degrees. As such, CA will ultimately be informed by studies examining the retention of physics students to degree and the general retention of college students. See Chapter 2 for a literature review of student retention in college and physics.

7.1.3 Curricular Analytics

Student retention research focuses on the improvement of student graduation rates and retention throughout an undergraduate program. En route to graduating, students must successfully traverse their program's curriculum. This progression through a program's curriculum is fundamental to a student's overall academic success. Delays in their progression through the curriculum such as failing a class or changing majors generally will delay their

graduation and increase the risk of leaving college. Retention research often focuses on interventions designed to improve retention which affect student progression through their program's curriculum. Heileman *et al.* [152] proposed a quantitative framework called Curricular Analytics (CA) for analyzing the structure of a program's curriculum to improve understanding of the progression to degree. These analytic methods are used to quantify the effect of retention interventions on curricular structure and complexity. This approach to analyzing the sequence of courses and its effect on student retention is not unique to CA. Other methods that are similar in scope and design have been used to explore student progression through degree programs and are a growing area of STEM education research [153].

Curricular Analytics is a quantitative method to analyze curricula so as to inform decisions regarding curricular reform in a way to make curricula more equitable while retaining quality. The specifics of CA will be explored later in the chapter but, in brief, CA quantifies the structural complexity of a curriculum. This complexity is based on the prerequisite structure of classes in the curriculum and the sequence of classes that a student must follow, with a small contribution from the total number of required classes. The structural complexity of a curriculum is part of a curriculum's overall complexity; instructional complexity, the instructional practices applied in the courses in the curriculum, also contributes to curricular complexity. Heileman *et al.* argue that as curricular complexity is decreased, student completion rate of the curriculum will increase; a less complex curriculum will be more equitable with less chance of delay of graduation as students progress through the curriculum.

Curricular Analytics has been used to understand the structural effects of successful curricular innovations. Klingbeil and Bourne [154] introduced a curricular modification de-

signed to aid the progression of incoming engineering students through the Calculus 1 and 2 sequence. Many students enter the university not ready to enroll in Calculus 1. These students require several semesters to complete additional mathematics classes before they can enroll in their first engineering course. This is a common problem in physics and engineering programs where students must complete the introductory calculus sequence before entering their program-centered classes. While maintaining ABET standards in the engineering program at the university, an introductory Engineering Mathematics (ENGR 101) course was introduced. This course focused on hands-on approaches to the most important mathematics methods that are used in engineering courses. Successful completion of ENGR 101 allowed students to advance to program-centered engineering courses such as the introductory physics sequence, engineering mechanics and statics, and computer programming sequences before completing the traditional calculus prerequisites for these courses. This change nearly doubled graduation rates while narrowly improving average GPA. Students from historically marginalized communities, including women and minorities, experienced the largest increase in graduation rate. This change reduced the effect of the introductory calculus sequence, allowing students to take the introductory calculus sequence at the same time as their program centered courses. Heileman *et al.* [152, 155] showed that this change reduced the curricular complexity for students unprepared to take Calculus 1 upon entering the program, supporting their argument that less complex curricula lead to increased graduation rates.

These types of curricular changes and their effects were investigated by Slim *et al.* [156]. In their study, Markov decision processes were used to quantify the relationship between program complexity and graduation rate, and were then used to model how curricular changes

affect graduation rates. Decreasing the complexity of the curriculum increased graduation rates.

To further support the benefit of less complex curricula, a study compared the curricular complexity of Electrical Engineering programs with the ranking of that program to determine if higher ranking programs were more or less complex than lower ranking programs [157]. Program ranking was taken from the US News rankings of graduate engineering programs. Programs with higher rankings had less complex curricular structures than schools with lower ranking. For clarity, a school ranked 5th is considered to have a higher ranking than a school ranked 95th. This implies that higher-ranking schools had less complex paths to completion of an Electrical Engineering degree than lower-ranking programs. A similar study compared program complexity and ranking within Computer Science programs finding similar results [158]. The relationship between complexity and ranking in disciplines other than Electrical Engineering and Computer Science has yet to be established.

Other studies have applied CA to analyze the complexity of transfer student pathways to degree completion, with the result that transfer student pathways are more complex than standard program pathways [159, 160]. Similarly, one study looked at the complexity of the suggested path of study that an institution advises students to take and found that the actual paths that students followed to degree completion were less complex than the suggested path [161]. The study recommends that universities adjust the complexity of their suggested paths of study to reflect the least complex curricular structures possible. Other applications of CA include the use of the structural complexity of a program as a variable in a geometric probabilistic model that was used to predict graduation rates of students in different academic programs [162]. Other variables included ACT/SAT scores, HSGPA, and

course completion rates. The geometric model predictions were within 3 percentage points of the true 4 year graduation rates.

7.2 Methods

7.2.1 Sample

Curricular complexity was compared across three tiers of physics programs in the US. Following prior studies in Electrical Engineering and Computer Science, these tiers were selected using program rankings from the 2022 U.S. News and World Report College Rankings [163] for graduate physics programs. Although it was the undergraduate programs that were analyzed, we hypothesized that graduate rankings would largely mirror undergraduate rankings with some slight variation. Each of the programs in the ranking offer a doctoral degree, and there are 188 programs in the ranking. Twenty schools were randomly selected from the first two deciles in the rankings to make up the upper tier. Each decile of the ranking contained 19 programs, thus the top two deciles contain programs ranked from 1-38. These deciles included schools such as Harvard, the University of Washington, and the University of Texas at Austin (the schools listed here were not necessarily schools included in the analysis; they are simply representative of the schools in the first two deciles). The middle tier consisted of 20 schools randomly selected from the fourth and fifth deciles of the rankings and included schools ranked from 75 to 113. These deciles included schools such as the University of Nebraska-Lincoln, Brigham Young University, and the University of Oregon. The lower tier was made up of 20 schools from the ninth and tenth deciles, which included schools ranked from 150 to 188 such as Portland State University, Utah State University,

and the University of Alabama-Birmingham. In the random sampling within each tier, if an institution was selected that did not have a clear, publicly available, delineation of the requirements to complete their undergraduate physics program, a different institution was randomly selected. Institutions that operate on a quarter system were also excluded from the sampling as it was unclear how to modify the complexity of a program in a quarter system to be comparable to a program in a semester system.

To answer research questions 2 and 3, we focused on the physics curriculum of a single university from the second tier. A recent study explored the physics retention patterns of this program [164]; this study is discussed in Chapter 4. The institution is a large public land-grant with an overall undergraduate population of 20,500. The general undergraduate demographic composition in fall 2019 was 82% White, 4% Black or African American, 4% Hispanic/Latino, 4% non-resident alien, 4% two or more races, with other groups 2% or less. The 25th to 75th percentile range of ACT composite scores range was 21 (59%) to 27 (85%) for the 25th percentile to the 75th percentile of students scores [119]. Thirty-one percent of undergraduate students met the eligiblity requirements for Pell grants. This institution will be referenced as Middle Tier Public University (MTPU) in this study.

7.2.2 Curricular Analytics

Curricular Analytics is a method of quantifying the complexity of of an academic program's curriculum, developed with the purpose of quantifying the impact of curricular reform. For a full treatment of the methods and theories of CA, see the study in which CA was introduced [152].

The primary metric of CA is the overall curricular complexity. This is composed of

two components: a structural component and an instructional component. The instructional component is defined to be a function of a vector of factors of all the instructional properties of a curriculum. Similarly, the structural component is a function of the vector that contains all of the structural characteristics in a curriculum. The instructional properties consist of the instructor quality, course support services such as tutoring and office hours, and any other property of the instruction. Structural properties include the prerequisite and corequisite structure of courses, course credit hour totals, etc. The overall complexity of curriculum c is given by a functional f of the instructional complexity function and the structural complexity function:

$$\psi_c = f(\alpha_c, \gamma_c) \tag{7.1}$$

with ψ_c as the overall complexity, α_c as the structural complexity function, and γ_c as the instructional complexity function. The primary assertion of CA is that as overall complexity increases, the completion rate of curriculum c decreases:

$$\psi_c \uparrow \implies \beta_c \downarrow \tag{7.2}$$

where β_c is the completion rate of the curriculum. The inverse is also assumed, that if curricular complexity is decreased, the completion rate will increase. Decreasing the overall complexity can be accomplished in two ways: by improving (decreasing) the instructional complexity or by lowering the structural complexity.

The structural complexity of a curriculum is quantified by examining the prerequisite structure of the curriculum. This prerequisite structure is visualized by using a directed

acyclic graph (DAG), where individual courses are nodes and the edges connecting nodes are prerequisite or corequisite requirements. This is called a curriculum graph. An example of a curriculum graph is shown in Fig. 7.1. A program's curriculum graph contains all of the pertinent characteristics of the structural complexity of that program. Heileman *et al.* defined five characteristics of a program's structure: the delay factor, the degrees of freedom, the blocking factor, the reachability factor, and the centrality factor. In the present work, only the delay and blocking factors, which are required to calculate the structural complexity, and the centrality factor are discussed. For a full treatment of each factor refer to Heileman *et al.* [152].

Required courses in a curriculum are generally part of a required course sequence, where each course in the sequence must be completed before advancing to the next course in the sequence. Some courses may be part of several sequences. The delay factor, d_n, of a course n is defined as the number of courses (or nodes on the curriculum graph) that are included in the longest sequence that contains course n. For example, in Fig. 7.1 the delay factor of General Physics 1 would be 6 resulting from the path traversing nodes, Calculus 1, General Physics 1, General Physics 2, Introductory Modern Physics, Quantum Mechanics 1, and Quantum Mechanics 2. Often the longest sequence includes courses that act as gateway courses; courses that are a prerequisite course to many other required courses.

The blocking factor, b_n, of course n is the number of courses or nodes for which n is a prerequisite or equivalently the total number of courses that follow after n in all the course sequences that include n. For example, the blocking factor of General Physics 1 in Fig. 7.1 is 13; the classes blocked by General Physics 1 are shaded in grey in the figure.

The overall curricular delay factor is the sum of the delay factors of each of its con-

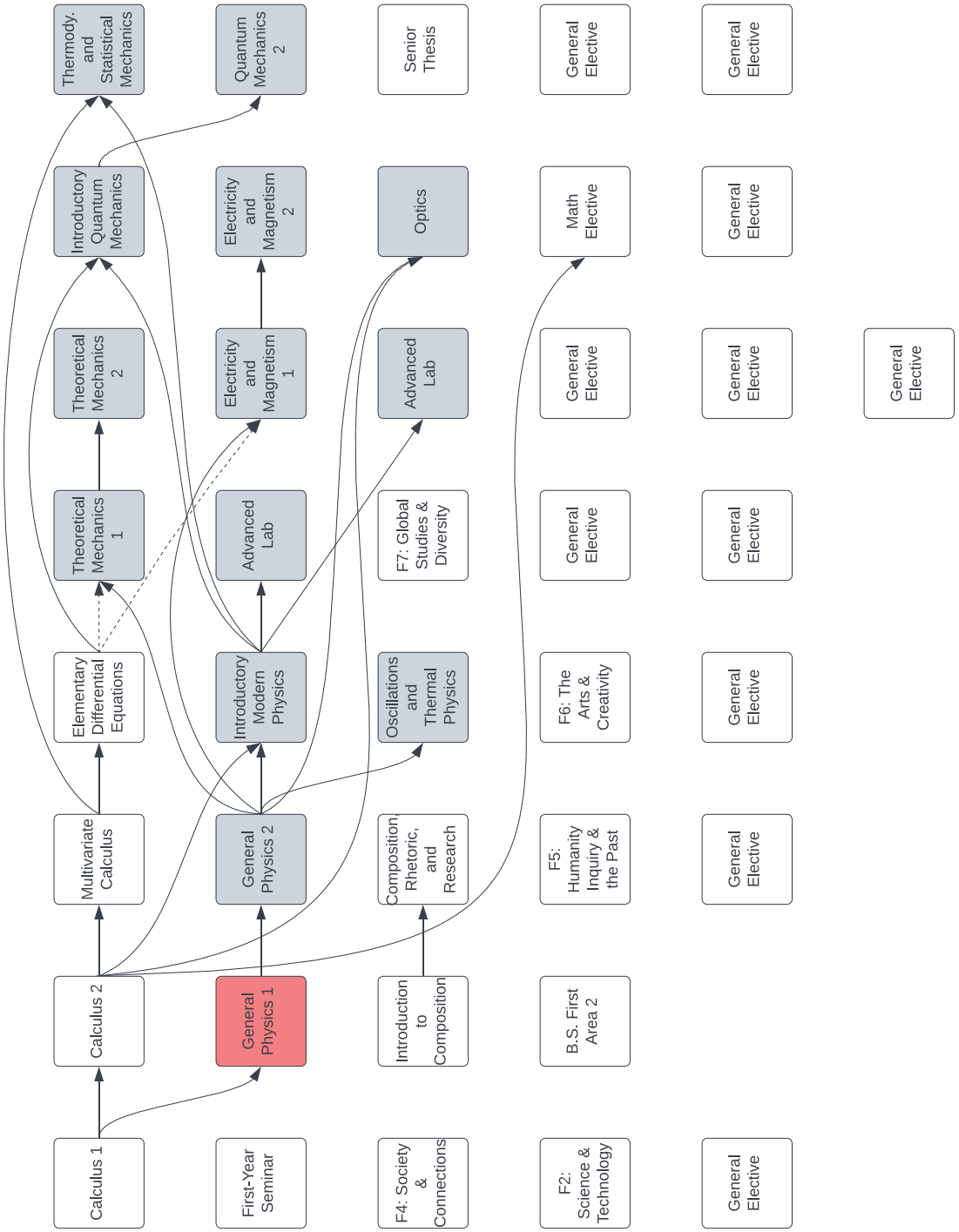


Figure 7.1: Example curriculum graph for a mid-tier institution. General Physics 1 is shaded red and the courses it blocks are shaded gray. The count of the gray courses is the blocking factor of General Physics 1. The delay factor of General Physics 1 can be found by counting the number of courses in its longest path, in this case 6. The university divides general education requirements into seven categories labeled F1 to F7 in the figure.

stituent courses and similarly the overall curricular blocking factor is the sum of each course's blocking factor. For each course in the curriculum, the delay factor and the blocking factor are added to give the individual course complexity v_n such that $v_n = d_n + b_n$. The structural complexity of a program, α_c, is given by summing the course complexity of each course in the curriculum as shown in Eq. 7.3.

$$\alpha_c = \sum_n v_n. \tag{7.3}$$

The structural complexity can also be calculated by adding the overall curriculum delay factor and the overall curriculum blocking factor.

The course centrality factor identifies courses that have several important prerequisites which are also prerequisite for many required courses. The course centrality factor attempts to measure how critical the progression through a course is for the completion of a curriculum. The course centrality factor of course n is calculated by summing the length of all the complete paths p that contain the course n. For a path to be included in the summation, course n must be an interior node to p, or in other words it cannot be a terminal node on a path. The most central course to a curriculum, or the course with the highest centrality factor, is defined to be the course with the most long course sequences. Although the course centrality factor plays no direct role in calculating the structural complexity, it gives information as to what courses are especially crucial to successful program outcomes and as such is of interest to student retention research studies.

The instructional complexity component of a curriculum's complexity is more difficult to quantify than the structural complexity. Instructional characteristics are qualitative in

nature; it is challenging to consistently quantify their effects on student outcomes. Heileman *et al.* suggest the use of course grade outcomes or pass/fail rates as an estimation for instructional complexity [152]; however, this is far from a complete measure. This proposal originates from the observation that, as students progress through a curriculum, any failing grade in a class delays their progression in the curriculum. Classes with higher failure rates, therefore, are problematic courses and increase the complexity of the curriculum and decrease the completion rate of students. The current work focuses on the structural complexity of physics programs, and reserves the instructional component for future studies.

To calculate the curricular complexities of the curricula analyzed in this paper, the CA website was used [165]. This site automates the calculation while providing a rich graphical representation of the relations in the curriculum.

7.3 Results

7.3.1 Curricular analytics across multiple institutions

The physics program requirements for 60 institutions in the U.S. were analyzed using CA. These institutions were separated into three tiers based on their graduate physics program rankings [163]: the upper tier, middle tier, and lower tier. For each program, the structural complexity was calculated and the central course identified. To find the prerequisite structures of each program, the institution's catalog was examined for the program requirements. Most programs included in the analysis have several different degree tracks available to physics majors. In each case, the degree track that was suggested for students planning to continue their physics education in graduate school was selected. The program

requirements consist of a set of core required classes that all students must take and then a number of physics or mathematics electives with a list of course offerings which fulfill the elective. To maintain consistency, similar courses were selected for each program's elective requirements when possible. For each institution, the first math class required for the major was Calculus 1; the curriculum was designed for students who were prepared to take Calculus 1 in their first enrolled semester. The effect on the curricular complexity of a student not being ready to take Calculus 1 in their first semester is discussed in Sec. 7.3.2.

Tier	Mean	SD	SE	95% CI
Lower	239	39	9	(220, 257)
Mid	224	38	8	(206, 242)
Upper	237	46	10	(215, 259)

Table 7.1: Summary of the structural complexities of each tier. The table presents the mean, standard deviation (SD), standard error (SE), and 95% confidence interval.

The summary statistics of each tier are reported in Table 7.1. There was not a large difference between the mean structural complexity of the tiers, with only a 13 complexity point difference between the upper and middle tier and 15 points between the middle and lower tiers. The 95% confidence interval is also reported and the intervals for each tier substantially overlap. The range of complexity scores in each tier is similar, with the upper tier spanning 160 complexity points, the middle tier spanning 152, and the lowest tier spanning 157. There was some variation of the distribution of each tier, and this variation is illustrated in Fig. 7.2 where the shaded boxes of each tier represent the 25% to 75% range of that tier. The dark vertical line is the median and the two light horizontal lines, called whiskers, span the first and fourth quartile. The full range of the data points in each tier are contained between the tips of the whiskers. The probability density plot of the tiers is overlaid on the box and whisker plot for each tier.

Figure 7.2 shows the range of each tier is similar. Each tier has one of the three most complex structures and one of the three least complex structures. The most complex curricular structure is within the lower tier, the second most complex structure is in the middle tier, and the third most complex structure is in the upper tier. The least complex structure is in the upper tier, the second least complex is in the lower tier, and the third least complex structure is in the middle tier.

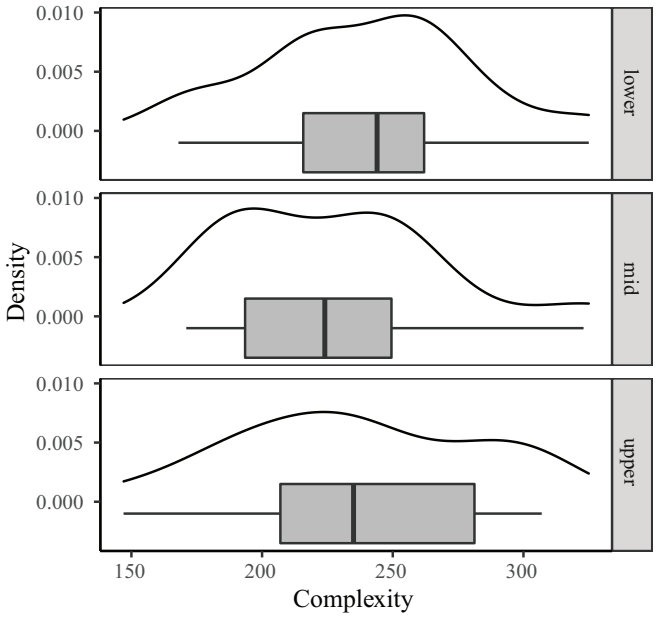


Figure 7.2: Distribution of curricular complexity for physics programs with different rankings.

The means of the structural complexity of each tier were compared using analysis of variance (ANOVA) and tested against the null hypothesis that there is no difference between the means of the tiers. The traditional threshold of significance for the F-statistic in ANOVA is a value above 3.15, which corresponds to a p-value of 0.05. The ANOVA analysis resulted in a F-statistic of 0.75, well below the threshold of 3.15, as such the null hypothesis is not rejected. This sampling of possible institutions does not indicate that there is a difference

between the means of structural complexity of the three tiers of physics programs.

The central course, the course with both a number of prerequisites and a number of courses for which it is prerequisite, is the most important course to progress to and through for successful matriculation through the curriculum. For 38 institutions, the central course was Calculus 2; General University Physics 2 was the central course for 17 institutions, and General University Physics 3 was the central course for 5 institutions. General University Physics 2 was the introductory, calculus based, electricity and magnetism course. General University Physics 3 had a different description at each of the 5 institutions where it was the central course. At each institution, it had some coverage of wave mechanics and introductory quantum physics; at two of the institutions, it had some coverage in relativity. One of the institutions also included basic thermodynamics as part of its description.

7.3.2 The role of math readiness

To investigate the effect of math readiness on curricular complexity (this section) and the effect of degree tracks on curricular complexity (next section), we focus on one of the middle tier institutions, called Middle Tier Public University (MTPU). This institution is situated in a small eastern state with high levels of poverty and low levels of academic achievement. Its student body is moderately prepared for college based on ACT score ranges and the institution is not very selective, accepting 90% of its applicants [119]. As such, MTPU represents an interesting laboratory to study the effect of curricular changes on complexity particularly for students not ready to enroll in Calculus 1 upon entering college.

Most physics programs offer a suggested plan of study outlining one or more paths a student could take to earn a physics degree. For physics, these suggested plans of study

generally assume that incoming students are ready to take Calculus 1 upon entry; however, this is often not the case for many students. At MTPU, 41% of students who enter enrolled as physics majors are not ready to take Calculus 1 [164] and are considered to be not math-ready. The more prepared of these non-math-ready students enroll in a two-course stretch Calculus sequence, Calculus 1a/b with Precalculus. This sequence replaces the more common Precalculus to Calculus 1 sequence at many institutions. Students are allowed to progress to Physics 1 after completing only the first of the two courses in the sequence which then allows students to enroll in their program specific classes earlier. These students, however, must take an additional mathematics class not taken by math ready students which may affect their curricular complexity. Some incoming students are not prepared to take the stretch calculus sequence and must take additional mathematics classes, usually College Algebra and Plane Trigonometry before enrolling in Calculus 1a/b with Precalculus further increasing curricular complexity. The number of additional math courses will vary from institution to institution. Some students enter college with credit for Calculus 1 either through Advanced Placement or a similar program or by transferring college credit earned in high school. We only consider Calculus 2 as a potential first mathematics class in this study, but more advanced classes are possible.

Figure 7.3 shows the curricular complexity for various levels of math readiness using the degree track selected by students planning to attend physics graduate school. This is plotted with the first mathematics class in which a students enrolls as a freshman on the horizontal axis; the student must also take all mathematics classes to the left of this class on the axis to complete their degree. The more mathematics courses a student must complete before taking Calculus 1, the higher the curricular complexity. The additional chain of prerequisite

math courses increases both the longest path of many courses and their blocking factor. The additional math courses also, generally, shift the central course of a curriculum. For most institutions examined in Sec. 7.3.1, for math ready students, the central course was Calculus 2. The central course shifts depending on math readiness to usually the second course in the math sequence. For example, a student who begins mathematics in the College Algebra course at MTPU will follow the sequence of 1) College Algebra, 2) Plane Trigonometry, 3) Calculus 1a with Precalculus, 4) Calculus 1b with Precalculus. In this sequence, Plane Trigonometry would become the central course of the curriculum.

This additional complexity affects a student's time to degree. If a student is ready for Calculus 1 upon entering college, there is enough flexibility in the MTPU physics curriculum for the student to finish in 4 years even if they fail a course. As math readiness decreases, that flexibility to traverse a curriculum in the target 4 year period also decreases. At MTPU, if a student must enroll in College Algebra upon entering, it is not possible for the student to graduate in 4 years or 8 semesters (assuming he or she does not take summer classes) because of the prerequisite requirements of the courses. The minimal sequence requires 4.5 years, 9 semesters, assuming the student begins in the fall semester, and they do not take any electives which have a prerequisite they cannot take until their final term. Failing a course will generally cause the time to degree to increase. Beyond the effect on time to degree and program complexity, a student who begins mathematics in College Algebra will not be able to enroll in Physics 1 until their fourth semester, assuming no delays arise in traversing the curriculum. MTPU finds it very hard to retain physics students who do not enroll in their first real physics class until the end of the sophomore year (they do take a 1-credit freshman seminar class their first semester) [164].

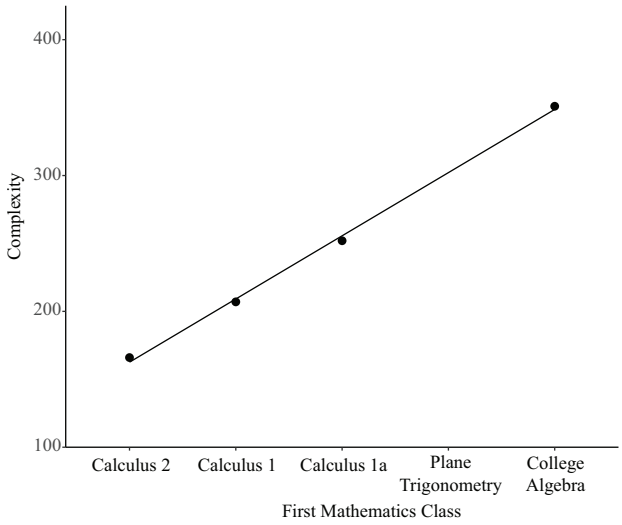


Figure 7.3: The complexity of the graduate-intending degree track plotted against the first mathematics class in which the students enrolls.

7.3.3 The effect of degree tracks

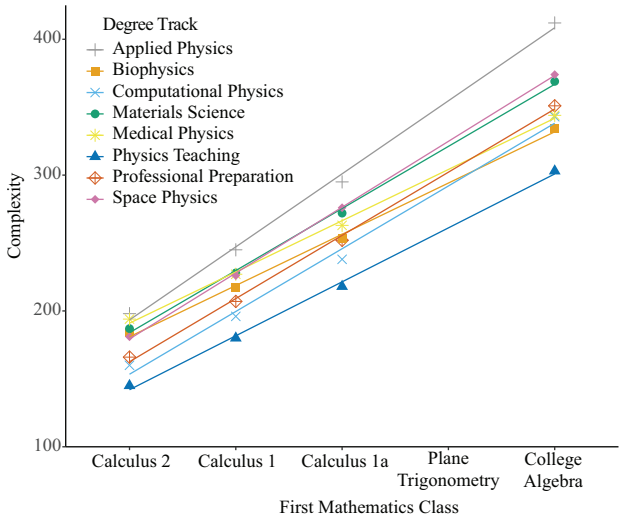


Figure 7.4: The complexities of various degree tracks versus the first mathematics class the student takes in college.

Many institutions offer degree tracks, sometimes called a concentration area or an area of emphasis, as part of their plan of studies. Some institutions may require a student to select a degree track, while others list them as options for students who wish to pursue a specialization within physics. Each degree track contains different required courses, which

have different prerequisite requirements. A typical general plan of study for physics usually involves multiple upper-level physics and mathematics requirements, often allowing the student to select from a number of options (electives). Should a student choose to pursue a degree track, those open-ended electives are typically replaced by a set of courses which are required in order to fulfill the learning outcomes of the track. Some of these courses may be courses offered outside the physics department, such as mathematics, engineering or chemistry courses. These additional requirements affect the overall curricular complexity, either positively or negatively, depending on the required courses. Figure 7.4 plots curricular complexity of each degree track at MTPU against the first mathematics class in which the student enrolls. The professional preparation track is the track selected by students planning on attending graduate school. This was the degree track used, when possible, for the calculations in Sec. 7.3.1.

The complexity changed substantially by degree track depending on the kinds of additional courses required. At MTPU, tracks which require more engineering, chemistry, and/or biology (i.e., Applied Physics or Materials Science) tend to have higher complexities than tracks which require more computer science and mathematics (i.e., Computational Physics or Physics Teaching). Engineering courses tend to have similar mathematics requirements to courses in a physics curriculum and require the same introductory physics courses. While the mathematics requirements of chemistry courses, particularly introductory courses, vary across institutions, at MTPU there are not additional mathematics requirements that add to the complexity; however, the lab grading structure contributes to the higher complexity. The chemistry courses and accompanying laboratory sections are graded independently; a passing grade of both the lecture component and the laboratory component are required to

progress to the next course in the sequence. This is the main factor increasing the complexity of degree tracks which require chemistry, creating a much longer delay factor. A similar effect occurs in the biology courses, as well as the engineering courses, which contain a laboratory component. The mathematics and computer science courses do not have these compounding effects. The laboratory component of the introductory physics classes is integrated with the lecture section and graded as a single class.

7.4 Discussion

7.4.1 Research Questions

This work explored three research questions which will be addressed in the order proposed.

RQ1: Is there a correlation between program ranking and program curricular complexity across physics programs in the US? ANOVA showed that there was no significant difference between curricular complexity of the three tiers of institutions. Further, the means of each tier were within a 15 point spread, and the 95% confidence intervals of the means of each tier overlapped strongly as shown in Table 7.1. There does not appear to be a correlation between program ranking and program curricular complexity for physics programs. Physics programs in the US across a broad range of national rankings have fairly similar curricular structures, as indicated by the similarity of means between tiers.

The greatest difference between the complexity of any two programs in the analysis is 178. In the study by Heileman *et al.*, where electrical engineering programs were compared across tiers [157], the difference between the least and most complex curriculum analyzed

was over 400 complexity points. All of the programs analyzed in the electrical engineering study were ABET accredited programs and thus their curriculum was constrained by external requirements. Physics programs have no such external constraints, and yet have similar curricula across institutions with substantially different national rankings. Most of the 60 physics programs analyzed require Calculus 1 through Differential Equations, a 2-course introductory, lab-based sequence in physics followed by modern physics and core advanced classes in classical mechanics, electromagnetism, and quantum mechanics. These advanced classes are supplemented with a form of advanced laboratory, often multiple forms. The advanced classes produce most of the differences in curricular structure. These differences are often in the number of elective courses required beyond the core requirements for a physics degree; however, some programs had extra intermediate required physics or mathematics classes such as linear algebra, a second modern physics class, wave mechanics, or mathematical methods in physics. Table 7.2 presents some characteristics of the first and last complexity quartile of the institutions studied aggregating all 60 institutions. Institutions with a complexity in the first quartile have on average 5 fewer required physics, math, or science courses than the institutions with a complexity in the last quartile.

Quartile	Mean Complexity	Avg. Required Physics and Math courses	Avg. Longest Path
First Quartile	182	19	6
Final Quartile	287	24	8

Table 7.2: Comparison between the first and last quartiles (ignoring tier placement) of the institutions included in the study

There was substantial variation present in all tiers of institutions; however, the variation is slightly larger for the upper tier than the middle and lower tiers. This may reflect different

approaches to student preparation. One approach is to increase the number of required courses and electives in an effort to increase the coverage of a student's physics education. This approach would give students a broader insight into different specialities in physics and perhaps better prepare them to choose an area of research in post-baccalaureate studies. It also makes the program more complex, limiting the possible ways a student could traverse the requirements in a reasonable time and increasing the chance of students dropping out of physics. The other approach is to require just the most basic core classes in physics and allow students to pursue additional courses which fit their goals and interests. This approach may not have the consistent coverage of different areas of physics but it allows students more freedom in their undergraduate education, allowing them the room to explore other fields and become more well-rounded students, while also increasing the likelihood that they complete the physics degree.

If there is not a correlation between program ranking and complexity, then why not lower the complexity of the curriculum in an effort to retain more physics students? Some may argue that lowering the complexity will decrease the quality of the education students receive. To refute this, note that the least complex curriculum analyzed is in the upper tier. This institution is a private university with an admissions ACT inter-quartile range of 33-35. It is consistently ranked in the top ten universities in the U.S. and internationally for general undergraduate and graduate education. The result that there appears to be no significant difference between the mean complexities of the tiers suggests that the more complex structures of some programs are unnecessary.

The purpose of CA is to allow departments and universities to make informed decisions on curriculum and pedagogical change based of the quantitative metrics so as to increase

student retention while maintaining the desired learning outcomes. Physics departments want to retain and graduate more students. Lowering the curricular structural complexity can facilitate this goal while maintaining program quality.

RQ2: How does a student's math readiness affect the complexity of their possible degree plan in physics? The analysis of math readiness (along with the analysis of degree tracks) showed the overall structural complexity increased as the number of required math courses increased. Figure 7.4 shows a linear trend of increasing complexity per additional math course. These additional math courses, which form a chain of prerequisites required to enroll in Calculus 1, not only add additional complexity to the curriculum, but also delay a student's entry into the introductory mechanics course, delaying the point where the student actually begins taking physics classes. This was evaluated at one institution (MTPU); this trend should hold for other institutions. The complexity or the delay added at other institutions will depend on their respective mathematics prerequisite structure, as well as the requirements for entering Physics 1 (i.e., whether Calculus 1 is required before enrollment or if it can be taken concurrently to Physics 1).

All institutions should consider the effect of math readiness. There are several factors which contribute to an incoming student's math readiness. Some students come from disadvantaged backgrounds and may not have access to college preparatory high school mathematics classes needed to prepare them for a math-heavy field such as physics. Institutions should examine possible solutions to ease the math transition of those students who require additional mathematics to complete their physics degree.

RQ3: How do different physics degree tracks alter the curricular complexity? Is the effect of math readiness different in some tracks than other tracks? While general physics

curricula are of similar complexity across a range of institutions, many institutions offer degree tracks to give students the opportunity to specialize in a sub-field of interest in physics or related fields, such as engineering or computer science. Altering the general curriculum to accommodate these degree tracks influences the overall structural complexity. At MTPU, some degree tracks require courses outside of the physics department, which have varying effects on complexity. Engineering and chemistry courses tend to add more complexity, especially if they have lab-based courses. These courses often grade the lab separately from the lecture part of the course requiring the student to pass the lab independently from the lecture. Math and computer science courses tend to add less complexity.

It is not uncommon for physics students to seek minors and/or a second major in a related field, such as mathematics, engineering, or computer science. Physics curricula which have a higher complexity of the physics portion of the curricula not only affect factors such as time-to-degree for their physics degree, but also make it much more difficult for students to pursue opportunities outside of the physics programs, such as a minor or additional major. Should a student suffer a setback in their trajectory, it may become more difficult for a student to pursue a degree track without jeopardizing their time-to-degree. This effect can be magnified particularly within smaller programs, as in many cases courses may only be offered once a year, or once every other year. Increased complexity combined with a decrease in the availability of course offerings can make it difficult to traverse through the general physics curriculum, and often more difficult to traverse through a degree track.

7.4.2 Other Observations

To illustrate how a department might decrease their curricular complexity, a semi-quantitative comparison of the least complex programs and the most complex programs included in the analysis is provided. The 15 programs with the lowest complexity scores, regardless of ranking tier, make up the first quartile of the data, and the 15 programs with the highest complexity scores make up the final quartile of the data. This comparison is found in Table 7.2. The difference between the means of the quartiles is 105 complexity points, which is largely explained by the increased number of required physics and math courses. These are the courses specifically required by the physics program and exclude the institution's general education requirements. The institutions in the final quartile require 5 more courses than those in first quartile. The impact of these additional courses is that they increase the delay factor of many of the required courses; essentially, they elongate the paths that a student must complete within the curriculum as shown in the Average Longest Path column, which presents the average of the longest paths present in the programs in each quartile. The length of the longest path should not be confused with the minimum number of semesters required to complete the program. The length of the longest path is a count of all the courses in the longest path. Courses that are corequisites and are completed in the same semester both count toward the longest path, and so the minimum number of semesters to complete a program and the longest path in that program are not always equal. Some of the programs in the final quartile have longest paths of length 8; students who arrive at the university ready to take Calculus 1 and who never fail or retake a class can graduate in 4 years. Any misstep or scheduling conflict will extend their time to degree.

Several of the programs in the final quartile have longest paths of length 9, and one program has a longest path length of 10. Students in these programs must complete at least two courses as corequisites to be able to complete the program in 4 years, and any course failure or scheduling conflict will extend their time to degree. Shortening the longest paths in a curriculum is a straightforward solution to decreasing the structural complexity. This can be done in two ways: by decreasing the total number of required physics and math courses, and by reorganizing the prerequisite structure of the curriculum. To reorganize the prerequisite structure, academic faculty should analyze the required prerequisite knowledge of a course to determine if the prerequisite course is necessary; an example is provided in Nash *et al.* [166]. Other tools could also be utilized, such as the Markov decision processes in [156], to model what effect changes in prerequisite structure will have on graduation rates.

7.5 Simplifying Curriculum by Making Prerequisite Adjustments

This section presents an example of how a physics department could rearrange the prerequisite structures of their program to reduce complexity. We created a curriculum consisting of 20 physics and math courses that are representative of common requirements for an undergraduate physics degree. This curriculum is not from a specific institution, but rather contains common structures that are present in many of the curricula we analyzed. The initial curriculum had 20 required courses and an overall structural complexity of 290. After changing the prerequisite structures the less complex curriculum had 20 required courses and an overall structural complexity of 222, a reduction of 23% percent. Most of the changes made were changing the prerequisite math course of a physics class to an earlier math course.

For example University Physics 1 had a prerequisite of Calculus 2; this was changed so the prerequisite was Calculus 1. All of the prerequisites in the initial curriculum can be found in various curricula from the institutions we analyzed. Similarly all of the prerequisites in the less complex curriculum can also be found among the institutions we analyzed.

The example curriculum of 20 physics and math courses with complexity of 290 is presented in Fig. 7.5. The same curriculum is then presented with an adjusted prerequisite structure with a complexity of 222 is presented in Fig. 7.6. No courses were dropped from the curriculum to make this change. We will refer to these as Curriculum A and Curriculum B.

7.5.1 Curriculum A

The courses in Curriculum A are fairly typical of physics curricula that were analyzed in the study, and 20 courses is about average for all the programs analyzed in the study. All of the prerequisite structures used in this curriculum are present in several of the curricula analyzed in the study, though none of the studied curriculum are an exact copy of the example curriculum here. Curriculum A has a maximum delay factor of 8; the longest course sequence in the curriculum contains 8 courses. There are two 8-course sequences.

7.5.2 Curriculum B

Curriculum B contains the same courses as Curriculum A, but the prerequisites of the courses have been adjusted to shorten the longest course sequences, resulting in a reduced structural complexity. All of the adjusted prerequisite structures are present in several of the analyzed curricula in the study. Curriculum B has a delay factor of 6; there are two 6-course

sequences. The most straightforward change made to Curriculum A to create Curriculum B was to shift the math prerequisite forward for several classes. In Curriculum A, the prerequisite to take Introductory Physics 1 is Calculus 2. This is not an uncommon requirement, though most of the analyzed curricula in the study have Calculus 1 as a prerequisite for Introductory Physics 1. Curriculum B reflects this, and has Calculus 1 as the prerequisite for Introductory Physics 1. This change also shifted the math prerequisite for several other classes. These changes and others are detailed in Table 7.3.

Curriculum A		**Curriculum B**	
Course	Prerequisite/Corequisite	Course	Prerequisite/Corequisite
Calculus 1		Calculus 1	
Calculus 2	Calculus 1	Calculus 2	Calculus 1
Calculus 3	Calculus 2	Calculus 3	Calculus 2
Differential Equations	Calculus 3	Differential Equations	Calculus 2
Linear Algebra	Calculus 3	Linear Algebra	Calculus 3
Partial Differential Equations	Differential Equations	Partial Differential Equations	Differential Equations
Introductory Physics 1	Calculus 2	Introductory Physics 1	Calculus 1
Introductory Physics 2	Calculus 3, Introductory Physics 1	Introductory Physics 2	Calculus 2, Introductory Physics 1
Wave Mechanics	Differential Equations, Introductory Physics 2	Wave Mechanics	Introductory Physics 2
Modern Physics	Linear Algebra, Introductory Physics 2	Modern Physics	Introductory Physics 2
Classical Mechanics	Differential Equations, Introductory Physics 2, Math Physics	Classical Mechanics	Differential Equations, Introductory Physics 2
Math Physics	Differential Equations, Linear Algebra	Math Physics	Differential Equations, Linear Algebra
Electricity and Magnetism	Wave Mechanics, Math Physics	Electricity and Magnetism	Differential Equations, Introductory Physics 2
Electricity and Magnetism 2	Electricity and Magnetism	Electricity and Magnetism 2	Wave Mechanics, Electricity and Magnetism
Quantum Mechanics	Wave Mechanics, Modern Physics, Classical Mechanics	Quantum Mechanics	Linear Algebra, Wave Mechanics, Modern Physics, Classical Mechanics
Quantum Mechanics 2	Quantum Mechanics	Quantum Mechanics 2	Quantum Mechanics
Thermal Physics	Modern Physics, Math Physics	Thermal Physics	Differential Equations, Linear Algebra, Modern Physics
Computational Physics	Modern Physics	Computational Physics	Modern Physics
Advanced Physics Lab	Computational Physics	Advanced Physics Lab	Modern Physics
Advanced Physics Lab 2	Advanced Physics Lab	Advanced Physics Lab 2	Advanced Physics Lab

Table 7.3: Two example curricular structures with differing complexity.

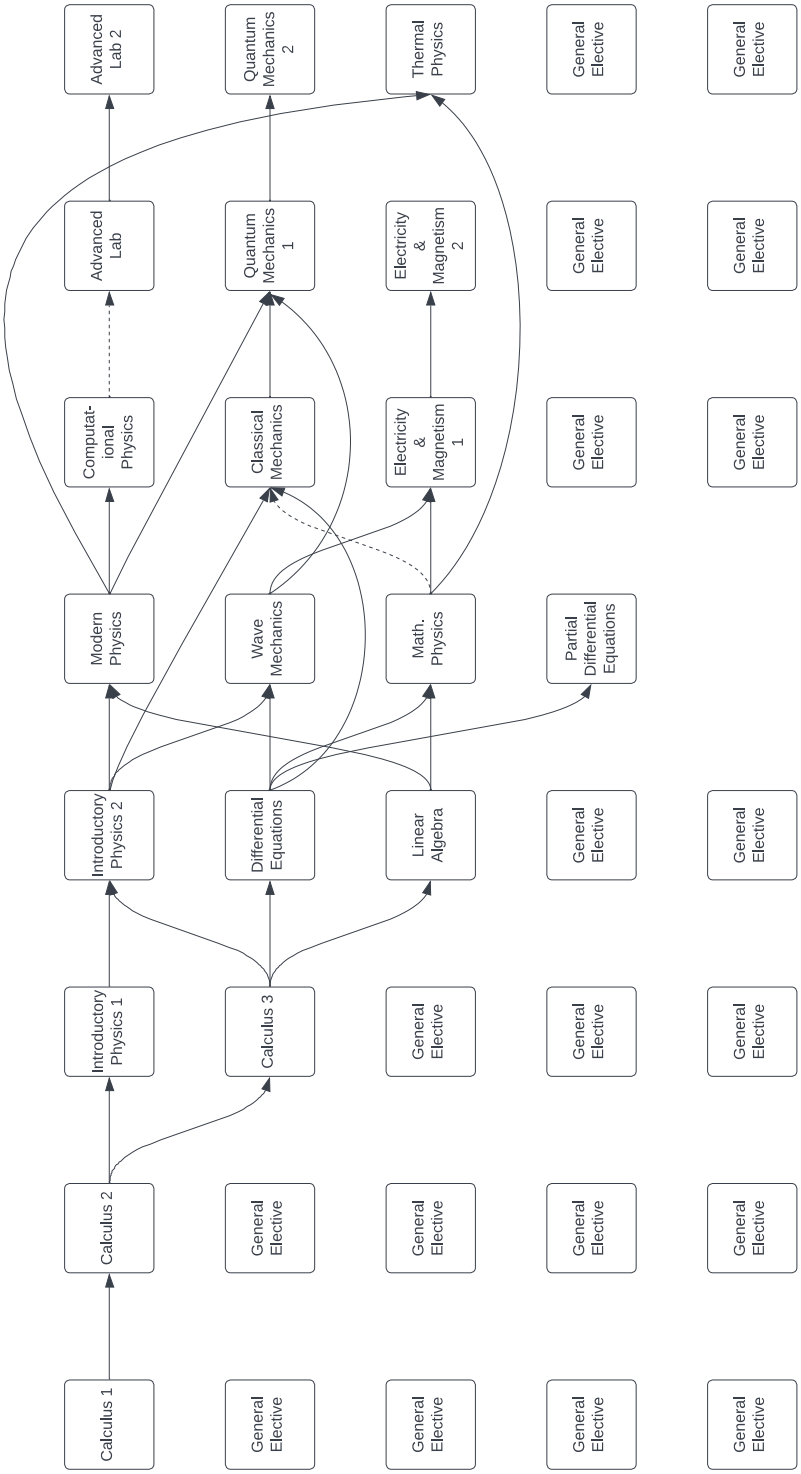


Figure 7.5: Curriculum A, with 20 required physics and math courses, and a structural complexity 290.

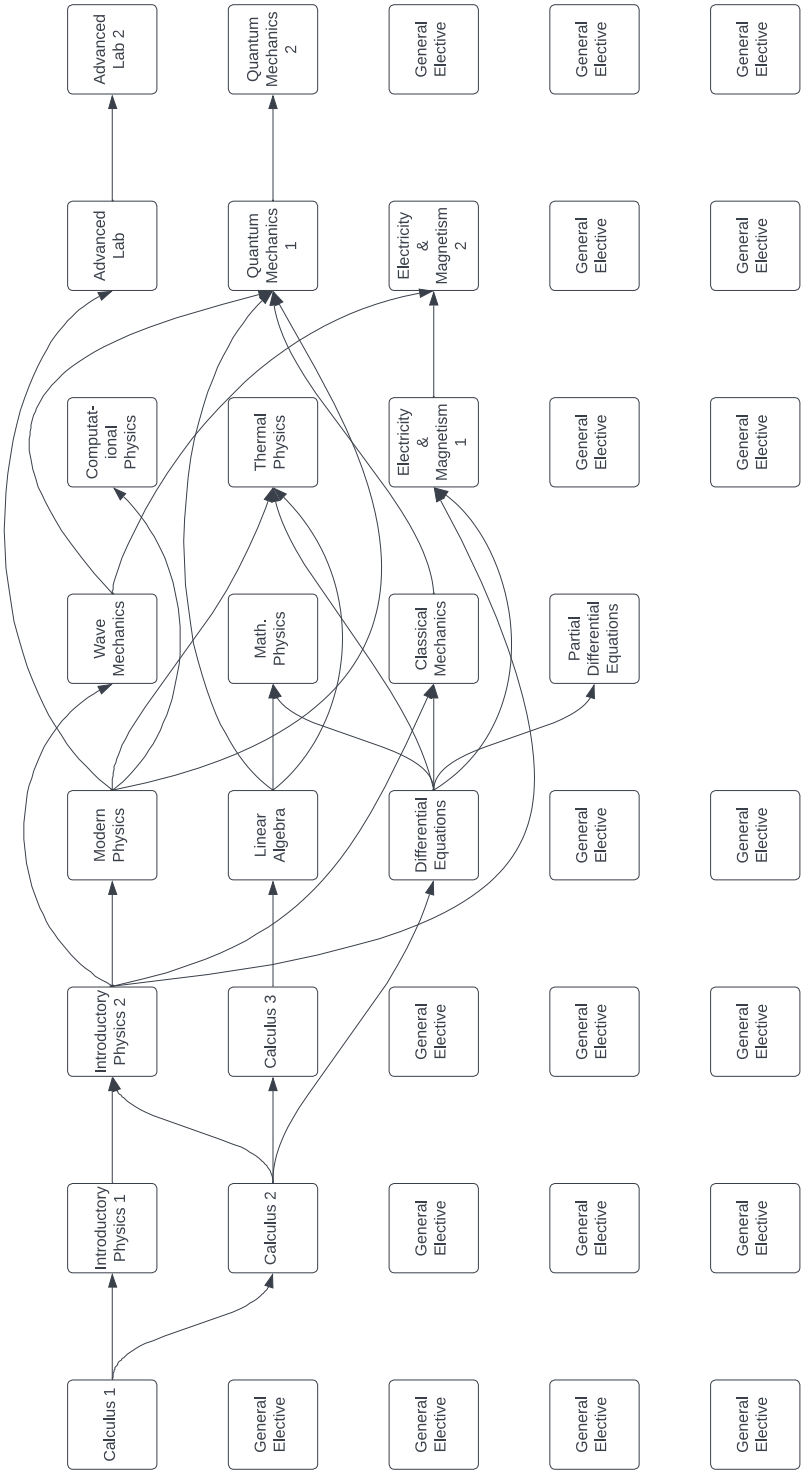


Figure 7.6: Curriculum B, the adjusted curriculum, with 20 required physics and math courses, and a structural complexity of 222.

7.6 Implications

Curricular Analytics provides all academic units with quantitative metrics to characterize the complexity of a curriculum. Physics departments can use these tools along with others tools to identify problematic course structures and to evaluate potential changes to program requirements quantitatively. These tools can help ensure all students complete their academic programs in a timely manner and help mitigate the risk that an academic misstep will cause a student to leave the program or not graduate.

The average structural complexity of each of the three tiers was approximately the same for the physics curricula taken by graduate-intending students. This implies that the additional complexity of some programs may not be necessary. As there was no significant correlation between program complexity and tier ranking, we recommend that physics departments use the tools provided by CA to determine their program complexity and then make changes to the curriculum structure that simplify students' paths to completion. In making a program's curriculum less complex, departments should not lose their identity; rather they should look at the desired outcomes for students who complete the program and then trim any course or unnecessary prerequisite that does not directly contribute toward those outcomes.

Because complexity increases linearly with decreasing student math-readiness, departments should investigate what changes can be made accelerate the progression of non-math-ready students into physics classes. Students from historically marginalized communities often do not have access to advanced high school college preparatory course options [122], and so this problem of increased complexity for non-math-ready students becomes a problem

of equity. A model where students who are not ready to take calculus are allowed to take the introductory, algebra-based physics courses instead of the introductory calculus-based physics courses so they can complete the math requirements while taking physics courses [164] may be one solution. These students could then be given credit for the calculus-based introductory classes when they have completed Calculus 1 and some advanced physics class such as Modern Physics. Another model could be that employed by Klingbeil *et al.* [154]. An introductory physics course could be created that teaches the basic math skills required in introductory physics using active learning methods. This course would serve as the pre-requisite to introductory physics instead of Calculus 1, and students could enter physics courses before or while they are completing the required calculus sequence. While these are not the only solutions, they reflect a type of solution that makes the curricular structure less complex while creating a more equitable path to completion for students with different levels of college preparation.

The complexity of different degree tracks should also by analyzed. While degree tracks will have differing complexities due to a difference in elective courses and their prerequisites, there should not be a large disparity between the complexities of different degree tracks. Degree tracks allow students to specialize in a particular sub-field of physics, perhaps in preparation for specific careers or specific areas of research in graduate school. If one degree track's complexity is significantly greater than the others, then students seeking to enter that sub-field are at a disadvantage compared to their peers. Any large disparities in the complexity of degree tracks should be addressed through curricular reform.

The goal of the present work was to introduce CA to the physics community particularly the PER research community and to replicate the work of Heilman *et al.* [152] in physics.

A program's curricular complexity is only part of the structural features influencing student success; each student must fit the curricular requirements into an 8-semester degree plan. The semester in which a class is offered (spring or fall) and the frequency the class is offered (every semester, every year, every other year) can further impact time to degree. Transfer students and students who were not math ready often cannot follow the typical degree plan prescribed by the department, and are often considered "off sequence". Required classes that are offered infrequently (once a year or once every two years) are especially detrimental to students who are off-sequence, and they often have longer times to degree due to the necessity of waiting until a required class is offered again. The overall difficulty of each semester (measured by rate students pass courses and the total credit hours in the semester) can affect the student's likelihood of successfully passing all courses in a semester. These effects will be investigated in a future work.

7.7 Limitations

The rankings of the physics programs in tiers were taken from the 2022 US News rankings of the best physics graduate schools [163]. These rankings are the product of a survey conducted by US News that asked department chairs and department directors of graduate studies to rank schools with physics PhD programs from 1 (marginal) to 5 (outstanding). The response rate for this survey in physics was 27.9%. If a school received less than ten ratings, it was not included in the rankings. This is not a scientific determination of hierarchy among physics programs in the US, but rather is a ranking based upon popular opinion and public perception. We feel this is still a useful tool, and that most would

agree that the groups of randomly selected institutions in the upper, middle, and lower tiers are approximately in the same general order as would be accomplished by a more rigorous classification system.

This study also focuses on structural complexity while ignoring instructional complexity. Instructional complexity may alter the relation of curricular complexity to student success. The developers of Curricular Analytics recommend using course completion rates as an estimate of instructional complexity. Instructional complexity encompasses all aspects of a course's delivery and environment, and course completion rate may be an oversimplification of a complex metric. Development of robust metrics of instructional complexity will be explored in future work.

7.8 Conclusions

Curricular Analytics (CA) is a quantitative framework for characterizing the complexity of college curricula with the goal of increasing student success. Physics departments could benefit from applying this framework to optimize course requirements thus giving every student the greatest possibility of successfully earning a physics degree.

This study applied CA to compare undergraduate physics programs at 60 academic institutions in the US, separated into three tiers based on the US News and World Report rankings. There was no significant relationship between program ranking and program complexity. This suggests that the increased complexity of some programs may be unnecessary; physics departments should consider making their curricula less complex to improve student retention. The most straightforward way to reduce the complexity of a curriculum is to

minimize the delay factors in the curriculum, by shortening the longest paths in a curriculum. This can be done by reducing the number of required courses, or by rearranging the prerequisite structures of courses.

One of the 60 institutions, MTPU, was selected to determine the relationship between curricular complexity and the level of the mathematics course in which a student first enrolls in college; there was a linear relationship between the number of math courses taken before Calculus 1 and the curricular complexity. This was the case for each degree track at MTPU indicating that students who arrive on campus not ready to take Calculus 1 must traverse a more complex curriculum than students who are ready to take or have already taken Calculus 1. Physics departments should be aware of the effect that student math-readiness has on the curricular complexity of their programs and make changes that make their programs more equitable for students who did not have the opportunity to take college preparatory mathematics courses.

At MTPU, degree tracks containing increased numbers of engineering and chemistry courses were more complex than degree tracks containing more mathematics and computer science courses. Physics departments should be aware of the difference in complexity between degree tracks and ensure that each track has a reasonable complexity; the additional requirements of a degree track should still allow graduation in four years.

Chapter 8

Exploring Student Knowledge Structures in the BEMA as measured by MIRT

*

*This chapter was published in "Hansen, J., & Stewart, J. (2021).*Multidimensional item response theory and the Brief Electricity and Magnetism Assessment.* Physical Review Physics Education Research, **17(2), 020139.**"

The Brief Electricity and Magnetism Assessment (BEMA) was developed to measure students' qualitative understanding of basic concepts in electricity and magnetism [167, 168]. The BEMA and the Conceptual Survey of Electricity and Magnetism (CSEM) [9] have been used in the majority of Physics Education Research (PER) studies of conceptual understanding of electricity and magnetism. Both were developed after Halloun and Hestenes demonstrated that students leave traditional physics classes with little change in their conceptual understanding [80]. This observation lead to the development of the broadly applied Force Concept Inventory (FCI) [7] which measured conceptual understanding of Newtonian mechanics. Using the FCI, Hake demonstrated that the failure of traditional instruction to foster conceptual learning gains was common to physics classes at many institutions [1]. The introduction of the FCI, CSEM, and BEMA as well as the Force and Motion Conceptual Evaluation (FMCE) [8] begin an extensive research strand in PER studying student understanding with multiple-choice conceptual instruments [15].

8.1 The Brief Electricity and Magnetism Assessment

The BEMA is a 31-item multiple-choice instrument that covers electricity and magnetism topics [167, 168]. It includes items covering electrostatics, electric potential, magnetostatics, and magnetic induction. This study used the version available from PhysPort [169]. Unlike the CSEM, the BEMA also includes 6 items involving electric circuits and 4 items asking the students to select responses involving quantitative formulas. The items present students with a variable number of possible responses with some items using up to 10 responses. Most responses include either a "none of the above" response or a response

that is zero; these types of responses have been shown to cause psychometric problems in other instruments [170].

The instrument contains multiple "item blocks" where multiple items refer to a common item stem or a common description of the physics system. Items {1, 2, 3}, {4, 5}, {8, 9}, {14, 15, 16}, {21, 22}, {26, 27}, and {28, 29} are blocked. Multiple studies have shown that the practice of item blocking can generate correlations between the blocked items that make them difficult to interpret [11–13].

The version of the BEMA at PhysPort [169] suggests a scoring rubric which accounts for some of the relations between the items. Item 3 is to be graded as correct if it is answered correctly based on the response to item 2 (both involve the forces on two point charges). Item 16 is to be graded as correct if it is consistent with item 14 and if the answer to item 15 is zero. Items 14 to 16 ask about the potential difference between different points in a uniform electric field. Items 28 and 29 are to be graded together; the student receives one point if both are correct, zero otherwise. By grading items 28 and 29 as a group, the total score on the instrument is reduced from 31 to 30.

The BEMA contains 5 items which are nearly identical to items on the CSEM only differing by the number of responses. Items 1, 2, and 3 are very similar to CSEM items 3, 4, and 5. These items are blocked in both instruments. BEMA items 30 and 31 are likewise similar to CSEM items 31 and 32, again differing by the number of responses.

8.1.1 Research Questions

This work is the fourth of a series of papers applying Multidimensional Item Response Theory (MIRT) to widely used physics conceptual assessments. As in the prior work, MIRT

will be applied both as an exploratory method and as a confirmatory method by constraining the MIRT models to a theoretical model developed from expert solutions.

This study seeks to answer the following research questions:

RQ1: What relations between BEMA items are identified by exploratory analyses? What do these relations imply for the interpretation of the results of applying the BEMA?

RQ2: What is the model of student knowledge measured by the BEMA identified by constrained MIRT? What insights can this model provide into the structure of the instrument?

RQ3: How is the model of the BEMA related to the models of other conceptual inventories?

8.2 Item Response Theory

Item Response Theory (IRT) represents a rich set of statistical models which describe the probability a student selects a certain response in a multiple-choice instrument. Many IRT models have been used to explore physics conceptual inventories: the Rasch model [171–173], the 2-parameter logistic (2PL) [174, 175], the 3-parameter logistic (3PL) [176, 177], nested-logit model [178], the nominal model [179], and MIRT [180, 11–13]. The statistical properties of each model are reviewed in Sec. 8.5.2.

8.2.1 Prior constrained MIRT studies

Multidimensional IRT was applied as both an exploratory and confirmatory analyses method to popular physics conceptual inventories. These studies will be referenced as Studies 1, 2, and 3 in this work.

Study 1 - FCI

The use of constrained MIRT was first introduced by Stewart and Zabriskie to examine the FCI [11]. The general structure of the instrument was investigated using correlation analysis, partial correlation analysis, and exploratory factor analysis (EFA) as exploratory methods. These analyses showed that a substantial part of the factor structure and partial correlation structure of the FCI could be explained by the practice of blocking items into item groups all referring to a common stem or where later items in the group directly referenced prior items. This study then applied MIRT as a confirmatory method constraining the parameter matrix to a theoretical model of the principles needed to solve each item. This model was developed from expert solutions. Principles are fundamental reasoning steps in the solution of the item. Constrained MIRT was then used to explore theoretically motivated modifications to the initial model to identify the model of best fit. The best-fitting model revealed that there were four groups of isomorphic items requiring very similar solution structure: items {4, 15, 16, 28}, {5, 18}, {6, 7}, and {17, 25}. These isomorphic items explained the factor structure not explained by the item blocks. The best-fitting MIRT model was far better fitting than the original model of the FCI proposed by its authors.

Study 2 - FMCE

The same methods as in Study 1 were then applied to the FMCE by Yang *et al.* [13]. The FMCE makes much heavier use of blocking than the FCI, CSEM, or BEMA with all but one item included in an item block. Correlation analysis and MIRT EFA showed that these item blocks and combinations of the item blocks explained much of the structure of

the instrument. Confirmatory MIRT was then used to develop a best-fitting model which showed the items in the item blocks were generally isomorphic. As such, it was impossible to determine if the similar solution structure or the practice of blocking resulted in these items being identified in the same factors by EFA. The confirmatory analysis was then used to show that the FMCE covered far fewer principles than the FCI and that the principles covered were used differently with the FMCE containing many items using a single principle while the FCI generally used items mixing a number of principles.

Study 3 - CSEM

Zabriskie and Stewart applied MIRT to two CSEM datasets drawn from different institutions [12]. Study 3 identified 3 isomorphic item groups: items {6, 8}, {16, 17}, and {21, 27}. These isomorphic groups were less important to the exploratory factor structure with only {21, 27} loading strongly on the same factor. This work also fit a general model of the instrument using the overall categories: mechanics, electrostatics, electric potential, magnetostatics, magnetic induction, and superposition. Like the FCI, this general model was not as well fitting as the best-fitting constrained model; however, unlike the mechanics instruments, some fit statistics suggested the general model was superior. The best-fitting models extracted for the two institutions were very similar. Model parameters for the different institutions were different, but still related. This suggests the best-fitting models extracted may have some generality.

These works have been productively employed by other studies because they produced a detailed mapping of the concepts measured by the instrument and each demonstrated the central role of the practice of blocking items in determining the factor structure of the

instrument [83, 181, 182, 179].

8.3 Prior Studies of the BEMA

The BEMA was introduced in 1997 [168] and has been used in several studies as an assessment to measure gains in electricity and magnetism conceptual knowledge [183–185]. A study conducted by Ding *et al.* [10] explored the reliability of the BEMA as an assessment tool examining the reliability of the instrument as a whole and of the individual items. The study looked at five statistics: item difficulty index (the score of each item), item discrimination index (a measure of how well an item discriminates between high-ability and low-ability students), point biserial-coefficient (a correlation between a student's score on an individual item and their score on the entire test), Kuder-Richardson reliability index (a measurement of a test's self-consistency) and Ferguson's delta (a measurement of the discrimination of an entire test). Each statistic indicated that the BEMA was a reliable instrument with sufficient discrimination between high-ability and low-ability students. A later study by Ding [186] used Rasch theory to test the construct validity of the BEMA, and found that the BEMA does measure a unidimensional construct even though the items cover a broad range of topics in electricity and magnetism.

Kohlmyer *et al.* [184] used the BEMA to test the knowledge level of students in two different introductory electricity and magnetism courses: a traditional electricity and magnetism course and the second semester of the Matter and Interactions (MI) curriculum [36]. Students enrolled in the MI course had significantly higher post-test scores than the students enrolled in the traditional electricity and magnetism course at each of the four

institutions studied.

Ding [173], using a dataset that was collected from students in parallel traditional and MI electromagnetism courses at the same institution, found five BEMA items with different averages in the two courses; two were higher in the MI course (items 5 and 7) and three were higher in the traditional electricity and magnetism course (items 17, 22, and 25). BEMA items 9 and 17 were also shown to be problematic because of low discrimination. Item 9 asks about current flow in an ionic channel and requires an answer with mathematical formula unlike most items in the instrument. Item 17 tests the electric potential in an open circuit.

A recent study by Xiao *et al.* [177] found that some conceptual instruments, including the BEMA and CSEM, could be shortened without diminishing the validity and reliability of assessment. This was done using item response theory. The latent constructs of student learning in electricity and magnetism that are measured by the BEMA were shown to be measured with similar reliability by a shortened BEMA assessment.

8.3.1 Studies comparing the BEMA and the CSEM

Xiao *et al.* [177] also showed that student scores on the BEMA and CSEM can be compared after linking the assessment scales and appropriately transforming them. This supports prior work done by Pollock [187]. Pollock compared the CSEM and BEMA and found them to be fairly equally effective in assessing conceptual understanding [187]; however, the instruments have somewhat different coverage. Eaton *et al.* [188] used item response theory (IRT) and classical test theory (CTT) on the BEMA and CSEM to show that the assessments were nearly equal in overall difficulty. Any differences found between the two tests were minimal and potentially caused by differences in the test samples. The circuit

questions on the BEMA were poorly correlated with other concept areas on the assessment.

Some of the differences in coverage of the BEMA and CSEM were evident in an EFA comparing the BEMA and the CSEM by Eaton *et al.* [189]. They concluded the two instruments cover nearly the same conceptual content, with the exception of a few factors. The CSEM had an EFA model of six factors while the BEMA had a five-factor model. This study did not use all BEMA items, removing several items due to low Kaiser-Meyer-Olkin (KMO) test values which measure how well a sample loads onto different factors.

8.4 The Structure of Knowledge

The current work built a detailed model of the BEMA involving 50 principles of electromagnetic theory. This model shares many features with earlier models of physics problem solving constructed using the paradigm of cognitive research introduced by Simon and Newell [190]. This paradigm dominated research into problem solving for 30 years and is reviewed by Ohlsson [191]. The paradigm built exceptionally detailed, computationally functional models of the problem solving process. These models could then be run on computers to reproduce the problem solving sequence of participants. The technique was used to understand expert-novice differences in problem solving in physics and many other fields [192, 193]. This paradigm ultimately lost favor because it was difficult to explore general features of complex problem solving; however, in Physics Education Research (PER) it is often the goal to understand specific features of the physics problem solving process. As such, the detailed models produced by this method may be productive. Reif and Heller also produced a fine-grained model of physics problem solving, but this model did not meet the test of

being computationally functional [194]. These models involved identifying the fundamental transformations, called principles, needed to navigate the problem space. These principles are closely related to the principles identified in the theoretical MIRT model; the MIRT principles take their name from these earlier works.

8.5 Methods

8.5.1 Sample

The sample for this study was collected at a large western land-grant university in the United States serving 34,000 students. Fifty percent of the undergraduate student population had ACT scores in the range 25 to 30. The demographic composition of the general undergraduate population was 67% White, 12% Hispanic, 6% Asian, 6% two or more races, 6% International, 2% Black with other races less than 1% [195].

The aggregate dataset was drawn from 22 semesters of an introductory, calculus-based, electricity and magnetism class. It contains 9666 BEMA post-test records. Any record that contained one or more missing responses was removed, as well as records that had suspicious response patterns, e.g., "A" repeated or "ABCDE" repeated.

8.5.2 Item Response Theory

Item response theory (IRT) encompasses a broad collection of statistical models of the response patterns to multiple-choice instruments. These models estimate the probability of either selecting the correct response or each response in terms of a latent student-level trait called the ability. This latent trait represents the general facility of each student with the

material tested by the instrument. Unidimensional IRT, estimating a single latent ability, has been used in many PER studies of the FCI, FMCE, and CSEM [176, 171, 196–198, 174, 199–201]. These studies are summarized in detail for the individual instruments in Studies 1 to 3.

Multidimensional IRT (MIRT) is a generalization of unidimensional IRT which estimates multiple latent abilities for each student. It was used as both an exploratory and confirmatory method in Studies 1 to 3. MIRT was also used by Scott and Schumayer [180] to perform an exploratory factor analysis of the FCI. MIRT provided similar, but not identical, results to an earlier work on the same dataset using traditional factor analysis [202].

An exploratory analysis allows the model to be deduced from the data without the input of a theoretical model. A confirmatory analysis begins with a theoretical model and seeks to determine how well a set of data is described by the model. Studies 1 to 3 and 50 years of social science research [203, 204] argue that purely exploratory analyses are susceptible to misinterpreting random fluctuations in the data as real effects.

MIRT estimates the probability π_{ij} that student i will answer correctly on item j. For each item, MIRT estimates a parameter d_j related to the overall difficulty of the problem. Items with larger d_j are answered correctly more often. More difficult problems have smaller d_j, easier problems larger d_j. MIRT also estimates K discrimination parameters a_{jk} for each item and K ability traits θ_{ik} for each student. The discrimination and the ability can be written as K element vectors, $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_i$. The MIRT probability model is shown in Eqn. 8.1.

$$\pi_{ij} = \frac{\exp[\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i + d_j]}{1 + \exp[\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i + d_j]}, \tag{8.1}$$

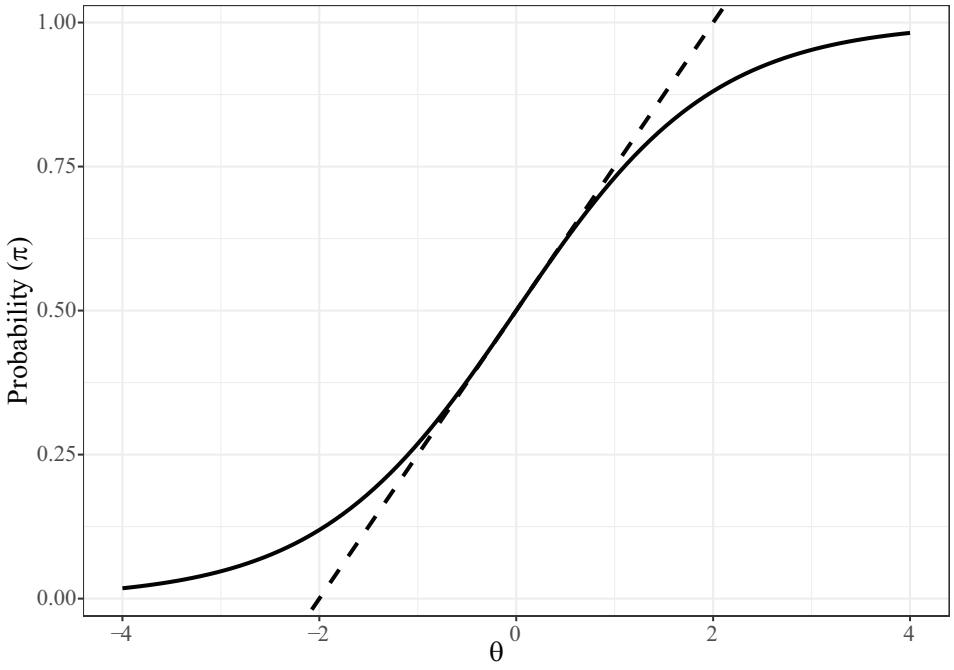


Figure 8.1: Probability of selecting the correct response, $\pi(\theta)$, versus ability θ using $d = 0$ and $a = 1$. The dashed line represents the slope at $\theta = 0$ and has slope $a/4 = 0.25$.

Some qualitative understanding of the features of the probability function are helpful when interpreting the MIRT models. For this discussion, consider a model with one discrimination parameter ($K = 1$). If $a > 0$, the probability curve has the characteristic S-shape shown in Fig. 8.1. The figure shows the probability curve drawn with $d = 0$ and $a = 1$. With this choice of parameters, the probability of answering correctly is 0.5 at $\theta = 0$. In general, the $\theta_{1/2}$ where the probability is 0.5 occurs when the argument of the exponential is zero, $\theta_{1/2} = -d_j/a_j$; therefore, a combination of a_j and d_j determine the ability at which a student has a 50% chance of answering the problem correctly. The slope of the probability at $\theta_{1/2}$ is $a_j/4$; therefore, the discrimination a_j is related to how fast the probability is increasing when the students have a 50% chance of answering correctly. If a_j is larger, the transition from low probability to high probability is faster, the item discriminates between low and high ability students more strongly. If $a_j = 0$, the probability curve is flat, low and high ability

students have equal chances of answering correctly, a characteristic of a problematic item. More problematic are items with $a_j < 0$; for these items the S curve inverts and students with low ability have a higher probability of getting the item correct than students with high ability.

The MIRT models were fit using the "mirt" package [205] which is part of the R software system [120]. Models were fit using the Metropolis-Hastings Robbins-Monro (MHRM) algorithm [206] which uses stochastic methods to maximize the likelihood function. Maximum likelihood estimation does not require the assumption of an underlying normal distribution.

8.5.3 Model Fit Statistics

The parameters in a MIRT model are estimated using maximum likelihood (ML) methods where the parameters are selected to make the observed response pattern the most probable using Eqn. 8.1. Maximum likelihood methods calculate the likelihood function, L, the probability the observed response pattern occurred given the MIRT probability model and a set of parameters. The parameters are modified until L is maximized. A broad collection of model fit statistics have been developed to characterize and compare ML models. This study reports Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). These statistics are explored in detail in Study 3 and are summarized below.

AIC (Eqn. 8.2) and BIC (Eqn. 8.3) measure the relative information lost between the model fit and the true model; better fitting models lose less information and thus minimize AIC and BIC. Both penalize for the addition of parameters with BIC penalizing additional

parameters more strongly.

$$AIC = 2k - 2\ln(L), \tag{8.2}$$

$$BIC = k\ln(n) - 2\ln(L), \tag{8.3}$$

where k is the number of parameters estimated and n is the sample size. Both AIC and BIC depend on the logrithm of the likelihood, so small changes in either measure large changes in likelihood. Raftery provided criteria for the effect size of differences in BIC: $\Delta BIC \leq 2$ as "weak," $2 < \Delta BIC \leq 6$ as "positive," $6 < \Delta BIC \leq 10$ as "strong," and $\Delta BIC > 10$ as "very strong" [207]. The definition of AIC and BIC are very similar; therefore, this work also adopts Raftery's convention for AIC.

RMSEA, CFI, and TLI are measures of model fit or misfit derived from the chi-squared (χ^2) statistic. For a N-item dichotomously scored instrument, there are $C = 2^N$ possible response sequences. To calculate chi-squared, the probability of each possible response sequence, P_c, is compared to the observed frequency of the sequence, O_c, $\chi^2 = n\sum_0^C (O_c - P_c)/O_c$ where n is the number of observations. For the BEMA with $N = 31$ items and for most multiple-choice instruments of reasonable length, it would require an enormous amount of data to estimate χ^2 accurately. As such, MIRT uses an approximation to χ^2 called M_2 to approximate χ^2 [208, 209].

RMSEA (Eqn. 8.4) characterizes badness of model fit on a scale of 0 to 1 using χ^2 normalized by the number of degrees of freedom (df) [210]; models with larger RMSEA represent worse fitting models. RMSEA less than 0.05 represents good model fit; RMSEA above 0.10 represents poor model fit [211].

$$RMSEA = \sqrt{\frac{(\chi^2/df) - 1}{n - 1}} \tag{8.4}$$

CFI (Eqn.8.5) and TLI are incremental goodness-of-fit statistics which characterize how much the model differs from a null model [210]. The null model used by MIRT constrains the discrimination matrix to zero, $\vec{a}_j = 0$, and fits the model containing only d_j. CFI and TLI values above 0.95 represent good model fit [212].

$$CFI = 1 - \frac{\chi^2 - df}{\chi^2_{null} - df_{null}} \tag{8.5}$$

The equation for TLI contains a slightly modified combination of the null and fitted models.

Hu and Bentler recommend using multiple fit statistics to compare models [212]. As such, a superior model has AIC and BIC at least 20 lower than other models, RMSEA near zero, and CFI and TLI near one.

The relation of RMSEA, CFI, and TLI to the number of parameters fit is complicated. All three statistics involve the ratio of an effective chi-squared statistic to the number of degrees of freedom. As more parameters are fit, generally χ^2 decreases, but the number of degrees of freedom also decreases. Eventually, the decrease in χ^2 is not enough to compensate for the decrease in the degrees of freedom and the statistics begin to increase as the models become sufficiently complex.

8.6 Results

The BEMA was first examined with two exploratory analyses: correlation analysis and exploratory factor analysis. The instrument was then examined with a confirmatory analysis fitting a model based on expert solutions to the instrument.

8.6.1 Exploratory Analyses

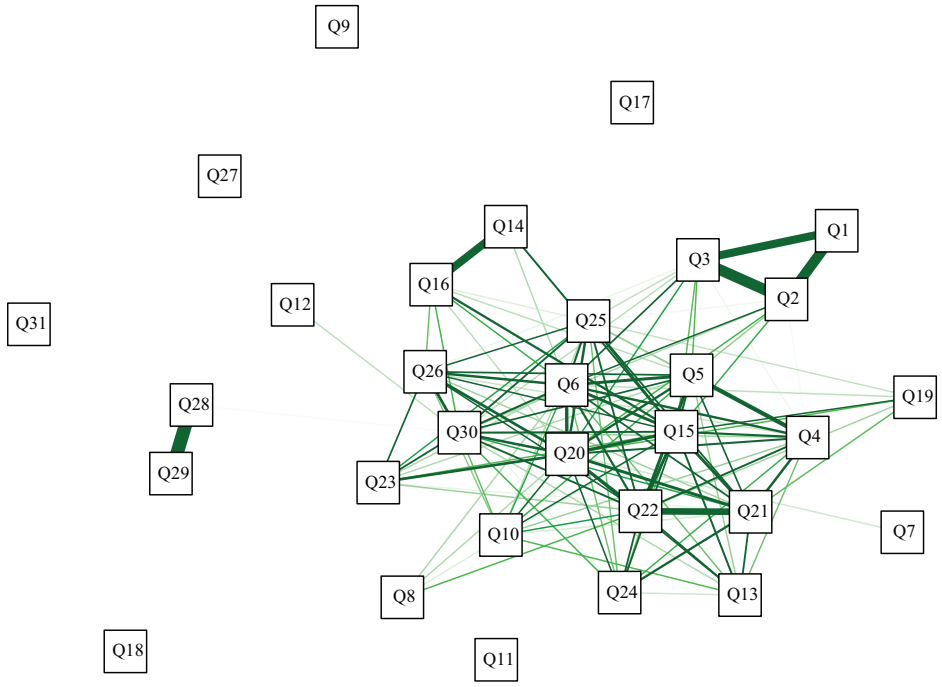


Figure 8.2: Correlation matrix. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations. Thicker lines represent larger correlations.

The BEMA was first examined using the the correlation and partial correlation matrices. The correlation matrix is presented in the Fig. 8.2. The partial correlation matrix, which corrects for correlations resulting from overall BEMA scores, is shown in Fig. 8.3. The partial correlation matrix shows five groups of items that are substantially positively

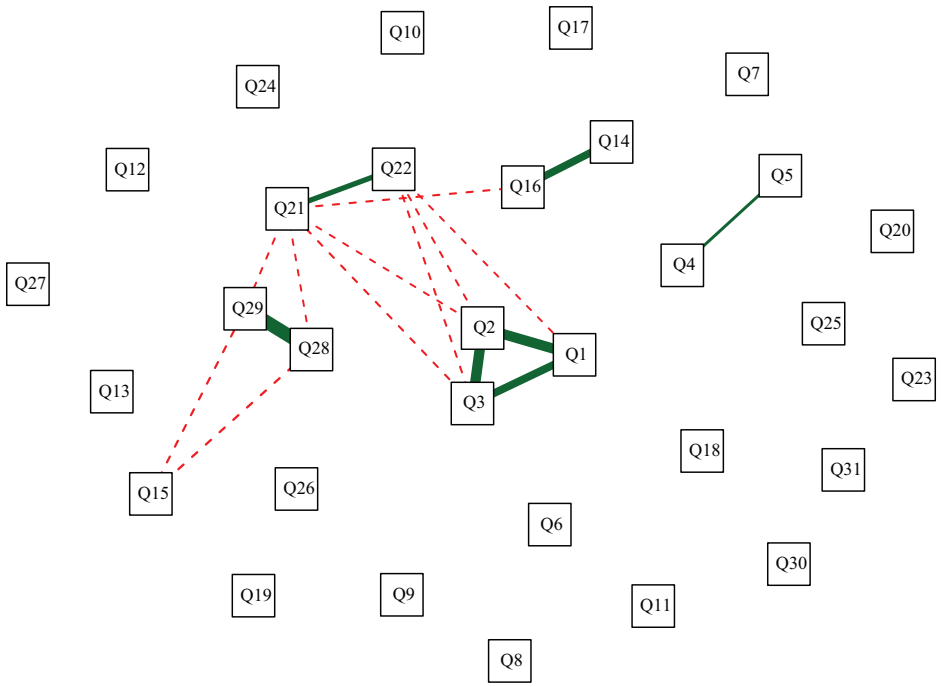


Figure 8.3: Partial correlation matrix. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations. Thicker lines represent larger correlations.

correlated after correcting for overall BEMA score: {1, 2, 3}, {4, 5}, {14, 16}, {21, 22}, and {28, 29}. All groups are part of item blocks. Item 15 is not present in the {14, 16} group. This group of items asks about the potential difference between various points in a uniform electric field. Item 15 asks about the potential difference along an equipotential unlike the other two items.

An EFA was performed using MIRT. To use MIRT as an exploratory method, the discrimination matrix $\boldsymbol{a}_j$ is allowed to vary freely. MIRT was used to extract from 1 to 10 factors; the fit statistics for each model are shown in Table 8.1. Consistent with Studies 1 to 3, the fit statistics do not clearly identify a single best-fitting model. The 6-factor model minimizes AIC and BIC while the 5-factor model minimizes RMSEA, and maximizes CFI

Factors	AIC	BIC	RMSEA	TLI	CFI
1	332,570	333,015	0.07	0.80	0.80
2	325,941	326,601	0.05	0.89	0.91
3	321,556	320,413	0.02	0.95	0.96
4	319,343	320,412	0.02	0.98	0.98
5	318,895	320,158	0.02	0.99	0.99
6	318,657	320,107	0.06	0.85	0.90
7	318,667	320,296	0.06	0.84	0.91
8	318,872	320,674	0.06	0.85	0.92
9	318,694	320,713	0.06	0.84	0.92
10	318,694	320,818	0.06	0.85	0.93

Table 8.1: MIRT fit statistics for an Exploratory Factor Analysis of the BEMA.

and TLI. The relatively poor RMSEA, CFI, and TLI of the 6-factor model strongly indicates the 5-factor model is the superior model. The factor structure of the 5-factor model is shown in Table 8.2. As in prior studies, the factor structure is dominated by the blocked items; block items form the highest loadings on factors 1 to 4. This is consistent with the correlation analysis which shows only blocked items are more correlated with each other than with the total instrument score. This supports the work of Eaton *et al.* [189] who also reported a 5-factor model as optimal.

8.6.2 Confirmatory Analyses

For a confirmatory analysis, one first develops a theoretical model and then determines how well the data fit the model. One can also propose a small number of theoretically motivated modifications to the model. Ideally, these should be proposed before the model is initially fit. By constraining the analysis to a theoretical model, confirmatory methods are less likely to erroneously identify structure resulting from random effects.

BEMA Item	FC1	FC2	FC3	FC4	FC5
1		0.92			
2		0.97			
3		0.90			
4	0.46				
5	0.44				
6	0.36				0.38
7					
8					
9					
10					0.31
11					
12					
13	0.40				
14				0.87	
15	0.44			0.40	0.36
16				0.88	
17					
18					
19	0.31				
20	0.44				0.43
21	0.89				
22	0.79				
23					0.35
24	0.40				
25	0.34				0.34
26	0.31				0.42
27					
28			0.97		
29			0.98		
30	0.39				0.31
31					

Table 8.2: Factor structure for the five-factor model. Only loadings greater than 0.3 are shown. The factors are labeled FC1 to FC5.

Label	Derived From	CSEM Principle	BEMA#	Principle
			Mechanics	
L1			26, 27	Newton's 1st law.
L2		×		Newton's 2nd law.
L3		×	2(2)	Newton's 3rd law.
C1	L2	×	6, 23	If a particle is turning in some direction, there is a force in that direction.
			Electrostatics	
L4		×	1, 2(1), 3, 7(2)	Coulomb's law for the electric force ($\vec{F} = \frac{kq_1q_2}{r^2}\hat{r}$).
L5		×	4(1), 5(1)	Coulomb's law for the electric field ($\vec{E} = \frac{kq}{r^2}\hat{r}$).
LM1	L4	×	4(1), 5(1)	Opposite charges attract/likes repel.
DF1		×	4(1), 5(1), 6, 26, 27	Definition electric field ($\vec{F} = q\vec{E}$)
LM2	L5		4(1), 5(1), 7(2)	Electric field weakens as distance increases.
C2			4(2), 5(2)	Electric dipole field shape.
C3	F1, LM2		7(1)	Charged object attracts a neutral object.
F1			7(2)	An insulator polarizes in an external field.
L6			18	Gauss's law ($\oint_S \vec{E} \cdot \hat{n} dA = \frac{Q}{\epsilon_0}$).
DF2			18	Definition of electric flux ($\Phi = \int_S \vec{E} \cdot \hat{n} dA$).
F2			19	Electric field is zero in a conductor.
			Electric Potential	
DF3		×	19	Definition of electric potential ($\Delta V = \frac{W_{ext}}{q} = -\int E dx$).
LM3	DF3	×	14, 16	Electric field points to lower potential.
LM4	DF3		14, 16	Potential difference in uniform field is ($\lvert\Delta V\rvert = \lvert Ed\rvert$).
LM5	DF3		15, 16	Potential difference is zero perpendicular to the field.
LM6	DF3		16	Total potential difference is the sum of ΔV over paths.
			Magnetostatics	
L7		×	24(2), 25(2)	Biot-Savart law ($d\vec{B} = \frac{\mu_0}{4\pi}\frac{Id\vec{\ell}\times\hat{r}}{r^2}$).
L15				Ampere's law ($\oint \vec{B} \cdot \vec{dl} = \mu_0 I$).
LM7	L15		31(1)	Magnetic field is proportional to current.
L8		×	23, 25(2), 25(3), 26, 27, 30	Lorentz force ($\vec{F} = q\vec{v} \times \vec{B}$ or $d\vec{F} = Id\vec{\ell} \times \vec{B}$).
LM8	L8	×	20	The magnetic force on a stationary charge is zero.
LM9	L7, L8, DF4	×	25(1)	Like currents attract/opposites repel.
F3			21, 22, 24(3)	Magnetic dipole field shape.
DF4		×	23, 24(2), 26, 27, 30	Right-hand rule for the cross-product.
DF5		×	27	Magnitude of the cross product ($\lvert\vec{A} \times \vec{B}\rvert = \lvert\vec{A}\rvert\lvert\vec{B}\rvert \sin\theta$).
C5	L7, DF4		24(1), 25(3)	Right-hand rule for a wire.
DF6			24(3)	Right-hand rule for a magnetic moment.
C6	L8, DF4		30	Conductor moving in a magnetic field experiences a potential difference.
			Induction	
L9		×	28, 29, 31(1)	Faraday's law ($emf = -\frac{d\Phi}{dt}$).
DF7		×	28, 29, 31(1)	Definition of magnetic flux ($\Phi = \int_S \vec{B} \cdot \hat{n} dA$).
L10			28, 29	Lenz' law.
C7	L9, L10		28, 29	Right hand rule for changing flux.
C8	L9, L10		31(2)	Mutual inductance ($emf = -M\frac{d\Phi}{dt}$).
			Superposition	
L11		×	4(1), 5(1), 24	Electric and magnetic fields add as vectors.

Table 8.3: Theoretical model tested by the BEMA. An × indicates that the principle is used in the CSEM.

Theoretical Framework

Label	Derived From	CSEM Principle	BEMA#	Principle
				Electric Circuits
F4			8, 9	Battery produces current flowing from - terminal to + terminal.
DF8			9	Positive current is in the direction of flow of positive charge or opposite the direction of flow of negative charge.
F5			10	Ammeters have negligible resistance.
C9			10	Current same in series.
F6			11	Brighter light bulb indicates more current.
L12			10, 11, 17	Ohm's law ($\Delta V = IR$).
C10			11	Parallel elements have the same potential difference.
C11			11	Resistance adds for resistors in series.
L13			12	Ohm's law for the electric field ($\vec{J} = \sigma \vec{E}$).
F7			17	Complete circuit required for current flow.
L14			17	Kirchhoff's Loop Rule.
DF9			13	Definition of capacitance ($C = \frac{Q}{\Delta V}$).
F8			13	RC circuits decay.

Table 8.4: A continuation of Table 8.3

A model of the knowledge structure measured by the BEMA is shown in Table 8.3. This model was developed in the same way that models of knowledge structure were developed in Studies 1, 2, and 3. Content experts including members of the research team and instructors of introductory, calculus-based physics courses at the institution where the analysis was performed were asked to complete the BEMA and write the reasoning used to solve each problem. These responses were decomposed to the sentence or phrase level. Sentences and phrases representing the same fundamental reasoning process were grouped; these groups were called "principles." As in Studies 1 to 3, the principles were then classified as a laws (L) representing physical laws such as Gauss' Law, definitions (DF) introducing a new quantity, and facts (F) representing physical knowledge that was not as general as a law. From these primary principles, which define the core physical knowledge tested by the instrument, secondary principles were derived. The secondary principles included corollaries (C) and lemmas (LM). Corollaries are important secondary results of the laws, definitions,

and facts. Qualitative statements in the solutions that interpreted laws, definitions, and corollaries were called lemmas. Some secondary principles were derived from primary principles that were not included in the expert solutions. These principles were inferred and included in the model in Table 8.3. Some of the principles in Table 8.3 are characterized with a bold font. These principles are those that are retained in the best-fitting principle model (M13) found through constrained MIRT. Finally, broad subtopics were introduced; mechanics, superposition, electrostatics, electric potential, magnetostatics, magnetic induction, and electric circuits.

Several principles in the electric circuits subcategory are secondary principles that could be derived from a primary principle. For example, C9 (current is the same throughout a series circuit) is derived from the law of conservation of charge. None of the expert solutions used these primary principles and it seemed unlikely that a student would use the primary principle. Such principles were not included in Table 8.3 and were not explored in the MIRT analysis.

Some BEMA items had multiple expert solution paths including items 4, 5, 7, 24, 25, and 31 (item 2 also had a secondary solution path, but was eliminated from the analysis because of blocking). In Table 8.3, the principles necessary for secondary and tertiary solutions paths are presented in parentheses with the solution path number within the parenthesis. For example, the first solution path of item 4 uses L5, L11, LM1, DF1, and LM2; the second solution path uses only C2. These different solution paths were explored using MIRT to determine which solution path was the most important to model student thinking.

Some principles were always used together. Borrowing the terminology of factor analysis, when a principle is used in an item it is said to "load" onto that item. In the BEMA,

Faraday's law and the definition of magnetic flux are used together to solve items 28, 29, and 31, but are not used in other items. As such, MIRT cannot resolve them as separate principles. In the MIRT analysis, these principles were combined as a single principle labelled L9-DF7. Similarly, Gauss' law (L6) and the definition of electric flux (DF2) load together on item 18 as L6-DF2. Many other combined principles are shown in Table 8.3.

Not all items were retained in the analysis. There are several problem blocks where multiple problems refer to a common physical system or refer to the same image. Studies 1 to 3 showed that blocked items can exhibit correlations unrelated to the physical reasoning needed to solve the item. Each item block was examined to determine if the items in the block were fairly independent. Items 2 and 3 depend on the response to item 1 and were removed from the analysis. Responses to items 4 and 5 do not depend on each other and were retained. Item 16 depends on the responses from items 14 and 15 and was removed. The responses to items 21 and 22 are fairly independent and were retained. Item 27 depends on the response from item 26 and was removed. The responses to items 28 and 29 are independent and were retained. The removed items still appear in Table 8.3 but were not included in the analysis.

Table 8.3 also indicates whether the principle was also tested by the CSEM to facilitate the comparison of the two instruments.

Model Transformation Plan

In a confirmatory analysis, the theoretical model is fit, then a series of theoretically motivated model transformations are performed to possibly improve the initial model fit. Table 8.3 represents the initial model and was fit first. Items 4, 5, 7, 24, 25, and 31 have

multiple solutions paths. For each of these items, the first solution path, shown in Table 8.3 with a "(1)," was fit in the initial model. For example, principle L4 is used in solution path 1 of item 2 which is shown as "2(1)" in Table 8.3. The second solution path for each item was then fit, followed by the third solution paths for items 24 and 25. These model fits were compared with the original model and any model that was an improvement was retained.

The granularity of student knowledge was then explored to determine if the secondary principles were needed to understand student thinking. Models were constructed which removed the secondary principles– the lemmas (LM) and corollaries (C)– by replacing them with the primary principles from which they were derived. For example, LM1 (opposites attract/likes repel) can be derived from L4 (Coulomb's force law). To test whether LM1 was needed in addition to L4, all items that were set to load on LM1 in the initial model were set to load on L4 in the transformed model. This model was fit and fit statistics compared with the original model. This process was called "collapsing" LM1 into L4. Seven models were transformed in this way; C1 was collapsed into L1; LM1 was collapsed into L4; LM2 was collapsed into L5; LM3, LM4, LM5, and LM6 were collapsed into DF3; and LM8 was collapsed in L8. In the case of LM3, LM4, LM5, and LM6, any item loading onto one of the principles was set to load onto DF3. Two additional models were constructed which required a somewhat more complex transformation. The other two models were slightly more complex in that the secondary principle was not derived from a single primary principle but rather from several primary principles. In M15, LM9 was set to load onto L10, DF7, and DF4; in M16, C7 was set to load onto L9, L10, DF7, DF4. These models were fit; models with improved fit were retained.

The definition of electric potential (DF3) loads onto item 19 with fact F2 (the electric

field inside a conductor is zero). These two principles were not combined as a single principle in the MIRT analysis because during the third set of transformations LM3, LM4, LM5, and LM6 were all collapsed into DF3. Because of this transformation, F2 and DF3 no longer load on an item exclusively together, and so to maintain a nested model sequence they were not combined in the initial model.

The final set of transformations collapsed the principles and items onto the general topics of electricity and magnetism which form the divisions in Table 8.3. These models are called "topical models." The first transformation, M17, collapsed each principle onto the general topics of electricity and magnetism: mechanics, electrostatics, electric potential, magnetostatics, magnetic induction, superposition, and electric circuits. To form M18, the principles involving mechanics and superposition were removed so as to include only topics specific to electricity and magnetism. This resulted in each item loading onto a single topic with the exception of item 26 which loaded onto both electrostatics and magnetostatics (a constant, uniform electric field and a constant, uniform magnetic field are both acting on a charged particle).

Constrained MIRT

In the exploratory work in Sec. 8.6.1, the MIRT discrimination matrix $\mathbf{a_j}$ was allowed to take on any value. To apply MIRT as a confirmatory method, elements of the discrimination matrix which can not theoretically be involved in solving an item are constrained to zero. For example, L1 (Newton's 1st law) is only used in items 26 and 27; $a_{j,L1}$ was only be allowed to be non-zero for items 26 and 27. This constraint means that abilities associated with the application of principles, not theoretically required for the solution of the item, do

not influence the probability of answering the item correctly. In this way, the theoretical model in Table 8.3 is mapped onto the MIRT discrimination matrix. This analysis proceeds with the 27-item instrument removing some of the blocked items. The reduced instrument contains items: 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 12, 13, 14, 25, 26, 28, 29, 30, and 31.

The transformation plan was carried out in panels: first testing each type of transformation independently to identify those which improved model fit, then combinations of the transformations which improved model fit were investigated. The results of carrying out the transformation plan are shown in Table 8.5. The "Transformed Model" column shows the model number of the model after the transformation has been applied to a prior model (the "Original Model"); M0 is the initial model which implements the model in Table 8.3. The "Transformation" column summarizes the transformation applied to the original model to form the transformed model. The fit statistics are then presented for the transformed model and the best fitting of the two models identified by the fit statistics and indicated in the "Superior Model" column. The first set of transformations tried alternate solution paths identified in the expert solutions. Many of these alternate solutions produced superior models. For item 24, both solution paths 2 and 3 improved model fit with M4 producing superior fit; of these two transformed models, only M4 was used in further models. The next set of transformations tried combinations of the transformations in the first stage that produced superior models. All these combinations failed to improve on model M4. As such, the only modification to the initial model M0 that was retained was using the 3rd solution path to item 24. Item 24 asks about the magnetic field at the center of two parallel loops of wire; solution path 3 solved the item using the dipole moments of the loops.

Transformed Model	Transformation	Original Model	AIC	BIC	RMSEA	TLI	CFI	Superior Model
			Full Model					
M0		-	289,783	290,536	0.016	0.987	0.990	-
			Explore Alternate Solution Paths					
M1	Solution path 2 for items 4, 5	M0	289,793	290,496	0.016	0.987	0.990	M0
M2	Solution path 2 for item 7	M0	289,766	290,533	0.016	0.987	0.990	M2
M3	Solution path 2 for item 24	M0	289,767	290,528	0.016	0.987	0.990	M3
M4	Solution path 3 for item 24	M0	289,727	290,488	0.016	0.988	0.991	M4
M5	Solution path 2 for item 25	M0	289,877	290,645	0.016	0.988	0.990	M0
M6	Solution path 3 for item 25	M0	289,752	290,513	0.015	0.988	0.991	M6
M7	Solution path 2 for item 31	M0	289,758	290,505	0.015	0.987	0.990	M7
			Combine Alternate Solution Path Models					
M8	Combine M2 and M4	M4	289,721	290,496	0.016	0.988	0.991	M4
M9	Combine M6 and M4	M4	289,860	290,628	0.016	0.988	0.990	M4
M10	Combine M7 and M4	M4	289,759	290,512	0.016	0.988	0.990	M4
			Collapse Lemma into Primary Principles					
M11	Combine LM1 with L4	M4	289,746	290,507	0.016	0.988	0.990	M4
M12	Combine LM2 with L5	M4	289,732	290,486	0.016	0.988	0.990	M4
M13	Combine LM3, LM4, LM5, LM6 to DF3	M4	289,684	290,445	0.015	0.990	0.992	M13
M14	Combine LM8 with L8	M4	289,739	290,492	0.016	0.987	0.990	M4
M15	Combine LM9 with L7, L8, DF4	M4	289,826	290,603	0.016	0.988	0.991	M4
M16	Combine C7 with L9, L10, DF7,DF4	M4	289,725	290,500	0.016	0.988	0.991	M4
			Topical Models					
M17	Collapse all principles into main topics	M13	289,471	290,117	0.015	0.989	0.991	M17
M18	Collapse all items into main topics	M17	289,435	290,024	0.016	0.989	0.990	M18

Table 8.5: Model transformation table. Each entry presents the result of modifying a prior model (the original model) with one of the planned transformations to produce a modified model (the transformed model). These two models are compared and the model with superior fit statistics identified (the superior model).

The next set of transformations investigated whether the lemmas were required to model student thinking. Lemmas are qualitative principles derived from the generally quantitative laws and definitions. Studies 1 to 3 found lemmas were retained in the best-fitting model to a varying degree. In Study 1 of the FCI, all lemmas were removed together which improved model fit. In Studies 2 and 3, the lemmas were removed individually and some lemmas were retained in the best-fitting model. Lemmas play an important role in electromagnetism with LM1 (opposites attract/like repel) often used as part of the solution to both qualitative and quantitative problems. Only M13 was superior to M4; M13 combined lemmas LM3, LM4, LM5, and LM6 into DF3. DF3 is the definition of electric potential and

the lemmas are different properties of electric potential such as the electric field points to lower potential (LM3) or the potential difference is zero perpendicular to the field (LM5). As such, most lemmas were important to modeling student reasoning about electromagnetism with only lemmas involving electric potential removed from the optimal model.

The final set of transformation tested more general models of student thinking using the general electromagnetic topics which form the divisions in Table 8.3. In M17, the principles were set to load on the subtopic containing the principle; as such, some items loaded onto an electromagnetism subtopic and also onto the topics of mechanics and superposition. In M18, the subtopics of mechanics and superposition were removed and items were only loaded onto the electromagnetism subtopics. The fit statistics of M18 were superior to all other models. This result was diametrically opposite to that of Studies 1 to 3 where the MIRT models involving the individual principles were superior to models using general topics. Possible reasons for this difference are explored as part of RQ3.

With M18 having superior fit statistics over more detailed models, one might consider whether the instrument is simply unidimensional with no substructure. A model containing only the a_0 discrimination is equivalent to the 1-factor model in Table 8.1. Comparison of the 1-factor model and M18 shows that M18 is a substantially superior model with consistently superior fit statistics.

As such, M13 was the best-fitting model of student reasoning based on the granular model in Table 8.3 involving reasoning principles. The principles contained in this model are bolded in the table. M18 was the overall best fitting model. We note both M13 and M18 have exceptional RMSEA, CFI, and TLI, and as such, both can provide useful insights into student thinking. The difference in AIC and BIC likely results from the penalty these

statistics place on the addition of parameters. The structure of M13 is further explored in Sec. 8.6.4, the structure of M18 in Sec. 8.6.3.

8.6.3 Topical Model

	Electrostatics	Electric Circuits	Electric Potential	Magnetostatics	Magnetic Induction
Items	1, 4, 5, 6, 7, 26	8, 9, 10, 11, 12, 13, 17	14, 15, 19	20, 21, 22, 23, 24, 25, 26, 30	28, 29, 31
Mean ± SD	0.59 ± 0.24	0.66 ± 0.30	0.48 ± 0.21	0.57 ± 0.27	0.22 ± 0.30
Cronbach's α	0.51	0.38	0.37	0.69	0.54

Table 8.6: Subscale scores for each topic. The mean ± the standard deviation (SD) are shown. The mean calculates the average fraction of item in the subscale answered correctly by the students

The overall best-fitting model, M18, involved only the general electromagnetic topics which suggests these topics may be used as subscales, coherent measures of the topic. This model was called the "topical model." Table 8.6 shows the average fraction of students answering the items in the general topical groups (subscales) correctly. The electrostatics, magnetostatics, electric potential, and electric circuits subscales all have fairly similar averages differing by a maximum of 18%. Table 8.7 presents the item-level score (fraction of students answering the item correctly), d_j, general discrimination a_{j0}, and subscale discrimination a^s_{jk}, where j indexes the item and k the subscale. Items within these subscales have a broad range of average item scores. The parameter d_j is related to the probability of answering the item correctly for students of average ability ($\vec{\theta} = 0$); items with larger d_j are answered correctly with higher probability by average students of average ability. The magnetic induction subscale has a much lower average score than the other subscales. It also contains two items, 28 and 29, with the lowest average score and the most negative d_j of any items in the instrument. Items with negative d_j are answered correctly less often

than an item with average d_j. In general, the range of discriminations a_{j0} was more narrow

Item #	Item Score	Principle Model (M13) Principles	a_{j0}	d_j	Topical Model (M18) Topic	a^s_{jk}	a_{j0}	d_j
1	0.83	L4(0.19)	0.64	1.87	electrostatics	0.08	0.58	1.70
4	0.76	DF1(0.21) LM1(0.24) LM2(0.11) L5(0.26) L11(0.19)	1.35	2.00	electrostatics	0.79	1.35	1.97
5	0.54	DF1(0.21) LM1(0.21) L5(0.37) L11(0.12)	1.37	0.28	electrostatics	0.58	1.26	0.24
6	0.57	C1(0.10) DF1(0.28)	1.36	0.44	electrostatics	0.15	1.24	0.38
7	0.49	C3(0.18)	0.72	−0.03	electrostatics	−0.01	0.53	−0.04
8	0.76	F4(0.08)	0.72	1.32	electric circuits	0.01	0.70	1.28
9	0.30	F4(0.10) DF8(0.19)	0.13	−0.95	electric circuits	0.04	0.11	−0.85
10	0.60	F5-C9(0.18) L12(0.22)	0.97	0.55	electric circuits	0.19	0.87	0.50
11	0.43	L12(0.16) F6-C10-C11(0.22)	0.60	−0.35	electric circuits	0.10	0.53	−0.31
12	0.21	L13(0.18)	0.75	−1.62	electric circuits	0.09	0.69	−1.51
13	0.75	DF9-F8(0.15)	1.02	1.48	electric circuits	0.10	0.93	1.38
14	0.45	DF3(0.28)	0.62	−0.22	electric potential	0.29	0.63	−0.23
15	0.75	DF3(0.50)	1.80	1.99	electric potential	0.48	1.77	1.94
17	0.34	L12(0.07) L14-F7(0.17)	0.36	−0.76	electric circuits	0.10	0.33	−0.71
18	0.55	L6-DF2(0.33)	0.20	0.23	electrostatics	0.01	0.18	0.20
19	0.76	F2(0.15) DF3(0.08)	0.77	1.43	electric potential	0.07	0.73	1.34
20	0.58	LM8(0.28)	1.66	0.52	magnetostatics	0.04	1.45	0.45
21	0.84	F3(1.43)	2.55	4.25	magnetostatics	1.71	2.82	4.77
22	0.66	F3(0.94)	1.76	1.25	magnetostatics	0.84	1.64	1.17
23	0.49	L8(0.09) DF4(0.08) C1(0.04)	0.83	−0.04	magnetostatics	−0.06	0.85	−0.05
24	0.68	F3(0.11)L11(0.06) DF6(0.11)	0.92	0.94	magnetostatics	0.11	0.87	0.89
25	0.56	LM9(0.13)	1.04	0.31	magnetostatics	0.04	0.97	0.29
26	0.39	DF1(0.18) L8(0.13) DF4(0.16) L1(0.20)	1.23	−0.67	electrostatics magnetostatics	−0.01 −0.06	1.13	−0.59
28	0.18	L9-DF7(2.14) C7-C8(3.21)	2.25	−7.51	magnetic induction	5.53	2.44	−7.95
29	0.18	L9-DF7(2.14) C7-C8(3.82)	2.63	−9.24	magnetic induction	5.76	2.40	−8.22
30	0.40	L8(0.18) DF4(0.11) C6(0.18)	1.31	−0.62	magnetostatics	−0.01	1.16	−0.55
31	0.30	L9-DF7(0.07) LM7(0.12)	0.29	−0.92	magnetic induction	(0.06)	0.29	−0.87

Table 8.7: Best-fitting principle and topical MIRT models. The first column shows the item number (#). Not all items of the BEMA were modelled. The discrimination for principle k on item j, a_{jk}, is given by the number in parentheses following the principle label. The overall discrimination of item j on a knowledge of electromagnetism is given by a_{j0}. The difficulty of each item is related to d_j; items with larger positive d_j are easier, items with more negative d_j, harder. The discrimination of the item on the subtopics of the topical model is given by a^s_{jk}.

than the range of d_j. The discrimination is related to the slope of the probability curve with respect to θ at $\vec{a}_j \cdot \theta_i + d_j = 0$ where the probability of selecting the c correct response is 0.5;

larger discriminations represent probability curves that are more steeply sloped at this point and a transition between a low probability of answering correctly and a high probability over a more narrow range of θ. The item discriminates between low ability and high ability students more strongly than lower discrimination items. All overall discriminations are positive, indicating items are generally well functioning. A negative discrimination would indicate the items was more likely to be answered correctly by lower ability students. The largest discriminations are associated with the two hardest items (items 28 and 29) involving magnetic induction and the easiest item (item 21) which asks about the direction of the magnetic field of a bar magnet. There are windowing effects relating d_j and discrimination; an item with either very high or very low d_j has a narrow range of θ to transition from low to high probability leading to high discrimination. Items 15 and 22 both have discriminations of about 1.5 with moderate item scores; item 15 asks about the electric potential difference along an equipotential and item 22 is blocked with item 21 and asks about the direction of the magnetic field of a bar magnet. Because of their moderate d_j and high discrimination, these two items are probably the most effective for discriminating between high and low ability students. We note, in this context, and throughout this work, ability is narrowly defined as the facility to answer conceptual electromagnetism questions as presented in the BEMA.

The d_j and overall discrimination of M13 was very similar to that of M18 and, therefore, the above discussion can be extended to this model as well.

The subscale discrimination, shown as a^s_{jk} in Table 8.7, represents the amount the item discriminates on the subscale over its overall discrimination. Most subscale discriminations were fairly small; three of the largest discriminations were for items 15, 28, and 29 which

also have large overall discrimination. Item 22, which is blocked with item 21, also has a comparatively large discrimination. The only two other items that stand out are items 4 and 5 within the electrostatic subscale; the items are blocked and ask about the electric field direction at two points of an electric dipole. These two items do appear to more synthetically test for a knowledge of electrostatics than other items in the subscale. Item 26 requires a knowledge of both electrostatics and magnetostatics and has a subscale discrimination for both topics; both discrimination are small. This item largely discriminates on a students general facility with electromagnetism.

Characterizing the internal reliability or consistency of a subscale is a common problem in Classical Test Theory. One of the most used statistics for internal reliability is Cronbach's α which is also presented in Table 8.6 [213]. The α values vary widely and none reach the threshold of 0.7 required for low-stakes testing. As such, the subscales in Table 8.6 do not represent a coherent measurement of the subtopic, but rather represent the average of the student's knowledge on the individual items making up the subtopic. This is hardly surprising examining the broad set of reasoning represented by the principles in each subtopic.

8.6.4 Principle Model

The principle model, M13, contains items requiring from 1 to 4 principles for their solution. Principles that were combined because the MIRT model could not individually resolve them such as L9-DF7 were counted as a single principle. The overall d_j and discrimination a_{j0} of each item was very similar in M13 and M18, and were discussed in Sec. 8.6.3. Table 8.7 shows the principles used in M13 and the discrimination of each principle, in parentheses, as well as the overall discrimination a_{j0} of the item. Many principles had

discriminations which were small compared to the overall discrimination; these items test a general facility with the material measured by the BEMA more strongly than the individual reasoning required by the principle. Some items had discrimination approximately commensurate with the overall discrimination: items 9, 18, 28 and 29. These items discriminate more strongly on the application of the principle than an overall facility with the material. The largest principle discriminations (items 20, 21, 28, and 29) were generally associated with large overall discriminations. These items were discussed in the previous section. Very little stood out in the principle discriminations; most principles on the same item had similar discrimination and few items had one principle discrimination substantially different than the others.

Isomorphic items are items that are solved with the same process, items requiring the same principles for their solution. Item pairs {14, 15}, {21, 22}, and {28, 29} are isomorphic. All are also part of item blocks complicating their statistical interpretation. Items 21 and 22 ask the student about the magnetic field at two different points around a bar magnet. Items 28 and 29 ask about the induced electric field direction at two points around a solenoid whose current is increasing. The similarity of items 14 and 15 are less clear. Item 14 involves the electric potential difference along an electric field line; item 15 involves the potential difference along an equipotential. Each item was initially coded as requiring different lemmas. All lemmas associated with the definition of electric potential (DF3) were collapsed into DF3 to form M13 which improved model fit. To determine if collapsing all electric potential items simultaneously obscured differences in student reasoning on items 14 and 15, model M13 was modified to include LM5 (the electric potential difference is zero perpendicular to the field) as a separate principle. This transformation did not improve

model fit. The students do not differentiate these two principles above their general difference in overall difficulty and discrimination.

8.7 Discussion

8.7.1 Research Questions

This study investigated three research questions; they will be discussed in the order proposed. The results of the individual analyses were discussed in the previous section as these analyses were introduced. This section summarizes and synthesizes the results of these analyses.

RQ1: What relations between BEMA items are identified by exploratory analyses? What do these relations imply for the interpretation of the results of applying the BEMA? Correlation analysis using the partial correlation correcting for overall instrument score (Fig. 8.3) showed that items within item blocks were correlated with each other above the average level of correlation expected of items testing a general knowledge of electromagnetism. The larger topical subscales tested by the instrument shown as subdivisions in Table 8.3 were not substantially correlated controlling for overall BEMA score as shown in Fig. 8.3. The blocked items stand out as the strongest correlations in the correlation matrix as well (Fig. 8.2); however, substantial positive correlations exist between many items. There is little evidence that items in the general subtopics in the topical model (M18) are generally more correlated with each other than with other items in the instrument in either the correlation or partial correlation matrix.

Exploratory factor analysis supported the conclusion that the blocked items represent

the only statistically meaningful substructure of the instrument. Of the 5 factors in the best-fitting factor model (Table 8.2), the highest loadings in four of the factors were items within the same item block. The fifth factor had no item with a large loading. Items from all subtopics except magnetic induction had similar, but small, loadings on factor 5.

The prevalence of the blocked items in all the exploratory analysis strongly implies these items may be correlated more than would be the case if not blocked. This raises concerns about interpretation of the results of blocked items and suggests all items except the first in an item block be discarded. The grading rubric provided with the instrument at PhysPort [169] does suggest modified scoring rules for items 2 and 3 and items 28 and 29, all blocked items.

RQ2: What is the model of student knowledge measured by the BEMA identified by constrained MIRT? What insights can this model provide into the structure of the instrument? This work presented two models of the BEMA with excellent fit statistics: one featuring a detailed model of the instrument in terms of reasoning principles (M13) and one involving general electromagnetic subtopics (M18). Both of these models had similar and excellent fit statistics (RMSEA, CFI, and TLI). The topical model was better fitting measured by AIC and BIC probably because these measures penalize the additional parameters more strongly than RMSEA, CFI, and TLI. The Cronbach's α of the subtopics did not suggest they had strong internal consistency and the subtopics were not extracted as factors in factor analysis. As such, the principle model (M13), derived from a model of expert solutions of the instrument, may represent the best model of the instrument as a set of items that measure a broad set of fairly loosely related (in student thinking) pieces of electromagnetic reasoning.

The list of principles forming the initial model in Table 8.3 was extensive, larger than

that of the FCI, FMCE, and CSEM in Studies 1 to 3. The four models are compared in RQ3. Most principles, including secondary principles, were retained in the principle model (M13) indicating that student thinking about the material is composed of many disparate reasoning fragments. Many of these fragments were tested by single items making it difficult to explore student thinking in detail; for example, Gauss' law and the definition of electric flux are tested together by only a single item. There are a number of these combinations of principles that are only tested together which does not allow the instrument to determine if they are understood independently.

The sheer breadth of principles and their variety, combined with the failure to find evidence that principles in the same subtopic are generally correlated above correlations through overall test score or to find subtopics as factors suggest that the overall design of the instrument may need refinement. An instrument with a more top down design around the five subtopics which focused on testing the most important principles within each subtopic well might provide instructors with a superior tool to manage their classes.

Classical Test Theory (CTT) suggests that items with either very high or very low item scores (called "difficulty" in CTT) or items with very low discrimination be considered problematic [213]. The item scores of items 1, 21, 28, and 29 indicate that they may be problematic. Qualitatively, both IRT and CTT discriminations are similar measuring how well the items distinguishes between low and high performing students; however, they are not directly comparable quantitatively. As such, there is not a well established critical discrimination value for problematic MIRT items. Items with very small MIRT discriminations have fairly flat probability curves, so low and high ability students have similar probability of answer correctly. Items 9 and 18 have very small overall discrimination and should be

investigated further to determine if they are functioning correctly.

RQ3: How are the best-fitting models of the BEMA, CSEM, FMCE, and FCI similar? How are they different? This work sought to understand the physical principles tested by the BEMA. It is the fourth of four papers using constrained MIRT to investigate some of the most widely applied physics conceptual instruments. To answer this research question, a comparison of the similarity and differences of the four instruments is provided. The BEMA is most topically related to the CSEM and specific comparisons to this instrument are made when appropriate. All four studies investigated three general dimensions: (1) the exploratory structure found by correlation analysis and factor analysis, (2) the best-fitting principle model found by constrained MIRT and theoretically motivated modifications of an initial expert model, and (3) a comparison of the best-fitting principle model to a more general model of the instrument (the topical model in the case of the BEMA).

Exploratory analyses of the FCI, FMCE, and BEMA proceeded first with a partial correlation analysis. All studies then employed exploratory factor analysis using MIRT. Best-fitting factor models were selected by examining fit statistics. The partial correlation analysis showed strong correlation between many blocked items; however, not all items within each item block were strongly partially correlated suggesting that, while important, blocking was not the only feature affecting the correlation structure. This pattern continued in the BEMA where item 15 was not strongly correlated with the other items in its block, items 14 and 16. Exploratory factor analysis of the four instruments yielded best-fitting factor models with from 5 to 9 factors: FCI (9), FMCE (5), CSEM-1 (9), CSEM-2 (8), and BEMA (5); Study 3 presented two samples of CSEM data labeled CSEM-1 and CSEM-2. For all factor structures, the fit statistics did not clearly identify a single best-fitting model; different

models were selected by different statistics. For all models, the factor structure had a strong relationship to the blocking structure of the instrument, but was not fully explained by the blocked structure. This effect was weaker in the CSEM with only 1 of the 3 item blocks consistently loading on the same factor in either sample. The strong effect of blocking was clearly evident for the BEMA; blocked items form the largest loadings on 4 of the 5 of the factors. The fifth factor includes many items across disparate topics, all with fairly low loadings. It is unclear what this factor actually measures. Blocked items explained only a subset of the 9 FCI factors; many of the other factors were related to isomorphic items which were not blocked. All FMCE factors were related to blocking, but all FMCE items except one are blocked. Like the FMCE, all isomorphic BEMA items ({14, 15}, {21, 22}, and {28, 29}) are also in item blocks, so the two effects cannot be separated.

The best-fitting principle models for each of the four conceptual inventories (FMCE, FCI, CSEM, and BEMA) can be compared to develop a greater understanding of the relationship of these instruments. This comparison may be valuable to practicing instructors trying to choose a conceptual instrument or to researchers comparing results of studies applying different instruments. Each study made a number of decisions about the inclusion of items in the analysis; therefore, the best-fitting principle models generally do not include all items while the initial theoretical models generally do include all items. Both the CSEM and BEMA contained combinations of principles where the combination always loaded on the same items; these combinations were coded as a single loading in MIRT. To compare instruments, these combinations contribute the number of principles in the combination toward the principle count. Two samples of the CSEM were analyzed producing slightly different best-fitting principle models. For comparison, CSEM Sample 1 is used, because its principle

model is the most similar to that of the BEMA. The models of the two CSEM samples differ only in the handling of the lemmas associated with electric potential.

Table 8.8 presents a comparison of the BEMA and CSEM initial expert models which cover all items in the instruments. The principles are split into 3 groups: definitions and laws (DF, L) representing the most general coverage of the instrument, facts (F) representing specific knowledge needed to solve the instrument, and corollaries and lemmas (C, LM) representing qualitative and quantitative reasoning derived from the general principles needed to solve specific problems. The principles are also split between the subtopics introduced in Table 8.3. Examining the (DF, L) column shows the BEMA in general covers most of the general physics covered by the CSEM, but the reverse is not true with the BEMA covering 10 additional principles. The number of principles covered by the other instrument is shown in parenthesis. Half of this difference involves the coverage of electric circuits. The difference in the L and DF principles between the instruments are generally localized to only a few items. For this discussion, differences in the use of mechanics are not considered. The CSEM includes two items involving the behavior of net charge on conductors and insulators, which require the law of conservation of charge. The CSEM also requires the student to read an electric field map which requires the definition of an electric field line. The BEMA contains a single Gauss' law item requiring both the application of Gauss' law and the definition of electric flux. The general coverage of magnetostatics is even more similar with the only difference found in the BEMA in one item applying the right hand rule for magnetic moment along one solution path. The specific coverage of the instruments, captured in the number of F, C, and LM principles, is fairly different. These principles involve the less general patterns of reasoning required to solve specific individual items. The majority of these types of

	BEMA					CSEM				
Subtopic	Items	DF, L	F	C, LM	Total	Items	DF, L	F	C, LM	Total
Mechanics	0	3(2)	0(0)	1(1)	4(3)	0	3(2)	0(0)	2(1)	5(3)
Electrostatics	8	5(3)	2(0)	4(1)	11(4)	14	5(3)	1(0)	4(1)	10(4)
Electric Potential	4	1(1)	0(0)	4(1)	5(2)	6	1(1)	0(0)	4(1)	5(2)
Magnetostatics	9	6(4)	1(0)	5(2)	12(6)	9	4(4)	1(0)	2(2)	7(6)
Magnetic Induction	3	3(2)	0(0)	2(0)	5(2)	3	2(2)	0(0)	0(0)	2(2)
Superposition	0	1(1)	0(0)	0(0)	1(1)	0	1(1)	0(0)	0(0)	1(1)
Electric Circuits	7	5(0)	5(0)	3(0)	13(0)	0	0(0)	0(0)	0(0)	0(0)
Total	31	24(13)	8(0)	19(5)	51(18)	32	16(13)	2(0)	12(5)	30(18)

Table 8.8: Comparison of BEMA and CSEM. DF, L, R, C, and LM represent principles in each instrument. The number in parenthesis is the number of the principles also in the other instrument. The Items column refers to the number of items in the instrument grouped into the electricity and magnetism subtopics; Mechanics and Superposition are not subtopics specific to electricity and magnetism, so their Items columns are 0.

principles are not shared between the instruments, only 5 of the 39 principles are shared. As such, while the general coverage of the instruments is similar (except for electric circuits), the specific coverage is quite different. Many more of these specific principles were identified in the BEMA; the CSEM covers electricity and magnetism at a somewhat more general level. This has important implications for the generalizability of BEMA or CSEM results because specific pedagogical choices can affect the detailed coverage of a class, as well as its general coverage.

One of the benefits of building a detailed model of an instrument such as that in Table 8.3 is the facilitation of new qualitative and quantitative comparisons between instruments. The general complexity of items in the instrument can be characterized by the average number of reasoning steps per item. The degree to which the instrument measures a piece of reasoning with multiple items (providing generally higher reliability) can be characterized by the number of items per reasoning step.

Table 8.9 presents a general comparison of the best-fitting principle models of all four

instruments. The principle models did not fit all items (except in the FMCE), but do allow a more detailed comparison of the instruments on the items fit. The models of the FCI and FMCE involved two additional types of principles not found in the CSEM or BEMA: results (R) such as the 3-dimensional kinetic equations for motion under a constant force and reasoning steps (RS) such as reading a graph. Table 8.9 presents two measures of overall instrument length: the independent principles representing the number of unique principles needed to solve the instrument and the total principles representing the number of reasoning steps required to solve the instrument. Each independent principle may be required to solve multiple items and thus be counted multiple times in the total principles. As above, the BEMA involved applying more independent principles (an independent principle represents one of the rows in Table 8.3) than the CSEM, approximately 40% more independent principles per item. As such, a greater variety of physical knowledge is needed to solve each item. A larger fraction of the BEMA items were fairly complex requiring four or five principles for their solution. These two differences led to generally more complex items requiring 2.3 principles per item on average for the BEMA in comparison to 2.0 principles per item in the CSEM. As such, the BEMA generally involves applying longer, more complex, patterns of reasoning than the CSEM. This observation also implies that each principle in the BEMA is not as thoroughly measured as in the other instruments with only 0.71 items measuring each independent principle. The FMCE strongly stands out on this metric with each independent principle in the FMCE measured on average by over 5 items.

All four studies paid particular attention to the role of lemmas in the models. The other types of principles represent standard content that might be present in most textbooks; lemmas represent qualitative interpretations of these principles. All studies found that ex-

	BEMA	CSEM-1	FCI	FMCE
Items analyzed	27	25	20	43
DF, L, R	18	16	9	5
F	7	2	6	1
C, LM, RS	13	6	4	2
Ind. principles	38	24	19	8
Ind. pcpl. per item	1.41	0.96	0.95	0.19
Items per ind. pcpl.	0.71	1.05	1.05	5.38
1 principle items	10 (37%)	8 (32%)	4 (20%)	25 (58%)
2 principle items	5 (29%)	11 (44%)	7 (35%)	15 (35%)
3 principle items	6 (22%)	5 (20%)	7 (35%)	2 (5%)
4 principle items	5 (19%)	1 (4%)	1 (5%)	1 (2%)
5 principle items	1 (4%)	0 (0%)	1 (5%)	0 (0%)
Total principles	63	49	48	65
Total pcpl. per item	2.30	1.96	2.40	1.51

Table 8.9: Comparison of conceptual instruments. DF, L, R, F, C, LM, and RS represent principles in each instrument. Independent is abbreviated "ind" and principle "pcpl" when needed for spacing.

perts used many lemmas in their solutions; however, it was unclear whether these principles were needed to model student thinking. In Study 1, all lemmas were removed simultaneously which improved model fit. In the studies of the FMCE and CSEM, lemmas were removed in groups, as they were in the present study. The best-fitting model for the FMCE, CSEM, and BEMA all contained some, but not all, of the lemmas in the initial expert model. The lemmas remaining in the FMCE involved motion opposite the direction of acceleration and are associated with a type of problem particularly difficult for students. Many of the lemmas identified for the electricity and magnetism instruments represent principles central to solving qualitative (and quantitative) items such as "opposites attract - likes repel." As such, it would have been surprising if these principles were not found to be part of the model of student reasoning.

All studies explored a model more general than the best-fitting principle model. For the FCI, Study 1 fit a decomposition of the items of the instruments into topics that was

proposed with the original publication of the instrument [7]. The principle model improved AIC by 448 and BIC by 226 over the topical model, very strong changes. Study 3 fit a model similar to the best-fitting principle model of the current study (excluding the electric circuits topic); for both samples, the topical model had worse model fit than all of the principle models. For one of the samples, the best-fitting topical model did not meet the requirements of acceptable model fit [212]. The FMCE uses fewer principles and repeats the principles more often than the other instruments. As such, rather than proposing general topical principles, Study 2 grouped FMCE items into subscales. Confirmatory factor analysis was then performed to determine if this model fit the instrument well; it did not (CFI= 0.80, TLI= 0.79, and RMSEA= 0.080). So for the FCI, FMCE, and the CSEM, the more general topical model was substantially less well fitting than the best-fitting principle model. The results for the BEMA were different; the topical model was better fitting than the best-fitting principle model with difference in AIC and BIC similar to those observed for the FCI and larger than the differences observed for either CSEM sample. For context, the difference in the two model's AIC was 249, this means the topical model was $e^{249/2}$ times more probable than the principle model. The reason for this difference is unclear; perhaps the larger number of principles and the fairly weak interconnections of principles within items generates an instrument which is more a measurement of general topics than specific information within the topics. The BEMA topical model had similar RMSEA, CFI, and TLI to the principle model. Correlation and factor analysis also did not support the view of the instrument measuring independent subtopics; therefore, the best-fitting principle model may be a better general model of the knowledge measured by the BEMA.

8.7.2 Synthesis

Through the four studies applying constrained MIRT, some important themes have emerged. We attempt to encapsulate those themes in this section.

The general quality of the initial expert model

The studies of the FMCE, CSEM, and BEMA reported CFI, TLI, and RMSEA for each stage of the model transformation process. In general, the initial expert model had excellent fit statistics. These were improved only slightly through the transformation process. We revisited the models used for the FCI and a similar pattern of excellent fit throughout the transformation process was observed. As such, the initial expert models derived from observations of expert solutions were very good models of the material and could be constructed without the need to collect large datasets and without the application of MIRT. This observation opens the possibility of developing similarly detailed models of an entire domain such as introductory mechanics or electricity and magnetism. These models would allow one to quantitatively express the relationship between the conceptual instruments and the domain they profess to measure. The decisions about item selection and topical coverage used to construct the instruments could be evaluated by the PER community within this framework. Such a framework could also serve to allow more detailed description of instructional innovations by providing a mechanism to specify in detail any changes in topical coverage resulting from the innovation.

The negative effect of item blocking

All studies found that blocked items dominated both the partial correlation and exploratory factor structures. In some instruments, such as the FMCE and the CSEM, blocked items were often isomorphic. This was not the case in the FCI which strongly suggests that blocked items have correlations and other statistical properties that are the result of blocking, not the physical constructs the items were intended to measure. This and other work strongly suggests that item blocking should be discontinued in future instruments. It also suggests that exploratory factor structures extracted from instruments with item blocks including the FCI, FMCE, CSEM, and BEMA are strongly related to these blocks and is not a general measure of the substructure of student reasoning on the topic; conclusions drawn from these analyses should be interpreted with care.

The granularity of the knowledge measured by the instruments

In the three studies that removed lemmas independently, the best-fitting models contained some lemmas and corollaries, very specific reasoning pieces. Further, in the studies of the FCI, FMCE, and the CSEM a general topical model was not as well fitting as the model involving a decomposition into principles. The topical model was better fitting for the BEMA by AIC and BIC, but had very similar (and sometimes weaker) RMSEA, CFI, and TLI than the best-fitting principle model. There was very little evidence in the correlation or factor analysis to suggest the items within the subtopics were more related with themselves than with items in other subtopics. As such, all of these popular instruments measure a detailed set of reasoning skills as opposed to a general construct such as "Newtonian thinking."

This is supported by the analysis of the FMCE and the BEMA which found poor subscale internal consistency as measured by Cronbach's α. This has important implications for the general interpretation of the results of applying the instruments; the instruments may be susceptible to small changes in the coverage or focus of the courses studied.

The general dissimilarity of the instruments

The constrained MIRT models produced a very detailed picture of the four conceptual instruments. While the FCI and FMCE, as well as the CSEM and BEMA, cover the same general topics (Newtonian mechanics or electromagnetism), the pairs of instruments were quite different through the detailed lens of MIRT. The quantitative differences are explored in the discussion of RQ3; the qualitative differences are self-evident through a comparison of the expert models. As such, comparing studies using different instruments should be done with care and should consider how the detailed differences of the instruments might interact with the student population or any pedagogical differences between treatments.

8.7.3 Future work

All four MIRT studies identified the blocking of items as a potential problem, generating correlations between items not related to the physical reasoning needed to solve the items. Network analytic studies have also identified items where connections between correct and incorrect responses suggest students may be answering correctly for incorrect reasons [181, 214, 182]. Multiple authors have suggested alternate scoring rubrics for some of these instruments in response to these and other problems [215, 216, 169, 182]. Many classical test theory and item response theory studies have identified items within these in-

struments with performance outside the suggested range for good psychometric functioning [174, 175, 198, 199]. Substantial biases have also been identified in some of the instruments [174]. With the accumulation of evidence that these instruments at the very least should be revisited and revised, a model of a revised instrument in terms of principles grounded in a more general model of the domain measured could provide basis for a discussion within the research community of what should be assessed in introductory physics leading to a new generation of conceptual instruments.

8.8 Limitations

This work was performed using data drawn for a single institution. The models should be tested with additional student populations to determine if the conclusions are general.

8.9 Conclusions

This study investigated the structure of the BEMA using correlation analysis and exploratory factor analysis and, then, explored the models of student knowledge tested by the BEMA using constrained MIRT. Correlation analysis revealed that items within item blocks account for nearly all of the substructure of the instrument. Exploratory factor analysis identified a 5-factor model as having the best fit. The highest loadings in four of the five factors were items in the same item block, consistent with the correlation analysis. Two models of student knowledge were presented; one involved 28 detailed reasoning principles (M13) and the other contained five general electromagnetic subtopics (M18). Both models had excellent fit statistics. The five topics in M18 were investigated as subscales; however,

none of the subscales had a Cronbach's alpha of 0.7 suggested for low-stakes testing. As such, the model of student knowledge tested by the BEMA consists of a broad collection of loosely related reasoning pieces.

The best-fitting principle models of the FCI, FMCE, and CSEM had fewer principles than that of the BEMA. The best-fitting principle model of the BEMA also required more lemmas and corollaries than any of the other instruments' models. The coverage differences between the CSEM and BEMA were largely the result of the coverage of electric circuits in the BEMA and differences in the coverage of electrostatics. Quantitative comparison of the four conceptual instruments investigated using constrained MIRT identified substantial differences in terms of the number of principles and the number of principles per item. As such, while related, the FCI and the FMCE as well as the CSEM and the BEMA measure their conceptual domains with different coverage and with items with different intellectual complexity.

8.10 Acknowledgement

We thank Steven Pollock for the collection and curation of this exceptional dataset and his helpful commentary.

Chapter 9

Conclusions and Future Work

The work presented in this dissertation can be split into four main parts.

Physics Student Retention

Physics student retention was explored by applying the statistical methods of logistic regression, survival analysis, decision trees, and Bayesian networks to student retention and major progression at an university in the eastern U.S. These tools effectively identified high school GPA and math readiness as key predictive factors as to whether students will be retained in the physics program and progress towards graduation. Once a student's college performance was added to the retention model, pre-college academic factors were found to be less predictive.

Physics Course Grade Prediction

Bayesian Networks were also used to predict student outcomes in physics courses using prior course grades. Each course whose outcomes were predicted with Bayesian networks had balanced accuracies greater than 70%, but some courses were more predictable than others. The less predictable courses seem to be less well integrated in the program; this could be due to inconsistent grading in the courses, inconsistent instruction in the courses, or that the content in the less predictable courses does not productively build upon the knowledge and skills learned in prior courses.

Analysis of the Structure of Physics Curricula

The method of Curricular Analytics was used to analyze the physics curricula at 60 institutions in the U.S. Curricular Analytics provides a quantitative framework for analyzing

the course requirements of an academic program. The range in curricular complexity of the 60 physics programs was nearly 200 complexity points. The increased complexity of some programs may be unnecessary. A more complex curriculum may take more time to complete, and students who fail a course or are not math ready will have an extended time to degree. An extended time to degree may be detrimental to some students' retention.

Identifying the Structure of Knowledge Measured by a Conceptual Inventory

Multidimensional Item Response Theory (MIRT) was used to build a model of student knowledge tested by the BEMA. Two models of student knowledge were found that had excellent model fit statistics: one contained 28 detailed reasoning principles, and the other involved five general electromagnetic subtopics. The knowledge structure measured by the BEMA was compared with similar knowledge models developed through MIRT analysis of three other common conceptual inventories.

One purpose of the work in this dissertation was to apply these methods to the problem of student retention and report the results in the hope that physics departments will also apply these methods at their institutions to identify the factors that affect their students' retention and make changes to courses or program requirements that will improve student retention. Much of the work in this dissertation only used data from a single institution. It is likely that the specific findings herein are specific to the institution from which the data were gathered. For departments to make effective changes in their programs, some of these analyses should be replicated with local data.

Further research in student retention will be vital to the improvement of student success in physics in higher education. Future research should continue to identify factors that

influence student retention in physics, and also begin to test interventions that are designed to improve student retention. Some possible future projects include:

- Applying logistic regression and survival analysis at institutions with higher and lower rates of math readiness. Universities with more selective or less selective admissions requirements will likely have a different dependence on pre-college factors and student retention; identifying these differences would be of great interest to the PER community.

- Implement the prediction of course grades with Bayesian networks for use in physics student advising and measure the effect on student retention. Accurate grade predictions give advisors excellent information they can use to advise students on what classes they should take in a given semester and in what order they should take required courses. Measuring the effects of implementing Bayesian networks in advising would provide concrete evidence that grade prediction can be used to improve retention.

- Change the pre-requisite structures in a curriculum with the intent of improving the retention of physics students. The use of Curricular Analytics to influence curriculum changes has yet to be proven as a method that increases student retention. Making curriculum changes that reduce complexity, and then measuring the effect of that change on student retention support the central claim of Curricular Analytics that lower curricular complexity increases retention.

- Use MIRT to identify the structure of knowledge measured by a physics course. MIRT is an analysis tool specific to instrument analysis; however, it, in principle, could be used to identify the knowledge that an entire physics course measures. Identifying

the domain of knowledge that a physics course measures would be immensely useful in course reformation, as courses that don't actual teach what they purport to teach could be rebuilt to better support student learning.

Bibliography

[1] R.R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, 66:64–74, 1998.

[2] President's Council of Advisors on Science and Technology. Report to the President. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Executive Office of the President, Washington, DC, 2012.

[3] K. Cummings. A developmental history of physics education research. In *Second Committee Meeting on the Status, Contributions, and Future Directions of Discipline-Based Education Research.*, 2011. `http://www7.nationalacademies.org/bose/DBER\_Cummings\_October\_Paper.pdf`.

[4] A.B. Champagne, L.E. Klopfer, and J.H. Anderson. Factors influencing the learning of classical mechanics. *Am. J. Phys.*, 48(12):1074, 1980.

[5] J. Clement. Students' preconceptions in introductory mechanics. *Am. J. Phys.*, 50(1):66–71, 1982.

[6] L.C. McDermott. Research on conceptual understanding in mechanics. *Phys. Today*, 37:24–32, 1984.

[7] D. Hestenes, M. Wells, and G. Swackhamer. Force Concept Inventory. *Phys. Teach.*, 30:141–158, 1992.

[8] R.K. Thornton and D.R. Sokoloff. Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *Am. J. Phys.*, 66(4):338–352, 1998.

[9] D.P. Maloney, T.L. O'Kuma, C. Hieggelke, and A. Van Huevelen. Surveying students' conceptual knowledge of electricity and magnetism. *Am. J. Phys.*, 69(S1):S12, 2001.

[10] L. Ding, R. Chabay, B. Sherwood, and R. Beichner. Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. *Phys. Rev. Phys. Educ. Res.*, 2:010105, Mar 2006.

[11] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart. Multidimensional item response theory and the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 14:010137, Jun 2018.

[12] R. Henderson, C. Zabriskie, and J. Stewart. Rural and first generation performance differences on the Force and Motion Conceptual Evaluation. Physics Education Research Conference Proceedings 2018 (accepted)., 2018.

[13] J. Yang, C. Zabriskie, and J. Stewart. Multidimensional item response theory and the Force and Motion Conceptual Evaluation. *Phys. Rev. Phys. Educ. Res.*, 15(2):020141, 2019.

[14] J. Hansen and J. Stewart. Multidimensional item response theory and the Brief Electricity and Magnetism Assessment. *Phys. Rev. Phys. Educ. Res.*, 17(2):020139, 2021.

[15] J. L. Docktor and J. P. Mestre. Synthesis of discipline-based education research in physics. *Phys. Rev. Phys. Educ. Res.*, 10(2):020119, 2014.

[16] E. Mazur. *Peer Instruction: A User's Manual*. Prentice Hall, Upper Saddle River, NJ, 1997.

[17] C.H. Crouch, J. Watkins, A.P. Fagen, and E. Mazur. Peer instruction: Engaging students one-on-one, all at once. *Research-based reform of university physics*, 1(1):40–95, 2007.

[18] D. R. Sokoloff and R. K. Thornton. *Interactive lecture demonstrations*. John Wiley and Sons, New York, NY, 2004.

[19] L.C. McDermott and P.S. Shaffer. *Tutorials in Introductory Physics*. Prentice Hall, Upper Saddle River, NJ, 1998.

[20] M. Wittman, R. Steinberg, E. Redish, et al. *The Physics Suite, Activity Based Tutorials, Vol. 2-Modern Mechanics*. John Wiley and Sons, New York, NY, 2004.

[21] A. Elby. Helping physics students learn how to learn. *Am. J. Phys.*, 69(S1):S54–S64, 2001.

[22] N.D. Finkelstein and S.J. Pollock. Replicating and understanding successful innovations: Implementing Tutorials in Introductory Physics. *Phys. Rev. Phys. Educ. Res.*, 1(1):010101, 2005.

[23] E.F. Redish, J.M. Saul, and R.N. Steinberg. On the effectiveness of active-engagement microcomputer-based laboratories. *Am. J Phys*, 65(1):45–54, 1997.

[24] H.V. Mauk and D. Hingley. Student understanding of induced current: Using tutorials in introductory physics to teach electricity and magnetism. *Am. J Phys*, 73(12):1164–1171, 2005.

[25] R. T. Johnson and D. W. Johnson. *An overview of cooperative learning, creativity and collaborative learning*, volume 25. Baltimore, MD: Brookes Press., 1994.

[26] A. Van Heuvelen. Learning to think like a physicist: A review of research-based instructional strategies. *Am. J. Phys.*, 59(891), 1991.

[27] H. Brasell. The effect of real-time laboratory graphing on learning graphic representations of distance and velocity. *J. Res. Sci. Teach.*, 24(4):385–395, 1987.

[28] R. J. Beichner. The impact of video motion analysis on kinematics graph interpretation skills. *Am. J. Phys.*, 64(10):1272–1277, 1996.

[29] E. Etkina, S. Murthy, and X. Zou. Using introductory labs to engage students in experimental design. *Am. J. Phys.*, 74(11):979–986, 2006.

[30] N.G. Holmes, J. Olsen, J.L. Thomas, and C.E. Wieman. Value added or misattributed? a multi-institution study on the educational benefit of labs for reinforcing physics content. *Phys. Rev. Phys. Educ. Res.*, 13(1):010129, 2017.

[31] R.J. Beichner, J.M. Saul, D.S. Abbott, J.J. Morse, D. Deardorff, R.J. Allain, S.W. Bonham, M.H. Dancy, and J.S. Risley. The student-centered activities for large enrollment undergraduate programs (scale-up) project. *Research-based reform of university physics*, 1(1):2–39, 2007.

[32] J. W. Belcher. Improving student understanding with TEAL. *Fa. Newsl.*, 16(8), 2003.

[33] K. Cummings, P. W. Laws, E. F. Redish, and P. J. Cooney. *Understanding physics*. John Wiley and Sons, New York, NY, 2004.

[34] R. D. Knight. *Physics for scientists and engineers*. Pearson Higher Ed., Upper Saddle River, NJ, 2017.

[35] T. A. Moore. *Six ideas that shaped physics*. WCB/McGraw-Hill, New York, NY, 1998.

[36] R. W. Chabay and B. A. Sherwood. *Matter and interactions*. John Wiley and Sons, New York, NY, 2015.

[37] K. Perkins, W. Adams, M. Dubson, S. Finkelstein, N.and Reid, C. Wieman, and R. LeMaster. Phet: Interactive simulations for teaching and learning physics. *Phys. Teach.*, 44(1):18–23, 2006.

[38] E. Seymour and N.M. Hewitt. *Talking about Leaving: Why Undergraduates Leave the Sciences*, volume 34. Westview Press, Boulder, CO, 1997.

[39] S. Tobias. *They're not Dumb, They're Different*. Research Corporation, Tuscon, AZ, 1990.

[40] E. Seymour and A. Hunter. Talking about leaving revisited. *Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education*, 2019.

[41] B.L. Whitten, S.R. Foster, M.L. Duncombe, P.E. Allen, P. Heron, L. McCullough, K.A. Shaw, B. Taylor, and Heather M. Zorn. What works? Increasing the participation of women in undergraduate physics. *J. Women Minorities Sci. Eng.*, 9(3&4), 2003.

[42] K. Rosa and F.M. Mensah. Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life. *Phys. Rev. Phys. Educ. Res.*, 12(2):020113, 2016.

[43] M. Ong. Body projects of young women of color in physics: Intersections of gender, race, and science. *Social Problems*, 52(4):593–617, 2005.

[44] L.T. Ko, R.R. Kachchaf, A.K. Hodari, and M. Ong. Agency of women of color in physics and astronomy: Strategies for persistence and success. *J. Women Minor. Sci. Eng.*, 20(2), 2014.

[45] J.M. Aiken, R. Henderson, and M.D. Caballero. Modeling student pathways in a physics bachelor's degree program. *arXiv preprint arXiv:1810.11272*, 2018.

[46] J.P. Zwolak, R. Dou, E.A. Williams, and E. Brewe. Students' network integration as a predictor of persistence in introductory physics courses. *Phys. Rev. Phys. Educ. Res.*, 13:010113, Mar 2017.

[47] J. Forsman, R. Moll, and C. Linder. Extending the theoretical framing for physics education research: An illustrative application of complexity science. *Phys. Rev. ST Phys. Educ. Res.*, 10:020122, Sep 2014.

[48] L. Stiles-Clarke and K. MacLeod. Demystifying the scaffolding required for first-year physics student retention: Contextualizing content and nurturing physics identity. *Can. J. Phys.*, 96(4):29, 2018.

[49] P.A. Westrick, J.P. Marini, L. Young, H. Ng, D. Shmueli, and E.J. Shaw. Validity of the SAT® for predicting first-year grades and retention to the second year. *Coll. Board Res. Pap.*, 2019.

[50] Jeff Allen. Updating the ACT college readiness benchmarks. ACT research report series 2013 (6). *ACT, Inc.*, 2013.

[51] W.J. Camara and G. Echternacht. The SAT I and high school grades: Utility in predicting success in college. *Res. Not.*, 10, 2000.

[52] S.l Geiser and R. Studley. UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educ. Assess.*, 8(1):1, 2002.

[53] P.A. Westrick, H. Le, S.B. Robbins, J.M.R. Radunzel, and F.L. Schmidt. College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educ. Assess.*, 20(1):23, 2015.

[54] S. Geiser and M.V. Santelices. Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. Research & Occasional Paper Series: CSHE. 6.07. *Ctr. Stud. High. Educ.*, 2007.

[55] A. Nandeshwar, T. Menzies, and A. Nelson. Learning patterns of university student retention. *Expert Syst. Appl.*, 38(12):14984, 2011.

[56] V. Tinto. *College Student Retention: Formula for Student Success*. Greenwood Publishing Group, Santa Barbara, CA, 2005.

[57] V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Rev. Educ. Res.*, 45(1):89, 1975.

[58] V. Tinto. *Leaving College: Rethinking the Causes and Cures of Student Attrition.* University of Chicago Press, Chicago, IL, 1993.

[59] V. Tinto. *Completing College: Rethinking Institutional Action*. University of Chicago Press, Chicago, IL, 2012.

[60] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Phys. Rev. Phys. Educ. Res.*, 8(2):020104, 2012.

[61] National Science Board. *Revisiting the STEM workforce: A companion to science and engineering indicators 2014*. National Science Foundation VA, 2015.

[62] X. Chen. STEM attrition: College students' paths into and out of STEM fields. Statistical Analysis Report. *NCES*, 2013.

[63] K. Rask. Attrition in stem fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Econ Educ Rev*, 29(6):892–900, 2010.

[64] E. J. Shaw and S. Barbuti. Patterns of persistence in intended college major with a focus on stem majors. *NACADA Journal*, 30(2):19–34, 2010.

[65] A. V. Maltese and R. H. Tai. Pipeline persistence: Examining the association of educational experiences with earned degrees in stem among us students. *Sci Educ*, 95(5):877–907, 2011.

[66] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *J Eng Educ*, 93(4):313–320, 2004.

[67] B. F. French, J. C. Immekus, and W. C. Oakes. An examination of indicators of engineering students' success and persistence. *J Eng Educ*, 94(4):419–425, 2005.

[68] R. M. Marra, K. A. Rodgers, D. Shen, and B. Bogue. Leaving engineering: A multi-year single institution study. *J Eng Educ*, 101(1):6–27, 2012.

[69] C. W. Hall, P. J. Kauffmann, K. L. Wuensch, W. E. Swart, K. A. DeUrquidi, H. O. Griffin, and S. C. Duncan. Aptitude and personality traits in retention of engineering students. *J Eng Educ*, 104(2):167–188, 2015.

[70] B.L. Christe. The importance of faculty-student connections in STEM disciplines. *J. STEM Educ. I. R.*, 14(3):22, 2013.

[71] Z.S. Wilson, L. Holmes, K. Degravelles, M.R. Sylvain, L. Batiste, M. Johnson, S.Y. McGuire, S.S. Pang, and I.M. Warner. Hierarchical mentoring: A transformative strategy for improving diversity and retention in undergraduate STEM disciplines. *J. Sci. Educ. Technol.*, 21(1):148, 2012.

[72] M. Dagley, M. Georgiopoulos, A. Reece, and C. Young. Increasing retention and graduation rates through a STEM learning community. *J. Coll. St. Ret. R. T. P.*, 18(2):167, 2016.

[73] C.T. Belser, D.J. Prescod, A.P. Daire, M.A. Dagley, and C.Y. Young. Predicting undergraduate student retention in STEM majors based on career development factors. *Career Dev. Q.*, 65(1):88, 2017.

[74] C.T. Belser, M. Shillingford, A.P. Daire, D.J. Prescod, and M.A. Dagley. Factors influencing undergraduate student retention in STEM majors: Career development, math ability, and demographics. *Prof. Couns.*, 8(3):262, 2018.

[75] K. Koenig, M. Schen, M. Edwards, and L. Bao. Addressing STEM retention through a scientific thought and methods course. *J. Coll. Sci. Teach.*, 41(4):41, 2012.

[76] K.A. Wingate, A.A. Ferri, and K.M. Feigh. The impact of the physics, statics, and mechanics sequence on student retention and performance in mechanical engineering. In *2018 ASEE Annual Conference & Exposition*, Washington, DC, 2018. American Society for Engineering Education.

[77] Y.J. Xu. Attention to retention: Exploring and addressing the needs of college students in STEM majors. *J. Educ. Train. Stud.*, 4(2):67, 2016.

[78] A. Sithole, E.T. Chiyaka, P. McCarthy, D.M. Mupinga, B.K. Bucklein, and J. Kibirige. Student attraction, persistence and retention in STEM programs: Successes and continuing challenges. *High. Educ. Stud.*, 7(1):46, 2017.

[79] D.E. Meltzer and R.K. Thornton. Resource letter ALIP–1: Active-learning instruction in physics. *Am. J. Phys.*, 80(6):478–496, 2012.

[80] I.A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *Am. J. Phys.*, 53(11):1043–1055, 1985.

[81] I.A. Halloun and D. Hestenes. Common sense concepts about motion. *Am. J. Phys.*, 53(11):1056, 1985.

[82] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen. Surveying students' conceptual knowledge of electricity and magnetism. *Am. J. Phys.*, 69(S1):S12–S23, 2001.

[83] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler. Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis. *Phys. Rev. Phys. Educ. Res.*, 15:020122, Sep 2019.

[84] S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, and M.Pat. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *P. Nat. Acad. Sci. USA*, 111(23):8410–8415, 2014.

[85] C.M. Schroeder, T.P. Scott, T.Y. Tolson, H.and Huang, and Y.H. Lee. A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *J. Res. Sci. Teach.*, 44(10):1436, 2007.

[86] S. Salehi, E. Burkholder, G.P. Lepage, S. Pollock, and C. Wieman. Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 15:020114, Jul 2019.

[87] R. Henderson, J. Stewart, and A. Traxler. Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 15:010131, May 2019.

[88] J. Stewart, G.L. Cochran, R. Henderson, C. Zabriskie, S. DeVore, P. Miller, G. Stewart, and L. Michaluk. Mediational effect of prior preparation on performance differences of students underrepresented in physics. *Phys. Rev. Phys. Educ. Res.*, 17(1):010107, 2021.

[89] Z. Hazari, R.H. Tai, and P.M. Sadler. Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Sci. Educ.*, 91(6):847–876, 2007.

[90] C. Romero, S. Ventura, P.G. Espejo, and C. Hervás. Data mining algorithms to classify students. In R.S. Joazeiro de Baker, T. Barnes, and J.E. Beck, editors, *Proceeding of the 1st International Conference on Educational Data Mining*, Montreal, Quebec, Canada, 2008.

[91] A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Syst. Appl.*, 41(4):1432–1462, 2014.

[92] A.M. Shahiri, W. Husain, and N.A. Rashid. A review on predicting student's performance using data mining techniques. *Procedia Comput. Sci.*, 72:414–422, 2015.

[93] P. Baepler and C. J. Murdoch. Academic analytics and data mining in higher education. *Int. J. Scholarsh. Teach. Learn.*, 4(2):17, 2010.

[94] R. S. J. D. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *J Educ Data Mining*, 1(1):3–17, 2009.

[95] Z. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *J Educ Tech and Society*, 17(4), 2014.

[96] A. Dutt, M. A. Ismail, and T. Herawan. A systematic review on educational data mining. *IEEE Access*, 5:15991–16005, 2017.

[97] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[98] R. Alkhasawneh and R.H. Hargraves. Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. *J. STEM Educ. I. R.*, 15(3):35, 2014.

[99] N. Misiunas, M. Raspopovic, K. Chandra, and A. Oztekin. Sensitivity of predictors in educational data: A Bayesian network model. In *2015 INFORMS Workshop on Data Mining and Analytics*, Catonsville, MD, 2015. CIP, The Institute for Operations Research and the Management Sciences.

[100] A. McGovern, C.M. Utz, S.E. Walden, and D.A. Trytten. Learning the structure of retention data using Bayesian networks. In *2008 38th Annual Frontiers in Education Conference*, page F3D, Piscataway, NJ, 2008. IEEE.

[101] A. Sharabiani, F. Karim, An. Sharabiani, M. Atanasov, and H. Darabi. An enhanced Bayesian network model for prediction of students' academic performance in engineering programs. In *2014 IEEE Global Engineering Education Conference (EDUCON)*, page 832. IEEE, 2014.

[102] C. Lacave, A.I. Molina, and J.A. Cruz-Lemus. Learning analytics to identify dropout factors of computer science studies through Bayesian networks. *Behav. Inf. Technol.*, 37(10-11):993, 2018.

[103] R. Torabi, P. Moradi, and A.R. Khantaimoori. Predict student scores using Bayesian networks. *Procd. Soc. Behv.*, 46:4476, 2012.

[104] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart. Using machine learning to predict physics course outcomes. *Phys. Rev. Phys. Educ. Res.*, 15:020120, Aug 2019.

[105] J. Yang, S. DeVore, D. Hewagallage, P. Miller, Q.X. Ryan, and J. Stewart. Using machine learning to identify the most at-risk students in physics classes. *Phys. Rev. Phys. Educ. Res.*, 16(2):020130, 2020.

[106] G.J. Privitera. *Essential statistics for the behavioral sciences*. Sage publications, 2017.

[107] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, NY, 1977.

[108] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.*, 31(4):337–350, 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/pdf/10654_2016_Article_149.pdf.

[109] R. J. Calin-Jageman and G. Cumming. The new statistics for better science: Ask how much, how uncertain, and what else is known. *Am. Stat.*, 73(sup1):271–280, 2019.

[110] G. Cumming. The new statistics: Why and how. *Psychol. Sci.*, 25(1):7–29, 2014.

[111] M. Kubsch, I. Stamer, M. Steiner, K. Neumann, and I. Parchmann. Beyond p-values: Using Bayesian data analysis in science education research. *PARE*, 26(1):4, 2021.

[112] C. Spearman. General intelligence, objectively determined and measured. *Am. J. Psych.*, 100(3):697, 1987.

[113] D. A. Sass and T. A. Schmitt. A comparative investigation of rotation criteria within exploratory factor analysis. *Multivar. Behav. Res.*, 45(1):73–103, 2010. https://www.researchgate.net/publication/232890796_A_Comparative_Investigation_of_Rotation_Criteria_Within_Exploratory_Factor_Analysis.

[114] L. Chen, P. Chen, and Z. Lin. Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278, 2020.

[115] A.L. Traxler, X.C. Cid, J. Blue, and R. Barthelemy. Enriching gender in physics education research: A binary past and a complex future. *Phys. Rev. Phys. Educ. Res.*, 12:020114, Aug 2016.

[116] I. Rodriguez, E. Brewe, V. Sawtelle, and L.H. Kramer. Impact of equity models and statistical measures on interpretations of educational reform. *Phys. Rev. Phys. Educ. Res.*, 8(2):020103, 2012.

[117] S. Nicholson and P.J. Mulvey. Roster of physics departments with enrollment and degree data, 2020. American Institute of Physics, College Park, MD, 2020.

[118] T. Hodapp. The back page: The economics of education: Closing undergraduate physics programs. *APS News*, 20(11):8, 2011.

[119] National Center for Education Statistics. https://nces.ed.gov/collegenavigator.

[120] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[121] J.C. Brunson and Q.D. Read. ggalluvial: Alluvial plots in "ggplot2", 2020. R package version 0.12.3.

[122] P. R. Aschbacher, E. Li, and E. J. Roth. Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *J. Res. Sci. Teach.*, 47(5):564–582, 2010.

[123] D. Niedermayer. An introduction to Bayesian networks and their contemporary applications. In *Innovations in Bayesian networks: Theory and applications*, pages 117–130. Springer, 2008.

[124] M. Scutari and J. B. Denis. *Bayesian networks: with examples in R*. CRC press, 2021.

[125] M. Scutari. Understanding Bayesian networks with examples in r. University lecture, University of Oxford.

[126] R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian networks in r*. Springer, 2013.

[127] S. Beretta, M. Castelli, I. Gonçalves, R. Henriques, D. Ramazzotti, et al. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, 2018, 2018.

[128] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. Dynamic Bayesian networks for student modeling. *IEEE T. Learn. Technol.*, 10(4):450–462, 2017.

[129] H. M. Seffrin, G. L. Rubi, and P. A. Jaques. A dynamic Bayesian network for inference of learners' algebraic knowledge. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 235–240, 2014.

[130] I. Uglanova. Model criticism of Bayesian networks in educational assessment: A systematic review. *PARE*, 26(1):22, 2021.

[131] M. J. Culbertson. Bayesian networks in educational assessment: The state of the field. *Appl. Psych. Meas.*, 40(1):3–21, 2016.

[132] Y. Tseng, C. Yang, and B. Kuo. Using SVM to combine Bayesian networks for educational test data classification. *Int. J. Innov. Comput. I.*, 12(5):1679–1690, 2016.

[133] J. Martin and K. VanLehn. Student assessment using Bayesian nets. *Int. J. Hum-Comput. Int.*, 42(6):575–591, 1995.

[134] W. Xing, C. Li, G. Chen, X. Huang, J. Chao, J. Massicotte, and C. Xie. Automatic assessment of students' engineering design performance using a Bayesian network model. *J. Educ. Comput. Res.*, 59(2):230–256, 2021.

[135] J. C. Dunn. *Bayesian networks with expert elicitation as applicable to student retention in institutional research*. PhD thesis, Georgia State University, 2016.

[136] N. Misiunas, M. Raspopovic, K. Chandra, and A. Oztekin. Sensitivity of predictors in educational data: A Bayesian network model. In *2015 INFORMS Workshop on Data Mining and Analytics*.

[137] M. Al-Luhaybi, L. Yousefi, S. Swift, S. Counsell, and A. Tucker. Predicting academic performance: a bootstrapping approach for learning dynamic Bayesian networks. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 26–36. Springer, 2019.

[138] P. Arcuria. *Applying Academic Analytics Developing a Process for Utilizing Bayesian Networks to Predict Stopping Out Among Community College Students*. PhD thesis, Arizona State University, 2015.

[139] H. Dissanayake, D. Robinson, and O. Al-Azzam. Predictive modeling for student retention at St. Cloud State University. In *Proceedings of the International Conference on Data Science (ICDATA)*, page 215. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2016.

[140] M. Scutari. Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.*, 35(3):1–22, 2010.

[141] D. Margaritis et al. *Learning Bayesian network model structure from data*. PhD thesis, School of Computer Science, Carnegie Mellon University Pittsburgh, PA, USA, 2003.

[142] D. Heckerman. A tutorial on learning with Bayesian networks. In *Innovations in Bayesian networks: Theory and applications*, pages 33–82. Springer, 2008.

[143] R. Daly, Q. Shen, and S. Aitken. Learning Bayesian networks: approaches and issues. *Knowl. Eng. Rev.*, 26(2):99–157, 2011.

[144] M. A. Sanders and J. B. K. Advising in higher education. *Rad. Sci. Educ.*, 22(1), 2017.

[145] M. D. Hale, D. L. Graham, and D. M. Johnson. Are students more satisfied with academic advising when there is congruence between current and preferred advising styles? *Coll. Stud. J.*, 43(2):313–325, 2009.

[146] A. Anthony and M. Raney. Bayesian network analysis of computer science grade distributions. In *ACM Tech. Symp. Comp. Sci. Educ.*, pages 649–654, 2012.

[147] A. Dekhtyar, J. Goldsmith, H. Li, and B. Young. The Bayesian advisor project i: Modeling academic advising,. Technical report, University of Kentucky, 2001.

[148] J. Nissen, R. Donatello, and B. Van Dusen. Missing data and bias in physics education research: A case for using multiple imputation. *Phys. Rev. Phys. Educ. Res.*, 15(2):020106, 2019.

[149] A. Miles. Obtaining predictions from models fit to multiply imputed data. *Sociol. Method. Res.*, 45(1):175–185, 2016.

[150] N. Friedman. The Bayesian structural EM algorithm. *arXiv preprint arXiv:1301.7373*, 2013.

[151] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29:131–163, 1997.

[152] G. L. Heileman, C. T. Abdallah, A. Slim, and M. Hickman. Curricular Analytics: A framework for quantifying the impact of curricular reforms and pedagogical innovations. *arXiv preprint arXiv:1811.09676*, 2018.

[153] R. Molontay, N. Horváth, J. Bergmann, D. Szekrényes, and M. Szabó. Characterizing curriculum prerequisite networks by a student flow approach. *IEEE T. Learn. Technol.*, 13(3):491–501, 2020.

[154] N. W. Klingbeil and A. Bourne. The Wright State model for engineering mathematics education: Longitudinal impact on initially underprepared students. In *2015 ASEE Annual Conference & Exposition*, pages 26–1580, 2015.

[155] A. Slim, G. L. Heileman, C. T. Abdallah, A. Slim, and N. N. Sirhan. Restructuring curricular patterns using Bayesian networks. In *EDM*, 2021.

[156] A. Slim, H. Al Yusuf, N. Abbas, C. T. Abdallah, G. L. Heileman, and A. Slim. A Markov decision processes modeling for curricular analytics. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 415–421. IEEE, 2021.

[157] G. L. Heileman, W. G. Thompson-Arjona, O. Abar, and H. W Free. Does curricular complexity imply program quality? In *2019 ASEE Annual Conference & Exposition*, 2019.

[158] G. L. Heileman, H. W. Free, J. Flynn, C. Mackowiak, J. W. Jaromczyk, and C. T. Abdallah. Curricular complexity versus quality of computer science programs. *arXiv preprint arXiv:2006.06761*, 2020.

[159] D. Reeping, D. M. Grote, and D. B. Knight. Effects of large-scale programmatic change on electrical and computer engineering transfer student pathways. *IEEE T. Educ.*, 64(2):117–123, 2020.

[160] G. L. Heileman, C. T. Abdallah, and A. K. Koch. The transfer student experience: It's a lot like buying a used car. *arXiv preprint arXiv:2203.00610*, 2022.

[161] A. M. DeRocchis, L. E. Boucheron, M. Garcia, and S. J. Stochaj. Curricular complexity of student schedules compared to a canonical degree roadmap. In *2021 IEEE Frontiers in Education Conference (FIE)*, pages 1–5. IEEE, 2021.

[162] A. Slim, G. L. Heileman, M. Hickman, and C. T. Abdallah. A geometric distributed probabilistic model to predict graduation rates. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[163] US News & World Report: Education. US News and World Report, Washington, DC. `{h}ttps://premium.usnews.com/best-colleges`. Accessed 7/23/2022.

[164] J. Stewart, J. Hansen, and E. Burkholder. Visualizing and predicting the path to an undergraduate physics degree at two different institutions. *Phys. Rev. Phys. Educ. Res.*, 18:020117, Sep 2022.

[165] Curricular Analytics. https://curricularanalytics.org/. Accessed 5/30/2022.

[166] J. Nash, L. E. Boucheron, and S. J. Stochaj. A correlative analysis of course grades as related to curricular prerequisite structure and inter-class topic dependencies. In *2021 IEEE Frontiers in Education Conference (FIE)*, pages 1–5. IEEE, 2021.

[167] R. Chabay and B. Sherwood. Qualitative understanding and retention. *AAPT Announcer*, 27:96, 1997.

[168] The BEMA itself was never published in an archival journal. Early references to the instrument use Chabay and Sherwood (1997) (Ref 1) to cite the instrument. This issue of the AAPT Announcer is not available electronically. The citation references the program to the Summer 1997 American Association of Physics Teachers meeting then published in the Announcer. The page referenced contains Chabay and Sherwood's contributed talk abstracts about research applying the instrument. Interestingly, Maloney, O'Kuma, Van Heuvelen, and Hieggelke discussed challenges to developing an electricity and magnetism instrument in the same session which lead to the CSEM.

[169] Physport. https://www.physport.org. Accessed 8/8/2017.

[170] S. DeVore, J. Stewart, and G. Stewart. Examining the effects of testwiseness in conceptual physics evaluations. *Phys. Rev. Phys. Educ. Res.*, 12(2):020138, 2016.

[171] M. Planinic, L. Ivanjek, and A. Susac. Rasch model based analysis of the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 6:010103, Mar 2010.

[172] L. Ding. Applying Rasch theory to evaluate the construct validity of the Brief Electricity and Magnetism Assessment. In *2011 Physics Education Research Conference Proceedings*, volume 1413, pages 175–178, New York, 2012. AIP, AIP Publishing.

[173] L. Ding. Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis. *Phys. Rev. Phys. Educ. Res.*, 10:010105, Feb 2014.

[174] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell. Gender fairness within the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 14:010103, Jan 2018.

[175] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell. Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 14(2):020103, 2018.

[176] J. Wang and L. Bao. Analyzing Force Concept Inventory with Item Response Theory. *Am. J. Phys.*, 78(10):1064–1070, 2010.

[177] Y. Xiao, J.C. Fritchman, J.Y. Bao, Y. Nie, J. Han, J. Xiong, H. Xiao, and L. Bao. Linking and comparing short and full-length concept inventories of electricity and magnetism using item response theory. *Phys. Rev. Phys. Educ. Res.*, 15(2):020149, 2019.

[178] Youngsuk S. and Daniel B. Nested logit models for multiple-choice item response data. *Psychometrika*, 75(3):454–473, September 2010.

[179] J. Stewart, B. Drury, J. Wells, A. Adair, R. Henderson, Y. Ma, A. Pérez-Lemonche, and D. Pritchard. Examining the relation of correct knowledge and misconceptions using the nominal response model. *Phys. Rev. Phys. Educ. Res.*, 17(1):010122, 2021.

[180] T.F. Scott and D. Schumayer. Students' proficiency scores within multitrait item response theory. *Phys. Rev. Phys. Educ. Res.*, 11:020134, Nov 2015.

[181] J. Wells, R. Henderson, A. Traxler, P. Miller, and J. Stewart. Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis. *Phys. Rev. Phys. Educ. Res.*, 16:010121, April 2020.

[182] C. Wheatley, J. Wells, R. Henderson, and J. Stewart. Applying module analysis to the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 17:010102, Jan 2021.

[183] S. J. Pollock. Longitudinal study of student conceptual understanding in electricity and magnetism. *Phys. Rev. Phys. Educ. Res.*, 5(2):020110, 2009.

[184] M.A. Kohlmyer, M.D. Caballero, R. Catrambone, R.W. Chabay, L. Ding, M.P. Haugan, M.J. Marr, B.D. Sherwood, and M.F. Schatz. Tale of two curricula: The performance of 2000 students in introductory electromagnetism. *Phys. Rev. Phys. Educ. Res.*, 5(2):020105, 2009.

[185] M.W. McColgan, R.A. Finn, D.L. Broder, and G.E. Hassel. Assessing students' conceptual knowledge of electricity and magnetism. *Phys. Rev. Phys. Educ. Res.*, 13(2):020121, 2017.

[186] L. Ding and R. Beichner. Approaches to data analysis of multiple-choice questions. *Phys. Rev. Phys. Educ. Res.*, 5:020103, Sep 2009.

[187] S.J. Pollock. Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA. *AIP Conf. Proc.*, 1064:171–174, 2008.

[188] P. Eaton, K. Johnson, B. Frank, and S. Willoughby. Classical test theory and item response theory comparison of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 15(1):010102, 2019.

[189] P. Eaton, B. Frank, K. Johnson, and S. Willoughby. Comparing exploratory factor models of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 15(2):020133, 2019.

[190] A. Newell and H.A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[191] S. Ohlsson. The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. *J. Prob. Solving*, 5(1):7, 2012.

[192] J. Larkin, J. McDermott, D.P. Simon, and H.A. Simon. Expert and novice performance in solving physics problems. *Science*, 208(4450):1335–1342, 1980.

[193] J.H. Larkin, J. McDermott, D.P. Simon, and H.A. Simon. Models of competence in solving physics problems. *Cognitive Sci.*, 4(4):317–345, 1980.

[194] F. Reif and J.I. Heller. Knowledge structure and problem solving in physics. *Educ. Psychol.*, 17(2):102–127, 1982.

[195] US News & World Report: Education. US News and World Report, Washington, DC. `https://premium.usnews.com/best-colleges`. Accessed 4/30/2017.

[196] L. Chen, J. Han, J. Wang, and Y. Tu. Comparisons of Item Response Theory algorithms on Force Concept Inventory. *Res. Edu. As. Learn.*, 2(02):26–34, 2011.

[197] S. Osborn Popp, D. Meltzer, and M.C. Megowan-Romanowicz. Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics. In *2011 American Educational Research Association Conference*, Washington, DC, 2011. American Education Research Association.

[198] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig. Dividing the Force Concept Inventory into two equivalent half-length tests. *Phys. Rev. Phys. Educ. Res.*, 11:010112, May 2015.

[199] P. Eaton, K. Vavruska, and S. Willoughby. Exploring the preinstruction and postinstruction non-Newtonian world views as measured by the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 15:010123, Apr 2019.

[200] P. Eaton and S. Willoughby. Identifying a preinstruction to postinstruction factor model for the Force Concept Inventory within a multitrait item response theory framework. *Phys. Rev. Phys. Educ. Res.*, 16(1):010106, January 2020.

[201] T.I. Smith, K.J. Louis, B.J. Ricci IV, and N. Bendjilali. Quantitatively ranking incorrect responses to multiple-choice questions using item response theory. *Phys. Rev. Phys. Educ. Res.*, 16(1):010107, 2020.

[202] T.F. Scott, D. Schumayer, and A.R. Gray. Exploratory factor analysis of a Force Concept Inventory data set. *Phys. Rev. Phys. Educ. Res.*, 8(2):020105, 2012.

[203] L.J. Cronbach and P.E. Meehl. Construct validity in psychological tests. *Psychol. Bull.*, 52(4):281, 1955.

[204] L.A. Clark and D. Watson. Constructing validity: Basic issues in objective scale development. *Psychol. Assessment*, 7(3):309, 1995.

[205] R.P. Chalmers. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Soft.*, 48(6):1–29, 2012.

[206] L. Cai. A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4):581–612, 2010.

[207] A.E. Raftery. Bayesian model selection in social research. *Sociol. Methodol.*, 25:111–163, 1995.

[208] A. Maydeu-Olivares and H. Joe. Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71(4):713, 2006.

[209] A. Maydeu-Olivares. Goodness-of-fit assessment of item response theory models. *Measurement*, 11(3):71–101, 2013.

[210] R.E. Schumacker and R.G. Lomax. *A Beginner's Guide to Structural Equation*. Routledge, New York, NY, 2016.

[211] R.B. Kline. *Principles and practices of structural equation modeling, fourth edition*. Guilford Publications, New York, NY, 2016.

[212] L. Hu and P.M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Modeling*, 6(1):1–55, 1999.

[213] L. Crocker and J. Algina. *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, Mason, OH, 1986.

[214] J. Yang, J. Wells, R. Henderson, E. Christman, G. Stewart, and J. Stewart. Extending modified module analysis to include correct responses: Analysis of the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 16(1):010124, 2020.

[215] R.C. Hudson and F. Munley. Re-score the Force Concept Inventory! *Phys. Teach.*, 34(5):261, 1996.

[216] R.K. Thornton, D. Kuhl, K. Cummings, and J. Marx. Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 5(1):010105, 2009.