THE PAST, PRESENT, AND FUTURE OF GRADUATE ADMISSIONS IN PHYSICS

By

Nicholas T. Young

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics – Doctor of Philosophy
Computational Mathematics, Science, and Engineering – Dual Major

2021

ABSTRACT

THE PAST, PRESENT, AND FUTURE OF GRADUATE ADMISSIONS IN PHYSICS

By

Nicholas T. Young

While graduate admissions in physics directly affects only a small number of people on an annual basis, the number of people indirectly affected is many orders of magnitude greater. Those who complete graduate degrees in physics will go on to beyond leaders in industry, government, and academia, with the latter educating the next generation of leaders in science and engineering. Given the possibly enormous consequences of our decisions in physics graduate admissions, care should be taken to ensure that the process is working effectively. However, the evidence suggests otherwise. Many inequities exist in the admissions process, unfairly keeping potentially great scientists from even pursuing graduate studies. This thesis then seeks to understand what those inequities might be and how we might address them. First, I study the admissions process in the physics department at a Midwestern, public university using the random forest algorithm, a machine learning method, to understand what drives the admissions process. After finding that test scores and grades drive the process, I investigated whether one of those tests, the physics GRE, gives applicants an outsized advantage that it is claimed to provide, which it did not. Given the components that drove the admissions process contain inequities, the second half of the thesis explores whether a rubric-based holistic admissions process might be able to address those inequities. Preliminary evidence suggests that it does. Finally, to ensure that the methods used in the previous chapters were appropriate, the thesis concludes with a simulation study, finding that the methods used might lead to false negatives in the conclusion. Overall, this thesis suggests that the current graduate admissions process in physics contains inequities and that rubric-based admissions might be able to address them. By addressing those inequities, everyone can be given a fair shot in the admissions process and physics as a discipline can work toward becoming more representative of the population. Failure to act only perpetuates the inequities that have and will continue to keep people out of physics.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of too many folks to mention. However, I will attempt to do so here.

First, I would like to thank my dissertation committee for all the feedback and direction they have provided over the past few years in making this thesis as strong as possible.

Second, I would like to thank my advisor Danny Caballero who consistently provided support and challenged me to be the best researcher I could throughout this process. With Danny, whenever I brought up a new research idea or direction, his question was never 'why do you want to do that?', but 'how can I help you achieve it?' This thesis would not have been possible without his trust in allowing me to take this thesis in the directions I wanted.

Third, I would like to thank the many members of PERL who have provided feedback through group meetings, practice presentations, manuscript reviews, and general conversations and provided a sense of community. I'd especially like to thank Rachel Henderson who has served as my unofficial co-advisor for providing insights and new directions that helped me grow as a researcher.

Additionally, I'd like to thank the many members outside of PERL who have also provided feedback and inspiration that appear in this thesis, especially Devin Silvia whose data visualization class had a significant impact on the figures appearing in this thesis and Odd Petter Sand whose own data visualizations inspired those that appear in the introduction.

Fourth, I would like thank the many people who provided and curated the data that appears in this thesis including Scott Pratt, Kirsten Tollefson, and Remco Zegers for providing data about Michigan State's graduate program and Julie Posselt and Casey Miller for providing data collected by the Inclusive Graduate Education Network. In addition, I'd like to thank Nicole Verboncoeur and Tabitha Hudson who spent many hours of their summer reading through applications to extract the necessary data.

Fifth, I would like to thank Kim Crosslan. My experience in the graduate program would not have gone as smoothly as it did without all of the support and assistance Kim provided. Regardless

of what issue I encountered or problem I faced, the solution was always only an email or phone call to Kim away.

Sixth, I would like to thank my family for all their support over the years, both prior to and throughout graduate school.

Seventh, I would like to thank the world's best two dogs, Kali and Xavier. Over the course of doing this thesis, they have (literally) been by my side providing emotional support in between asking for belly rubs and play time.

Finally, I would like to thank my partner, Sarah. This thesis would not have been possible without her. From the daily support she provided to her feedback and suggestions, she has helped me grow into a better person and a better researcher. I'm excited to see where our journey will go next.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE PAST, PRESENT, AND FUTURE OF GRADUATE ADMISSIONS IN PHYSICS: AN OVERVIEW

People like us who believe in physics know that the distinction between past, present, and future is only a stubbornly persistent illusion -Albert Einstein

1.1 Establishing the need to study graduate admissions in physics

At first glance, studying graduate admissions in physics may seem like a small, trivial problem. After all, only a relatively small number of people are directly affected by the physics graduate admissions process. Using the number of test-takers of a commonly required applicant exam in physics, the physics GRE, as a proxy for the number of applicants, approximately 7,000 students apply to physics graduate programs annually [1]. In comparison, first-year physics courses that are often the focus of physics education research [2] enroll nearly 425,000 annually [3].

However, the number of people indirectly affected by physics graduate admissions is orders of magnitude larger. The applicants who are admitted to programs will go onto to become leaders in academia, industry, and government in fields as diverse as energy, technology, national defense, and medicine [4]. In addition, some of the admitted applicants will go on to beyond faculty who will train the next generation of scientists, engineers, medical doctors, and science teachers.

Furthermore, graduate admissions has economic consequences for both students and taxpayers. Applicants who are admitted and earn their PhD have higher average salaries than those with only a bachelor's degree [5], meaning that success in graduate admissions influences earning potential later in life. In addition, as many graduate students are indirectly supported by tax payers via grants awarded by governmental agencies, departments have a duty to use the tax payer money wisely by admitting applicants who will be successful in their programs. When taking into account tuition, stipend, and overhead, training a single graduate student can cost tax payers between a quarter and half a million dollars. For the department of physics at Michigan State, the estimated cost is

$80,000 a year per graduate student. However, if admitted students are not supported throughout their programs, neither applicants nor tax payers will see the benefits that can be afforded through graduate study.

Yet, given the potentially large impacts of graduate admissions in physics, not everyone is given a fair chance in the process. Physics remains largely white and male even as the United States population becomes increasingly less white and higher education is becoming less male-dominated [6, 7]. To stay in touch with the demographics of the country and the broader economy, physics as a discipline needs to reevaluate who is allowed to participate. Failing to do so risks physics losing out on the diversity of opinion and perspectives needed to advance as a discipline and do so ethically.

Graduate admissions is but one small part of that process. In this thesis, I will explore the historical approaches to graduate admissions in physics as well as offer a possible route toward achieving those goals of diversity and equity in the process.

1.2 How this thesis contributes to the field of physics education research

Unfortunately, there does not exist a consensus about the subfields of physics education research (PER) and various attempts have been made over years, both broadly [2, 8–10] and for specific populations and topics [11–13]. My work fits most naturally into Russ and Odden's framework [2] so I will map onto that.

Russ and Odden classified education research along seven dimensions: disciple, phenomenon, population, context, methods/methodology, theoretical/conceptual framework, and epistemology. This thesis contributes to PER in the dimensions of population, context, and methods/methodology so I will focus on those. A visual representation is shown in Fig. 1.1.

First, this thesis contributes to PER by focusing on graduate students and the graduate admissions process. Traditionally, the population of PER studies has been undergraduate students in introductory physics courses [2, 13]. More recently however, PER has expanded to focus on upper-division undergraduate students [14–19], non-physics majors [20, 21], graduate students [22–26],

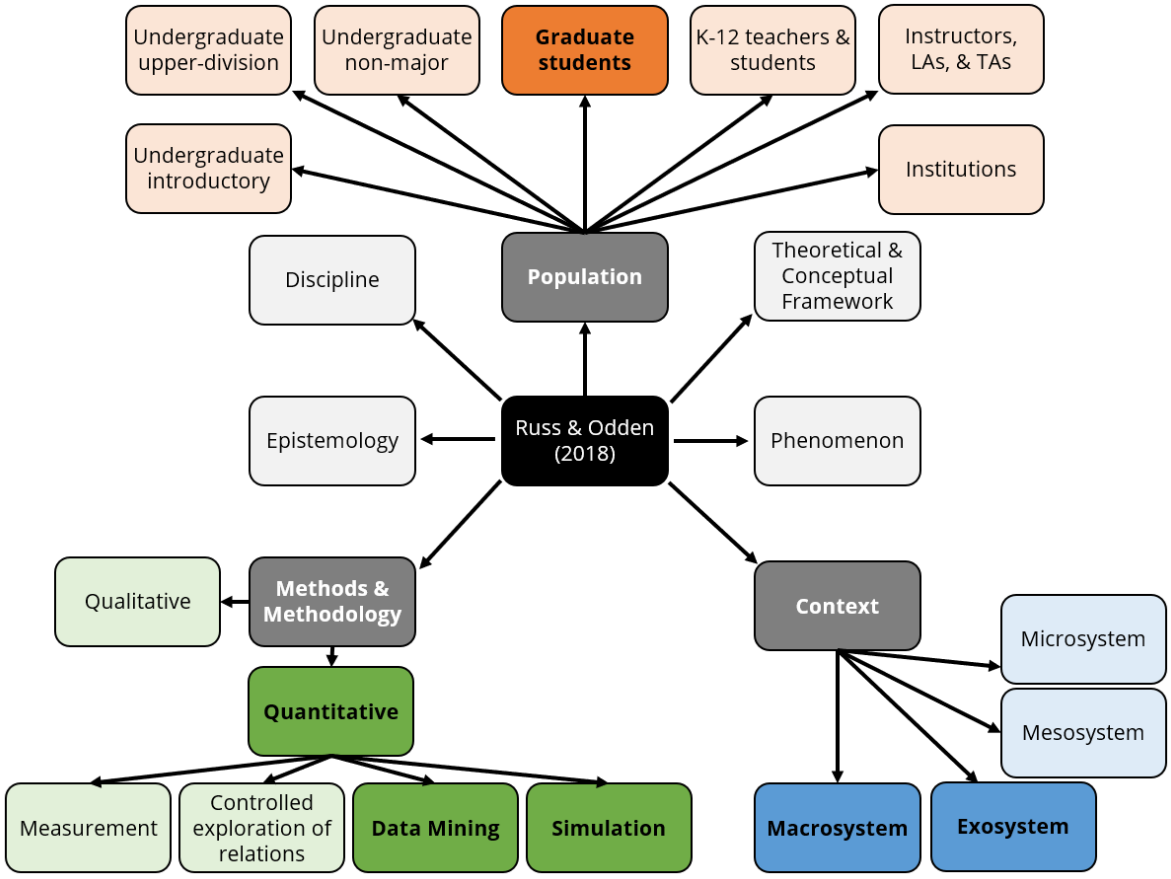

Figure 1.1: Visual representation of the framework presented in Ross and Odden [2] with methods and methodology broken down according to Ding's genres [12]. Dimensions that are in bold and expanded represent areas this thesis advances in PER based on their framework, including population, context, and methods and methodology.

K-12 teachers and students [27–33], instructors, teaching assistants, and learning assistants [34–42], and even institutions themselves [43,44]. This thesis then adds to the growing body of work regarding graduate students in physics. Studying graduate admissions specifically has only become an area of inquiry recently, with most of the relevant research published within the last five years [45–56].

Second, this thesis contributes to PER by focusing on contexts that have received less attention in the literature. When thinking about context, Russ and Odden leverage Bronfenbrenner's ecological systems theory [57] to describe the environment that the research study takes place. Ecological systems theory envisions individuals existing in a set of systems that describe the contexts and interactions that may inform the development of the individual. For example, the microsystem is the

individual's direct environment, the mesosystem is connections across microsystems, the exosystem is the indirect environment of the individual and the macrosystem is the societal norms and cultural values relevant to the individual. Given that PER has traditionally focused on undergraduate students in introductory physics classes, its context is typically the microsystem [2]. Thinking in terms of Nair's mapping of ecological systems theory onto a physics classroom [58], graduate admissions can be thought of as an exosystem and macrosystem phenomenon. Students do not actively participate in the process but are clearly affected by it (exosystem) and cultural norms about who can do sciences as well as beliefs about what is required to do physics (e.g. innate brilliance [59]) might affect faculty's decisions (macrosystem).

Finally, this thesis contributes to PER by using data analysis techniques that have only recently been incorporated into PER studies such as machine learning, Tomek Links, and simulation. Traditionally, PER has been divided into qualitative research and quantitative research [2]. Ding [12] then subdivided quantitative research into three genres: measurement, controlled exploration of relations, and data mining, which is the predominant genre in this thesis. As Romero and Ventura [60] and Cope and Kalantzis [61] note, data mining relies on data that has already been collected and hence, the research quesions that can be addressed may be limited.

However, instead of thinking of genres of quantitative research, we can think in terms of the specific methods used. Russ and Odden claim that basic statistical techniques and large-N summary and frequency analyses are the most common in PER with network analysis and methods for large data set analysis becoming increasingly common [2]. To this, I will add the umbrella term of modeling in which researchers try to develop some quantitative representation of the phenomenon of interest.

Modeling can then be further broken down into two broad categories: explanatory modeling and predictive modeling [62]. While explanatory modeling is a traditional PER method (Theobald et al. [63] provides an overview of these methods), predictive models are an emerging area, often using machine learning techniques. Machine learning, a method used throughout this thesis, especially is an emerging area of PER, with most studies utilizing it having been published in the last five

years [64–69].

Finally, as a result of modeling methodologies, a new genre of quantitative PER has started to emerge: simulation. In this genre, researchers create artificial data to study the methods of PER themselves. Simulation studies are still rare in PER, with only a few such studies published to date [70–72]. This thesis increases that number by one.

1.3 How this thesis contributes to computational mathematics, science, and engineering

As computational mathematics, science, and engineering is unique to Michigan State University, a research-based framework to describe the types of work conducted under the umbrella of the department does not exist. Instead, I will use the "triple junction" of computation on which it was formed [73].

The department defines "triple junction" of computation as algorithm development and analysis, high performance computing, and applications to scientific and engineering modeling and data science. This thesis focuses on the final of those.

Data science techniques such as machine learning, feature engineering, and simulation have seen limited use in physics education research. This thesis then brings data science tools into a new field, adding additional tools to the physics education researcher's tool kit. In addition, the interdisciplinary nature of this thesis of combining computational techniques to answer educational questions aligns with the broader goals of the department.

1.4 Summaries of the remaining chapters

This thesis considers the past, present, and future of graduate admissions. Chapters 2 and 3 focus on the past and present, Chapters 4 and 5 focus on the present and future, and Chapter 6 extends across dimensions of time, seeking to understand the limitations of models used in the previous chapters and to be used in future analyses.

In Chapter 2, I analyze 4 years of admissions records to Michigan State University's graduate physics program using a machine learning method known as random forest. I find that consistent

with surveys of admissions committees and observations of committees, quantitative parts of the application such as GRE scores and undergraduate GPA hold the most weight in the process. In fact, knowing only the applicant's undergraduate GPA, physics GRE score, and quantitative GRE score was sufficient to predict with 75% accuracy whether an applicant would be admitted.

In Chapter 3, I then explore the physics GRE in more depth and show that a common argument for keeping the physics GRE in the admissions process despite its documented issues, that it helps applicants who might be otherwise missed "stand out," is not supported by evidence. To reach this conclusion, I analyzed admissions records from five universities with both large-N frequency analysis and mediation and moderation models. Whether I defined applicant who might be missed as having a low GPA, graduating from a less selective undergraduate school, or graduating from a smaller undergraduate program did not affect the conclusion.

Given that the traditional methods of admissions have not resulted in substantial changes in the demographics of physics at the graduate level and those methods contain inequities in and of themselves, graduates admissions do not need to be revised, but rather rethought. Chapters 4 and 5 provide one such approach: rubric based holistic admissions.

In chapter 4, I introduce rubric-based holistic admissions, defining them as an approach to admissions where reviewers consider a broad range of applicant characteristics such as academic achievement, research experience, fit with the program, test scores, and noncognitive competencies, and rate applicants on those categories according to a pre-determined scoring rubric. I then consider the department of physics at Michigan State University's revised admissions process as an example in practice. I compared the distribution of faculty ratings by admission status, sex, and undergraduate background. The results indicate that the rubric does not show any unexpected inequities based on the applicant's sex or undergraduate background. It did however detect systematic issues such as test score differences and differing service-work expectations based on sex.

In chapter 5, I consider whether our rubric-based holistic admissions is actually a rethinking of graduate admissions or just a revision. That is, does switching from the traditional admissions to

rubric-based admissions fundamentally change the process. I again use the random forest method to analyze the admissions data after the implementation of the rubric as well as introduce a new technique to PER, Tomek Links. I first compare the two admissions processes using the same data extracted from applications and then consider what additional insights might exist by looking at the rubric-ratings data. Across four sets of analysis, the results suggest that rubric-based holistic admissions is a rethinking of graduate admissions, though additional work is needed to provide greater confidence in the results.

Finally, in chapter 6, I consider the modeling techniques I've used in the previous chapters, in addition to others, and how they perform under various data distributions encountered in PER and the previous chapters. Across the techniques, I find that the more a binary variable is imbalanced (e.g. 50/50 split vs 80/20), the less likely a technique is to find that variable predictive or explanative of an outcome. I then show that the effect is also found in real PER data by focusing on the admissions processes at 3 institutions.

1.5 Key conclusions in this thesis

The results of the chapters then suggest three overarching conclusions.

First, the traditional graduate admissions process in physics is metrics heavy and dominated by the physics GRE, whose value in the admissions process should be questioned. (Chapters 2 and 3)

Second, rubric-based admissions might offer a possible path forward in terms of making the process more equitable. While the components of the rubric seem to be equitable, we were not able to produce definitive evidence that rubric-based admissions sufficiently departed from the traditional admissions structure. However, the current results are promising. (Chapters 4 and 5)

Finally, modeling techniques in PER have biases and may affect what features are counted as statistically significant or predictive. If these modeling techniques are going to be used more broadly to make educational policy decisions (e.g. in graduate admissions), we need to really understand the models, including how they work and their limitations and caveats. (Chapter 6)

1.6 Key recommendations as a result of this thesis

As a result of work done in this thesis, I propose six key recommendations, three targeted toward departments and graduate admissions committees and three targeted toward researchers.

1.6.1 Recommendations for departments and graduate admissions committees

First, if the physics GRE is being used to identity applicants who might otherwise be missed, I recommend against that. My study in chapter 3 did not find evidence that the physics GRE allows that to happen in practice.

Second, departments should rethink their graduate admissions process in terms of what applicants are evaluated on, how those are evaluated, and who does the evaluating. Rubric-based admissions, introduced in chapter 4, seems to offer one way to do so. However, implementing rubric-based holistic admissions requires the department to take an active role in rethinking their process. Departments must decide how to address those three points as answers will depend on specifics of the department.

Finally, departments should engage in regular self-study of their processes and share the results so that the physics community has an idea of what works and what does not work. Currently, data about admissions practices is either reported in aggregate or individually for a limited number of programs. Greater reporting from many programs will allow for a better picture of how admissions processes are conducted and how they might be made more equitable.

1.6.2 Recommendations for researchers

First, researchers should be transparent about the data that goes into their models and how the distribution of those features may affect the results. For researchers and practitioners to evaluate the conclusions presented in papers, they need to understand the data itself. My work in chapter 6 showed the split of a binary feature can affect whether an algorithm finds it statistically significant or predictive of an outcome. As a result, it is possible that the literature contains false negatives,

where potentially important variables were missed.

Second, researchers should consider more modern techniques for analyzing data such as machine learning and penalized regression. Machine learning techniques are becoming more common in PER but are still a niched area. Penalized regression techniques have seen limited use in PER but appear to handle data as well as if not better than standard logistic regression. In addition, researchers should follow the recommendation of Aiken et al. and compare different models to produce the best fitting model [74].

Finally, researchers should conduct more simulation studies to examine the methods used in PER. Data in PER is often a mix of binary, categorical, and continuous features and hence, algorithms developed in fields that analyze primarily continuous data may not perform as expected. Simulation studies would be able to verify if this is the case and if researchers should be concerned.

1.7 Questions remaining unanswered

While this thesis extends the field's understanding of graduate admissions in physics, many questions are unanswered, specifically around rubric-based admissions and equity in graduate admissions, providing an avenue for future work.

First, in order to reach a more definitive conclusion as to whether our department's admissions process, future work could consider alternative approaches to analyzing the data such as mixed methods. On the quantitative side, these alternative approaches could be more traditional methods in PER like logistic regression or clustering-type methods. The latter may be able to tease out whether there are different "types" of successful applicants and possibly provide evidence as to whether the process became more holistic.

On the qualitative side, future work could include interviewing faculty on the admissions committee or observing deliberations in real time. Such data would provide greater insight into the process, especially regarding individual applicants. Quantitative methods try to summarize and simplify the data, causing us to potentially miss data-rich discussions of applicants. Qualitative methods might allow us to see these cases, especially discussions around borderline applicants who

might expose what faculty are really valuing in an applicant.

Second, in order to understand how the results may generalize, future work should extend the analysis done in this thesis to other physics graduate programs. Additionally, future work should consider programs in different geographical areas with different applicant pools and of different research intensities. Michigan State University is a predominantly white institution with a highly-ranked subdiscipline and hence, the applicant population might not be representative of the broader population who applies to physics graduate school.

Third, future work should examine other aspects of the graduate admissions process such as who is invited and able to apply to graduate school. Prior work has examined barriers current graduate students experienced when they applied [49]. However, studies such as these ignore students who wanted to apply but for one reason or another, did not. Therefore, future work should explore the barriers these students face in applying and how departments might address them. These might include supports students need in applying, how students find out about programs, and how departments advertise their graduate programs.

Fourth, future work should continue along the path outlined above and consider undergraduate physics students more broadly in graduate admissions. The evaluation of applications is the final step in the graduate admissions process that could be argued to begin as soon as a student declares a physics major. Between those two steps, potential applicants must decide whether they are interested in attending graduate school, research potential programs and advisors, complete an application, and submit the application. Even after departments have evaluated the applications and made offers, students must decide which program to attend or what to do if they are not accepted to any. All of these are ripe for future study.

Finally, to truly make an impact on diversity and equity in physics longer-term, future work needs to extend beyond just the admissions process and consider graduate school as a whole. Departments need to address diversity by examining who is encouraged and invited to apply to graduate school, equity by considering how they evaluate applicants applying to their program and current students in their program and what they are doing to retain the students they did admit, and inclusion by

intentionally addressing the climate in their department, being transparent with decisions being made, and creating support structures for students from underrepresented backgrounds. Simply making the admissions process more equitable and admitting more diverse students will not make an impact if corresponding efforts are not made to retain such students and prevent them from being pushed out. Possible areas of future work include studying qualifying and comprehensive exams, student-advisor relationships, mental health, and departmental support for students. Only when diverse students are not only actively welcomed to the academy but also actively retained will real change happen.

CHAPTER 2

IN THE BEGINNING: USING MACHINE LEARNING TO UNDERSTAND PHYSICS GRADUATE SCHOOL ADMISSIONS

The following chapter is adapted and expanded from its published version in the 2019 Physics Education Research Conference [75]. The published version includes Marcos D. Caballero as the second author. Following the Contributor Roles Taxonomy (CRediT) [76], my roles for this project include conceptualization, formal analysis, methodology, software, validation, visualization, and writing the original draft.

2.1 Introduction

Despite other science, technology, engineering, and mathematics (STEM) fields becoming more diverse over the past few decades, physics has lagged behind with only 20% of bachelor's degrees awarded to women and only 11% awarded to racial minorities [77]. These numbers do not improve when considering graduate degrees where 20% of doctoral degrees are granted to women and 7% are granted to racial minorities [77]. While this underrepresentation has both enrollment and retention causes, this chapter will focus on the factors that may affect enrollment in physics graduate programs.

When considering enrollment in physics PhD programs, prior work has found that minority students in physics are less likely to apply to programs if they feel that they will not be admitted based on low GPA or GRE scores, or lack of research experience [49]. Further, given that many graduate programs have application fees, financial concerns might prevent students from applying to graduate programs that they believe they will not be admitted to. Therefore, it is important to understand what matters when applying to physics graduate programs while acknowledging that many factors that are not easily quantifiable matter.

Previous research into graduate admissions in physics has tended to take a broad approach, characterizing the graduate admissions process across the United States, both for master's and PhD

programs [47, 50] or focusing on a specific subset of universities such as elite universities [46]. These studies find that faculty consider numerical measures such as undergraduate GPA and GRE scores most important in the admissions process and have been conducted by either observing the admissions process or by surveying faculty about what they believe to be most important in the admissions process. More recent work in physics graduate admissions has explored applicant perceptions of the various components of the application [51]. Missing in this analysis is an investigation of the actual applications of prospective physics graduate students. To our knowledge, there has only been one such study [54].

Given that applications to graduate programs consist of numerical data such as GPA and GRE scores, categorical data such as gender, race, and ethnicity, and open-ended data such as letters of recommendation and personal statements, graduate admissions is an ideal target for machine learning. Indeed, machine learning approaches to understanding graduate admissions have been employed in computer science to study self-reported admissions data [78] and to streamline the review process [79]. Machine learning methods have also been employed more broadly in higher education admissions to predict which admitted students will accept an offer to attend a small liberal arts school [80, 81] and to predict which students are likely to be admitted and to complete their MBA [82].

The goal of this work is to further the study of graduate admissions in physics by analyzing the applications using a machine learning approach. Specifically, we ask what features of an application to this physics graduate program are predictive of admission.

Unlike other studies in physics graduate admissions, this work represents a case study of a single institution rather than a broad look at the graduate admissions landscape. However, because physics is regarded as a high consensus discipline, that is, there is large agreement about what counts as legitimate admissions practices [83], we expect our results can generalize to similar doctoral programs.

2.2 Methods

2.2.1 Data

The data used in this study comes from the admissions records of 512 domestic applicants to the physics and astronomy graduate program at Michigan State University between 2013 and 2016 and would have enrolled between fall 2014 and fall 2017. Domestic and international applicants do not undergo the same review process and hence we only analyze applications from domestic students. Here, domestic student is defined to be a U.S. citizen or permanent resident. The admissions process is unique at this university in that the applications are not only reviewed by a central committee but also members of the subdiscipline in which the student expresses interest. The data include the applicant's undergraduate institution and grade point average (GPA), their general and physics GRE scores, and their physics subdiscipline of interest. Per a ballot initiative in the state of Michigan, Michigan State University and the other Michigan public universities are explicitly prohibited from discriminating against or granting preferential treatment to individuals based on race, sex, color, ethnicity, or national origin in education [84]. To comply with this law, our university's admissions system collects limited demographic data and our department chose not to record the information that was available when evaluating applicants. As such, demographics are not available to us. Overall, 48% of the domestic applicants were offered admission into the program.

2.2.2 Describing Undergraduate Institutions

Because the name of the undergraduate institution in itself does not provide useful information to an algorithm, we created new factors to describe characteristics of the institutions. To describe the overall institution, we classified each institution as public or private, whether it is a minority serving institution (MSI), the region of the country it is located in (such as Northeast, Southwest, etc.), and the Barron's selectivity of the institution, which describes how selective the undergraduate program is. We assume that selectivity serves as a proxy for prestige. Classifications for the first three categories were taken from the most recent Carnegie Rankings [85] while the Barron's

Table 2.1: Variables used in our model and their scale of measurement

Factor	Measurement Scale
Undergraduate GPA	Continuous
Verbal GRE score	Continuous
Quantitative GRE score	Continuous
Written GRE score	Continuous
Physics GRE score	Continuous
Proposed research area	Categorical
Application year	Categorical
Barron's selectivity	Categorical
Region of applicant's undergraduate institution	Categorical
Type of physics program at applicant's undergraduate institution	Categorical
Size of undergraduate physics program at applicant's undergraduate institution	Categorical
Size of doctoral physics program at applicant's undergraduate institution	Categorical
Applicant attended a minority serving institution	Binary
Public or Private	Binary
Output variable: admitted status	Binary

classification came from Barron's *Profiles of American Colleges*. Because the overall reputation of the applicant's undergraduate university might not describe the physics program at that university, we also included factors related to the physics program such as the highest physics degree offered at the university and the size of the undergraduate program and PhD program if applicable. The size of the undergraduate and PhD programs were determined by the median number of graduates of the program between the 2012-2013 and 2015-2016 academic years (i.e. the years that applicants applied to the program). The programs were then classified as small, medium-small, medium-large, or large based on which quartile they fell into. We used the Roster of Physics Departments with Enrollment and Degree Data to collect this data [86–89]. All factors appearing in our model are shown in Table 2.1 and include the scale of measurement.

2.2.3 Justifying our choices of institutional factors

Prior work has documented university pedigree is often considered in the application process because institutional quality is assumed to be a proxy for student quality [46, 90]. Here, we

measure institutional quality by Barron's selectivity and public or private status, with the assumption that physics faculty view private universities as more prestigious than public universities. We include region of the applicant's undergraduate university to account for the fact that the institution being studied is a public university and might therefore show a preference for students from the surrounding region.

Prior work has also found faculty exhibit a tendency to admit students like themselves, though it is more common among academics who graduated from elite institutions [46]. Therefore, it is not unreasonable to expect that faculty may prefer to admit students who followed similar paths as they did, meaning students from large, doctoral institutions might be more likely to be admitted than students from smaller institutions. Additionally, we use the size of the undergraduate and PhD programs as proxies for the perceived prestige of the physics department, assuming a more prestigious physics department attracts more students and hence graduates more students.

2.2.4 Random Forest Model

To analyze our data, we used the conditional inference forest algorithm, a variant of the random forest algorithm [91] shown to be less biased when the data includes both continuous and categorical variables [92] such as those used in our model (see Table 2.1). Random forest models in general are ensembles of individual decision trees, which use binary splits of the input features in order to make a prediction. The predictions are then averaged over the individual trees to obtain the overall prediction of the random forest.

While there are multiple metrics used to assess random forest and other machine learning models, two of the most common are the accuracy and the area under the curve (AUC). The accuracy is simply the proportion of correct predictions made by the model. To ensure that the accuracy isn't inflated by overtraining, only a fraction of the available data is used to construct the model while the rest is used to test the predictive power. It is this remaining data that is used to calculate the accuracy of the model.

The AUC is defined as the area beneath the receiver operator curve of the model, which

visualizes the false positive rate against the true positive rate and varies between 0.5 and 1, with values greater than 0.7 signifying an acceptable model [93]. The area describes the proportion of positive cases that are ranked above negative cases in the data set by the model. For example, for our data, the AUC would represent the proportion of all random pairs of admitted and not-admitted applicants in which the admitted applicant is classified as admitted and the not-admitted applicant is classified as not-admitted.

In addition to making predictions, the random forest algorithm can determine the importance of each feature to the model, referred to as the feature importance. For this analysis, we use two importance measures. First we used the AUC permutation feature importance [94] as it is claimed to be less biased than the accuracy based permutation importance when input features differ in scale (as do our factors listed in Table 2.1) and when the predicted variable is not split evenly between the two outcomes. Under this approach, each feature is randomly permuted and then passed through the model to make a prediction. The AUC is then recorded and the difference between this value and the original AUC is computed. As permuting a feature with more predictive information should result is a worse model than permuting a feature with less predictive information, a larger difference between the original AUC and the AUC with a permuted feature suggests that that feature contained more predictive information. These differences can then be used to create a relative ordering of features.

However, if the features are correlated, it is possible that the orderings may be biased or that permutations of one feature might result in unrealistic combinations of features and hence would cause the model to extrapolate performance [95]. For example, if all students who earned perfect scores on the physics GRE also had high GPAs, permuting GPA could cause there to be cases where a perfect physics GRE score goes with a low GPA, which would be outside of the region learned by the model. To prevent that, a conditional importance measure has been proposed in which features are permuted within a subset of similar cases [96]. Because of the correlations between various sections of the GRE, we also used this conditional approach to compute feature importances.

Feature importances are derived from the data and hence, are not assumed to follow any

statistical distribution. Therefore, there is no simple way to apply the idea of statistical significance to feature importances, though Chapter 6 provides some suggestions. We instead applied the recursive backward elimination technique described in Díaz-Uriarte and Alvarez de Andrés [97] to determine which features are predictive of admission and which are not. When using this technique, the features are ordered according to their importance. A model is then built using all the features and the accuracy is computed. A set fraction of the features with the smallest importances are then removed and a new model is built and the accuracy computed. This process continues until only 2 features are left. The model with the fewest number of features while maintaining an accuracy within a standard error of the highest accuracy across all models built in this process is then the selected model. We will refer to the features used in this selected model as the *meaningful* features and interpret them as the features that are predictive of the outcome. For more information about random forest models, biases, and feature importance measures, see APPENDIX.

To perform the analysis, we used R [98] and the `party` package [92,96,99] to create a conditional inference forest model. We used 70% of our data to train the model, 500 trees to build our forest and used $\sqrt{p}$ as the number of randomly selected features to use to build each tree, with p being the total number of features in the model. These values follow recommendations of Svetnik et al. [100]. We ran our model 30 times, randomly selecting 70% of our data for training each time, and averaged the feature importances over runs so that the resulting distribution of individual feature importances would be approximately normal. As the conditional inference forest algorithm has routines built in to handle missing data [101], applicants with missing information were not removed from the data set. However, the conditional importance approach requires there to be no missing values so we used the MICE algorithm [102] to fill impute the missing data in that case, following Nissen et al.'s recommendation for PER [71]. The imputation results were pooled using Rubin's Rules [103].

In addition, to determine if our model was dependent on our choice of hyperparameters, we also varied the fraction of data to train the model, the number of trees in the forest, and the number of randomly selected features to use to build each tree. We set the training fraction to be either 0.5, 0.6, 0.7, 0.8, or 0.9, the number of trees in the forest to be 50, 100, 500, 1000, or 5000, and

the number of features used for each tree to be 1, $\sqrt{p}$, $p/3$, $p/2$, or p for a total of 125 possible combinations (124 new and the original model). These choices are based off findings in Svetnik et al. [100]: namely that the error rates level off once the number of trees is on the order of 10^2 and their choices of the number of features in each tree. In addition, increasing the training fraction may improve performance as there is more data for the model to learn from. For each combination, we repeated the procedure in the previous paragraph. Due to the computational cost of the conditional permutation approach, we only calculated the AUC-permutation importance.

To determine if the changing the hyperparameters affected our models, we computed the minimum, median, and maximum value of each metric over the 125 hyperparameter combinations and relative ordering of the features in each model. We chose the minimum, median, and maximum instead of the mean and standard error because 1) we are looking across different models rather than getting repeated measurements of the same things so we cannot assume the results will be normally distributed and 2) we are interested in the best and worst performance achieved under hyperparameter tuning to get a sense of the possible values we can achieve which wouldn't be possible using the mean and standard error. If our model is largely unaffected by the choice of hyperparameters, we would expect the metrics to show minimal variation and the relative ordering of the features to be largely unchanged.

2.3 Results

Across the 30 runs, the average accuracy of our model predicting on the held-out data was 75.6% ± 0.6%, the average training AUC was 0.849 ± 0.002, and the average testing AUC was $0.756 \pm .006$. As our model's accuracy is significantly higher than the null accuracy of 52.7%, the percent of students who were not accepted, and our testing AUC is above 0.7, our model can be considered an acceptable model of the data.

The feature importances averaged over the 30 runs are shown in Fig. 2.1. We find numerical factors such as the applicant's score on the physics GRE, the applicant's score on the quantitative GRE, the applicant's undergraduate GPA, the applicant's verbal GRE score, and their proposed

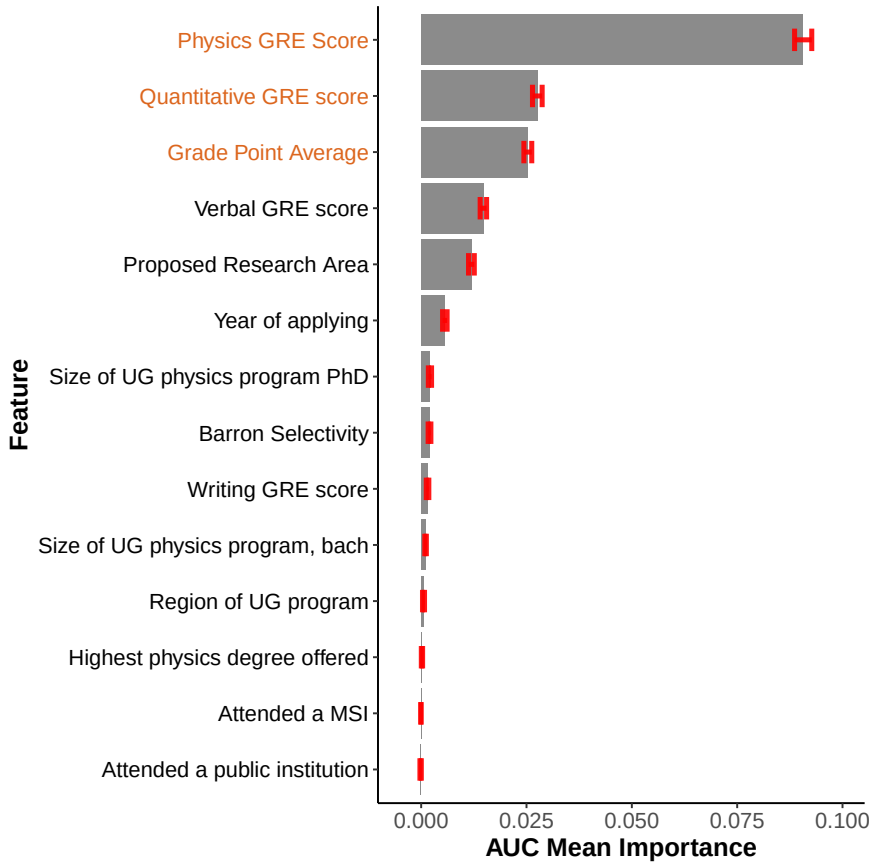

Figure 2.1: Averaged AUC feature importances over 30 trials. Physics GRE score, Quantitative GRE score, and undergraduate GPA, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted.

research area to be more important in the application process than any factor describing the applicant's undergraduate institution. Using recursive backward elimination to determine the meaningful factors, we find the applicant's physics GRE score, quantitative GRE score, and their undergraduate GPA to be the only meaningful factors.

To verify that the applicant's physics GRE score, quantitative GRE score, and undergraduate GPA were indeed the only meaningful factors, we then reran our random forest model 30 times using only these three factors as the predictors. Our average testing accuracy was then $75.4\% \pm 0.6\%$ and our testing average area under the curve was 0.754 ± 0.006, which are not statistically different from the values we found using all fourteen factors shown in table 2.1.

When we instead used MICE and the conditional importances, and the metrics were slightly

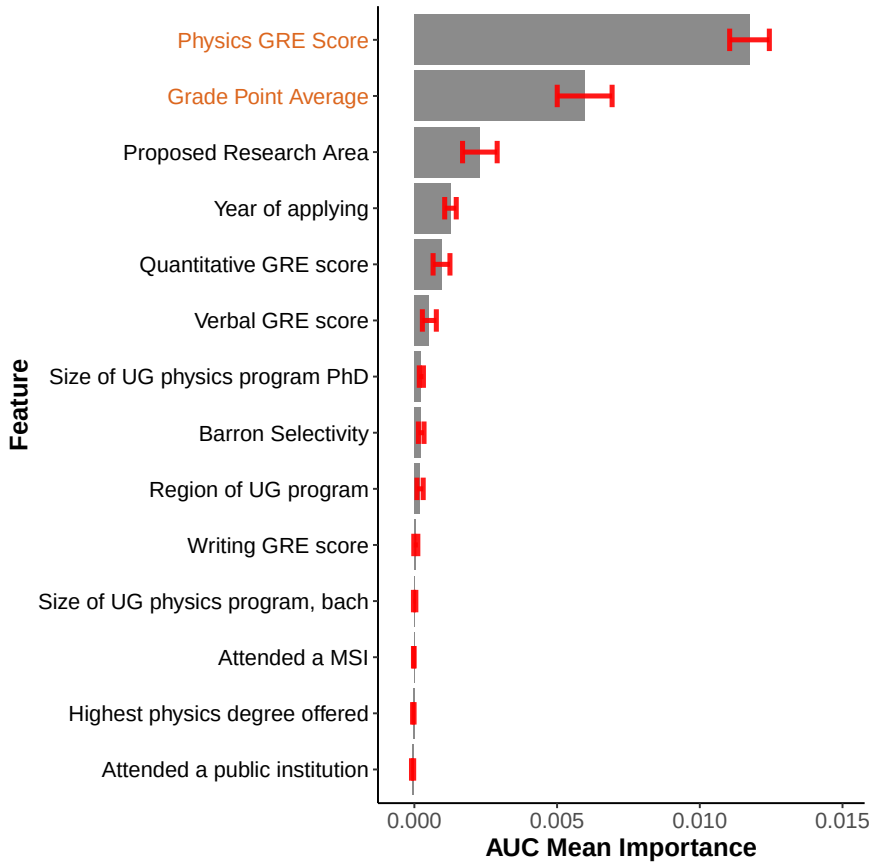

Figure 2.2: Averaged conditional feature importances over 30 trials. Physics GRE score and undergraduate GPA, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted when adjusting for correlations among the features.

higher, likely because the imputing the missing values provided more data for the algorithm to learn from. Specifically, the testing accuracy was $77.1\% \pm 0.1\%$ and the testing AUC was 0.770 ± 0.001.

The conditional feature importances are shown in Fig. 2.2. Compared to Fig. 2.1, we notice that the verbal and quantitative GRE scores are ranked lower than they were when we did not take correlations into account and proposed research area and year of applying are ranked higher than when we did not take correlations into account. The physics GRE and GPA are still rankly highly however, even after taking correlation into account.

Performing the recursive backward elimination, we find that physics GRE score and GPA are meaningful features and quantitative GRE score no longer is. Using only these two features to create a conditional inference forest on the imputed data, we find that the testing accuracy is $75.7\% \pm 0.7\%$

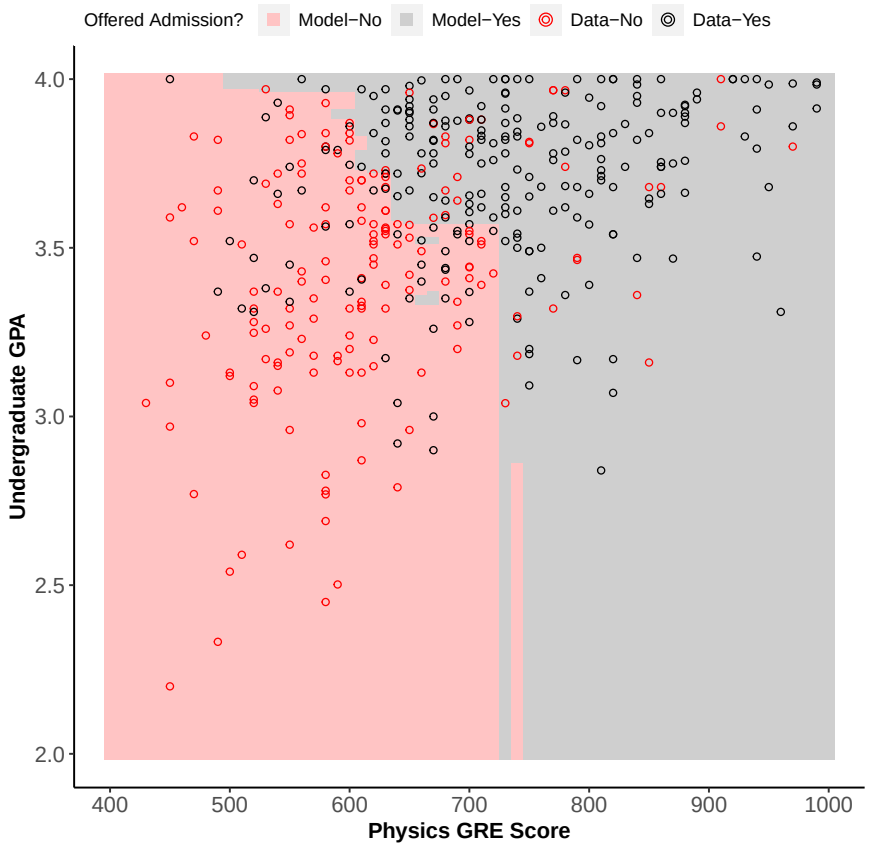

Figure 2.3: Plot of all applicant's physics GRE scores vs their undergraduate GPAs. The background coloring expresses the prediction of the model had an applicant had that score.

and the testing AUC is 0.757 ± 0.007, which are consistent with the full model.

As there are only two meaningful features, we can plot the features and see if there does appear to be a separation between admitted and not admitted applicants. To do so we generated all possible pairs of undergraduate GPA and physics GRE scores and ran them through our model to find the predicted admissions decision. The result is shown in Fig 2.3. We see that there does appear to be a boundary between admitted and non-admitted students around a physics GRE score of 700, which drops toward 650 for applicants with GPAs above 3.5, providing further evidence that physics GRE score and GPA are predictive of admission at this program.

When we test the various hyperparameter combinations, we find similar results. Looking at the metrics (Table 2.2), we see that the testing accuracy varies by 3.3 percentage points between the minimum and maximum values and the testing AUC varies by 0.034 between the minimum and

Table 2.2: Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations

metric	min	median	max
Train AUC	0.824	0.848	0.853
Test AUC	0.726	0.749	0.760
Test Accuracy	0.727	0.750	0.760
Null Accuracy	0.521	0.527	0.556

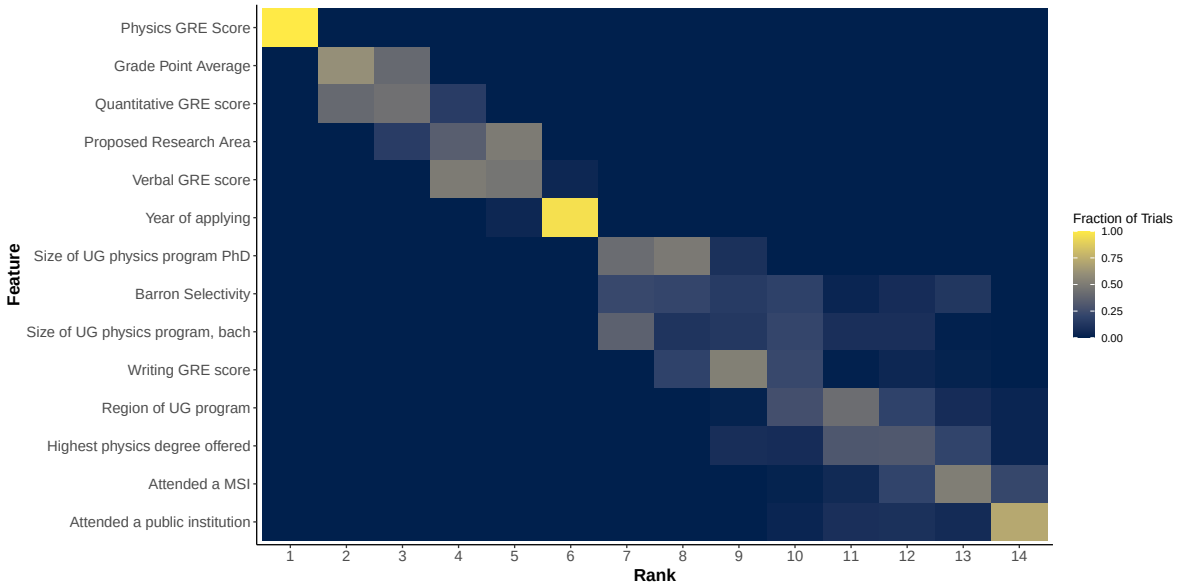

Figure 2.4: Proportion of the 125 hyperparameter combinations in which each feature had a given rank. Notice that there is a block of features that range between 1 and 5 and a block of features that rank between 7 and 14.

maximum values. As the variation is limited and these metrics are still within the acceptable range, the results suggest that our choice of hyperparameters has limited impact on the metrics.

When we look at the ranks of the features used each hyperparameter combination, we also see limited variation. In Fig. 2.4, we notice the plot is mostly diagonal and the presence of two blocks. First, we see that physics GRE score, GPA, quantitative and verbal GRE scores, and proposed research area are always the top five features, regardless of the hyperparameters. Second, we see that the institutional features never rank above a 7, meaning that no combination of hyperparameters can create a model where these features are predictive of admission.

Looking at the first block, we notice that physics GRE is always the top ranked feature followed by either GPA or quantitative GRE score, with GPA being the more common selection. Furthermore,

GPA never ranks lower than third while the quantitative GRE score ranks between second and fourth. For certain choices of hyperparameters, the applicant's proposed area of research ranks higher than the quantitative GRE score.

2.4 Discussion

Perhaps unsurprisingly, we find numerical measures are the most important factors for determining whether a domestic applicant will be accepted into this physics graduate program, consistent with findings that graduate programs with a large number of applicants use numerical measures as a first pass to evaluate applicants [46, 47]. Looking across our analyses, we find physics GRE score and undergraduate GPA are consistently found to be predictive of admission while quantitative GRE score is sometimes found to be predictive based on the choice of hyperparameters. However, once we take correlations among the features into account, the quantitative GRE score is no longer found to be predictive.

While we find no evidence of a minimum physics GRE score, we do find evidence of a "rough cutoff" as described in Potvin et al. around 700 [47]. Nevertheless, some students who scored significantly above this threshold were not admitted. While we do not know the reasons why these students were not admitted, Posselt noted that faculty might not admit superior applicants if they do not believe the applicant will actually enroll in their program [46].

Overall, our findings of the meaningful factors for admission to a physics graduate program are consistent with Potvin et al.'s findings obtained by surveying physics graduate admissions directors. Notably, we also find that the physics GRE score, undergraduate GPA, quantitative GRE score, and proposed research area are more important than other factors while the undergraduate institution, GRE written score, and proximity/familiarity are less important factors.

While the verbal GRE score was not found to be a meaningful feature, the program studied here appears to place more emphasis on it than the average program. This may be because our study only looked at domestic students while Potvin et al.'s looked at all applicants. Because international students also take the TOEFL while domestic students do not and admissions directors ranked the

TOEFL as more important the verbal GRE, the TOEFL may take the place of the verbal GRE and hence lower the perceived value of the verbal GRE relative to other factors.

Despite prior work suggesting institutional characteristics play an important role in graduate admissions, we did not find institutional or departmental characteristics to be meaningful to our model. Our result could be due to differences in methodology or due to institutional effects being influential but not dominant factors [104]. Indeed, Posselt suggests institutional factors might be used to differentiate applicants with similar GPAs and GRE scores [46]. Therefore, we might not have found institutional factors to be meaningful because they are used when primary factors such as GPA and physics GRE scores do not sufficiently separate applicants.

While we did not have access to other criterion included in the Potvin et al. such as application essays, research experiences, and recommendation letters, we were still able to create a model that correctly predicted whether an applicant would be admitted with approximately 75% accuracy based solely off the applicant's undergraduate GPA and physics GRE score (and highly higher if we also included the quantitative GRE score). While undergraduate GPA is a significant predictor of completing a physics PhD, physics GRE is not, as those scoring near the top of the physics GRE only have a 7% higher probability of completing their PhD than those scoring near the bottom [52]. As the GRE is not associated with completing a doctoral degree and is known to favor persons from majoritized groups in science [52, 105], the outsized role of the GRE in the admissions process should be questioned. Indeed, the American Association of Physics Teachers and American Astronomical Society have released recommendations against using the physics GRE in graduate admissions [106, 107].

2.5 Limitations

There are a few limitations to our study. First, the data we used to make our model was not all the data that would be available to a faculty member evaluating an application. In addition, our model did not contain demographic information about the applicants that could also impact the results given the barriers women and people of color face in physics. Therefore, it is possible

that meaningful features other than GPA and physics GRE score could lie in the data that was unavailable to us.

Second, this study was done at a primarily white institution (PWI). While Kanim and Cid note that having a relatively homogeneous research sample can be valuable for reducing variability, especially in early studies, they also note that exploring the effects of variability can lead to new results and a greater understanding of the results [13]. Thus, while our result might generalize to many physics graduate programs, it might also hide important differences in features predictive of admission for applicants of different demographics groups and institutions with different demographics than our own.

2.6 Future Work and Conclusion

Our work adds to the broader literature about graduate admissions and the process by which applicants are judged. Because minoritized students might not apply to graduate programs if they do not think they will be accepted, elucidating the factors that determine whether an applicant will be accepted is crucial. Simply increasing the number of applicants from currently and historically underrepresented groups in physics will not increase their representation unless corresponding efforts are made to admit these students. While our result that test scores and GPA are the most predictive parts of an application in terms of admission aligns with prior work, these results represent only one institution and might not be representative of all United States physics PhD programs. Given the unique structure of the admissions process at this university, graduate programs with a more traditional admissions process might assign different weights to the various parts of an application. Therefore, future work should investigate what features of the applicant drives the admissions process at other institutions.

Furthermore, our university has recently moved to a rubric-based admissions format, designed to take into account non-cognitive competencies and program fit in addition to the more traditional admissions criteria such as GPA and GRE scores. Our future work will examine how including these new criteria may change the factors that are most predictive of an applicant being admitted to

the program. The results of such analyses are presented in Chapters 4 and 5.

CHAPTER 3

FURTHER EVIDENCE AGAINST THE PHYSICS GRE: IT DOES NOT HELP APPLICANTS "STAND OUT"

The following chapter was published in Physical Review Physics Education Research in 2021 [108]. The published version includes Marcos D. Caballero as second author. Following the Contributor Roles Taxonomy (CRediT) [76], my roles for this project include conceptualization, formal analysis, methodology, software, validation, visualization, and writing the original draft.

3.1 Introduction

While applying to graduate programs requires many components, perhaps none is as scrutinized as the Graduate Records Exam (GRE), and in physics, the physics GRE. Indeed, research into graduate admissions in physics suggests that the physics GRE is one of the most important components of the applications for determining which applicants will be admitted, based on both student and faculty perspectives [47, 51] and analysis of the admissions process [46, 75]. Despite its prominence in the admissions process, the physics GRE is known to be biased against women and people of color in physics [105], resulting in lower average scores compared to white and Asian males. At least one in three programs use a cutoff score [47], with 700 being a common choice [52], meaning applicants from groups already underrepresented in physics graduate programs can be further marginalized as they are less likely to achieve these scores. This is in addition to the observation that many physics students of color already see the GRE as a barrier to applying to graduate school [49, 55, 109].

Further, the physics GRE might not even be useful for determining which applicants will be successful in graduate school. For example, Miller et al. suggest that the physics GRE is not useful for predicting which applicants will earn their PhDs [52]. Additionally, Levesque et al. argue that using the common 50th percentile cutoff score for the physics GRE would have caused admissions committees to reject nearly 30% of students who would later receive a national prize postdoctoral fellowship, which can be viewed as a proxy for research excellence [110]. Yet

despite evidence suggesting the physics GRE does not predict these typical ways of measuring "success" in graduate school and calls from the American Astronomical Society and the American Association of Physics Teachers to eliminate the physics GRE from admissions [106, 107], most physics graduate programs still require applicants to submit their physics GRE scores. Currently, nearly 90% of physics and astronomy graduate programs still accept the physics GRE, with over half requiring or recommending submitting a score [106]. Of those that do not accept physics GRE scores from applicants, all of the programs are solely astronomy graduate programs or joint physics and astronomy graduate programs. While it is uncertain where removing the physics GRE affects any measure of graduate school success (e.g. completion rate), initial work by Lopez suggests that removing the physics GRE does increase the diversity of applicants [111].

Given these documented issues with the physics GRE, why do departments continue to use it? First, given that many programs are seeing a larger number of applicants, the physics GRE provides a quick way to filter the applications down to a more reasonable number for faculty review. Unlike in undergraduate admissions, graduate admissions tend to be decentralized and done at the departmental level by a faculty committee. Hence, faculty are asked to review applications in addition to their regular teaching and research duties and thus, might not have the time to read the letters of recommendation and applicant essays for every applicant.

Second, some faculty view GRE scores as measures of innate intelligence [46, 48] or ability to become a PhD-level scientist [105]. After all, they and other faculty likely had high GRE scores in order to be admitted to graduate school, and may exhibit a survivorship bias, believing that a high GRE score is needed to succeed. Further, physics is seen as a "brilliance-required" field, where innate intelligence is required for success [59].

A third argument, and the most interesting one in terms of the scope of this paper, is that standardized tests such as the physics GRE can help students stand out [112]. The ETS, the creator of the GRE and physics GRE, claims that subject GREs "can help you stand out from other applicants by emphasizing your knowledge and skill level in a specific area" [113]. For example, a student with an average grade point average (GPA) might be able to stand out from other applicants

if they did exceptionally well on the physics GRE.

In addition, applicants from smaller universities or universities that are not known to the admissions committee might benefit from performing well on a standardized measure. For example, the ETS claims that the GRE provides a "common, objective measure to help programs compare students from different backgrounds" [114] and physics admissions committees worry that removing the GRE would limit their ability to compare applicants from different backgrounds [115]. Anecdotally, some faculty claim that a good physics GRE score could aid students from small liberal arts colleges in the admissions process [116].

We already know that GPAs are interpreted in context of the applicant's university. Posselt has shown that among more prestigious graduate programs, the applicant's GPA is viewed in the context of their undergraduate institution with high GPAs from prestigious institutions seen favorably, low GPAs from an unknown school as unfavorably, and high GPAs from unknown schools and middle GPAs from prestigious institutions in the middle [117]. Therefore, a standardized test such as the physics GRE could provide an assumed equal comparison for an admissions committee and might allow the applicant from an unknown school to stand out or have a similar chance of admission as an applicant from a more well-known school.

Finally, graduate admissions have been documented to be "risk-adverse," where admissions committees select applicants most likely to complete their program [46, 48]. As applicants from smaller universities may be judged based on how previously enrolled students from their university did in the program [117], a risk adverse admissions committee might be less likely to admit applicants from small universities whose students have previously struggled in their program. However, perhaps a high standardized test score could overcome these perceptions and signal that the applicant might indeed be successful in the program.

Our goal then is to focus on the third argument. Does the physics GRE help applicants "stand out" in the admissions process in practice? If that is the case, we would expect those disadvantaged in the admissions process, those who have low GPAs, attended a smaller institution, or identify as part of a group currently underrepresented in physics, to be admitted at similar rates as their more

advantaged peers with similar physics GRE scores. Specifically, we ask:

1. How does an applicant's physics GRE score and undergraduate GPA affect their probability of admission?

2. How are these probabilities of admission affected by an applicant's undergraduate institution, gender, and race?

As Small points out in his critique of admissions and standardized test studies [118], multiple variables rather than just a standardized test might best explain our results and therefore, a framework that allows for substitutions and trade-offs between variables is necessary. Therefore, we ask an additional research question:

3. How might the above relationships be accounted for through mediating and moderating relationships?

This paper is organized as follows: Sec. 3.2 provides an overview of mediation and moderation analysis. We then describe our data, how we determined what constitutes "standing out," and how we implemented mediation and moderation analysis in Sec. 3.3. In Sec. 3.4, we describe our findings and in Sec. 3.5, we use those findings to answer our research questions and explain our limitations and choices which may affect our results. Finally, we describe our future work in Sec. 3.6 and the implications of our work for graduate admissions in physics in Sec. 3.7.

3.2 Background

Before we can answer the third research question, it is important to describe what we mean by mediating and moderating relationships.

In a mediating relationship, two variables are only related because they are also related to some common third variable. For example, a student who played video games the night before an exam might do poorly because they stayed up playing video games too late and did not get enough sleep. Therefore, video games and doing poorly on the exam are only related due the common factor of lack of sleep. Lack of sleep is then a mediating variable.

In a moderating relationship, the strength of the relationship between two variables depends on some third variable. For example, the relationship between someone liking dogs and owning a dog likely depends on whether they are allergic to dogs. That is, we would expect someone who likes dogs but is allergic to dogs is less likely to own a dog than someone who likes dogs but is not allergic to dogs is. Being allergic to dogs is then a moderating variable.

Mathematically, suppose that some input X has an effect on output Y. We would say that some other input M mediates the relationship between X and Y if X only has an effect on Y because X has an effect on M and M has an effect on Y [119]. For a simple case, we can represent these relationships as

$$Y = i_1 + cX \tag{3.1}$$

$$M = i_2 + aX \tag{3.2}$$

$$Y = i_3 + c'X + bM \tag{3.3}$$

where i represents the intercepts. These relationships are visually shown in Fig. 3.1.

Using this representation, the direct effect of X on Y is represented by c' and the indirect effect is represented by ab. The total effect is then $c' + ab$ which for a linear regression models, is equal to c. Equivalently, in the case the linear regression, the indirect effect is $c - c'$.

However, if Y is binary, linear regression is not appropriate and logistic regression should be used instead. In this case, Rijnhart et al. recommend using ab as the indirect effect as their simulation studies found the ab estimate of the indirect effect exhibited less bias than the $c - c'$ estimate [120].

To determine if the indirect effect is statistically significant, a common approach is to use a Sobel test. However, simulations suggest that the Sobel test is underpowered and that bootstrapping is a good alternative [121]. Specifically, those simulations find that using the percentiles of a bootstrapped estimate of the indirect effect to estimate the confidence interval is a good compromise

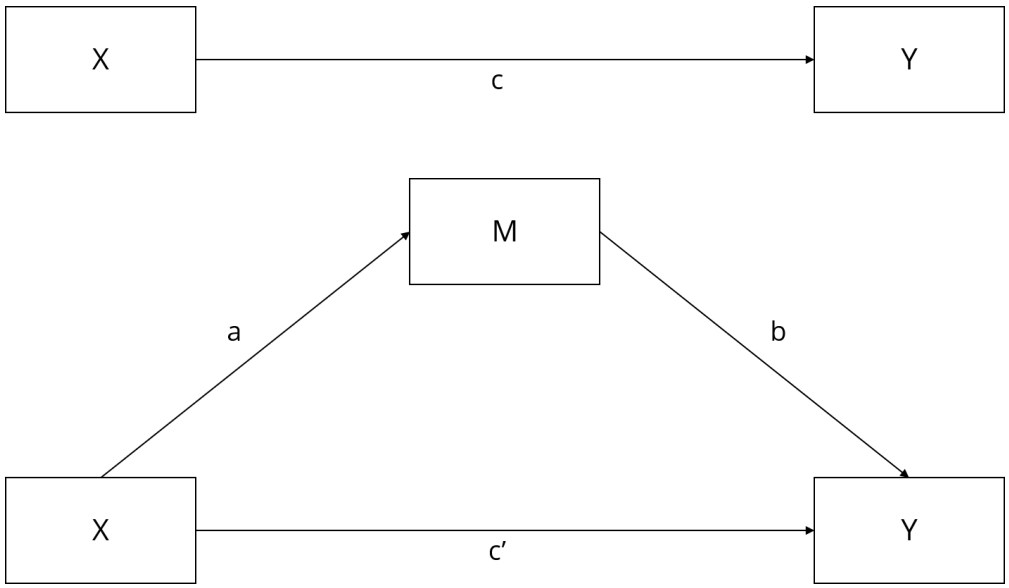

Figure 3.1: Visual representation of eqs. (3.1) to (3.3). The top graphic shows eq. (3.1) while the bottom graphic shows eqs. (3.2) and (3.3).

between avoiding type I errors while maintaining statistical power. In their approach (which has also been used in PER studies before, e.g., [122]), if ab is different than zero, then there is some degree of mediation.

More specifically, there are three cases.

1. If $ab \neq 0$ and $c' = 0$ then M fully mediates the relationship between X and Y.

2. If $ab \neq 0$ and $c' \neq 0$, then M partially mediates the relationship between X and Y. In that case, we can estimate the amount of mediation as the fraction of the total effect attributed to the indirect effect, $\frac{ab}{ab+c'}$ [123, 124].

3. If $ab = 0$, then M does not mediate the relationship between X and Y.

This approach can also be adapted to multiple mediators and these mediators can be predictors of other mediators. An example of this serial mediation case with two mediators is shown in Fig. 3.2. Equations (3.1) to (3.3) can then be modified to be

$$Y = i_4 + cX \tag{3.4}$$

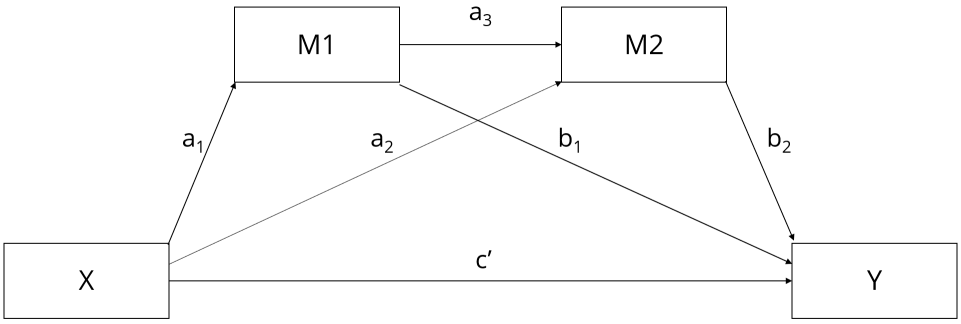

Figure 3.2: Visual representation of eqs. (3.5) to (3.7) showing serial mediation with two mediators.

$$M_1 = i_5 + a_1 X \tag{3.5}$$

$$M_2 = i_6 + a_2 X + a_3 M_1 \tag{3.6}$$

$$Y = i_7 + c' X + b_1 M_1 + b_2 M_2 \tag{3.7}$$

In this case, there are three indirect effects. First, there are the indirect effects of the mediators individually, $a_1 b_1$ and $a_2 b_2$, and second, there is the indirect effect of the mediators together $a_1 a_3 b_2$. The total indirect effect is then $a_1 b_1 + a_2 b_2 + a_1 a_3 b_2$ [125].

More generally, for N mediators, we can generate $N + 1$ equations where the first N are of the form

$$M_n = i_n + a_n X + \sum_{j=1}^{n-1} a_j M_j \tag{3.8}$$

and the final equation is of the form

$$Y = i_y + c' X + \sum_{j=1}^{N} b_j M_j \tag{3.9}$$

So far, we've assumed that the relationship between the mediator M and the output Y does not depend on any other variables. However, it is possible that the relationship between M and Y could also depend on X or some other variable, meaning there is a conditional indirect effect (see Preacher et al. [126]). In the case that the relationship between M and Y depends on X, we would say that X moderates the relationship between M and Y. Practically, this means we must add an interaction term to eq. (3.3), which then becomes [126]

$$\begin{aligned} Y &= i_3 + c'X + b_1 M + b'_1 XM \\ &= i_3 + c'X + (b_1 + b'_1 X)M \end{aligned} \tag{3.10}$$

We use the prime on b coefficients to denote an interaction coefficient for a mediator while an unprimed b coefficient is a coefficient of a mediator.

The conditional indirect effect is then $a(b_1 + b'_1 X)$. If $b'_1 = 0$, we would say that there is no moderation and the indirect effect is the standard ab.

In the case that there are multiple mediators, eq. 3.10 can be modified to include multiple mediators and interaction terms for all pairs of variables where moderation may be of interest.

In the special case that X is binary, eq. (3.10) reduces to $Y = i_{x=0} + b_1 M$ when $X = 0$ and $Y = i_{X=1} + (b_1 + b'_1)M$ when $X = 1$. Therefore, to test if there is moderation, we can simply regress M on Y given $X = 0$ and again given $X = 1$ and subtract the slopes to calculate b'_1 instead of including an interaction term in the model.

3.3 Methods

3.3.1 Data

Data for this study comes from the physics departments at five selective, research-intensive, primarily white universities. Four of these universities are public and part of the Big Ten Academic Alliance while the remaining university is a private Midwestern university. During the 2017-2018 and 2018-2019 academic years, graduate admissions committees at these five universities recorded all physics applicants' undergraduate GPA, GRE scores, undergraduate institution, and

demographic information such as gender, race, and domestic status. In addition, the universities recorded whether each applicant made the shortlist, was offered admission, and whether the applicant decided to enroll. Because our study includes all applicants rather than only admitted applicants, we are unlikely to suffer from the range restrictions noted in critiques of other admissions studies (e.g. [118, 127]). However, we do address a possible range restriction in the Limitations and Researcher Decisions section (sec. 3.5.2).

Due to different requirements and admissions processes for international students and domestic students (e.g. international students need to submit a test of English proficiency), we only include domestic students in our study. We then remove any applicant for whom a physics GRE and GPA were not recorded, leaving us with 2537 applicants. While we in theory could use multiple imputations to address the data as Nissen et al. recommends [71], faculty reviewing the applications do not, to our knowledge, do this and hence, we would be creating data that wasn't available in the admissions process. Distributions and analysis of the remaining physics GRE scores and GPAs appear in the appendix.

As the applicant's undergraduate university does not contain meaning in itself, we needed to categorize the institutions. We chose to categorize the institutions by their size and their selectivity. We then used the number of physics bachelor's degrees awarded per year as measure of the size of the university. We assume that universities with more graduates are more well known and hence, would likely be known to the admissions committees. In contrast, universities that produce fewer bachelor's degrees might not be known to the admissions committees and hence, might be unknown programs. It would then be these applicants from "smaller" programs who might need to "stand out." We acknowledge that some programs that produce a small number of physics bachelor's degrees each year might not be unknown to the admissions committees due to previous applicants from such schools or research collaborations or partnerships. However, there is no way in our data to know if this is the case.

To determine whether a university should be counted as a "small university", we used the undergraduate institution names to look up the number of typical physics bachelor's degrees from

Table 3.1: Summary of the comparisons we analyzed, which group needs to stand out and which does not, and the figure number showing the results

Variable	Group that tends to be privileged in admissions	Group that tends not to be privileged in admissions	Figure
Program size	Applicants from physics programs that rank in top 25% of programs based on yearly graduates	Applicants from physics programs that rank in bottom 75% of programs based on yearly graduates	Fig. 3.5
University selectivity	Applicants from universities ranked as most selective or highly selectively (Barron's Value of 1 or 2)	Applicants from any other university (Barron's Value of 3 or lower)	Fig. 3.6
Gender	Male applicants	Female applicants	Fig. 3.7
Race	Asian or white applicants	Black, Latinx, Multiracial, and Native applicants	Fig. 3.8

AIP's public degree data [128, 129]. As of this writing, degree data for the 2018-2019 academic year was not available, so we used data from the 2016-2017 and 2017-2018 academic years to quantify the number of bachelor's degrees. Additionally, this would have been the most recent data available when admissions committees would have reviewed applications and many of the applicants would be represented in the data as bachelor degree recipients

To account for the institution's prestige, we used Barron's Selectivity Index [130]. Barron's selectivity index is a measure based on the undergraduate acceptance rate of an institution as well as characteristics of its undergraduate incoming classes, such as mean SAT scores, high school GPAs, and class rank. We assume selectivity is a proxy for prestige as prestigious institutions tend to have low acceptance rates and high SAT scores and GPAs from incoming students. In contrast to the AIP data, Barron's selectivity index applies to the institution as a whole rather than only the physics department.

3.3.2 Probability of admission procedure

Determining whether an applicant is more or less likely to be admitted first requires computing admissions probabilities. To do so, we grouped applicants based on their GPAs and physics GRE scores. Prior work has found that the physics GRE score and undergraduate GPA are two of the

most important aspects of the applications [46, 47, 75]. Our previous work specifically found that the physics GRE score and undergraduate GPA were able to predict with 75% accuracy whether an applicant would be admitted to one public Midwestern physics graduate program.

In addition, physics is a "high consensus" discipline, meaning most programs agree on what consists of a successful applicant [46]. Therefore, despite many other components of the applications that affect whether an applicant will be admitted, we believe using the physics GRE score and undergraduate GPA provides a first-order overview of what admissions committees would use to admit applicants.

In order to ensure a reasonable number of applicants in each group to do meaningful analysis, we grouped applicants into bins based on their GPA and physics GRE score. We choose to use GPA bins 0.1 units in width and physics GRE bins 50 points in width. The GPA bins were selected to ensure that that GPAs with the same tenth digit were in a single bin. That is, 3.50 through 3.59 would be in a single bin. All GPAs were already reported on the 4.0 scale and physics GRE scores were reported using the standard 200-990 scale so we did not need to do any conversions.

We then computed the fraction of applicants in each bin who were admitted to the program they applied. As we are interested in applicants "standing out," we frame our results as whether applicants in a bin are admitted at a higher rate than the overall rate (all accepted applicants divided by all applicants). If applicants are admitted at a higher rate than the overall rate, it suggests that these applicants did in fact stand out to the admissions committee.

In our framing of "standing out," we are assuming that graduate admissions operate under a deficit model. That is, due to their privilege, some applicants had better resources, opportunities, or choices available to them and as a result, may appear as better candidates for the program compared to applicants who did not have those available to them. To our knowledge, those with less privilege and/or resources are not directly compared to those with more privilege and/or resources but instead compared to an ideal applicant who often resembles someone from a more privileged background. For example, Owens et al. found that faculty valued advanced course knowledge and programming skills in incoming graduate students [53], which may be more characteristic of applicants coming

from better resourced institutions. Using this framing, we created four groups of applicants who might or might not need to stand out, which are summarized in Table 3.1 and explained in detail below.

To take into account the size of the institution, we first used the AIP data to determine the national quartile each applicant's institution ranked in terms of all bachelor's degree recipients for each of the two years of data. Because not all institutions reported data in both years and the number of graduates could vary significantly between years, we conducted separate analyses first with the highest quartile an institution reached in the two years and second with the lowest quartile the program reached in the two years. For example, if an institution was ranked in the 3rd quartile the first year and the 4th quartile in the second year, our first analysis would use the 4th quartile and our second analysis would use the 3rd quartile. We then define the large programs as those in the 4th quartile and small programs as those in the 1st through 3rd quartiles. We address this choice in the discussion.

When using Barron's Selectivity Index to take into account the selectivity of the institution, we used Chetty et al.'s [131] five groupings (Ivy League +, remaining most selective institutions, highly selective institutions, selective institutions, and non-selective institutions) as a guide. As there was a single applicant from a non-selective institution, selective and non-selective were grouped into a single category. Because we are interested in smaller, less known programs compared to larger, well-known programs, we took the selective and non-selective group to be our "less selective institution" group and institutions in the first three of Chetty et al.'s categories as our "most selective institutions". This corresponds to grouping institutions with a Barron's Index of 1 and 2 together as the "most selective institutions" and all other values together as the "less selective institutions".

To understand how high physics GRE scores might help applicants identifying as part of a group currently underrepresented in physics, we compared women's admission probability to men's admission probability and applicants of color's admission probability to applicants not of color's admission probability. While it should be noted that gender is not binary [132], the data

Table 3.2: Counts of applicants by gender and race who provided both GPAs and physics GRE scores

	Race							
Gender	Asian	Black	Latinx	Multi	Native	White	Unreported	Total
Men	247	49	99	166	4	1410	112	2087
Women	56	2	19	26	0	308	28	439
Unreported	1	0	0	1	0	5	4	11
Total	304	51	118	193	4	1723	144	2537

the admissions committee recorded is only in terms of the male and female binary and hence, we cannot comment on how high physics GRE scores may impact applicants of other genders.

Furthermore, given the limited number of applicants identifying as part of a racial group underrepresented in physics, we combined all Black, Latinx, Multiracial, and Native applicants into a single category, which we will refer to as B/L/M/N following the recommendation of Williams [133]. We acknowledge that this may obscure important distinctions between groups, as Teranishi [134] and Williams suggest. We also acknowledge applicants identifying as a marginalized gender and race may face additional barriers and hence could stand out differently than an applicant identifying as either a marginalized gender or race. However, there are less than 50 applicants ($\sim$ 2% of the sample) identifying as a member of both a marginalized gender and marginalized race, limiting statistical power for analysis. Full demographics are shown in Table 3.2. For information about how race and ethnicity categories were constructed and standardized, see Posselt et al. [54] who previously used the 2017-2018 academic year application data from this study in their study.

3.3.3 Mediation and Moderation Procedure

Given that to some degree, both the physics GRE score and undergraduate GPA measure physics knowledge, we expect that these two measures will be correlated with each other. Therefore, we first tested whether the physics GRE has any mediating effects when predicting admission and whether GPA moderates the relationship between the physics GRE and admission; that is, is one only related to admission because it influences the other and that one influences admission or is the strength of

the relationship between one and admission affected by the other. Because admissions status is a binary outcome variable, we need to use logistic regression for eqs. (3.1), (3.3) and (3.10).

When taking an applicant's GPA and physics GRE score into account, we first centered and scaled both variables so they both have means of zero and variances of 1. As we are treating GPA and physics GRE scores as continuous, we can use linear regression for eq. (3.2).

To estimate the coefficients in eqs. (3.1) to (3.3) and (3.10), we generated 5000 bootstrap samples with replacement as was done in Hayes and Scharkow [121]. For each trial, we computed the indirect effect ab. To get the estimate of each parameter, we took the average of the 5000 bootstraps. To get the lower end of the 95% confidence interval, we used the value that corresponded to the 2.5th percentile of the values generated by the bootstrap. Likewise, to get the upper end of the 95% confidence interval, we used the value that corresponded to the 97.5th percentile.

For the institutional features, we treat institutional selectivity and institution size as binary input variables (most selective or less selective and larger institution or smaller institution) and for demographic features, we treat gender and race as binary variables. Again, we use B/L/M/N as one category for race and white and Asian as the other. The applicant's physics GRE score and GPA are again treated as continuous mediating and moderating variables.

Because the physics GRE score and GPA can both act as mediators and GPA may also influence the physics GRE score, we used a serial mediation model instead of the simple mediation model (eqs. (3.4) to (3.7)). While moderation by the independent variable X can occur for any of the relations between the other variables, only moderating relationships between GPA and admission and the physics GRE score and admission are within the scope of this work. Therefore, we only include those interaction terms in our models. For all of these analyses, we used the same bootstrapping process used for the simple mediation and moderation cases.

3.4 Results

3.4.1 Probability of admission results

When comparing the GPAs and physics GRE scores of all applicants, we notice that most applicants who are admitted have both high GPAs and high physics GRE scores (Fig. 3.3). Furthermore, while a near perfect GPA or physics GRE score resulted in the highest chance of admission, having either a high GPA or high physics GRE and a modest score on the other seemed to still offer an admission fraction around the overall average. However, having a low GPA or low physics GRE and a modest score on the other is usually grounds for rejection. Overall admissions fractions for a given physics GRE score or GPA are shown in the top and right margins of Fig. 3.3 respectively.

In regard to having a high physics GRE score despite a low GPA, we first note that only a small fraction of all applicants fall in this regime. Second, there appears to be no pattern in terms of higher than average fraction admitted for these applicants. Some combinations of low GPA and high physics GRE score result in a few applicants being admitted, and hence, an above average fraction of applicants being admitted, while other score combinations have no applicants being admitted, and hence, a below average change of admission. For example, having a GPA in the 3.3 bin and a physics GRE score in the 1000 bin resulted in an above average fraction admitted while having a GPA in the 3.4 bin and a physics GRE score in the 1000 bin did not result in an above average fraction admitted, despite the applicants having a higher GPA.

To further understand whether a high physics GRE score can highlight those with low GPA, we divided all students into either a high or low GPA and high or low physics GRE score bins, Fig. 3.4. Based on Fig. 3.3 in terms of admissions probabilities, a low GPA seems to be below a 3.5, while a high physics GRE score seems to be above 700. However, 700 is a common cutoff score which could explain why admissions probabilities increase after that score. Because hitting the minimum score might not catch the admission committee's eyes, we instead selected a higher score of 880 which represents the 80th percentile.

From Fig. 3.4, we notice two things. First, among applicants in the low GPA bin, less than

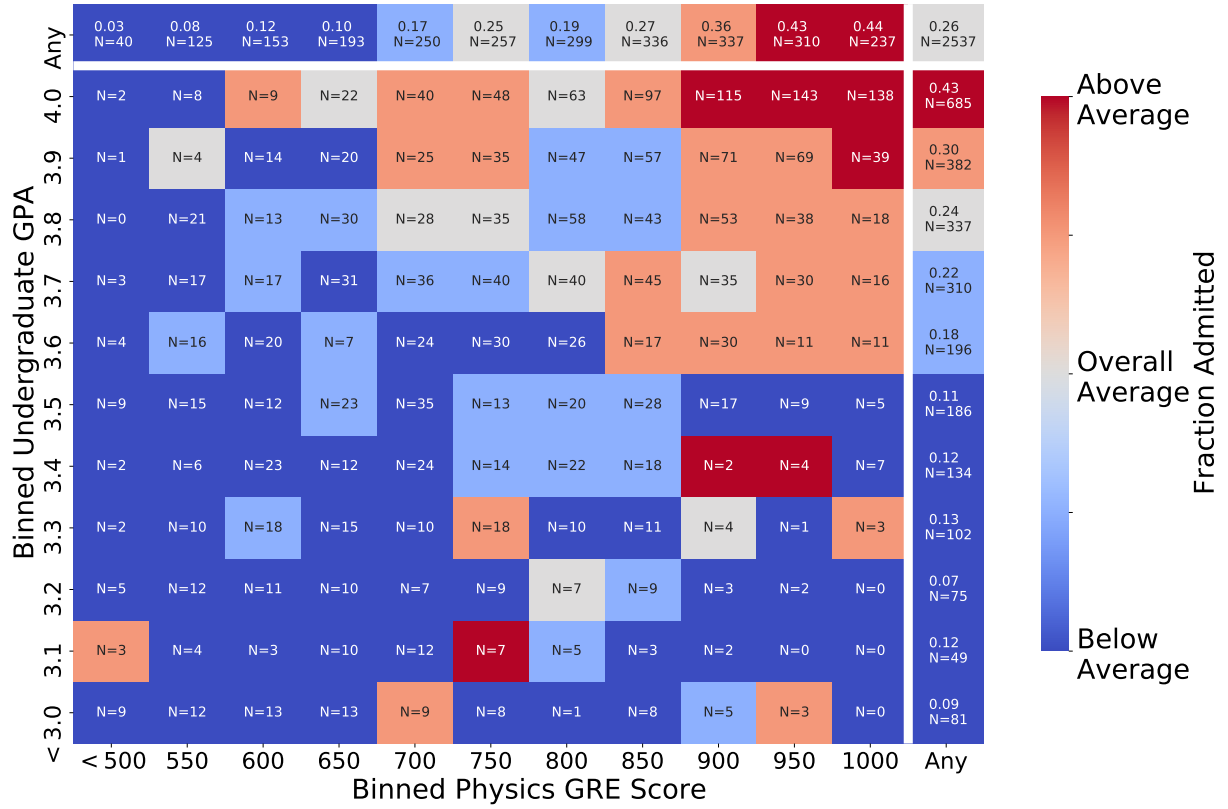

Figure 3.3: Fraction of applicants admitted by undergraduate GPA and physics GRE score. The number of students in each bin is also shown. 'Any' corresponds to the corresponding row or column totals. The bin label corresponds to the upper bound of values in the bin exclusive with the exception of the 4.0 GPA bin which includes 4.0. Values are colored based on whether they are above, below, or equal to the overall admissions rate. Admissions rates within 10% of the overall rate are colored the same as the overall rate. The above and below average colors are based on being above/below the midpoint between the max/min admission fraction and the overall average. These are based on raw numbers and not a statistical test.

half (44%) even make it above the typical cutoff score of 700 and less than 10% of those applicants with low GPAs score 880 or higher. These represent approximately 11% and 2% of all applicants respectively. Comparing the fraction of admitted applicants in each bin, applicants with high physics GRE scores and low GPAs are admitted at nearly the same rate as applicants with high GPA and low physics GRE scores.

Second, we notice that 16% of all applicants score in the high GPA but low physics GRE score bin. That is, more applicants could be penalized for having a low physics GRE score despite a high GPA than could benefit from a high physics GRE score despite a low GPA.

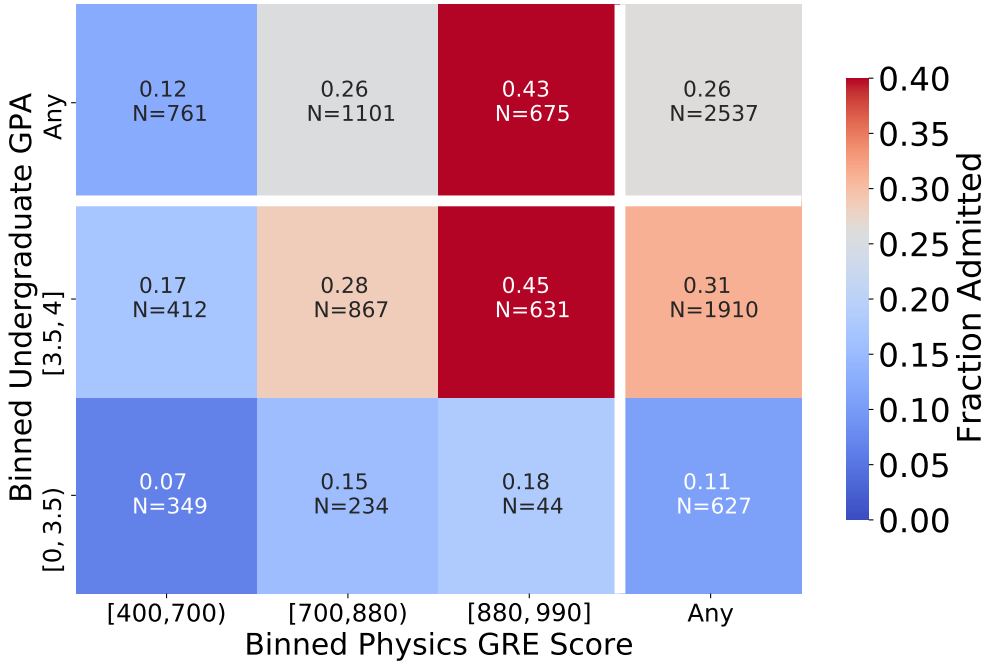

Figure 3.4: A condensed version of Fig. 3.3 showing the fraction of applicants admitted by undergraduate GPA and physics GRE score.

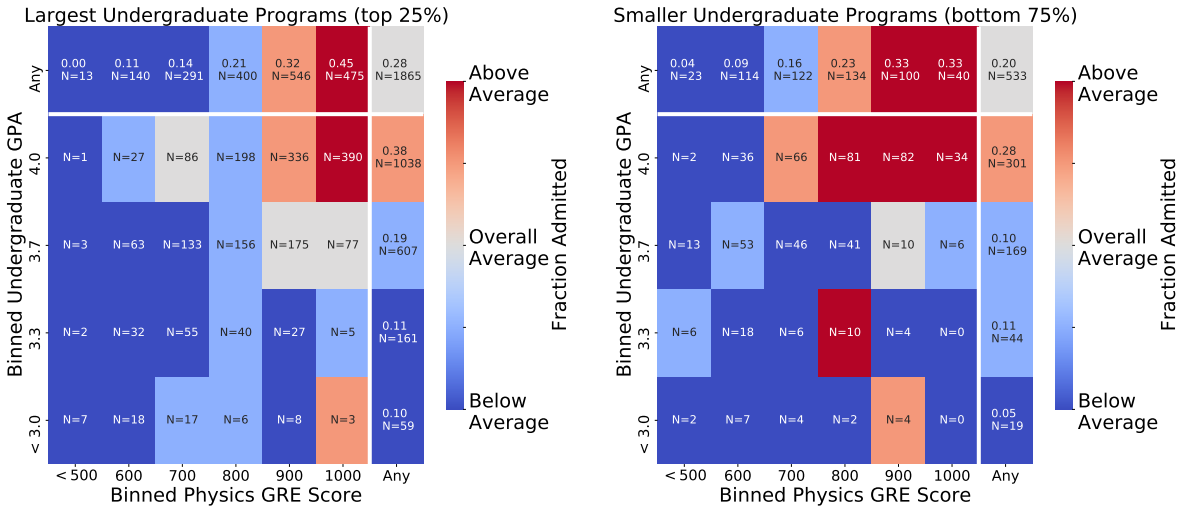

Figure 3.5: Fraction of applicants admitted by undergraduate GPA and physics GRE score and split by large or small undergraduate university

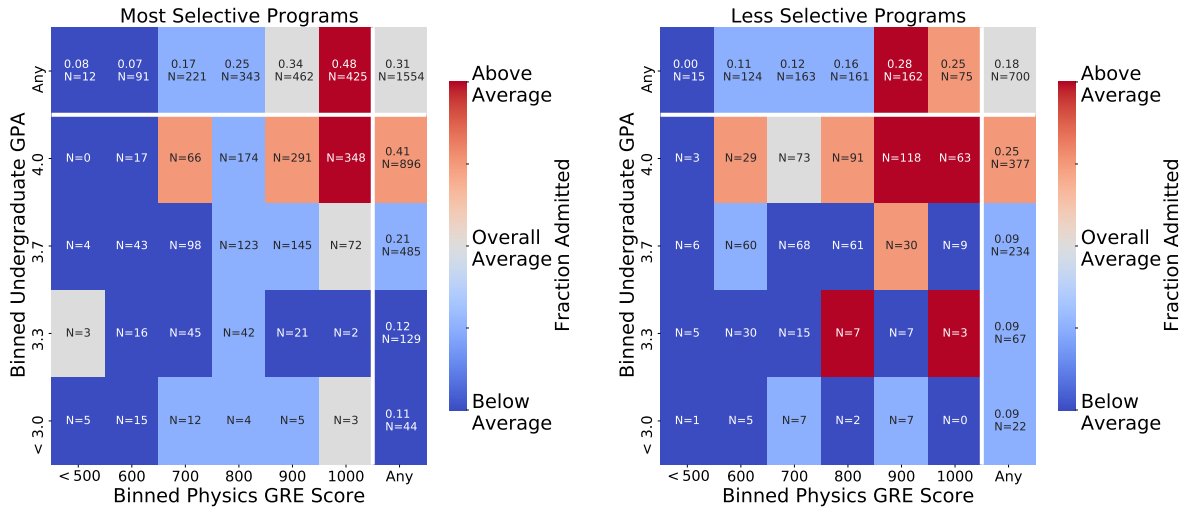

Figure 3.6: Fraction of applicants admitted by undergraduate GPA and physics GRE score and split by selective or non-selective undergraduate university.

Table 3.3: Distribution of applicants scoring in each Physics GRE range by size of institution. ETS only publishes overall score distributions and hence, we cannot report national scores from only domestic students.

Score	Large Schools	Small Schools	Selective Schools	Non-selective Schools	National
[400,500)	0.7%	4.3%	0.8%	2.1%	9%
[500,600)	7.5%	21.4%	5.9%	17.7%	19%
[600,700)	15.6%	22.9%	14.2%	23.3%	20%
[700,800)	21.5%	25.1%	22.1%	23.0%	19%
[800,900)	29.3%	18.8%	29.7%	23.1%	16%
[900,990]	25.5%	7.5%	27.3%	10.7%	17%

When taking the size of the applicant's undergraduate program into account, (large or small), using either the highest or lowest quartile of bachelor's graduates over the two year period did not substantially change the results. Therefore, we only present results from the highest quartile reached, which are shown in Fig. 3.5. Due to the much smaller number of applicants per bin, we reduce the number of GPA and physics GRE bins. We use bins of 3.0 or less, which corresponds to a B or lower, 3.0 to up 3.3, a B+, 3.3 up to 3.7, an A-, and 3.7 up to 4, an A under the standard 4.0 scale.

Overall, by looking at the bin in the 'Any' row and 'Any' column of Fig. 3.5 and Fig. 3.6, we see that applicants from the largest undergraduate programs are nearly 40% more likely to be

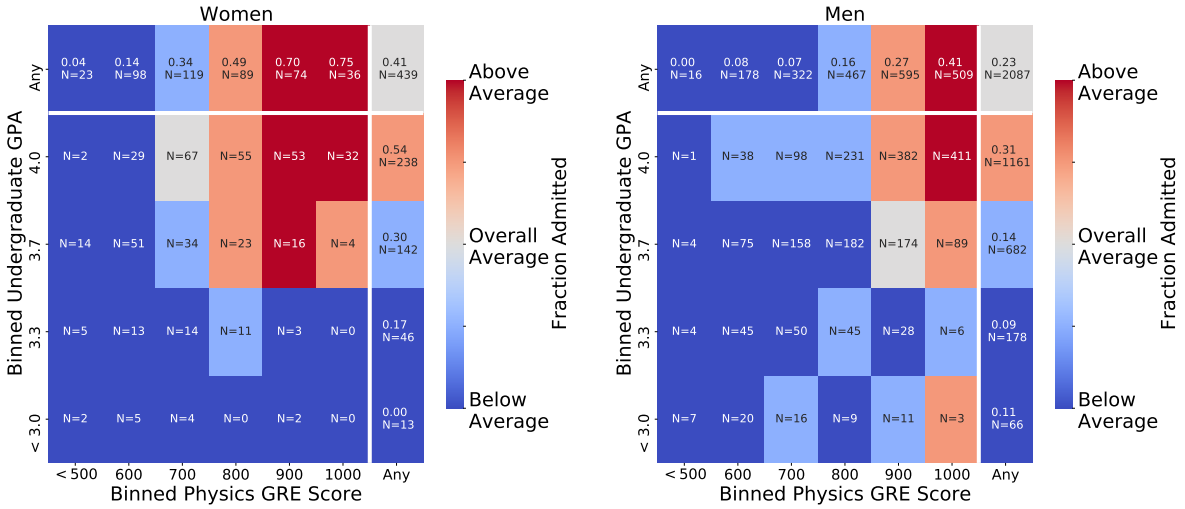

Figure 3.7: Fraction of applicants admitted by undergraduate GPA and physics GRE score and split by the applicant's gender.

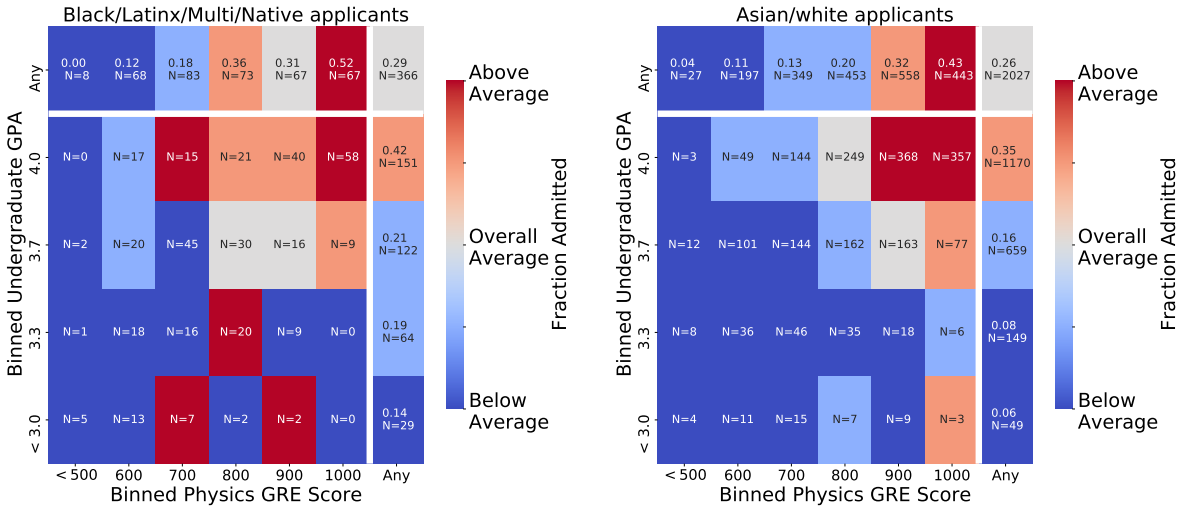

Figure 3.8: Fraction of applicants admitted by undergraduate GPA and physics GRE score and split by the applicant's race.

admitted (0.28 to 0.20) while applicants from selective institutions are nearly 70% more likely to be admitted (0.31 to 0.18). Looking at the individual admission fractions, there does not appear to be any advantage to applicants graduating from smaller institutions or less selective institutions. The physics GRE scores and GPAs where applicants are admitted at higher than average rates are the nearly same for large and small programs and selective and non-selective programs. Unsurprisingly, these tend to be higher physics GRE scores and higher GPAs. Outside of a few bins with a small number of applicants, no combination of low GPA (B+ or less) and high physics GRE score resulted in an above average admission fraction.

For the highest physics GRE scores, 900 and above, applicants from the largest or most selective universities seem to be admitted at a higher rate and a higher fraction of applicants from large or selective universities achieve these high scores compared to applicants from smaller universities. The fraction of applicants from both large universities, small universities, selective universities, and non-selective universities, as well as nationally, achieving each score is shown in table 3.3. Thus, it appears that even if higher scores did help applicants stand out, applicants from smaller and less selective schools most in need of standing out are less likely to achieve those scores in the first place.

Finally, the results from grouping by gender and race are shown in Fig. 3.7 and Fig. 3.8. Interestingly, we find that for most physics GRE scores, women are admitted at higher rates than men of equal score are. Likewise, we find that Black, Latinx, Multi-racial, and Native applicants are admitted at higher or similar rates as white or Asian applicants are for similar physics GRE scores. In addition, the same trend seems to hold for GPA as well. However, a high physics GRE score does not seem to help women with a low GPA. For B/L/M/N applicants, there appears to be a few places where applicants may stand out (such as the 800 physics GRE bin and 3.3 GPA bin). If these applicants were standing out due to their physics GRE score though, we would expect that pattern to continue for higher physics GRE scores but the same GPA. This does not appear to be the case, suggesting these applicants stood out for a reason other than their physics GRE scores. We address this in our discussion.

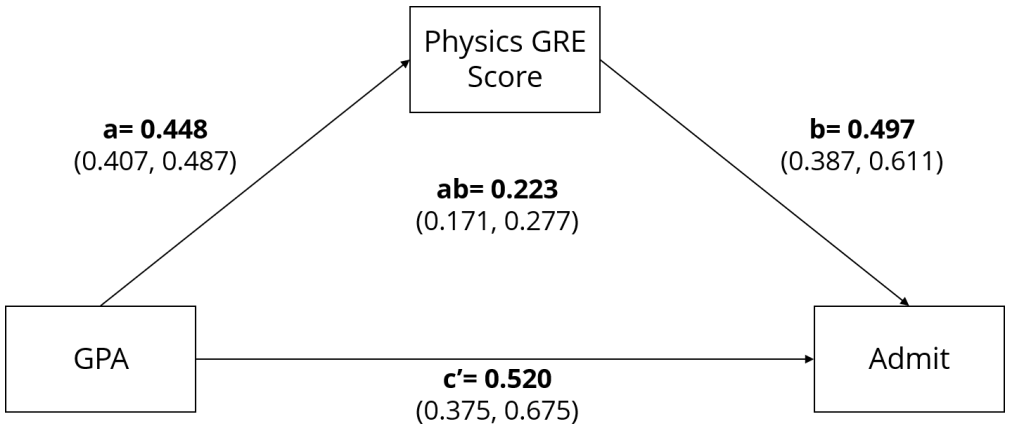

Figure 3.9: Visual representation of the bootstrapped coefficients in eqs. (3.1) to (3.3). We do find evidence of the physics GRE score mediating the relationship between GPA and admission status.

Because these applicants may have stood out for reasons other than their physics GRE score, we do not discuss any interactions between gender and race and selectivity and institution size. For completeness, plots showing these interactions are included in the supplementary material.

3.4.2 Mediation and moderation results

3.4.2.1 Physics GRE and GPA

A visual representation of our mediation results with the physics GRE score and GPA is shown in Fig. 3.9. We find that all coefficients are statistically different from zero.

From Fig. 3.9, we see that an applicant's physics GRE score and GPA have about the same effect on whether the applicant is admitted. Given that applicants who had either a high physics GRE score or a high GPA had about the same chance of being admitted, this is not a surprising result.

Second, we find that the indirect effect is not zero, meaning that there is partial mediation. That is, whether an applicant is admitted depends on their physics GRE score and their GPA. In terms of the amount of mediation, we find that the indirect effect accounts for nearly 30% of the total effect.

Finally, doing moderation analysis, we find that $b_1' = 0.024$ $(-0.114, 0.154)$. As zero is included in the confidence interval, we do not find evidence that GPA moderates the relationship

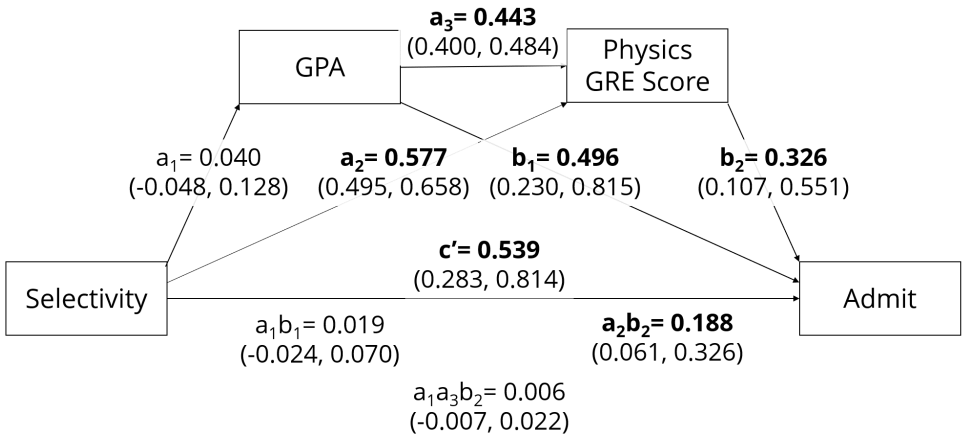

Figure 3.10: Visual representation of the bootstrapped coefficients in eqs. (3.5) to (3.7). We do find evidence of the physics GRE score mediating selectivity and admissions status but do not find evidence of GPA mediating selectivity and admissions status. We do not find evidence of a serial mediating relationship. Statistically significant coefficients are in bold.

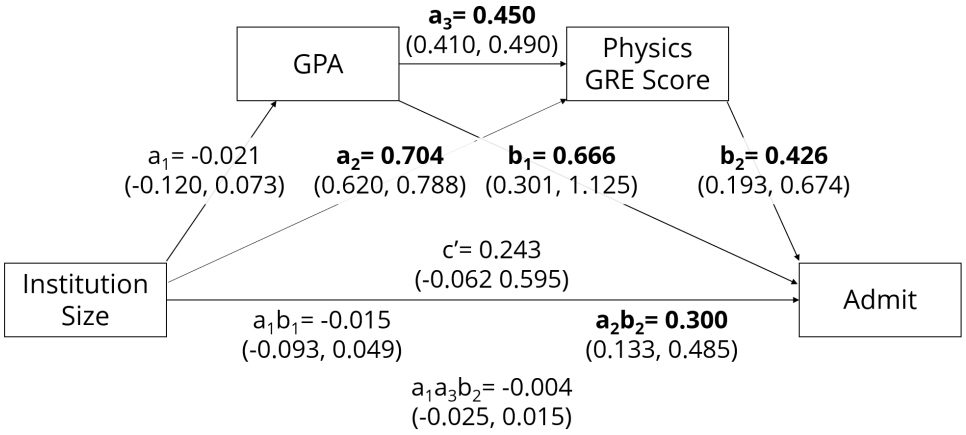

Figure 3.11: Visual representation of the bootstrapped coefficients in eqs. (3.5) to (3.7). We do find evidence of the physics GRE score mediating institution size and admission status but do not find evidence of GPA mediating institution size and admissions status. We do not find evidence of a serial mediating relationship. Statistically significant coefficients are in bold.

between an applicant's physics GRE score and whether they are admitted. That is, the relationship between an applicant's physics GRE score and whether they are admitted is not influenced by their GPA.

3.4.2.2 Institutional features

A visual depiction of our results is shown in Fig. 3.10 and Fig. 3.11. We find that the applicant's physics GRE score partially mediates the relationship between the selectivity of their undergraduate institution and whether they were admitted and fully mediates the relationship between their institution's size and whether they were admitted. The fractions of mediation due to the indirect effects from the physics GRE score were $\frac{a_2 b_2}{|a_1 b_1|+|a_2 b_2|+|a_1 a_3 b_2|+|c'|} = 0.25$ and 0.533 respectively.

In contrast, the applicant's GPA was not found to be a significant mediator in either case (zero was contained in the indirect effects' 95% confidence intervals), meaning that GPA is not a reason that there are differences in admission based on the applicant's undergraduate institution. Additionally, no serial mediation was observed for either case.

When looking at the results of the moderation analysis when the physics GRE is the mediating variable, we find that neither b'_2 value is statistically different from zero ($b'_{2,selectivity} = 0.136\ (-0.141, 0.411)$ and $b'_{2,size} = 0.080\ (-0.204, 0.361)$), meaning that the relationship between the physics GRE score and admission is the same regardless of the type of institution the applicant attended.

Likewise, we do not find evidence of moderation when GPA is the mediation variable. In those cases, $b'_{1,selectivity} = 0.052\ (-0.314, 0.394)$ and $b'_{1,size} = -0.143\ (-0.630, 0.267)$.

3.4.2.3 Demographic features

Our results are shown visually in Fig. 3.12 and Fig. 3.13. Because we chose woman to be "1" and B/L/M/N to be "1" in our logistic regression equation, some of the coefficients are negative. For example, the negative a coefficient for gender and physics GRE score means that women score lower on the physics GRE than men do. Because the sign depends on our choice of which category should be "1" and are in that sense arbitrary, we use the absolute values of c' and $a_i b_i$ to calculate the fraction of mediation.

We find that the applicant's physics GRE score partially mediates the relationship between gender and admission but not race and admission meaning that gender affects admission in part

because it affect physics GRE scores, which affect admission. The fraction of mediation for gender and admission due to the physics GRE score is $\frac{a_2 b_2}{|a_1 b_1|+|a_2 b_2|+|a_1 a_3 b_2|+|c'|} = 0.246$

For GPA, we find the opposite. GPA partially mediates the relationship between race and admission but not gender and admission. The fraction of mediation for race and admission due to GPA is $\frac{a_1 b_1}{|a_1 b_1|+|a_2 b_2|+|a_1 a_3 b_2|+|c'|} = 0.299$.

Likewise, we find a serial mediation effect for race and admission but not gender and admission. That is, admission is affected by race both because admission is related to GPA which is related to race and because admission is related to the physics GRE score which is related to GPA which is related to race.

When investigating whether any moderation effects exist, we do not find that to be the case. That is, we find that none of the interaction coefficients are statistically different from zero and hence, physics GRE scores and GPAs do not have a differential effect on admission based on the applicant's gender or race. Specifically,

- $b'_{2,pGRE,gender} = 0.154\ (-0.154, 0.462)$,
- $b'_{1,GPA,gender} = 0.209\ (-0.103, 0.538)$,
- $b'_{2,pGRE,race} = -0.007\ (-0.309, 0.314)$,
- $b'_{1,GPA,race} = -0.236\ (-0.586, 0.143)$.

All results and interpretations from the mediation and moderation analyses are summarized in table 3.4.

3.5 Discussion

Here, we address each of our research questions and possible limitations or confounding factors.

3.5.1 Research Questions

How does an applicant's physics GRE score and undergraduate GPA affect their probability of admission? We find that scoring highly on the physics GRE and having a high GPA results in the

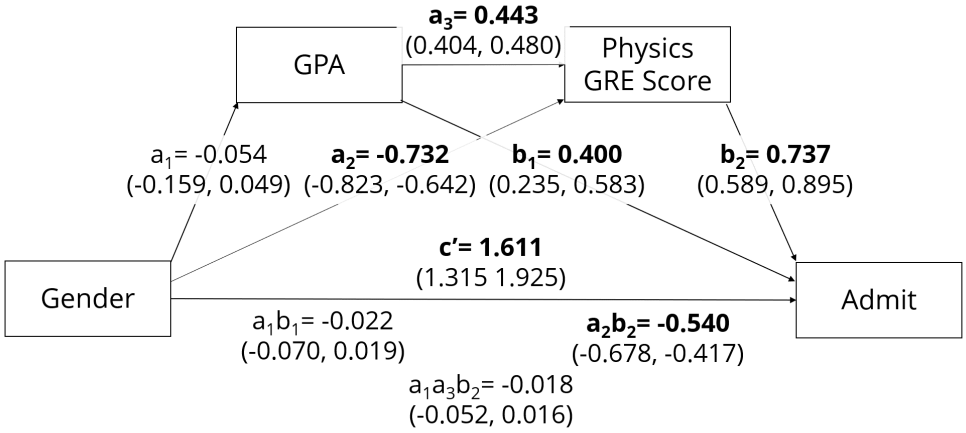

Figure 3.12: Visual representation of the bootstrapped coefficients in eqs. (3.5) to (3.7). We do find evidence of the physics GRE score mediating gender and admission status but do not find evidence of GPA mediating gender and admission status. We do not find evidence of a serial mediating relationship. Statistically significant coefficients are in bold.

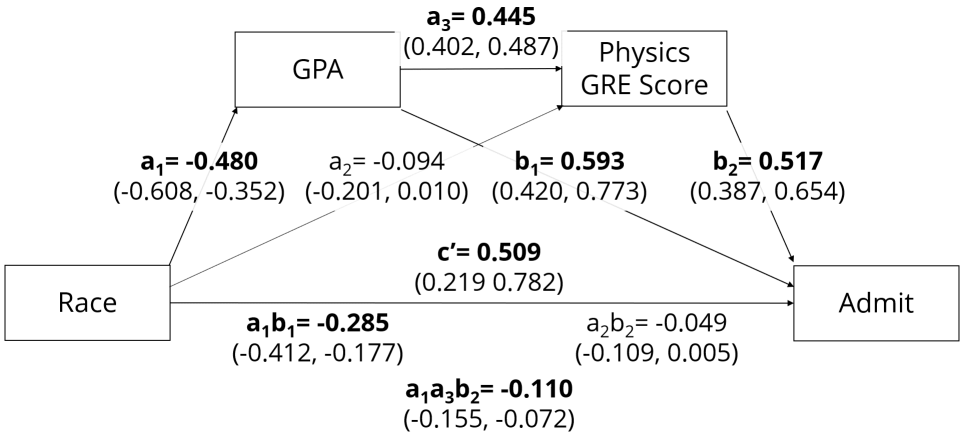

Figure 3.13: Visual representation of the bootstrapped coefficients in eqs. (3.5) to (3.7). We do find evidence of GPA mediating race and admission status and a serial mediation effect but do not find evidence of the physics GRE mediating race and admission status. Statistically significant coefficients are in bold.

Table 3.4: Summary of the mediating and moderation results. * signifies partial mediation is present, ** signifies full mediation is present, † signifies moderation is present. However no moderation effects were found.

Independent	Mediating	Indirect effect	Moderating effect
GPA	Physics GRE	0.223*	0.024
Selectivity	Physics GRE	0.188*	0.136
Selectivity	GPA	0.019	0.052
Selectivity	Serial	0.006	NA
Institution size	Physics GRE	0.300**	0.080
Institution size	GPA	-0.015	-0.143
Institution size	Serial	-0.004	NA
Gender	Physics GRE	-0.540*	0.154
Gender	GPA	-0.022	0.209
Gender	Serial	-0.018	NA
Race	Physics GRE	-0.049	-0.007
Race	GPA	-0.285*	-0.236
Race	Serial	-0.110*	NA

highest chance of admission (Fig. 3.4). Likewise, having a low physics GRE score and low GPA results in the lowest chance of admission. If either the applicant's physics GRE score or GPA is high while the other is not, the chance of admission is approximately equal, regardless of which one is high.

However, the number of applicants with high GPAs but low physics GRE scores is 9 times as large as applicants with low GPAs and high physics GRE scores (i.e. scoring above the 80th percentile; Fig. 3.4). Even if we consider meeting the minimum cutoff score as a high physics GRE score, the number of applicants who have high GPAs but low physics GRE scores is 1.5 times greater than the number of applicants with low GPAs but high physics GRE scores. Thus, many more high GPA applicants could be penalized by the physics GRE than low GPA applicants could stand out or benefit from a high physics GRE score.

Finally, we note that for low-GPA applicants with high physics GRE scores, they are all essentially admitted at the same rate, regardless of whether they scored in the 700-870 range or the 880-990 range. If these applicants were standing out, we would expect low GPA applicants scoring above 880 to be admitted at a much higher rate than low GPA applicants scoring between 700 and 870. Thus, it is hard to determine if these applicants actually stood out to the committee or if they

simply met the minimum physics GRE score needed for the committee to review the rest of the application.

How are these probabilities of admission affected by an applicant's undergraduate institution, gender, and race? First, we find that for most physics GRE scores, applicants from larger and smaller institutions are admitted at similar rates (Fig. 3.5). However, for the highest scores (above 900), applicants from larger universities are admitted at higher rates. Interestingly, for applicants from smaller programs, scoring above 900 does not appear to provide any additional benefit in terms of the fraction of applicants admitted compared to scoring between 800 and 900.

In contrast, applicants from less selective institutions are less likely to be admitted than applicants from more selective institutions for all physics GRE scores above the common cutoff score (Fig. 3.6). That is, the physics GRE does not seem to counteract any potential biases from admissions committees toward applicants from less selective institutions.

Overall, attending a large or selective institution and scoring highly on the physics GRE does result in a higher chance of admission than scoring highly on the physics GRE and attending a smaller or less selective institution.

It is important to note that there might be selection bias in our data because test-takers with high scores from smaller universities might not choose to apply to these schools. However, this seems unlikely because 1) these programs are highly regarded and hence, these would not be "safety schools" to high scoring applicants (as indicated by many high scoring applicants from large programs applying here) 2) while there is research suggesting students with low physics GRE scores might view their scores as barriers to applying [49], to our knowledge, there is no evidence that students with high scores do not apply to physics graduate programs. Given that students with low test scores might not apply, it is expected that our data is not representative of test-takers on the lower end of scores (as shown in table 3.3).

When looking at the demographic variables, we find that women are admitted at higher rates than men with similar scores (Fig. 3.7) and B/L/M/N applicants are also admitted at higher rates than white or Asian applicants (Fig. 3.8). As prior work has shown [105], women and B/L/M/N

test-takers tend to score lower than white men on the physics GRE and hence, scoring highly could cause these applicants to stand out to admissions committees.

How might the above relationships be accounted for through mediating and moderating relationships? Our mediation and moderation analysis further supports the results found through the probability of admissions procedure.

We find that the physics GRE score and GPA have similar regression coefficients when modeling admission, suggesting they have similar effects (Fig. 3.9) and that there is a mediation effect. In addition, we did not find any evidence of moderation. That means the relationship between GPA and admission is not different due to the applicant's physics GRE score. If a high physics GRE score did help a low-GPA applicant stand out, we would expect to see a moderation effect.

Combining the results of probability of admission analysis and the mediating and moderation analysis, we find that there is mediation but no moderation between an applicant's physics GRE score and their GPA when it comes to admission probability. In practice then, an applicant with a low GPA cannot simply overcome that low GPA by scoring highly on the physics GRE.

When we performed mediation analysis on the institutional factors, we found that the relation between institutional selectivity and admission was partially mediated by the applicant's physics GRE score and the relation between institutional size and admission was fully mediated by the applicant's physics GRE score (Fig. 3.10 and Fig. 3.11). Neither of these relationships was mediated by the applicant's GPA or serially however.

The results of the mediation analysis show that physics GRE scores seem to explain some of the differences in admission probability based on the applicant's undergraduate institution. Therefore an applicant from a smaller or less selective institution may be able to stand out by scoring highly on the physics GRE. However, looking at the fraction of applicants admitted by physics GRE scores, especially the highest scores, suggests that is not what happens in practice.

In terms of gender and race, we do find some mediating relationships, but no moderation relationships (Fig. 3.12 and Fig. 3.13). We find that the physics GRE partially mediates the relationship between gender and admission. We also find GPA and GPA plus the physics GRE score

partially mediates the relationship between race and admission. That is, some of the differences in admission rates between men and women can be explained by the differences in their physics GRE scores and some of the differences in admission rates between B/L/M/N applicants and non-B/L/M/N applicants can be explained by differences in their GPAs or physics GRE scores and GPAs.

These results then suggest that a female or B/L/M/N applicant may be able to stand out by doing well on the physics GRE. In practice, the probability of admission results do suggest that women and B/L/M/N applicants are admitted at higher rates than their male, white, or Asian peers are. However, as the five programs studied here were interested in increasing their diversity, our data does not allow us to disentangle "standing out" from highlighting. Therefore, our results should be interpreted with caution regarding any claims that the physics GRE may help applicants from groups underrepresented in physics stand out.

It should also be noted that women and B/L/M/N applicants are less likely to reach these higher scores than their male, white, and Asian peers. In our data, 75% of men and 72% of white or Asian applicants scored above 700 compared to 45% of women and 57% of B/L/M/N applicants. Thus, even if the physics GRE does allow these applicants to stand out, any potential benefit must be weighed against known scoring discrepancies.

Finally, it could be argued that even though we did not show that the physics GRE helps these applicants "stand out", doing well on the test could still provide some benefit for them in the admissions process. We would agree with that argument not because of any properties of the test but because of the structure of graduate admissions in physics. In theory, any part of the application could be weighted highly and therefore, doing well on that part would provide some benefit. Given that prior work has established that the physics GRE is weighted highly [46, 47, 51, 75], we would expect that good performance on the test would provide some benefit to applicants. Our goal, however, was not to determine whether a high physics GRE score benefits applicants in any capacity. Instead, our goal was to determine whether a high physics GRE score offers a disproportional benefit that would justify using it in graduate admissions given the disproportional

harms the physics GRE can cause, which we were unable to show in practice.

3.5.2 Limitations and Researcher Decisions

Data Biases As previously noted, applicants with lower physics GRE test scores may be less likely to apply, resulting in an over-representation of high scoring applicants. In addition, the programs in this study are well-regarded programs and there is likely a secondary bias toward applicants with high GPAs and high physics GRE scores applying overall. As a result, the results may not generalize to graduate programs whose applicants tend to have lower GPAs or low physics GRE scores.

In addition, it is possible that an applicant could be represented multiple times in the data set, as an applicant could have applied to more than one of the five universities in this study. However, each applicant applies to each program independently and thus, we can treat them as separate events for the admissions probabilities. On the other hand, results based on distributions such as table 3.3 and Fig. B.1 and Fig. B.2 would be affected by the duplicates.

To see if possible duplicates affected our results, we compared the distributions with and without possible duplicates. We assumed an application represented the same applicant and hence was a duplicate if two records had the same physics, verbal, written, and quantitative GRE scores, GPA, undergraduate university, and demographic features as the chance of all of these matching for a nonduplicate seems exceedingly low.

When we compare the distributions both with and without possible duplicates, Kolmogorov-Smirnov tests [135] suggest the distributions are not significantly different. Therefore, because we cannot actually determine which applicants are duplicates and excluding possible duplicates does not change our results, we did not remove possible duplicates.

Our choice of low GPA and high physics GRE While percentiles are available for the physics GRE, a "high score" is left to interpretation. Even among admissions committees, individual members may have different ideas of what a high score is. In our work, we have taken the common cutoff score of 700 as the minimum possible high score [52]. Even around this minimum score, the

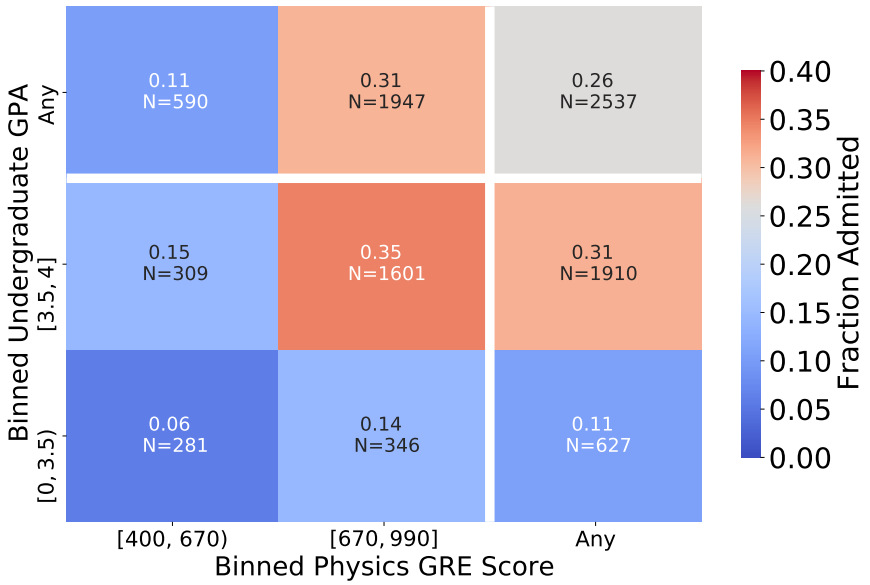

Figure 3.14: A revised version of Fig. 3.4 showing the fraction of applicants admitted by undergraduate GPA and physics GRE score when the cutoff score for a high physics GRE score is 670. Here, the number of applicants who could benefit from a high physics GRE score is approximately equal to the number of applicants who could be penalized by a low physics GRE score.

number of applicants with low GPAs who could benefit from scoring highly on the physics GRE is less than the number of high GPA applicants who could be penalized by having a score below the cutoff.

We find that the number of low GPA applicants who could benefit from a high physics GRE score is greater than the number of high GPA applicants who could be penalized by a low score when the high score cutoff is less than or equal to 670, which is lower than the typical cutoff score and is around the 43rd percentile (Fig. 3.14). Assuming a high score should be at least above the 50th percentile, our specific choice of a high score does not affect our result that more applicants could be penalized than could benefit.

The previous argument is also affected by what we consider a high GPA. We have chosen any GPAs less than 3.5 to be low based on the results shown in Fig. 3.3 where applicants with GPAs at or above 3.5 are nearly twice as likely to be admitted to as applicants with GPAs below 3.5. If we were to pick a lower threshold, there would be even fewer applicants in the low GPA-high physics GRE score group and more applicants in the high GPA-low physics GRE score group, meaning

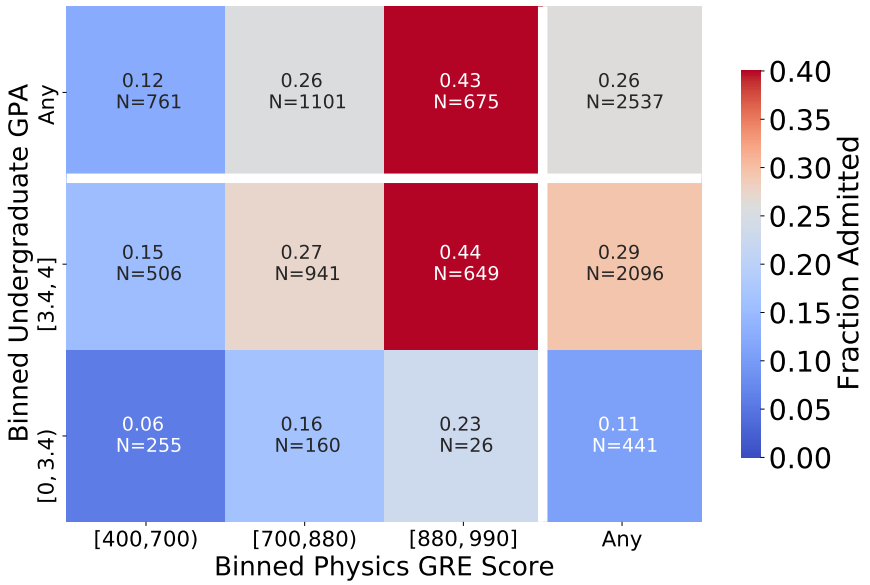

Figure 3.15: A revised version of Fig. 3.4 showing the fraction of applicants admitted by undergraduate GPA and physics GRE score when the cutoff score for a high undergraduate GPA is 3.4.

even more applicants would possibly be penalized rather than standout. Using a GPA cutoff of 3.4 instead of 3.5, the ratio of applicants who could be penalized compared to stand out changes from the original 9:1 to nearly 19:1 (Fig. 3.15).

If we instead picked a higher GPA such as 3.6, there would be more applicants who could potentially benefit, but even then, the number of applicants who could benefit is only greater than the number of applicants who could be penalized around a physics GRE score of 730, which is not a high physics GRE score (approximately 54th percentile) and does not significantly change our results (Fig. 3.16). If we were to pick an even higher GPA cutoff, we could be hard-pressed to justify why anything other than an 'A' GPA is considered a low GPA, especially because admissions committees seem to group applicants with GPAs between 3.5 and 3.6 more closely with applicants with GPAs between 3.7 and 3.8 than applicants with GPAs between 3.4 and 3.5 (based on the fraction of applicants admitted).

Based on our data and the fact that some universities use 3.5 as the only separation between 3.0 and 4.0, using 3.5 seems to represent the best option for separating high and low GPA students. Using any other choice either strengthens our claims or seems unrealistic to use as a cutoff.

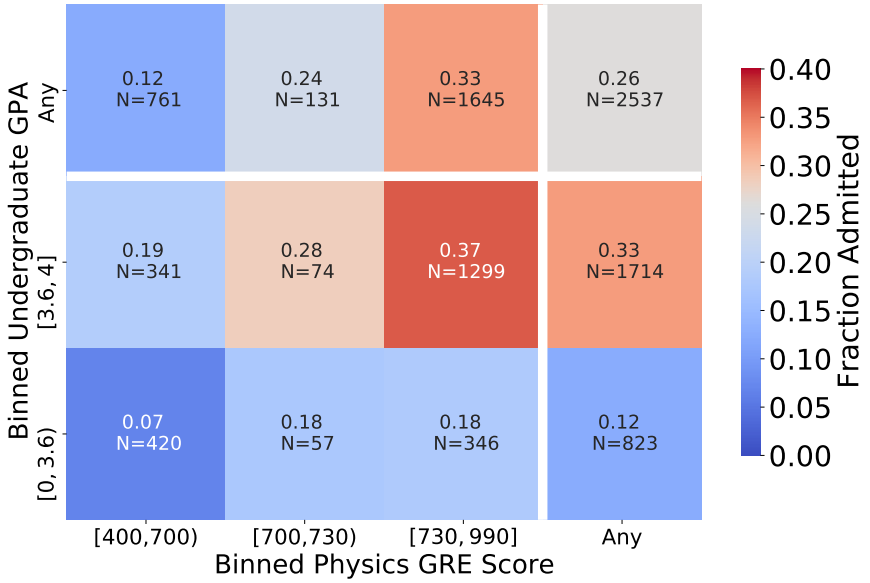

Figure 3.16: A revised version of Fig. 3.4 showing the fraction of applicants admitted by undergraduate GPA and physics GRE score when the cutoff score for a high undergraduate GPA is 3.6. Here, the number of applicants who could benefit from a high physics GRE score is approximately equal to the number of applicants who could be penalized by a low physics GRE score.

Our choice of non-selective school We choose to follow a modified version of Chetty et al.'s groupings of programs [131]. However, many large, state universities have a Barron's Selective Index of 3 and fall in Chetty et al.'s fourth group. For our analysis, we would have included these large, state institutions as part of the less selective programs. As we are concerned with whether the physics GRE helps applicants stand out, saying applicants from large, state universities (for example, the University of Colorado-Boulder, the University of Washington, and Michigan State University) may fall in the traditionally missed category may not be correct.

We reran the analysis with these large, state institutions as part of what we called the most selective programs. We find that the conclusions are then more aligned with the large vs small program results. Using this grouping, applicants from less selective programs are admitted at similar rates to applicants from more selective programs for most physics GRE scores. However, applicants from more selective institutions with physics GRE scores above 900 are still more likely to be admitted than applicants from less selective institutions with similar physics GRE scores.

In terms of the mediating and moderation analysis, our results would be strengthened under this

choice. While the physics GRE score would no longer mediate the relationship between selectivity and admission status, it would moderate the relationship ($b_2' = 0.311\ (0.015, 0.612)$). This positive moderation means that the physics GRE score has a greater effect on admission status for applicants from more selective programs. In terms of standing out arguments, the positive moderation result means that doing well on the physics GRE would provide more of a benefit to applicants from more selective universities and not to applicants from smaller programs who are the intended beneficiaries of the "standing out" argument.

Thus, even though the details change, the overall conclusion are not weakened by changing our groupings. In fact, changing the groupings may strengthen our conclusions instead.

Our choice of a "small" school We chose small schools to be any university not in the top quartile of yearly bachelor's degrees awarded. We acknowledge that using quartiles is an arbitrary decision. However, when we used halves instead of quartiles to divide large and small schools, our results were unchanged, both in terms of the probability of admission analysis and the mediation and moderation analysis. Using the bottom quartile as small schools and all other programs as large schools would not have yielded insightful results as less than 2% of applicants would have attended a small school under this choice.

Of the possible physics specific measures, the number of bachelor's degrees seems most appropriate because programs with more graduates are more likely to be known by admissions committees simply because there are more students to apply from those programs. For example, the programs in the top quartile by number of bachelor's graduates produce nearly two-thirds of all physics bachelor's graduates [128, 129]. In addition, we assume that programs with strong physics reputations attract more students and hence, produce more graduates. While this is likely to be more true at the graduate level, not all physics programs offer graduate degrees and hence, using the number of PhDs awarded would not be useful. Thus, we believe the number of bachelor's graduates serves as a rough proxy for physics reputation.

3.6 Future Work

While the five universities included in this study were interested in increasing their diversity and reducing inequities in their programs, their admissions processes still resembled the traditional metrics-based admissions model. Recently, many programs, including the ones studied here, have begun to employ holistic admissions, which looks at the overall application, taking into account non-cognitive competencies and contextualizes the accomplishments of the applicant in terms of the opportunities that were available to them [136, 137]. Often these holistic admissions use rubrics to weight the various components of each applicant (e.g. see [138, 139]). Evidence from biomedical science graduate programs suggests that the GRE can even be included in holistic admissions without reproducing its known gender and racial biases [140]. Furthermore, their two-tiered approach to holistic admissions did not significantly increase the workload of admission committee members. These findings could persuade faculty reluctant to remove GRE due to its ease and supposed ability to measure some innate quality to try holistic admissions. Whether these results would hold for decentralized admissions as is typical in physics and for the physics GRE though are still open questions.

Our future work will then examine how our results may be affected when a department uses holistic admissions. In theory, we should no longer see the discrepancies between admitted applicants from large and small programs and more selective and less selective universities. In addition, the sample rubric developed by the Inclusive Graduate Education Network (as shown in [138]) suggests ranking applicants by high, medium, or low on each part of their applicant. Therefore, we would expect to see a flatter distribution of admission fractions based on physics GRE scores because, for example, all scores within the 'high' range should be treated equally in the admissions process. Our future work will determine if this is indeed the case.

3.7 Conclusion and Implications

Our work suggests that, in practice, scoring highly on the physics GRE does not help applicants from small or less selective schools or applicants with a low GPA "stand out". Indeed, having a high

physics GRE and low GPA is no better than having a low physics GRE score and high GPA in terms of the fraction of applicants admitted. Similarly, for average physics GRE scores, the selectivity or size of the applicant's institution does not offer any advantage. For the highest scores though, attending a smaller or less selective institution does appear to result in an admissions penalty.

We also find that women and B/L/M/N applicants do have higher rates of admission based on physics GRE scores. However, given that the departments included in this study were actively trying to improve the diversity of their graduate student population [54], we are unable to attribute that standing out to the physics GRE.

While ETS's claim that the physics GRE can help applicants stand out from other applicants may be true in theory, we do not find evidence to support that claim in practice. In fact, our results suggest the opposite: the physics GRE may penalize applicants due to a low score rather than help applicants due to a high score.

As Small points out, facts and data do not unambiguously prescribe a course of action [118] and as other have noted, making such courses of action require a framework of assumptions and commitments [141]. Thus, we do not make a specific recommendation regarding whether the physics GRE should be kept or removed as a result of our work because the answer to that question depends on the priorities of the department. However, if departments are using the physics GRE to identify applicants who might be missed by other metrics to achieve their admissions priorities, we suggest against this practice as it does not appear to be backed by evidence.

CHAPTER 4

RUBRIC-BASED ADMISSIONS: A NEW APPROACH TO GRADUATE ADMISSIONS IN PHYSICS

This chapter is being drafted as a journal article. The working manuscript version includes K. Tollefson as the second author, Remco G. T. Zegers as the third author and Marcos D. Caballero as the fourth author. Following the Contributor Roles Taxonomy (CRediT) [76], my roles for this project include conceptualization, formal analysis, methodology, software, validation, visualization, and writing the original draft.

4.1 Introduction

Female and Black, Latinx, and Indigenous scholars have been and are underrepresented at all levels of physics. The percentage of physics degrees awarded to women has stagnated around 20% [142] while the percentage of physics degrees awarded to Black, Latinx, and Indigenous students has remained less than 10% despite these students making up a larger portion of the college population than in the past [143]. While there are numerous possibilities to address the systematic inequities these scholars face at all levels of academia that limit their participation [144–150], this paper will focus on graduate admissions. Specifically, if we treat graduate admissions as a four stage process similar to how O'Meara et al. treats faculty hiring as four-stage process [151], this paper focuses on evaluating applicants and making admissions offers stages of the process.

While physics departments may be interested in increasing their diversity, the dominant processes of evaluating applicants for graduate school do not support such aims. Prior work has found that diversity considerations are often secondary when evaluating applicants and are discussed after many diverse candidates have already been cut from the applicant pool [54, 152]. Therefore, increasing diversity and equity during the admissions process requires rethinking the process physics departments use to evaluate applicants.

One promising approach to rethinking the admissions process is holistic review, where a broad

range of candidate qualities are considered [137]. In physics, the use of rubric-based review to facilitate such holistic reviews has been gaining traction though the Inclusive Graduate Education Network [153]. Under this approach, applicants are rated on both traditional metrics such as GPA and test scores as well as noncognitive skills such as showing initiative and displaying perseverance according to a predefined rubric. Such an approach is claimed to ensure each applicant is treated fairly and biases by reviewers are checked [154], and hence, it could make graduate admissions more equitable.

To our knowledge however, few studies have examined how these rubrics work in practice and whether they fulfill such aims. Therefore, the goal of this paper is to empirically examine those claims in the context on our department's graduate program. Specifically, our paper addresses three questions related to rubric-based review in our department:

1. How do faculty assign rubric scores to applicants and how do those differ between admitted and rejected applicants?

2. How do the scores assigned by faculty differ by applicant's sex?

3. How do the scores assigned by faculty differ by the type of institution the applicant attended?

As Scherr et al. concluded in their study of graduate admissions practices in physics, many departments are unaware of what other departments do and hence, they might be willing to change their practices if they become aware of successful practices in use elsewhere [48]. Therefore, a secondary goal of this paper is to describe alternative admissions practices in physics and how departments may apply these alternative practices to their own admissions processes.

The rest of the paper is organized as follows. In Sec. 4.2, we provide an overview of holistic review, rubric-based review, and evidence from other fields about their potential for success. In Sec. 4.3, we describe how our department transitioned to rubric-based review, how we collected data relevant to evaluating our admissions process, and how we analyzed such data. In Sec. 4.4, we share results that suggest our rubric does support equitable admissions practices and in Sec. 4.5, we contextualize our results and examine how our choices as researchers may affect the results. In

Sec. 4.6 and Sec. 4.7 we examine the limitations of this study and suggest directions for future work. Finally, in Sec. 4.8, we provide recommendations for departments interested in adopting rubric-based review.

4.2 Background

4.2.1 A typical admissions process in physics

When applying to a physics graduate program in the United States, an applicant will typically submit their undergraduate transcripts, general and physics GRE scores, multiple statements addressing their background, prior preparation, and research interests, and letters of recommendation. A group of physics faculty, the admissions committee, then reviews the applications and offers admission to some of the applicants.

Historically, their have been two main approaches for admitting students: emphasizing research or emphasizing grades [45]. More recent work however has tended to find that programs, including the one studied in this paper, emphasize grades and test scores over research, both in terms of what faculty say they do [47, 51] and what faculty actually do [46, 75].

Yet, numerous potential equity issues emerge when admissions is focused around test scores and grades. First, there is evidence that GRE scores vary based on gender and race [52, 105] and the type of undergraduate university the test-taker attends [69]. When combined with the practice of using cutoff scores, which Potvin et al. estimate at least 1 in 3 departments do, despite the creators of the GRE and physics GRE recommending against it [47], applicants from underrepresented groups in physics may be more likely to not make the first cut.

Second, the tests themselves can be a financial burden for students [49]. The cost to take the General GRE is currently $205 in most parts of the world (and up to $255 in some regions) [1] and the cost to the take the physics GRE is $150 [155]. In addition, if the applicant applies to more than 4 programs, they must pay $27 per school to send their scores. As Owens et al. notes, some students also need to travel to a testing center, which may incur travel or lodging costs [55].

Third, grades vary by applicants' demographics and the type of university they attended.

Whitcomb and Singh found that wealthier, continuing-generation, white students earned higher grades and that even the most privileged racially-underrepresented students in physics earned lower grades than the least privileged white students [156]. Additionally, grades are not standardized measures across universities, with students at private universities tending to earn slightly higher grades than their peers at public universities [157].

Further, evidence has not necessarily supported these metrics as useful predictors of who will earn their PhD. For example, Miller et al. found that while grade point averages were useful to some degree for predicting completion, the physics GRE had limited use [52]. More recent evidence suggests that the physics GRE and undergraduate grade point average only have a relation to PhD completion because they are related to graduate grade point average, which is then related to PhD completion [56].

Given known issues with test scores and GPA, why do programs continue to emphasize them over the qualitative parts of the application. Perhaps the simplest answer is that comparing numbers is quick and convenient [46]. A more nuanced answer might be that qualitative parts of an application can contain substantial variability in what is addressed and these parts of an application can have their own inequities (see. Woo et al. for an overview [158]).

One possible conclusion is then that all application materials have inequities, after all they are produced in an inequitable society, so what is the point of changing anything. We instead adopt a pragmatic view that some parts of the admissions process are more inequitable than others and therefore, our goal is to develop methods to minimize or eliminate inequities to the best of our ability in an inequitable society.

4.2.2 Holistic Review

One possible approach to addressing inequities in the admissions process is holistic review, which Kent and McCarthy define as "the consideration of a broad range of candidate qualities including 'noncognitive' or personal attributes" [137]. Here, we will use holistic review to refer to the general process regardless of what tools or systems are used to conduct it. When talking about our

department's rubric-based process or similar processes, we will use rubric-based holistic review.

While the idea of holistic admissions is hardly new, its implementation is becoming more common due to both greater awareness that quantitative measures may not accurately predict success in graduate school [159–161] and institutions wanting to use the most predictive measures of success in their programs [137]. In addition, professional societies such as the American Astronomical Society (AAS) have called for programs to implement "evidence-based, systematic, holistic approaches" to graduate admissions [138].

Using holistic review has also been claimed to lead to beneficial outcomes for universities including increasing diversity and improving student outcomes (see [137]), though most of these studies have happened outside of physics and related fields. For example, Hawkins found that using holistic review increased diversity in a Doctor of Physical Therapy program [162] and in a literature review of predominantly medicine-related fields, Francis et al. found that holistic review generally increased racial and ethnic diversity [163]. For STEM fields, Wilson et al. found that using holistic review in a biomedical science program resulted in applicant assessments that were independent of gender, race, and citizenship status [140] and Pacheco et al. found that using a composite score that included GPA, test scores, research experience, and publications was correlated with earning a university fellowship and a shorter completion time while applicant's test scores and GPAs individually were not [164].

While holistic review shows promise, programs may have concerns about implementing it. For example, common concerns include limited faculty time to review applications, a lack of data correlating admissions criteria and student success, and limited resources to implement it [137]. In addition, there may be concerns that because the decisions can be more subjective than using a quantitative measure like a test score, there may be variability based on who reviews the application. However, a study of holistic admissions at the undergraduate level found that only 3% of reviews showed substantial variability in the overall score between reviewers [165], suggesting that in practice, variability in the overall rating between reviewers is limited.

4.2.2.1 Noncognitive skills

Regardless of the specifics of a holistic review process, most approaches include some examination of the applicant's noncognitive skills, which may also be referred to as soft skills, personality traits, character traits or socio-emotional skills depending on the discipline or context [166]. While there are multiple definitions of these (see [167]) we adopt Roberts' definition that noncognitive skills or personality traits are "the relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances" [167]. Often these have been operationalized as the Big Five which are openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism [166, 168], though other categorizations exist. For example, in higher education admissions, Sedlacek proposed eight noncognitive traits, which he defines as things not measured by standardized tests: positive self-concept, realistic self-appraisal, understands and knows how to handle racism (the system), prefers long-range to short-term or immediate needs, availability of strong support person, successful leadership experience, demonstrated community service, and knowledge acquired in or about a field [169].

In terms of their utility, noncognitive skills have been found to be predictive or correlated with academic success, though these studies have happened outside of the context of physics. At the undergraduate level, noncognitive skills in isolation and in concert with test scores have been found to be more predictive of success and graduation than test scores alone [170–172]. Likewise, at the graduate and professional levels, noncognitive skills have been found to be correlated with GPA and class rank [173, 174], clinical performance [175], and overall success in programs [176, 177] but were not found to be associated with doing well on a licensing exam [178]. Of the individual noncognitive skills, conscientiousness has been found to be mostly strongly and consistently associated with academic success [179].

In addition to their benefits related to academic success, noncognitve skills can be useful for promoting equity in admissions. For example, including noncognitive skills can increase diversity without harming validity [180, 181] as noncognitive measures have been shown to be just as valid for majoritized and minoritized groups [180, 182, 183]. While including noncognitive skills as part

of admissions may seem like a hard ask of faculty, many faculty already acknowledge the usefulness of noncognitive skills in graduate school [180], including in physics [53].

Yet, a pressing concern is how to measure such noncognitive skills accurately. While applicant self-reports or recommender ratings are typical approaches, such methods may result in inflated or skewed ratings [180]. A recent study suggests that even sharing descriptions of noncognitive skills and why they are useful for predicting later success can artificially inflate judgements [184]. Thus, how best to measure such skills is still an active area of inquiry [185].

4.2.2.2 Rubric-Based Review

One promising approach to implementing holistic review is rubric-based review. Under this approach, applicants are evaluated based on a set pre-defined criteria. By pre-selecting criteria, what is required for admission is clear to reviewers and provides a structure to assess all applicants [46, 154]. This explicitness has been shown to enhance both validity and reliability [158, 186, 187].

In addition, rubrics can help make the admissions process more equitable [46]. By explicitly laying out the review criteria and what is required to achieve each level of the rubric, all applicants can be judged fairly and individual reviewer's expectations can be mitigated [183]. From research into other areas of academic hiring, we know that gender and racial biases exist in the hiring process, including in physics [188, 189]. Specifically in graduate admissions, faculty, including astronomy and physics faculty, have been documented showing preferences to applicants with similar backgrounds as themselves or within the same research subfield of their discipline [46]. Thus, rubrics offer a possible route to counter those biases. Indeed, a recent study in admissions for a psychiatry residency program found that using rubric-based holistic review led to more underrepresented applicants receiving an offer to interview compared to the traditional approach [190] while a recent study of grade-school writing found that teachers rated writing attributed to a Black author lower than when it was attributed to a white author but did not find the effect when the teachers were instructed to use a clearly defined rubric [191].

As rubric-based approaches to admission are still relatively new, best practices are still in

development. Yet, a few recommendations do exist [154]. First, criteria should be selected before reviewing any applications with individual programs deciding what qualities are critical for success in their program [138]. Second, rubrics should be coarse-grained in that there are fewer possible scores for each construct such as low, medium, or high instead of 1-10 to limit disagreements over scores [183]. Third, each level of the rubric should be clearly defined so that a reviewer can easily determine which score an applicant should get on each construct. These levels should be picked so that each possible score will be received by many applicants [154]. Finally, these criteria and levels should allow for diverse forms of excellence to be counted as achievements so that applicants with non-traditional markers of excellence are not excluded [192].

While rubric-based approaches have received little research in physics, they have been successfully incorporated into larger physics graduate program initiatives. Two of the most well-known initiatives are the Fisk-Vanderbilt program, which graduates one of the largest classes of Black PhD physicists in the nation [193], and the APS Bridge Program, which has successfully admitted and retained graduate students of color at rates higher than the national average [143]. Even though rubrics in admission were one of many changes made, these programs suggest that rubric-based review has promise.

For a more in depth review about equitable admissions practices in STEM doctoral programs, we refer the reader to Roberts et al. [194]

4.3 Methods

4.3.1 Our Rubric and Applicant Evaluation Process

In 2018, the Department of Physics and Astronomy at Michigan State University introduced a rubric-based approach to evaluation applications to the graduate program in Physics, informed by the Council of Graduate Schools' 2016 report on Holistic Review in Graduate Admissions [137]. The main goal was to improve the identification of strong candidates for the program and to make the selection more equitable, thereby increasing the participation of students from underrepresented groups in the department. In preparation for the introduction of the rubric, Casey Miller and Julie

Posselt, the Inclusive Practice Hub Director and Research Hub Director, respectively of the National Science Foundation supported Inclusive Graduate Education Network [153], led a workshop with faculty who served at that time in the Graduate Recruiting Committee. This workshop resulted in a selection of five rubric categories, which each had several sub categories. Applicants are ranked with a score of either 0, 1, or 2, corresponding to low, medium, or high, for each subcategory, based on defined criteria for each score. The subcategory scores are then averaged per category and category scores summed (with weights as given below) to calculate the overall score. The categories, with subcategories in parenthesis, are:

- Academic Preparation, with a weight of 25% (Physics coursework, math coursework, other coursework, and academic recognition and honors)
- Research, with a weight of 25% (Variety and duration, quality of work, technical skills, and research disposition)
- Non-cognitive competencies, with a weight of 25% (Achievement orientation, conscientiousness, initiative, and perseverance)
- Fit with program, with a weight of 15% (Fit with research programs of the department, fit to research programs of specific faculty, (prior) commitment to participation in the department/school community, and advocacy for and/or contributions to a diverse, equitable, and inclusive physics community
- GRE scores, with a weight of 10% (General GRE scores, and Physics GRE scores)[1]

The choice of these categories and subcategories was based on the discussions in the workshop and advice from the workshop leaders, and included considerations based on experiences during previous recruiting cycles. Another consideration for the choice of the categories is a reasonably close alignment with criteria used at MSU for awarding fellowship packages to students. Therefore, the rubric scoring can also be used for selecting nominations for university fellowships. This is

[1]This category was not used in 2021, and will not be used in 2022 due to the impacts of COVID-19 on students ability to take these tests.

important because fellowship nominations are due shortly after the application deadline (January 1).

Applications for the graduate program are submitted to MSU's central application system. All folders with a complete or near-complete application package are reviewed. The applications are divided up into several groups, which each are reviewed by different members of the graduate recruiting committee. This committee has a rotating membership with representation from faculty in all major research directions present in the department. Committee members are instructed about the use of the rubric and provided with the criteria. As part of the review process, they also sort students by their interest in research area(s). The results from the rubric scoring are compiled by the Graduate Program Director. Students whose folders are near complete, but have a ranking for which an offer is not impossible, are contacted and ask to provide the missing information. If that additional information is provided, the rubric scoring is updated.

Subsequently, the spreadsheet is used by committee representatives from each major research area in the department to make a list of students they would like to make an offer to for a position in that specific research area. The number of students who are made an offer to depends on openings available per research area, the number of teaching assistant slots available, and the historical acceptance rates for each research area. Typically, the process results in a list of offers that will be made and a wait list for additional offers that can be made if recruiting targets are not met in the initial round of offers. In this stage of recruiting process, the match to available positions is revisited as committee members from specific research areas are better aware than general faculty members about the recruiting needs for that year. In spite of the instruction and criteria provided to reviewers, the scoring is still somewhat subject to differences in reviewing styles and interpretation of the criteria. This is, for example, apparent in the comparison of average summed scores per reviewer. Therefore, this second stage of the review process also allows for another comparison of application based on the rubric by a few faculty members in each research area. Because of these reasons, the list of students whom an offer will be made to, or who are put on a wait list, quite closely follows the original rubric scoring but modifications do occur.

The whole process is organized and overseen by the Graduate Program Director with support from the Graduate Program Secretary. The Graduate Program Director also serves as the point of contact for questions about the use and interpretation of the rubric, reviews applications of likely candidates, and leads the selection of nominations for fellowships.

The overall response from faculty who served in the recruiting committee and used the rubric has been positive, as it provides clear guidance for the review process and reduces the impact of different reviewing styles and biases to what are the most important skills applicants to a physics graduate program should have. On average, the time spent by individual committee members on reviewing the folders has not increased. Faculty reviewers have provided feedback that it would be better if applicants are first sorted by research area so that the review is done by several faculty from the relevant research areas in the first step. Given the large number of applications and the limitations of the current software used to manage applications, this could not easily be accomplished in the past. MSU is implementing new software for managing and reviewing applications, which will make presorting of applications by research area possible, leading to a considerable increase in the efficiency of the process.

4.3.2 Participants and Data Collection

Data for this study comes from compiled records from applicants to our physics graduate program for fall 2018, 2019, and 2020. Most admissions decisions for fall 2020 had already been made before coronavirus accommodations took effect, suggesting at most minimal effects on our data.

When applying to the university, applicants submit a general university application, transcripts, test scores, a personal statement, an academic or research statement, and letters of recommendation to a central system. As the current admissions system does not allow for records to be compiled across applicants, two researchers manually extracted relevant information for this study. The researchers independently extracted data from the first 20 applications and then compared results to ensure they were interpreting the applications the same and agreeing on any conventions for reporting the data. Afterwards, the researchers independently went through the rest of the applications.

Through this process, the researchers collected the applicant's demographics, grade point average, GRE scores, degrees earned or in progress, and previous institutions attended. Any information missing from the applications or entered into the application on a non-standard scale (e.g. a GPA on a non 4.0 scale or a GRE score outside of the current scoring range) was treated as missing data for the analysis.

As rubric scores are determined by faculty and are not part of the materials applicants submit, aggregated scores were then matched with individual applicants using the applicant IDs. Through this process, we collected data on 826 applicants, 511 of which were domestic applicants.

4.3.3 Analysis

Because of different application requirements and availability of institutional data for international and domestic students, we only include domestic students in our study. In addition, we only include applications sufficiently complete that faculty were able to rate and were included in the Graduate Program Director compiled records, leaving us with 321 domestic applicants for this study.

For our analysis, we were interested in how faculty rate applicants and hence, we computed the fraction of applicants in each level (low, medium, and high) of the rubric. In some cases (<5%), faculty used a rating that was in between levels (e.g. low-medium). Because of this, we performed all subsequent analyses by first rounding up (so low-medium would become medium) and then repeating the analysis by rounding down.

First, we computed the fraction of applicants in each level of the rubric for all applicants, all admitted applicants, and all non-admitted applicants.

Second, we compared applicants based on demographics by comparing the fraction of applicants in each bin of the rubric. While gender would be more appropriate, the application system only asks applicants about their sex and allows them to choose male or female. Thus we were only able to compare faculty ratings of males and females. We acknowledge that females is not the correct term to use but as being female does not automatically imply being a woman, we do not believe it is appropriate to assume that someone marking female as their sex is necessarily a woman.

In terms of race, the application system does not allow applicants to enter their race or ethnicity, so we are unable to compare applicants of different races.

Finally, we compared applicants from different undergraduate backgrounds because prior work suggests the applicant's background may influence faculty's perceptions of them. For example, faculty may prefer applicants with similar backgrounds as themselves [46] and may interpret grade point averages in the context of the applicant's undergraduate program, with high GPAs from more prestigious universities carrying more "weight" than a high GPA from a lesser known school [117]. In addition, graduate admissions in physics have been characterized as "risk-adverse" where faculty prefer to admit applicants who are likely to complete their program rather than take chances on someone who might not [46, 48]. As students from smaller program may be viewed as higher risk if previous students from that program struggled [117], it is possible faculty may be less likely to admit students from smaller undergraduate schools.

To characterize an applicant's undergraduate background, we used two measures. First, we used Barron's value, which is a measure of an institution's selectivity based on incoming students' SAT scores, GPA and class rank, and overall acceptance rates. While not equivalent to prestigious, we treat selectivity as a proxy for prestige based on that assumption that most selective institutions are also prestigious institutions. For our analysis, we defined institutions with Barron's values of "most competitive" or "highly competitive" as selective and all other institutions as not selective.

Second, we used the number of bachelor's degrees awarded by the physics department at the applicant's undergraduate institution to estimate the size and notoriety of the department, with the assumption that a department that grants more degrees is more likely to be known by an admissions committee member. Due to variability in yearly degrees, we used the median number of degrees over the 2016-2017, 2017-2018, and 2018-2019 academic years as the number of bachelor's degree awarded [3, 128, 129]. We then defined any program that was in the top quartile of physics bachelor's degrees awarded during that period as a large program and all other programs as smaller programs. For reference, the programs we classified as large produced nearly two-thirds of all physics bachelors degrees over the period.

Table 4.1: Percent of Missing Data by Rubric Construct

Rubric Construct	Percent Missing
Physics Coursework	20.0
Math Coursework	20.2
All Other Coursework	20.2
Academic Honors	22.1
Variety/Duration of Research	3.4
Quality of work	4.4
Technical Skills	4.1
Research Dispositions	4.7
Achievement Orientation	4.4
Conscientiousness	4.4
Initiative	4.0
Perseverance	4.4
Alignment of Research	7.2
Alignment with Faculty	32.1
Community Contributions	4.0
Diversity contributions	3.4
General GRE Scores	2.2
Physics GRE Score	2.5

To perform the comparisons in all cases, we used Fisher's Exact Test to examine whether the rubric score was associated with any of the metrics of interest (admission status, sex, institution selectivity, institution size). We used the standard choice of $\alpha = 0.05$ to judge claims of statistical significance. Because we did 18 comparisons for each metric of interest, it is likely that there would be at least one false positive. Therefore, we used the Holm-Bonferroni procedure to correct the p-values for multiple comparisons as it is less conservative than the traditional Bonferroni correction while maintaining statistical power [195].

For cases of missing data, we used pairwise deletion so that we could make the most use of the data we had. While Nissen et al recommends using multiple imputations for missing data in physics education research studies [71], the goal of this paper is to understand what faculty did as opposed to estimate a larger trend or predict an outcome. Therefore, we do not believe that using multiple imputations is aligned with the goal of this paper. The percent of missing data per rubric metric is shown in Table 4.1.

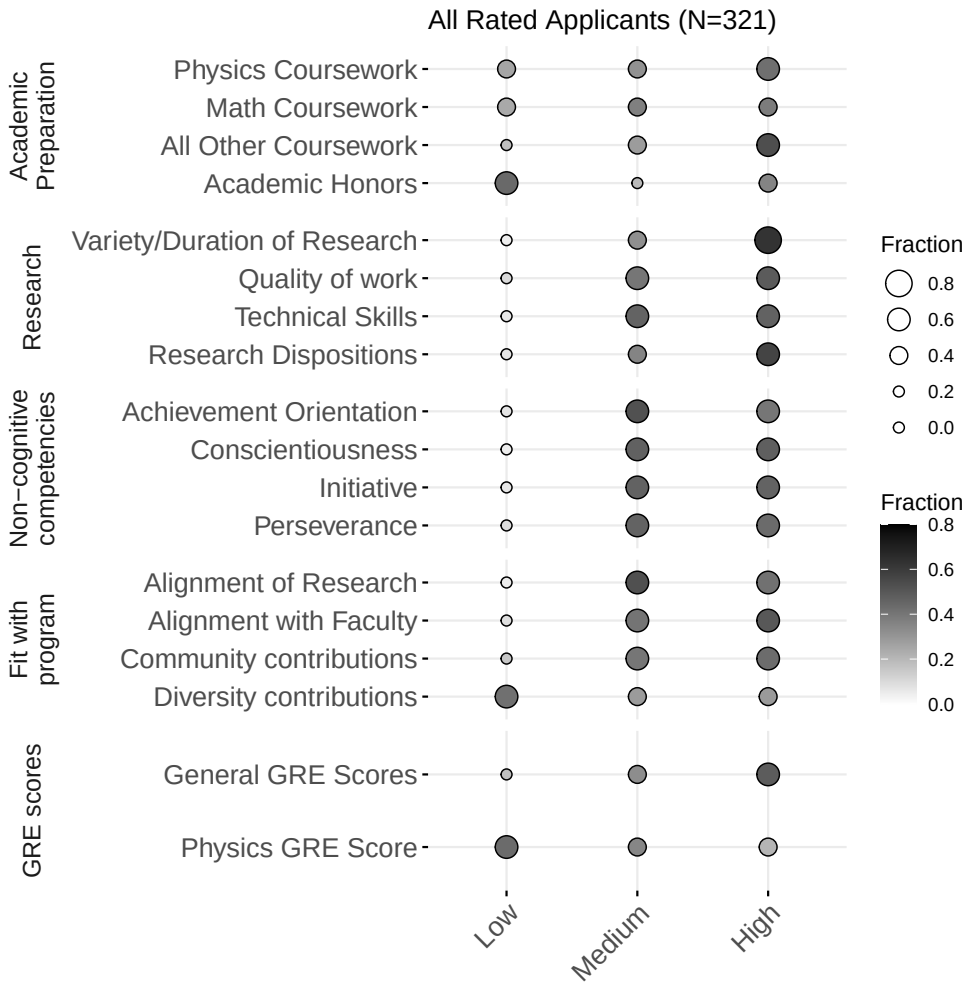

Figure 4.1: Faculty ratings of domestic applicants on 18 constructs. In the plot, a larger, darker circle means that more applicants are in that bin. While many applicants are in each level of the academic preparation and test score constructs, few applicants are in the "low" bin of the research, noncognitive skills, and program fit constructs.

4.4 Results

The results are largely unchanged based on whether we rounded up or rounded down when a faculty member gave a rating in-between levels of the rubric so we present only the round up results here.

When we examine the faculty's rating of all applicants in Fig. 4.1, we notice two overarching trends. First, for traditional measures of academic success such as grades and test scores, faculty tend to rate applicants using all three levels of the rubric. For the academic preparation constructs on the rubric, "high" is the most common rating given by faculty. However in terms of math and physics

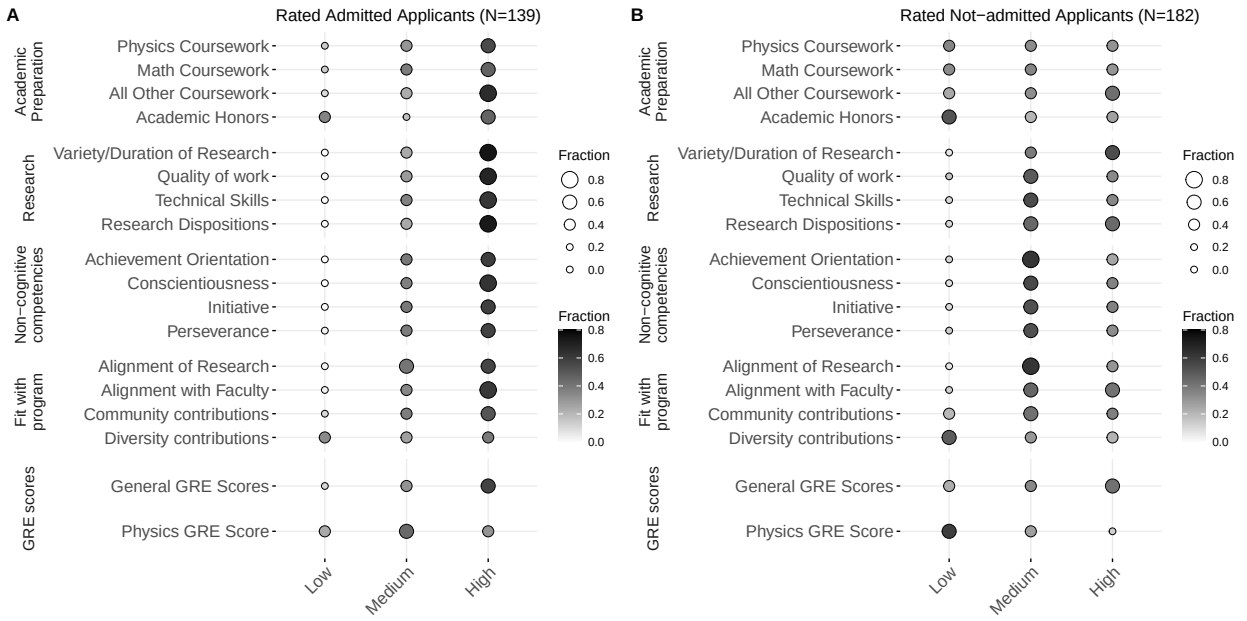

Figure 4.2: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant was admitted. The distribution of ratings of all constructs is statistically different for admitted applicants compared to non-admitted applicants. Overall, most admitted applicants were rated "high" while most non-admitted applicants were rated "medium."

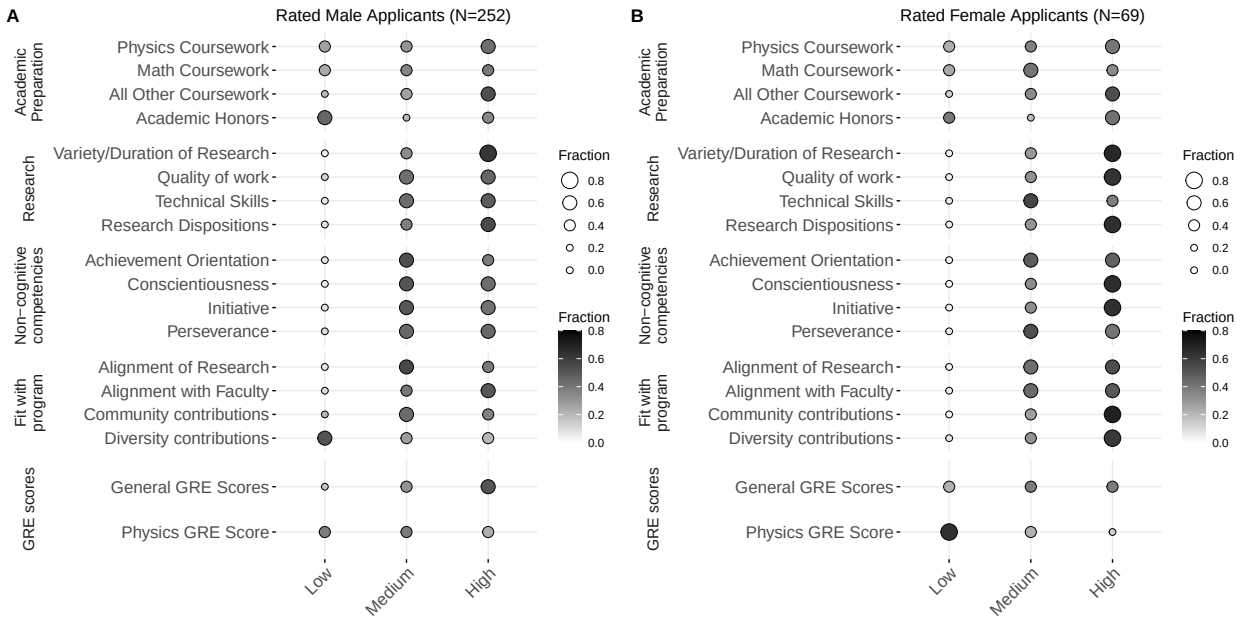

Figure 4.3: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant was male or female. Only three of the constructs showed differences between males and females: physics GRE score where males scored higher and community contributions and diversity contributions where females scored higher.

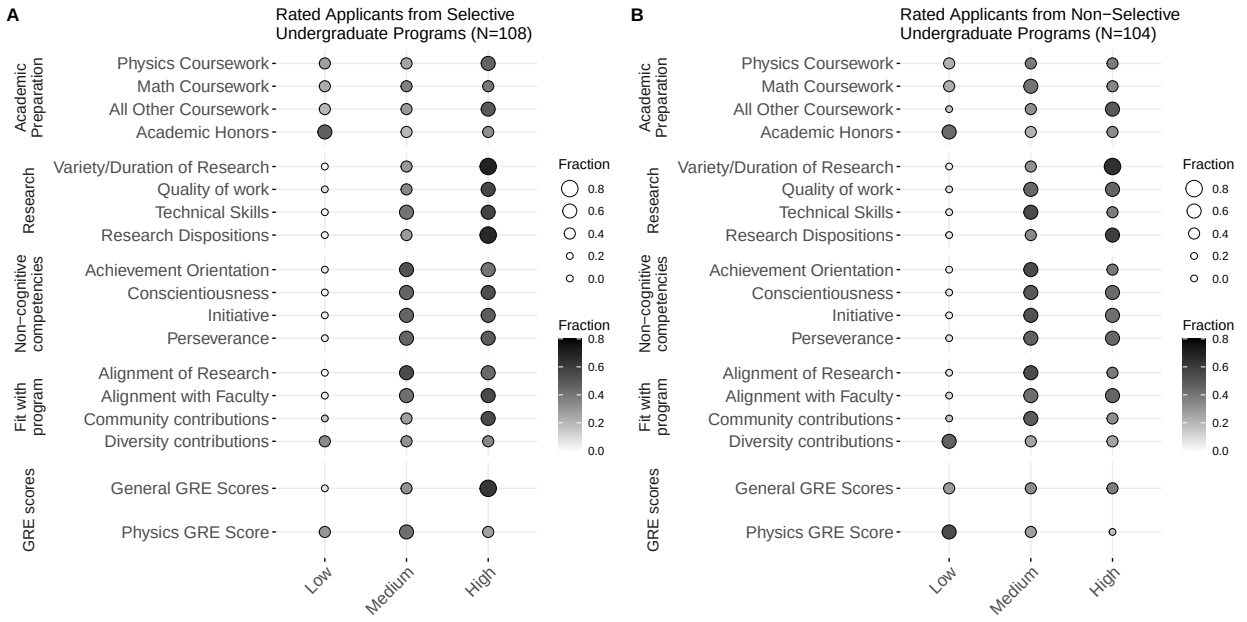

Figure 4.4: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant attended a more selective or less selective undergraduate university. Only the general GRE and physics GRE scores showed differences.

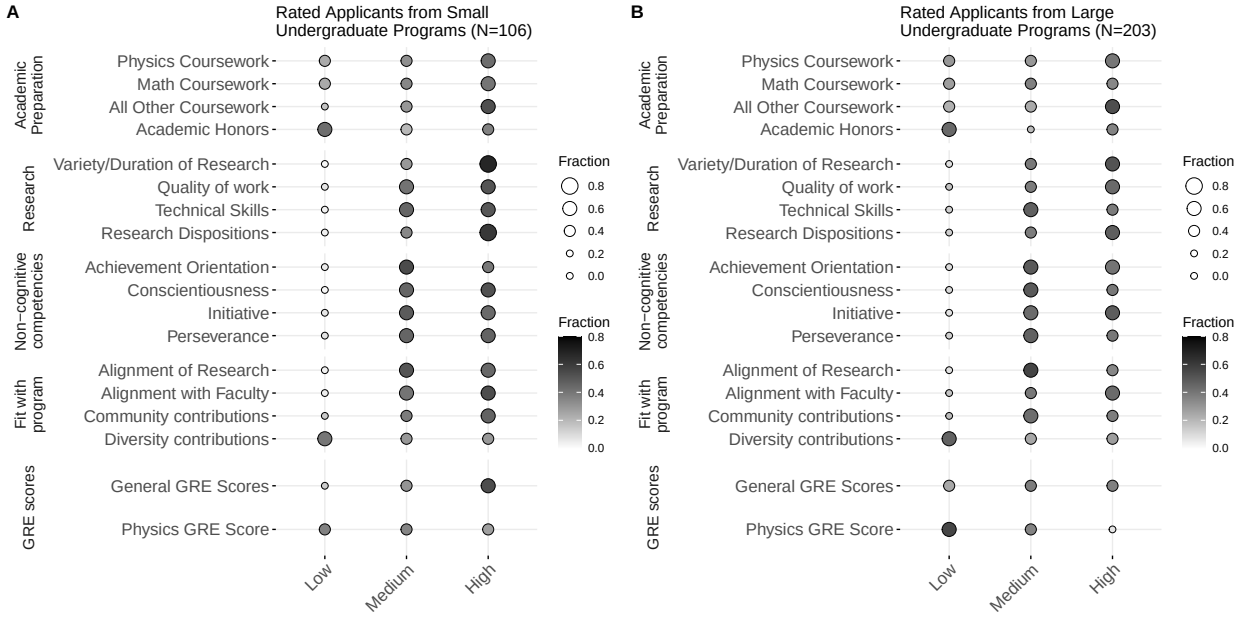

Figure 4.5: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant attended a university with a larger or smaller physics program. Only the physics GRE score and conscientiousness showed differences between the groups of applicants, with the latter dependent on how larger physics program is defined.

course grades, around 25% of applicants still scored in the low bin. Of the academic preparation constructs, academic honors follows a different structure than the others where faculty ratings are bi-modal, meaning that applicants either had no academic honors or had multiple academic honors.

Second, for the research, noncognitive, and fit constructs, faculty rarely used the "low" level of the rubric, with only three of the twelve constructs in those categories having more than 10% of applicants earning a "low." For research, the most common rating was "high" while for the noncognitive traits, the most common rating varied between "high" and "medium." In terms of the fit constructs, most applicants were rated as either "medium" or "high" for alignment of research, alignment with faculty, and community contributions. In contrast, for the diversity contributions construct, "low" was the most common rating, meaning that many applicants did not discuss how they promote or advocate for diversity in their applications.

When looking at how faculty rate applicants who would later be admitted compared to applicants who would not be admitted, we see statistically significant differences in the distribution of all ratings (Fig. 4.2). Overall, admitted applicants tended to be rated "high" on each construct while non-admitted applicants tended to be rated "medium" on each construct. There were a few exceptions to the general trend however. For academic honors, diversity contributions, and physics GRE scores, most admitted students were not rated as "High" and 25% of applicants received a "low" score while for all other course work, variety/duration of research, and general GRE scores, most non-admitted applicants were rated as high.

When looking at the ratings broken down by sex (Fig. 4.3), we notice that the results tend to follow the overall patterns of all three ratings on academic success and test scores and mainly "medium" and "high" ratings on research, noncognitive skills, and fit with the program for both males and females. Comparing ratings between males and females, we find that only physics GRE score, community contributions, and diversity contributions showed statistically significant differences. While males tended to score higher on the physics GRE score, females tended to score higher on community contributions and diversity contributions. As we elaborate on in the discussion, differences in these three constructs do not necessarily mean that faculty are rating

males and females differently but instead may be documenting inequities that already exist.

Likewise, when looking at the ratings broken down by the selectivity of the university where the applicant earned their bachelor's degree (Fig. 4.4) or the size of the department where they earned their bachelor's degree (Fig. 4.5), we may also be observing existing inequities reflected in the faculty ratings. For example, applicants from more selective universities only had statistically higher ratings than applicants from less selective universities on the general GRE and physics GRE scores. Similarly applicants from larger programs had statistically higher ratings on the physics GRE score than applicants from smaller programs did. However, applicants from larger program were also rated higher on conscientiousness than applicants from smaller programs, though this result is sensitive to how we define a large program.

While we could consider interactions between admission status and sex, institutional selectivity, or physics program size, we did not do so given the small sample sizes. For completeness, however, we include those plots in appendix D

4.5 Discussion

How do faculty assign rubric scores to applicants and how do those differ between admitted and rejected applicants? For academic achievement and test scores, faculty tended to use all three levels of the rubric when assigning scores to applicants. In contrast, faculty tended to use mainly "medium" and "high" when assigning scores to applicants in the research, noncognitive skills, and program fit categories. We argue that this result is more of a reflection of the rubric than it is of how faculty are using the rubric.

Given that grades and test scores are well defined via transcripts and test scores, rubric constructs measuring these tended to use quantitative measures to determine which score the applicant would receive. That is, a high score or high grades would correspond to a "high" rating while a low score or low grades would correspond to a "low rating. Additionally, as the courses required for a physics degree tend to be similar regardless of the specific program, most applicants will have taken the courses mentioned in the rubric and hence, faculty can rank applicants based on those grades.

In contrast, the research, noncognitive skills, and fit with department are less defined and instead depend on what applicants write in their statements and what information letters of recommendation contain. This means that not all constructs on the rubric may necessarily be addressed. For example, if an applicant takes quantum mechanics it will certainly appear on their transcript but if that applicant was also active in departmental service activities, it may not be reflected in any parts of the application. As a result, the rubric needs to take into account that applicants may not display a trait because either they do not exhibit it or because they did not mention it. Any display of the trait could then not fall into the "low" level of the rubric, which would then explain why faculty tended to use only the "medium" and "high" ratings.

A reasonable follow up is then whether combining "no evidence" with "evidence not presented" as a single level on the rubric represents an issue with the rubric. We argue that it does not, as it provides the best option given the data faculty have available. Applicants are asked to discuss certain topics in their statements that map broadly onto the rubric constructs but that does not necessarily mean they will. While interviews could be useful in separating "no evidence" cases from "evidence not presented" cases, we worry these would increase admissions committee members' work load.

In terms of comparing admitted and non-admitted applicants, all 18 rubric constructs showed statistically significant differences. Given the goal of the rubric is to aid faculty in determining who to admit, we would expect the rubric to show such differences. That all constructs show differences suggests that all parts of the rubric are useful for determining who to admit.

How do the scores assigned by faculty differ by applicant's sex? We found only three constructs on the rubric that showed sex differences: physics GRE score, community contributions, and diversity contributions. Given known scoring gaps on the physics GRE [52], it is not surprising that males are rated more highly than females on the physics GRE score. Given that females perform larger amounts of service work in academia [196], it is also not unexpected that constructs measuring these would show a difference between sexes. Because the constructs that show sex differences are related to effects documented in the literature, we believe that the rubric is reflecting inequities that already exist rather than creating additional ones. Therefore, we conclude that the

rubric is not providing an advantage to male or female applicants.

Additionally, the constructs of the rubric that do not show differences between sexes also align with what we would expect based on the literature. The result that physics and math GPA did not differ by sex aligns with the findings of [156] and the result that noncognitive skills did not differ by sex aligns with the general finding that noncognitive skills do not appear to depend on demographics [180, 183].

How do the scores assigned by faculty differ by the type of institution the applicant attended? When we compared applicants based on whether their undergraduate institution was a more or less selective institution, we found that the only constructs that showed differences were the general GRE and physics GRE scores. This result aligns with the results of our previous work investigating the physics GRE scores by undergraduate institution type [69, 197]. We note that if we instead define more-selective universities to include large state universities, such as Michigan State University, University of Colorado, Boulder, and University of Washington, our results are unchanged. This redefinition is equivalent to considering Barron's values of 1-3 as more-selective and everything else as less-selective compared to the our definition of more-selective as Barron's values of 1 and 2 in the Methods and Results sections.

The interpretation of the results when comparing applicants from larger or smaller physics departments is less straightforward because the results do depend on how we define "larger" and "smaller" departments. When we define larger programs as those that ranked in the top quartile of physics bachelor's degrees granted as measured by the median number of degrees awarded over the last three years of available data and rounded up in-between ratings, we find that the physics GRE score and conscientiousness showed differences between applicants from larger and smaller programs. However, if we rounded down on in-between ratings instead, only physics GRE score showed a difference between applicants from larger and smaller programs.

Furthermore, alternative definitions of "larger programs" also produced varying results. One could also have reasonably defined "larger" to mean 1) in the top half of physics bachelor's degrees granted as measured by the median number of degrees awarded over the last three years, 2) in

the top quartile of physics bachelor's degrees granted as measured by the total number of degrees awarded over the last three years, and 3) in the top half of physics bachelor's degrees granted as measured by the total number of degrees awarded over the last three years. When we also consider rounding up or rounding down in-between ratings, we could make various combinations of physics GRE score, general GRE score, physics coursework, and conscientiousness show a statistically significant difference. The only rubric construct that always showed a statistically significant difference regardless of how we defined "larger programs" was the physics GRE score. Therefore, the results suggest that applicants from larger physics programs score higher on the physics GRE than applicants from smaller program do, but the results are inconclusive as to whether other areas of the rubric might show differences based on the size of the physics program the applicant attended.

One area that unexpectedly did not show differences regardless of how we defined "larger program" was the research section. It is often assumed that students at larger programs have more opportunities to engage in research than students at smaller programs. Yet, even if that is true, it does not appear to be reflected in the rubric scores.

4.6 Limitations

Our study has three main limitations. First, our study does not include many disadvantaged groups in higher education who might not have the same opportunities as their more privileged peers and hence, may score lower on the rubric. While race is the most obvious due to the way our university interprets Proposal 2, our study does also not include a comparison of low-income applicants to higher-income applicants or first generation applicants to continuing generation applicants.

Additionally, the size of our study does not allow us to explore intersections and where possible inequities may lie. As Rudolph et al. note, using small sample sizes with sub-groups has insufficient statistical power and could lead to invalid inferences [138]. Hence, we refrained from performing such analyses in this paper.

Second, this study included only a single program. Under a more traditional graduate admissions system, physics has been called a "high consensus" discipline [46], meaning that physics faculty

tend to agree on what a "quality" applicant is and therefore, a single department's admissions process would be more or less representative of graduate admissions processes in physics. When switching to rubric-based admissions, we cannot necessarily make that same claim. As our rubric was created based on what faculty value, it is not unreasonable to assume that the results would generalize to other departments that also use rubric-based admissions. However, until such processes are evaluated at other departments, we cannot make sure a claim.

Third, as a result of using only one program, the applicants are likely not representative of the larger population. The data in this study comes from 1) people who applied to our program and 2) had a nearly complete application. Thus, if we consider those with an interest in attending physics graduate school as our population, we first selected on those who applied to graduate school, then selected on those who applied to our program, and finally selected on those who provided enough information in their applications for faculty to evaluate. At each step, we are excluding some of the larger population and thus our claims cannot necessarily be expected to hold for the larger population of potential applicants.

4.7 Future Work

As noted in the limitations, our study compared rubric scores of males and females and applicants from larger or more selective programs with applicants from smaller or less selective programs. Future work could then explore how rubric-based admissions may impact other historically and currently underrepresented groups in physics such as Black, Latinx, or Indigenous applicants. Racism, and specifically anti-Black racism, is still prevalent in physics [198–202] and therefore might be reflected in rubric-based admissions.

While physics faculty tend to think of diversity mainly in terms of race [46], we acknowledge that diversity is broader than race and studies of equity around the rubric should also consider first generation applicants, low-income applicants, disabled applicants, and veterans. Studies of undergraduate admissions suggest that when extracurriculars and subjective assessments of character and talent gleaned from essays and recommendations are added to the admissions process,

existing inequalities may increase [203] and these applicants may become further disadvantaged in the admissions process. Therefore, future work should ensure that rubric-based admissions do increase equity rather than just use a new tool to perpetuate existing inequities.

Second, future work should examine how the use of rubrics may affect what parts of an application drive the admissions process. In our prior work, we found that the physics GRE and grade point average were the main drivers of the admissions process [65]. Given the rubric is designed to emphasize more than just grades and test scores, we would hope to see these factors deemphasized under the rubric system. Such a result would suggest that the rubric is fundamentally changing how faculty are reviewing applicants. We explore whether that is the case in Chapter 5.

Third, future work could examine how rubric-based admissions may change the type of applicant admitted and student outcomes. Faculty skeptical of holistic admissions may worry that by deemphasizing grades and test scores, their program is admitting less academically prepared students. Future work can explore if these fears have any merit. Research at the undergraduate level on holistic admissions has found that adding noncognitive traits increased graduation rates, especially among those from disadvantaged backgrounds [204]. At the graduate level, a study of a materials science and engineering program found that after changing their admissions to include noncognitive skills, their incoming students won more university fellowships, though the authors cautioned they could not attribute the increase in fellowships solely to their changes in admissions [205]. Thus, evidence from outside of physics suggests that these fears may be unfounded, but we will not know for sure until physics specific studies are conducted.

Additionally, future work can examine noncognitive skills in physics more broadly. Physics has been characterized as a brilliance-dominated field [59] and hence, it is not surprising that most studies of success in physics have also focused on cognitive measures such as grades, exam scores, and standardized test scores. While such studies could be useful at all levels of physics, studies at the graduate level are especially importance given the limited number of studies exploring their useful for predicting success in graduate school. [138].

Finally, future work around equity in graduate admissions should investigate who is invited

to apply to graduate school in the first place, what barriers those who do not apply but wish to encounter, and how those barriers may be removed. In previous work, Cochran et al. investigated what barriers applicants to physics graduate school through the APS Bridge Program perceived, finding that GRE scores, lack of research experience, low GPA, program deadlines, and application costs were common concerns [49]. Unless we also work to make the application process more equitable, making the evaluation process more equitable will not result in large-scale changes in equity at the graduate level.

4.8 Recommendations for Departments

The results of this study suggest a general recommendation to implement rubrics in physics graduate school admissions. Rubrics can aid reviewing applications by standardizing the process and limiting bias and using rubrics does not appear to increase the time to review applications.

Of course, simply using a rubric will not result in changes unless it is implemented well. We therefore propose three more specific recommendations.

First, we recommend that admissions committees have multiple members review each application. For a well-constructed rubric, there should be limited uncertainty as to what rating an applicant will receive. However, for constructs that are more subjective in nature, faculty may have differing opinions about what counts as achieving each level. For example, for the quality of work construct on our rubric, what counts as "making significant contributions to the project" might vary based on the reviewer. Therefore, having multiple reviewers can reduce potential bias when reviewing applications.

Second, following the call of others [194, 206, 207], we recommend that members of the admissions committee should be of diverse backgrounds and representative of the applicant pool. To accomplish that, departments might also consider adding non-tenure stream faculty, post-docs, and current graduate students to their admissions committees, providing appropriate recognition and compensation as necessary. Prior work has shown that faculty may prefer to admit applicants like themselves [46] and therefore, a representative admissions committee is needed to ensure that

minoritized applicants are given equal consideration.

Finally, we recommend that departments conduct regular self-studies of their graduate admissions processes and share the results. While Rudolph et al. have previously called for departments to conduct self-studies of their admissions process [138], we believe it is equally important to share the results of those self-studies so that the physics community can know what is and what is not working. This collective knowledge of what is working and what is not working can then be used by all to improve graduate admissions in physics for everyone.

For the sharing of results to be impactful however, the results must be easy to access and easy to understand. While individual departments could post their results on their websites, we believe doing so adds an extra layer of complexity and makes the results harder to access. Instead, we advocate for a centralized system to be created so that departments can easily report their data in a standardized way and practitioners can easily see and compare results across programs. Such a system could be maintained by professional societies such the American Physical Society or the American Institute of Physics, or other organizations. A system like this has been designed for research-based assessments [208], but to our knowledge, there exists no such system for graduate admissions.

However, when conducting such self-study of what is working well and what is not working well, it is important to consider the question of "working well for whom?". As Razack et al. note, "working well" depends on one's social positioning [192] and therefore, a change that works well for applicants of one background may not be working for applicants of a different background. By considering the "for whom?", the physics community can ensure that changes made are for the benefit of all rather than as new methods to continue the existing exclusionary practices in graduate admissions.

4.9 Conclusion

In this paper, we demonstrated that rubric-based admissions are a promising avenue for increasing equity in graduate admissions. We showed that faculty ratings of applicant's grades, research

experiences, and noncognitive abilities do not differ based on the applicant's sex or undergraduate background. The differences we did observe in faculty ratings could be explained as observing known systematic issues in physics regarding test scores and service work expectations.

Based on the results of this study, we recommend that departments use rubric-based holistic review for their graduate admissions process. Multiple people should review each application and those people should be representative of the applicant pool to limit any bias in the review process. Finally, departments should engage in self-study to see how their graduate admissions process is working and share those results so that the physics community can collectively learn what is working and what is not working in making graduate admissions more equitable.

CHAPTER 5

A "NEW APPROACH" OR THE SAME APPROACH IN NEW PACKAGING?

This chapter is being drafted as a journal article. The working manuscript version includes Marcos D. Caballero as the second author. Following the Contributor Roles Taxonomy (CRediT) [76], my roles for this project include conceptualization, formal analysis, methodology, software, validation, visualization, and writing the original draft.

5.1 Introduction

Physics departments are increasingly interested in making graduate admissions more equitable and rubric-based holistic review has been gaining traction as a possible route to do so. Unlike the traditional approach that emphasizes test scores and grades (see Chapter 2), rubric-based admissions extend the criteria of interest to include noncognitive competencies, fit with the department and research accomplishments. In theory, all applicants are evaluated on the same set of explicit criteria so the process should be more fair and there should be less bias [183].

In Chapter 4, we introduced our department's approach to rubric-based admissions. The results suggest that our rubric is equitable among men and women and with respect to applicants from smaller or selective schools. However, just because the rubric is more equitable does not mean that it made our admissions process more equitable or even changed the factors that drive our admission process. That is, the rubric might just be a new tool to do the same process.

The goal of this chapter is then to go beyond comparing rubric scores for different applicants and instead, consider the admissions process as a whole. Specifically, we ask how did the introduction of the rubric change our program's admissions process? To operationalize that question, we ask two research questions:

1. How do admissions models before and after the implementation of the rubric differ in terms predictive ability and meaningful features when our models are based on the data contained

in applications?

2. How does using the data produced by faculty when rating applicants using the rubric affect our ability to create admissions models?

To answer these questions, we compare admissions models of the current process using data from both faculty ratings and the applications to the models we generated in Chapter 2 of the program's initial process. From Fig 2.3, we notice that there are cases where applicants have similar physics GRE scores and GPA yet one applicant is accepted while the other is not. Given that cases such as these might add additional challenges to modeling the data, removing such applicants might allow us to better characterize the general trends in the data. We therefore consider a new approach that detects similar applicants with different admissions outcomes and removes them from the data set: Tomek Links [209]. We then ask a third research question:

3. How does using Tomek Links affect our ability to model the admissions data, both before and after the implementation of the rubric?

5.2 Background

While admissions committees use common criteria for initially judging applicants, deliberations of borderline applicants under the traditional process might come down to subtle distinctions that were not used for other applicants [46]. Thinking in terms of a modeling perspective, this means that some applicants might be assessed according to additional criteria and hence, these borderline applicants might not be easily classified by a general model of the admissions process. As a result, including these borderline applicants might cause our model performance to suffer while excluding these applicants and instead focusing on a more typical applicant could improve model performance.

Unfortunately, whether an applicant is a borderline applicant is not included in faculty ratings of applicants and hence, we do not know who is a borderline applicant. To determine who might be a borderline applicant, let us assume there is a predictive model of a graduate admissions process that

perfectly separates those who are admitted and those who are not admitted in some n-dimensional application space. We could then say that those applicants who are near the n-dimensional boundary that separates the admitted applicants and not admitted applicants are borderline applicants. To differentiate borderline applicants in the admissions process from borderline applicants in the modeling process, we will refer to the latter as *boundary applicants*. Such a definition of *boundary applicants* is similar Hoens and Chawla's definition of borderline cases in classification, which are cases where a small change in the features would cause the classification boundary to shift [210].

However, such an approach assumes that those who are admitted and not admitted can be cleanly split in some n-dimensional space and are not intermixed. For a variety of reasons, an applicant with a stellar application might be rejected or an applicant with a weaker application might be admitted and hence an admitted applicant might fall on the not-admit side of the separating boundary or vice versa. While these applicants might not be borderline in the traditional sense, their admission decision likely would have required deliberation and hence, might have gone through a similar process as a borderline applicant. We should therefore also consider these applicants as borderline applicants in the sense of possibility hurting our model's performance. Perhaps more accurately, we should refer to these applicants as *noise applicants* following Hoens and Chawla's definition of noise cases, which are case that result from random variation and not are representative of the underlying pattern [210].

While we have operationalized borderline applicants in terms of a model as *boundary applicants* and *noise applicants*, we still need a method to determine which applicants these are before constructing any models. Tomek Links offers one possible method as it is a method of identifying the boundary or noise cases in the data [209].

To identify the Tomek Links in a data set, the distance between all cases in the data set are computed. Using the distances, the nearest neighbor of each case is computed. For two cases, e.g. case 1 and case 2, the cases are Tomek Links if and only if case 1 is the nearest neighbor of case 2, case 2 is the nearest neighbor of case 1, case 1 and case 2 are of different classes. The only way for these conditions to be fulfilled is if case 1 and case 2 are boundary cases or if case 1 or case 2

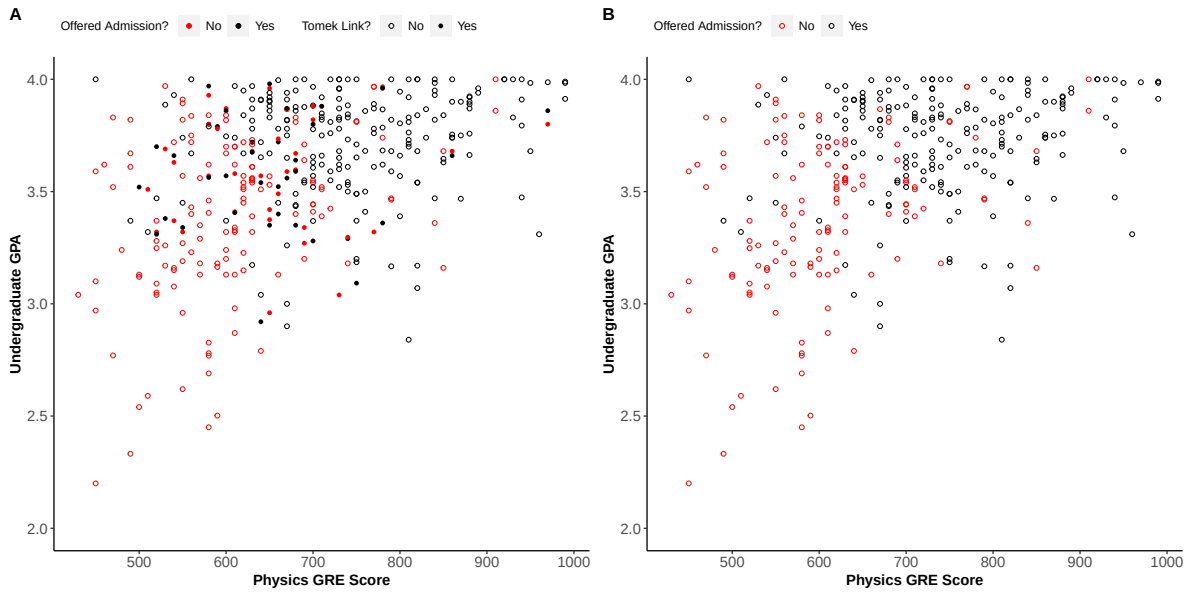

Figure 5.1: Plot A shows Fig 2.3 with the Tomek Links marked. Filled points represent Tomek Links. Plot B shows the same plot after the Tomek Links have been removed

is a noise case [210]. Therefore, Tomek Links allows us to identify *boundary applicants* and *noise applicants* in our data. An example of this approach in practice is shown in Fig. 5.1.

While Tomek Links have been successfully used in other contexts (e.g. see [211–213]), these approaches have tended to use data augmentation in conjunction with Tomek Links. While data augmentation approaches are valid from a modeling perspective, they might be questionable from a ethics and policy perspective. For example, altering the data set might lead to a model that is highly inaccurate of the underlying process [214]. For our data set, using data augmentation is analogous to creating applicants and thus our conclusions about how our admissions process might or might not have changed would be based on both real and imaginary applicants. For this reason, we will not use data augmentation.

As we note in our methods, we do impute our data. Readers may view this as a contradiction of the previous paragraph but we view data imputation and data augmentation as different. Data imputation is using the existing data to fill in the missing values. In the case of multiple imputations, which we use in this study, the filling in happens multiple times in multiple ways so that the results represent the average result across many possible ways the complete data set might have looked. In

contrast, data augmentation is using the existing data to create new data rather than fill in "holes" in the data. More generally, data imputation is estimating the results as if we knew the values of the missing data while data augmentation is creating estimating new data to simulate a bigger data set.

5.3 Methods

5.3.1 Data

Data for this study comes from applications to our graduate program to enroll in fall 2014 thorough fall 2020. Applicants submitted general and physics GRE scores, transcripts, a personal statement, a research statement, and letters of recommendation. Starting for the cohort to begin our program in fall 2018, the admissions committee used a rubric to rate applicants on 18 criteria. Those scores are also included in our data. Further details about the data from fall 2014 through fall 2017 are discussed in Chapter 2 and further details about the data from fall 2018 through fall 2020 and the rubric are discussed in Chapter 4.

For convention, I will refer to data collected before the implementation of the rubric (fall 2014 - fall 2017) as *data set 0* following the convention of using "naught" for initial time in physics and data collected after the implementation of the rubric (fall 2018 - fall 2020) as *data set 1*, following the convention of using "1" to be mean the next time the thing was measured. Furthermore, data in data set 1 that comes from the applications will be referred to as the *data set 1a* while data that comes from the faculty ratings using the rubric will be referred to as *data set 1b* data. These are summarized in Table 5.1

5.3.2 Modeling

To model our data, we used the `cforest` algorithm [92, 96, 99] in R [98]. As in Chapter 2, we used 70% of our data to train the model, 500 trees to build the forest and $\sqrt{p}$ of the p features to construct each tree. We ran each model 30 times, selecting a new training and test set each time and averaged the results over the runs. For each trial, we calculated the training AUC, testing

Table 5.1: The three models compared in this chapter and the data that went into each

Name	Data source and features	Where results are reported
Data Set 0	Information pulled from the applications before our department implemented a rubric (2014-2017). Features are shown in Table 2.1	Section 2.3
Data Set 1a	Information pulled from the *applications* after our department implemented a rubric (2018-2020). Uses the same features as model 1.	Section 5.4.1
Data Set 1b	Rubric ratings generated *by* faculty as they evaluated applications (2018-2020). Features are shown in Table 4.1.	Section 5.4.3

AUC, testing accuracy, null accuracy, and the permutation AUC importances. At this stage of the analysis, missing data was handled using the default `cforest` procedures [101].

For data sets 0 and 1a, the same features were used as in Table 2.1, with the size of the physics program factors updated with new data for the post-data models. For data set 1b, all features were treated as categorical (0, 1, or 2) and as in Chapter 4, any values between a rubric level were rounded up.

For both data sets 1a and 1b, we also varied our choice of hyperparameters to determine if our conclusions depended on our modeling choices. As in Chapter 2, we set the training fraction to be either 0.5, 0.6, 0.7, 0.8, or 0.9, the number of trees in the forest to be 50, 100, 500, 1000, or 5000, and the number of features used for each tree to be 1, $\sqrt{p}$, $p/3$, $p/2$, or p for a total of 125 possible combinations. Each model was grown using the same procedure listed above.

To account for correlations among the features, we also calculated the conditional importances for the both data sets 1a and 1b. We used the MICE algorithm with the default choices [102] to impute missing data, calculated the conditional importances, and then pooled the results using Rubin's Rules [103].

To compute the Tomek Links, we used the `TomekClassif` function in the `UBL` package [215]. We first used MICE to impute the data before calculating the Tomek Links using the function

defaults with the exception of the distance metrics. Following the recommendation of the package's documentation, we used the `HVDM` distance for data sets 0 and 1a because those data sets contain both categorical and continuous data and we used the `Overlap` distance for data set 1b because all features were categorical.

After removing the Tomek Links, we ran each model 30 times and averaged results. Results were then pooled using Rubin's Rules.

5.4 Results

5.4.1 Data Set 1a

Across the 30 runs, the average accuracy of our model predicting on the held-out data was 71.4% ± 0.6%, the average training AUC was 0.720 ± 0.004, and the average testing AUC was $0.626 \pm .006$. Our null accuracy was 66.0% which suggests that our model is only doing slightly better than if it were to predict everyone was not admitted to our program. The low testing AUC also suggests a poor model.

When looking at the feature importances (Fig. 5.2), we see that the physics GRE score, undergraduate GPA, quantitative and verbal GRE scores, and proposed research area are near the top while the institutional features near the bottom. Performing the backward elimination, we find that physics GRE score, undergraduate GPA, quantitative and verbal GRE scores, and proposed research area are the meaningful features predictive admission after the implementation of the rubric.

We can then plot the ranks of the features so that we can compare them to the ranks of the features in data set 0. The resulting slopeplot is shown in Fig. 5.3. We notice that the order of features is largely unchanged for the most predictive features, with only quantitative GRE score and GPA switching places. The major difference between the features predictive of admission in datasets 0 and 1a is the number of meaningful features.

When we take correlations among the features into account however, we the set of meaningful features shrinks. As shown in Fig. 5.4, only the applicant's physics GRE score and proposed

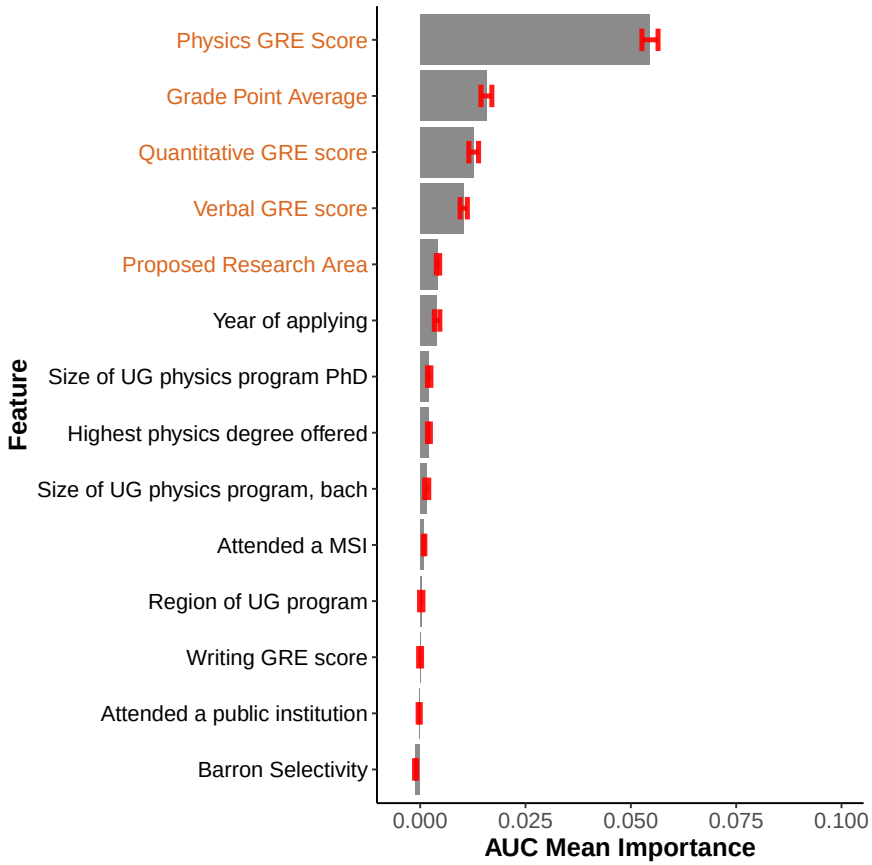

Figure 5.2: Averaged AUC feature importances over 30 trials. Physics GRE score, undergraduate GPA, Quantitative GRE score, Verbal GRE score and proposed research area, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted.

Table 5.2: Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations for models built from data set 1a

metric	min	median	max
Train AUC	0.602	0.735	0.749
Test AUC	0.549	0.633	0.676
Test Accuracy	0.679	0.712	0.732
Null Accuracy	0.645	0.661	0.666

research area were found to be predictive of admission. It is also important to note that the quantitative and verbal GRE scores are ranked lower once correlations are accounted for, suggesting that their initial importances were inflated.

Given the poor performance of our model, hyperparameter tuning might have improved the model. While it did to a degree, the testing accuracy was still only a few percentage points

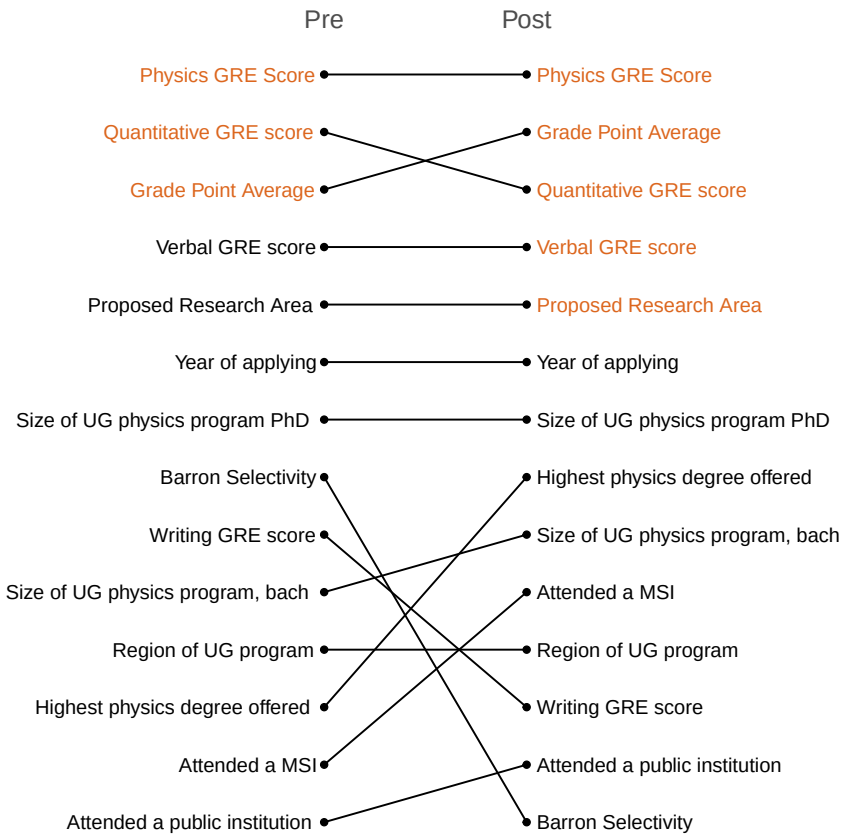

Figure 5.3: Slopeplot showing the ranks of each feature before the implementation of the rubric (left) and the after the implementation of the rubric (right) using data sets 0 and 1a respectively. Features toward the top of the plot are more predictive. Features in orange were found to be the meaningful features needed to predict whether the applicant was admitted in their respective model. Notice that the ordering of the more predictive features is largely unchanged. Plot adapted from [216].

above the null accuracy and the testing AUC was still below 0.7 (Table 5.2). Thus, even with hyperparameter tuning, the models of data set 1a were poor.

Finally, to see how the feature ranks varied based on the hyperparameters, we plotted the occurrence fraction of each rank for each feature (Fig. 5.5). We notice that across the 125 hyperparameter combinations, physics GRE score and GPA are almost always the top two features followed by quantitative and verbal GRE scores. In addition, none of the institutional features never rank in the upper half of the importances. These results are similar to what we found for data set 0 in Chapter 2.

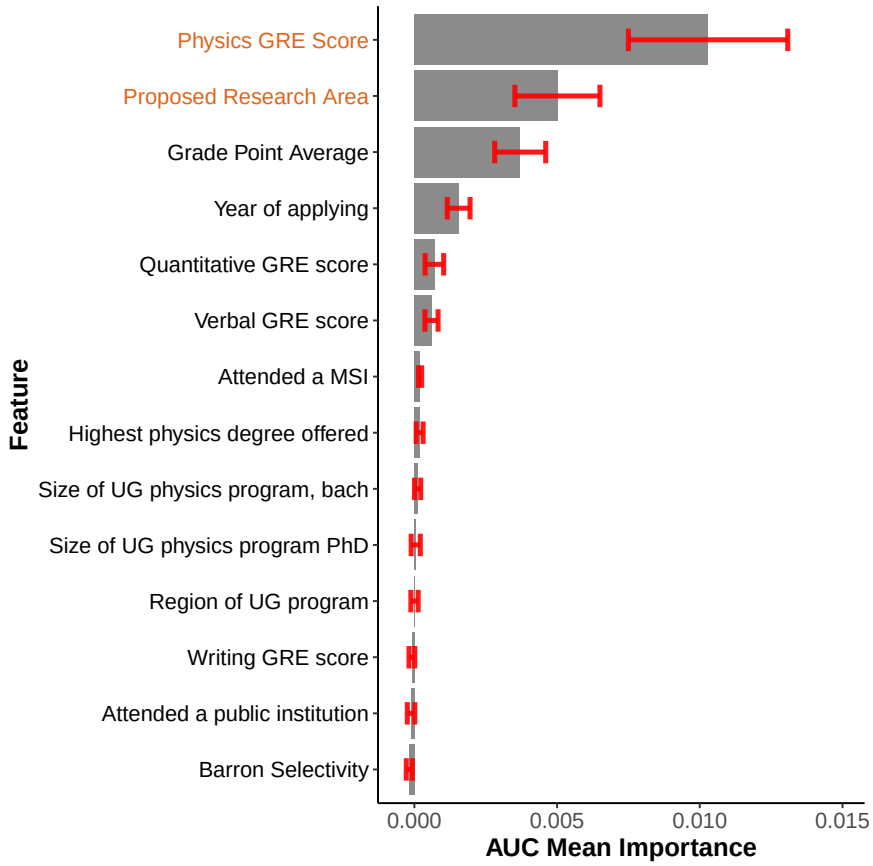

Figure 5.4: Averaged conditional feature importances over 30 trials. Physics GRE score and proposed research area, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted once correlations were accounted for.

5.4.2 Using a True Testing Set

In addition to looking at the feature order to determine if the admissions process changed, we can compare the performance of the models themselves. If the process didn't change, then a model built from data set 0 should perform equally well on a data set 0 testing set as on data set 1 and a model built from data set 1a should perform equally well on a data set 1a testing set as on data set 0. If the process did change, we would expect better performance on the test data pulled from the train/test split than other data set.

When looking at the results, which are shown in Figures 5.6 and 5.7, we see that the latter case better describes the data. In Figure 5.6A, we see that the data set 0 test AUC is larger than the data set 1a AUC, and in Figure 5.7A, we see that the data set 0 test accuracy is larger than the data set 0

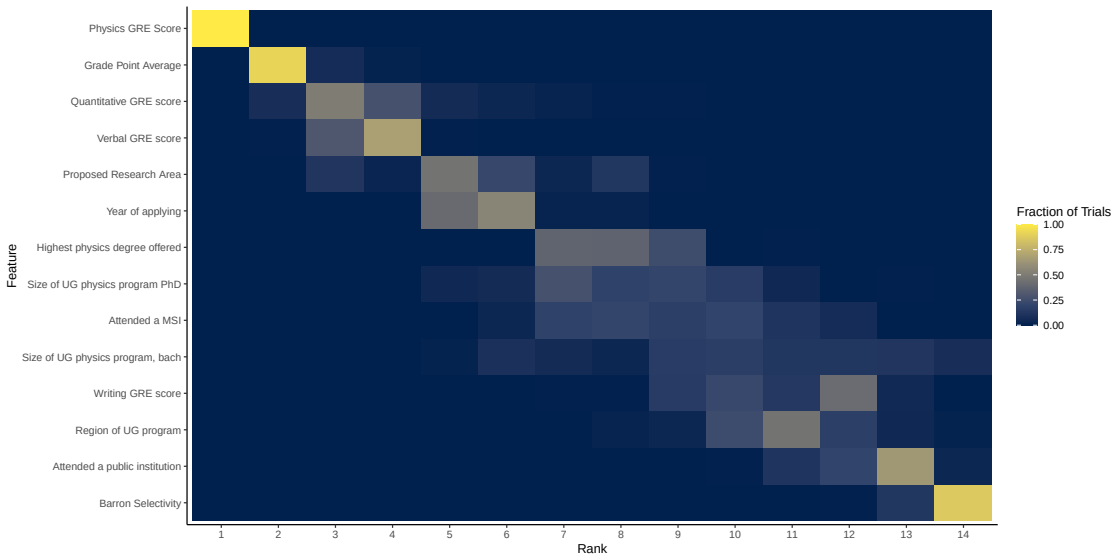

Figure 5.5: Proportion of the 125 hyperparameter combinations in which each feature had a given rank for data set 1a. Notice that the plot is mostly diagonal and that physics GRE score and GPA are almost always the top two features.

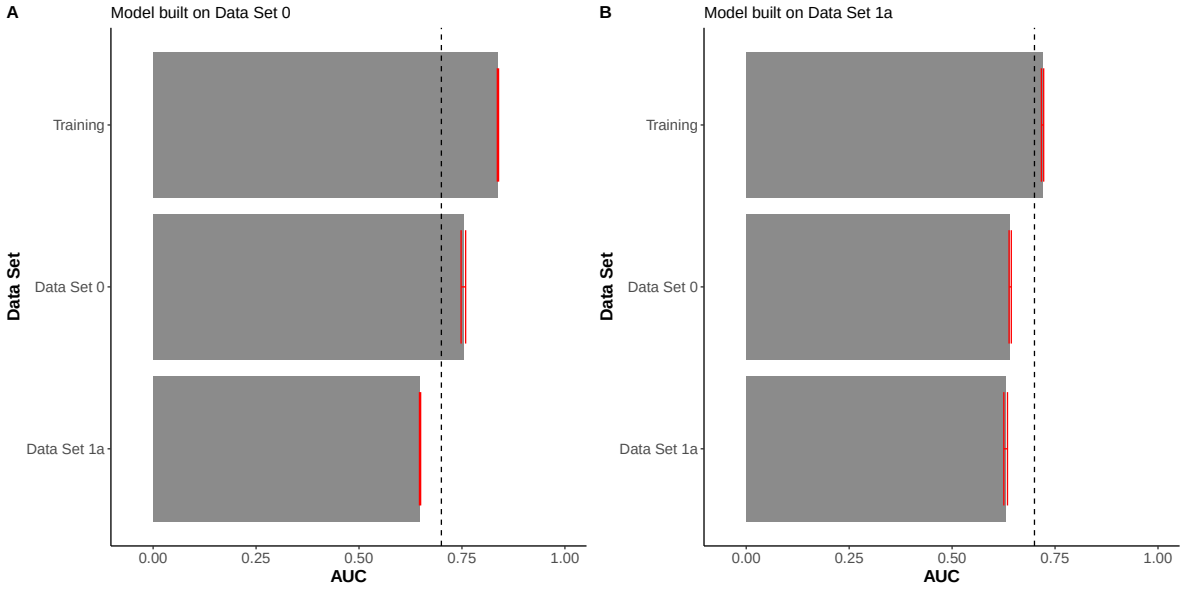

Figure 5.6: Comparison of the testing AUC when A) Data Set 0 is used to train the model and B) when Data Set 1a is used to train the model. Training refers to the training AUC for the model. All error bars are 1 standard error. Results were averaged over 30 trials.

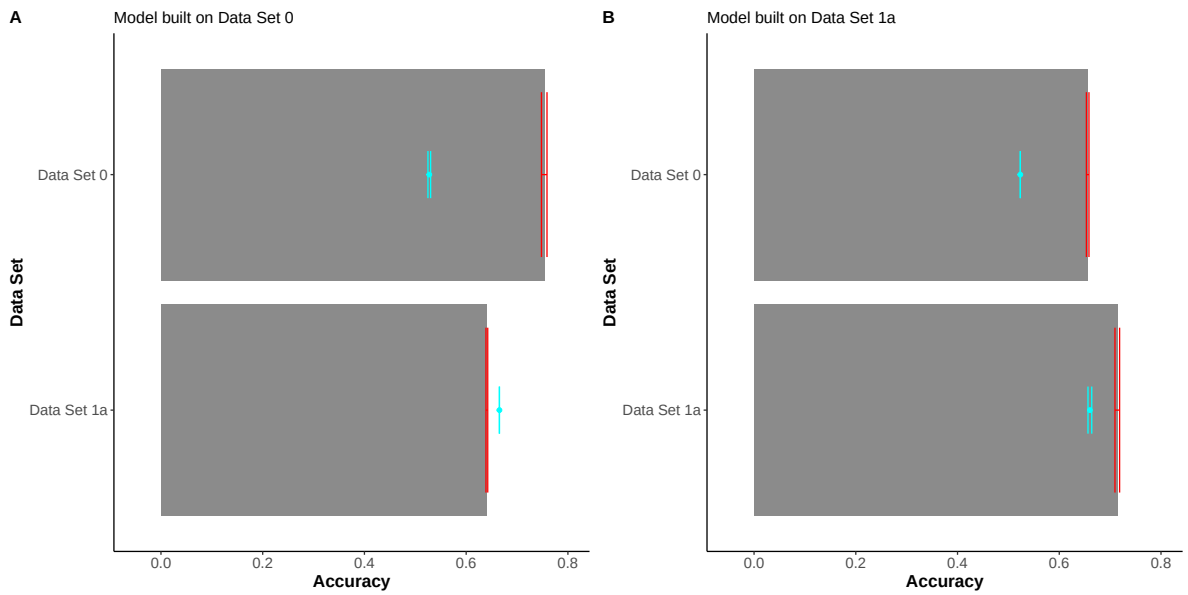

Figure 5.7: Comparison of the testing accuracy when A) Data Set 0 is used to train the model and B) when Data Set 1a is used to train the model. The null accuracy is shown in cyan with the shorter in height error bars. All error bars are 1 standard error. Results were averaged over 30 trials.

null accuracy while the data set 1a test accuracy is smaller than the data set 1a null accuracy. These metrics suggest that the data set 0 model fits data set 0 well but does not fit data set 1a well and therefore, that the process might have changed.

Looking at Figure 5.6B, we see that none of the metrics are especially good. The test AUCs are both in the poor range suggesting that the model built from data set 1a does not fit that well in the first place. It is then not surprising that the model does not predict data set 0 well. Given that the initial model did not fit the data well, we cannot use the result to make a claim about whether the process changed.

5.4.3 Data Set 1b

Given that after the implementation of the rubric applicants are rated on the rubric constructs, perhaps using the rubrics constructs instead of the application data in a model would lead to better performance. Yet, that wasn't the case. We find that the testing AUC was 0.664 ± 0.007 and the testing accuracy was 0.675 ± 0.007 (null accuracy 0.553 ± 0.006). Given that not all applicants had sufficiently complete applications to be reviewed by faculty and those with incomplete applications

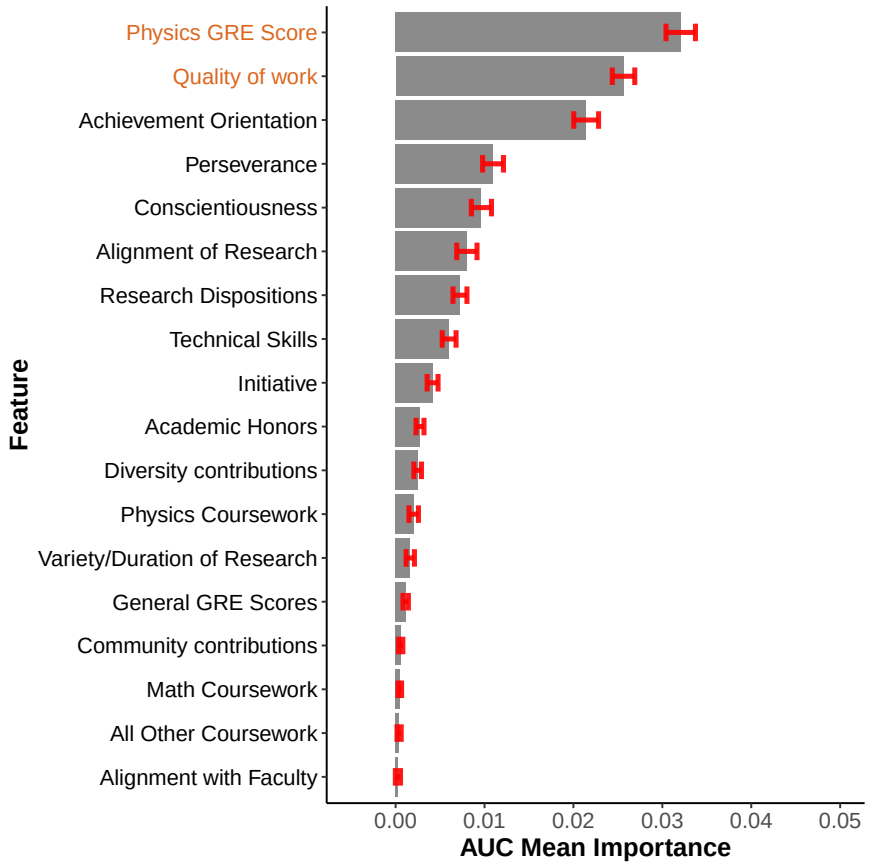

Figure 5.8: Averaged conditional feature importances over 30 trials for the models of data set 1b. Physics GRE score and quality of work, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted.

tended to be not admitted, the null accuracy is smaller for models of data set 1b than the models of data set 1a

When we looked at the feature importances, the results showed similarities to the importances from the models of data set 1a. From Fig 5.8, we notice that physics GRE score is still the top feature. However, measures of GPA such as physics coursework, math coursework, and all other coursework tended to be in the lower half of the rankings, alignment of research (the closest construct to proposed research area) was toward the middle of the rankings, and general GRE scores was toward the bottom despite GPA, proposed research area, and general GRE scores being top ranking features under the models of data set 1a.

From the figure, we also notice that measures related to research (quality of work, research

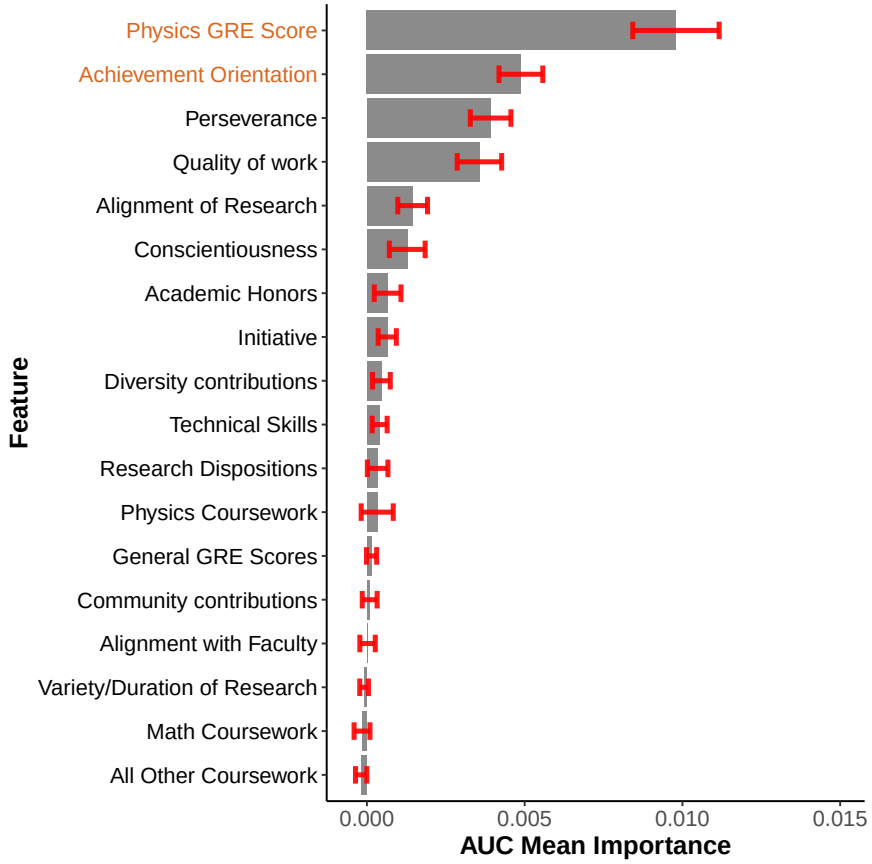

Figure 5.9: Averaged conditional feature importances over 30 trials for the models of data set 1b. Physics GRE score and achievement orientation, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted once correlations were accounted for.

dispositions, and technical skills) are ranked in the upper half as are measures of noncognitive skills (achievement orientation, perseverance, and conscientiousness) while measures of fit (diversity contributions, community contributions, and alignment with faculty) are ranked in the bottom half of features.

When performing the backward elimination, we find that only physics GRE score and quality of work are selected, suggesting that only these two features are needed to produce similar predictive performance as using all 18 features.

We then repeated the analysis taking correlations between features into account. The result is shown in Fig. 5.9. We notice that the top features are similar, though the rank of quality of work decreased to fourth. Now, physics GRE score and achievement orientation were found to be the

Table 5.3: Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations for the models of data set 1b

metric	min	median	max
Train AUC	0.711	0.767	0.791
Test AUC	0.654	0.669	0.686
Test Accuracy	0.660	0.678	0.696
Null Accuracy	0.559	0.561	0.586

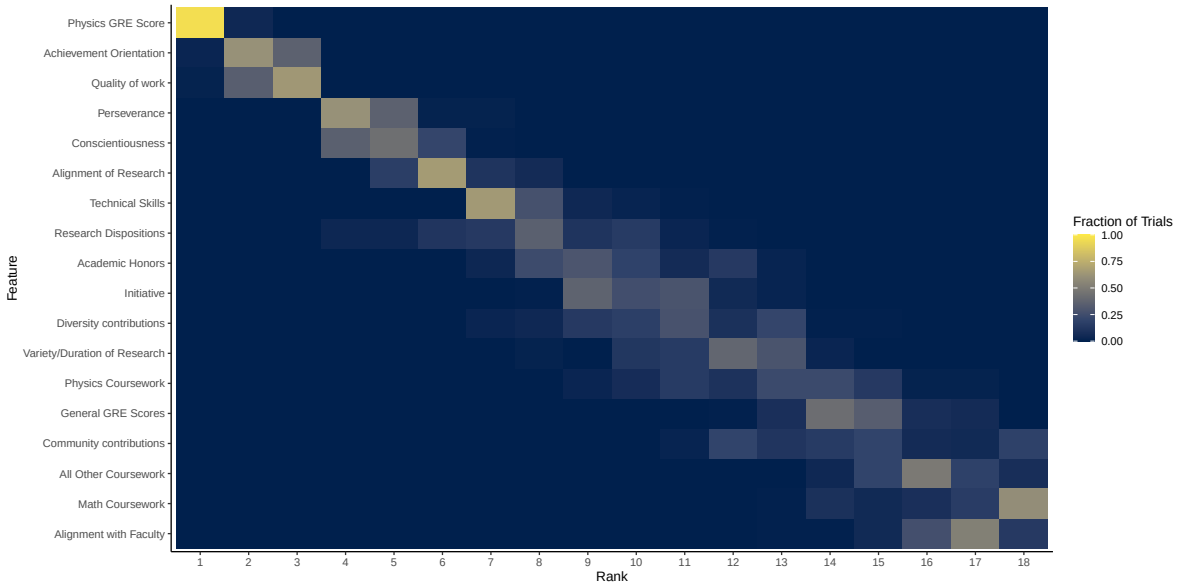

Figure 5.10: Proportion of the 125 hyperparameter combinations in which each feature had a given rank for models of data set 1b. Notice that the plot is mostly diagonal and that physics GRE score, achievement orientation, and quality of work are always the top three features.

meaningful and hence predictive features.

Finally, we performed hyperparameter tuning to determine if we could create a model with acceptable metrics. Unfortunately, we could not. Even the best AUC among the 125 hyperparameter tuning combinations did not exceed 0.7. The full results are shown in Table 5.3.

Looking at the feature ranks, we again see a diagonal pattern toward the upper left of the plot (Fig. 5.10, suggesting the same few features are selected as the most predictive. Regardless of our hyperparameter choices, the top three features are the physics GRE score, achievement orientation, and quality of work. However, the pattern becomes less diagonal toward the bottom right, suggesting that these features are more or less noise in the model.

Table 5.4: Metrics when using Tomek Links and MICE for each of the three data sets

	Data Set 0	Data Set 1a	Data Set 1b
Cases Dropped	11%-14%	15%-18%	12%-17%
Training AUC	0.880 ± 0.004	0.760 ± 0.015	0.779 ± 0.010
Testing AUC	0.809 ± 0.009	0.670 ± 0.015	0.704 ± 0.014
Testing Accuracy	0.806 ± 0.009	0.775 ± 0.012	0.717 ± 0.012
Null Accuracy	0.539 ± 0.006	0.699 ± 0.009	0.575 ± 0.010

5.4.4 Tomek Links

Given the limited ability of the conditional inference forest to model data sets 1a and 1b, we used Tomek Links to remove boundary cases. As we were removing cases, we did not compute importances and focused on the model metrics instead. The results are shown in Table 5.4. As MICE generates new values for each imputation and hence, affects which cases are nearest neighbors, the percent of cases dropped for each trial varies.

First, we notice that for data set 0, using Tomek Links increased the testing AUC and testing accuracy by 0.05 over original model reported in Chapter 2. In fact, the testing AUC is now about 0.8 which is considered "good" as compared to "fair" for the original model [93].

Likewise, using Tomek Links also results in an approximately 0.05 increase in the testing AUC and testing accuracy for data set 1a. However, the AUC is still in the poor range and the testing accuracy is only slightly better than the null accuracy.

For data set 1b, using Tomek Links increases the testing AUC and testing accuracy by approximately 0.04. This time, the increase to the testing AUC is enough for the model to be classified as "fair".

To better understand what Tomek Links were doing in the modeling process, we investigated how removing the boundary cases affected the decision boundary. In order to plot the results, we only used the physics GRE score and undergraduate GPA to make a simple model for data sets 0 and 1a. To compute the Tomek Links, we used MICE to create a complete data set first and then found the Tomek Links. As all the data in data set 1b was categorical, a 2D plot of the decision boundary would have yielded limited insight and hence, we did not do so. The results of a single

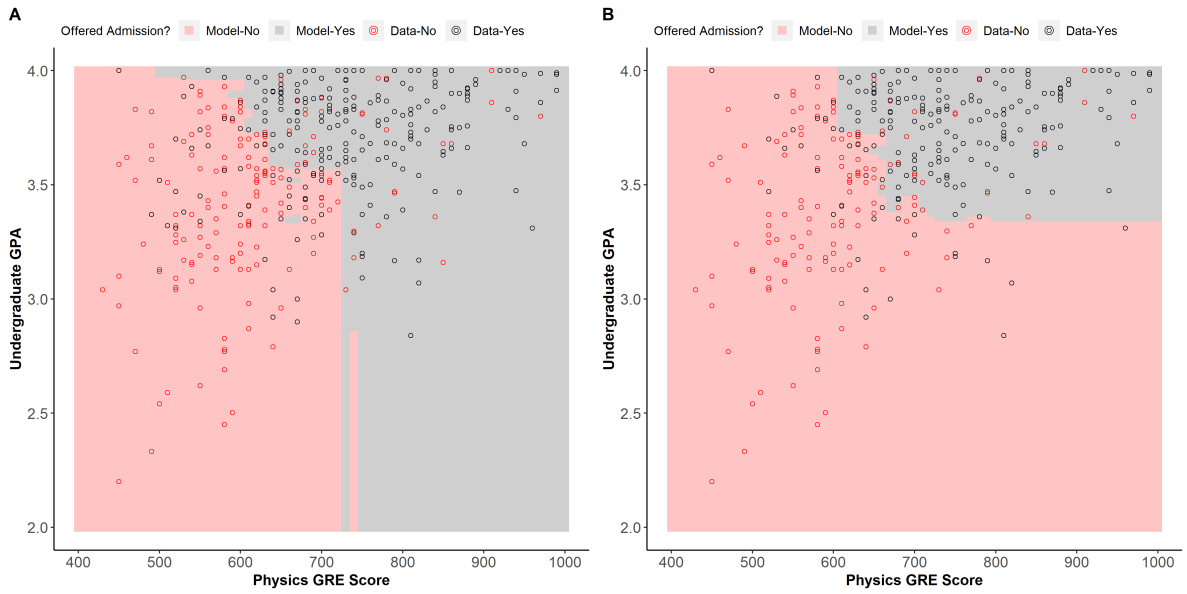

Figure 5.11: Plot A shows data set 0 with the decision boundary for a model with just the physics GRE score and undergraduate GPA (Fig 2.3) while plot B shows the data with the Tomek Links removed and the resulting decision boundary for the 2D model.

trial are shown in Figures 5.11 and 5.12.

From the figures, we see that removing the Tomek Links does affect the boundary. In the case of Figure 5.11, we see the area with limited data in the lower right switches to not admitted and in general, the overfitting is reduced. In addition, the decision boundary matches closer to what we might expect anecdotally and based on Chapter 3 in that having a higher physics GRE score and GPA is more likely to result in admission as opposed to having only one of those being stellar.

Likewise, in Figure 5.12, we again see reduced overfitting in the decision boundary. We also see that higher physics GRE scores and GPA are predicted to result in admission as was the case before the implementation of the rubric. However, the threshold for what counts as a high physics GRE score and GPA seems to be higher after the implementation of the rubric based on the decision boundaries.

5.5 Discussion

Here, we first provide answers to our research question and then use those answers to address the larger question of whether our department's admissions process changed.

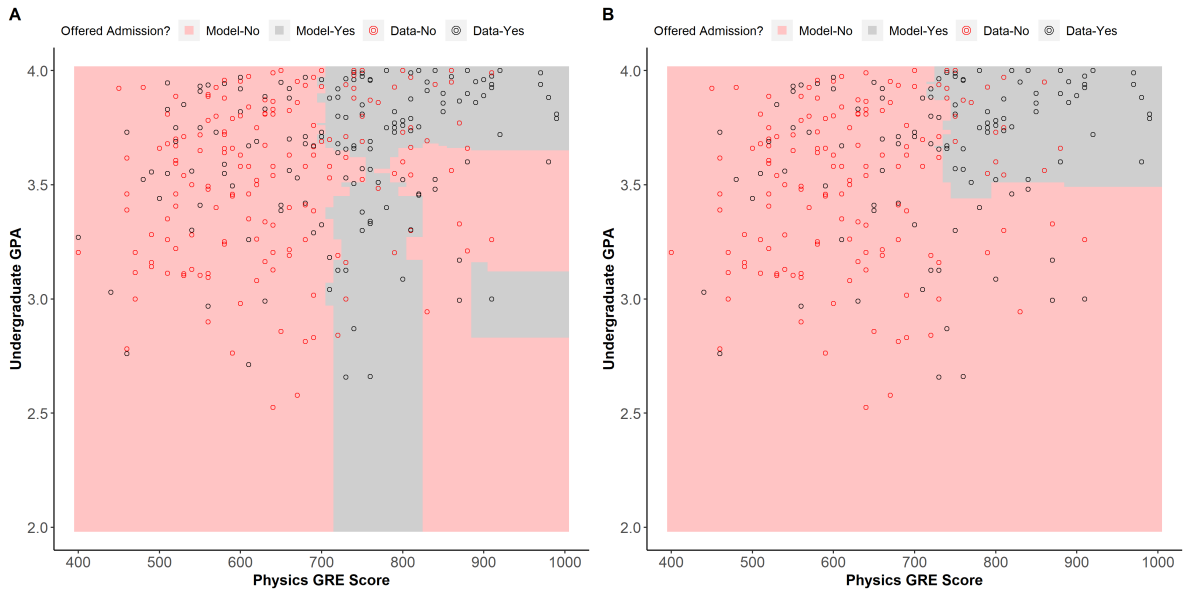

Figure 5.12: Plot A shows data set 1a with the decision boundary for a model with just the physics GRE score and undergraduate GPA. Plot B shows the data with the Tomek Links removed and the resulting decision boundary for the 2D model.

5.5.1 Research Questions

How do admissions models before and after the implementation of the rubric differ in terms predictive ability and meaningful features when our models are based on the data contained in applications?

While we were able to model the data before the implementation of the rubric to an acceptable degree, we were unable to do so for the data after the implementation of rubric. Even after hyperparamter tuning, we were unable to achieve a testing accuracy more than a few percentage points above the null accuracy or a testing AUC above 0.7, suggesting a poor model.

In terms of the meaningful features for data sets 0 and 1a, they were more or less the same. For data set 0 presented in Chapter 2, we found the applicant's physics GRE score, quantitative GRE score, and GPA to be the meaningful features while for data set 1a, we found the physics GRE score, GPA, quantitative GRE score, verbal GRE score, and proposed research area to be meaningful. After taking correlations into account only the physics GRE score and proposed research area were found to be meaningful. However, because conditional inference forests will always return

importance values regardless of how well the model fits, we should interpret the data set 1a results with a degree of caution.

Furthermore, if the top features were the same before and after the implementation of the rubric, we would expected a model trained either data set 0 or data set 1a would work equally well on the other. Yet, that wasn't what we found. Instead, we found that the model trained on data set 0 did not predict data set 1a well while the model trained on data set 1a did not predict either of the data sets well.

How does using the data produced by faculty when rating applicants using the rubric affect our ability to create admissions models?

While using the rubric features does result in increased metrics compared to the traditional features for the data collected after the implementation of the rubric, the metrics are still outside of the acceptable range. The testing AUC was still below 0.7 but the testing accuracy was greater than the null accuracy by a larger amount than the model created from data set 1a. However, that result may be explained by data set 1b having a less imbalanced outcome.

To see if that was the case, we created a model using the data in data set 1a that corresponded to the applicants in data set 1b. When we did so, we found that the metrics were comparable, but the original test data set 1b model slightly outperformed this new model (0.02 increase in testing AUC and accuracy). Thus, while some of the improvement in metrics might be attributable to the more balanced data set, using the rubric constructs also provided some benefit.

In terms of the features, we noticed some similarities and some differences. For the models of data set 1b, the physics GRE was still the top feature. However, measures of the GPA and general GRE scores were ranked in the lower half, suggesting they might not have been as important. Instead, measures of research ability and experience and noncognitive skills tended to be ranked towards the top. Again however, we should interpret the data set 1a results with a degree of caution as the model does not fit the data especially well.

How does using Tomek Links affect our ability to model the admissions data, both before and after the implementation of the rubric?

Using Tomek Links resulted in improved model performance for all three data sets. For data set 0, using Tomek Links increased the testing AUC over 0.8, which is considered "good," and for data set 1b, using Tomek Links increased the testing AUC over 0.7, which is considered "fair." However, while using Tomek Links for data set 1a did improve the testing AUC, it did not do so enough for the model to be considered acceptable.

When looking at the decision boundaries for data sets 0 and 1a with and without Tomek Links removed, we found that overfitting appeared to be reduced, suggesting that even if the metrics are not largely improved, there still may be benefits from using Tomek Links.

Thus, while the benefits were relatively small, these results suggest that Tomek Links are a promising technique for modeling PER data, especially for data sets where we expect many boundary cases or cases that go against the general trend. For example, if we were to predict who passes an introductory class, Tomek Links might allow us to remove students who earned exam scores around the minimum passing grade and thus might or might not have passed the course or anomalous students who did poorly on the midterms but managed to earn a high grade on the final to pass the class.

5.5.2 Addressing whether our process changed

Looking across the research questions, we can now address whether the introduction of the rubric changed our department's admissions process. Overall, the evidence points in the direction of the process changing.

In terms of evidence for the process changing, we find that the models of data sets 1a and 1b do not fit the data well. As we were able to fit the data set 0 models to an acceptable degree using the conditional inference forest algorithm but not the models of data sets 1a or 1b, this result seems to imply that there must be something different about the data sets. Because data set 0 and data set 1a used the same features, it is hard to explain why we could model one well but not the other unless the "true" models of the data were different and hence, the admissions process changed.

In addition, a model trained on data set 0 was better able to predict held-out data from data set

0 compared to data set 1a. If the process hadn't changed, we would have expected the predictive performance to be similar.

Finally, using Tomek Links to remove applicants who might have gone against the general trend resulted in minimal increases in the metrics for the models of data sets 1a and 1b. If the process did not change, we would expect that removing applicants who might have gone against the overall trend would have led to a better model because we were able to model the admissions data before the implementation of the rubric. Yet, that isn't what happened, suggesting again there must be something different about the data collected after the implementation of the rubric.

5.5.3 Limitations affecting our ability to address whether the process changed

Looking at the results, it is possible that someone could instead believe they suggest the process did not change. We address those here.

In terms of evidence for the process not changing, our results show that the most predictive features are similar regardless of which data set we used. When using data set 0, we found that the physics GRE, quantitative GRE, and GPA were most predictive of admission. Likewise, when looking at data set 1a, we found that the physics GRE, GPA, quantitative GRE, verbal GRE, and proposed area of research were most predictive. Using data set 1b showed the most differences in that the measures of grades and the general GRE scores were in the lower half of the rankings. However, the physics GRE was still the top ranked feature. Yet, both models of the data after the implementation of the rubric did not have acceptable testing metrics, suggesting that we should interpret the feature importance orders with caution. Conditional inference forest models will always produce feature importances regardless of how well the model fits the data. Because the metrics to assess fit are relatively poor, we should not trust the conclusion that the most predictive features are the same.

However, it is possible that the low metrics might be a result of the conditional inference forest method not being suited for the data we have. Recent work suggests that the conditional inference forest algorithm does not perform well with missing data [217]. However, when we used MICE to

impute the missing data, the models were still not able to produce testing metrics in the acceptable range, suggesting that the missing data was not the issue.

In addition, while conditional inference forests were designed to better handle categorical data than traditional random forests do, there could still be issues with categorical data. For example, for data set 1b, there are only three possible values for each feature. Therefore, the model can only split each feature 3 ways, which limits the depth of the trees and the fine tuning of the model. However, when we used the section total (which could take on any integer between 0 and 8), the results did not substantially improve, suggesting that the scale of the data may not be to blame.

Even if the number of categories does not matter, the fact that some of the categorical data are discretized continuous features ((e.g. physics GRE score, physics coursework)) could create problems. Prior work has shown that binning continuous features can lead to a loss of information and over- or under-estimation of effect sizes [218, 219]. It is possible that such an effect is present in our data. However, models built from data sets 1a and 1b both found the physics GRE score to be the top feature even though the physics GRE score was discretized in data set 1b. Because the model metrics were not great, this rebuttal should be treated with caution. On the other hand, the fact that models of data set 1a where discretization wasn't an issue suggests that it cannot fully explain the models' low metrics.

It is also possible that the low metrics are not a result of how we handled the data we had but rather what data we had. It is possible that committee members were using something not included in our data to evaluate applicants and if we had that data, our models of data set 1a and 1b would improve. While such an explanation seems possible for data set 1a, it seems unlikely for data set 1b because members of the department decided what qualities they wanted to evaluate applicants on and added them to the rubric.

Finally, it is possible that the low metrics might not be caused by the data or the model and instead, the low metrics could be caused by the admissions process itself. The goal of the rubric is rate applicants along multiple dimensions and hence in a holistic manner. If applicants were actually assessed holistically, we would expect that the model would not generalize well because

there is no single underlying process. Instead there might be multiple routes an applicant could take to gain admission and hence, the model might encounter difficulties modeling this process. The fact that hyperparameter tuning and Tomek Links did not increase the testing metrics to an acceptable range for models of data set 1a and barely did so for the models of data set 1b supports such an interpretation. However, claiming the process is more holistic based on these results alone is premature, especially given the relatively small number of applicants in data set 1b. Instead, results from other modeling attempts would either need to poor predictive ability or show evidence of multiple routes to admission to support such a claim.

5.6 Future Work

In order to provide better address the limitations and consider whether our admissions process became most holistic, future work should examine alternative techniques for analyzing the data.

First, instead of taking a predictive approach in our analysis, we could take an explanatory approach where we try to understand what inputs may have caused the outcome. Under this approach, whether a feature is related to the outcome is determined by statistical significance rather than its predictive ability [62]. Logistic regression is a common example of this technique in PER. The results of such future work would provide greater insight into why the models did not fit data sets 1a and 1b well.

Second, to determine if the process is more holistic, future work could analyze the data using cluster analysis or latent class analysis. While such methods are becoming popular for analyzing learning environments (e.g. see [220, 221]), to our knowledge, such methods are less common in studies of graduate admissions processes. To our knowledge, clustering-like techniques have only been used to understand admissions strategies based on surveys of faculty on admissions committees [45]. If the process is more holistic, such methods might be able to identify clusters of applicants who were admitted for similar reasons. For example, some applicants may be admitted due to stellar academic credentials, others may be admitted due to their research background, while others may be admitted based on which faculty members are seeking new students. Finding or not

finding such a result would provide greater clarity as to how the process may have changed. To do so however, would likely require a larger data set, especially if there are a large number of driving results for why an applicant is admitted.

Finally, future work could take a mixed methods approach by considering qualitative approaches to investigating how our admissions process might have changed. Such qualitative approaches could allow us to observe the admissions process itself (similar to the studies Posselt conducted as documented in [46]) and understand how faculty are evaluating and discussing applicants in real time. In addition, a qualitative approach would allow us to avoid many of the modeling limitations related to the scale of the data and metrics.

Alternatively, future work could directly ask faculty who have served on the admissions committee both before and after the implementation of the rubric about their perception of the process at each time. However, we must be careful of faculty's potential biases when recalling how things were done in the past (see Muggenburg for an overview [222]). For example, given the greater emphasis on diversity and equity in higher education now, faculty's recall may suffer from post-rationalization [223] where they justify their decisions using reasons that weren't available at the time but are consistent with their current self image or social desirability [224] where past events events may be distorted to conform to current attitudes and norms.

5.7 Conclusion

Overall, the results of this initial investigation are suggestive that our admissions process did change after the implementation of the rubric. We were able to model the data from before the implementation of the rubric to a sufficient degree but not the data after the implementation of the rubric. In addition, the model of the admissions process before the implementation of the rubric does not do well predicting the data collected after the implementation of the rubric and vice versa, suggesting that the underlying process did change. However, there are still numerous limitations that need to be addressed before we can make a definitive conclusion, including how we characterize the data and how we model the data.

Furthermore, the lack of good fitting models on the data post implementation of the rubric suggests that the process might be holistic. In order to make such a conclusion however, we would need either evidence in favor of holistic admissions is occurring or stronger evidence that the current admissions process is not easily modeled by known techniques. Such evidence could be obtained through a variety of quantitative or qualitative approaches.

In terms of the modeling approaches, Tomek Links seem like a promising technique for future PER studies. While their use was not enough to provide a more conclusive answer to the question of whether our admissions process changed, their use did provide evidence that the data collected after the implementation of the rubric may be modelable to an acceptable level, leaving open the possibility that other methods may be able to model the data and hence, should be explored.

Finally, to truly get a sense of whether admissions processes change after the implementation of a rubric or merely use a new tool to do the same process, studies such as these need to be completed in other physics departments. By doing so, we will have a better idea of how rubric-based admissions might change admissions processes and how well our results generalize to other programs.

CHAPTER 6

WHY WE CAN TRUST THE RESULTS IN THE PREVIOUS CHAPTERS: A SIMULATION STUDY

6.1 Introduction

When working with educational data, we often encounter imbalanced binary input and outcome features, by which we mean the variable is not equally split into its two categories. For example, demographics in science, technology, engineering, and mathematics (STEM) are often imbalanced due to historical and ongoing injustices. While data mining with imbalanced data has been studied extensively [225], less attention has been paid to the types of imbalanced data that appear in discipline-based education research (DBER) and educational data mining (EDM) studies.

For example, educational data sets might consist of a single course on the order of a hundred students (and hence a hundred data points) to the entire university or even multiple universities, resulting in hundreds of thousands of data points. Further, educational data often includes continuous, categorical, and binary variables. As a result, an educational data set might contain many features with different imbalances. For specific examples of these occurring in the DBER and EDM literature, we refer the reader to the following papers [65, 226–238].

In the context of logistic regression, which is a popular EDM technique [239] and was a common technique used in the previously cited studies, much work has focused around outcome imbalance and how to work with such data. When the outcome is imbalanced, the regression coefficients and the probabilities generated from the logistic regression model are biased [240]. To correct for these biases, various techniques such as Rare Events Logistic Regression [240], Firth penalized regression [241], and introducing a log-F distributed penalty [242] have been proposed, which we explain in depth in Sec. 6.2.

More recently, machine learning techniques have become popular in educational research. One example is random forest [91, 239]. Just as logistic regression has biases that might be relevant to

EDM and DBER data, random forest is also known to have such biases. In particular, random forest favors categorical features with many levels [92] and continuous features [243] when determining which features are most predictive of an outcome.

Most interesting for the context of this paper is a study by Boulesteix et al. [244] building on the work of Nicodemus [243]. In their paper, they systematically varied the amount of predictive information that each binary feature contained as well as the feature imbalance and then used random forest as well as a variant better suited for categorical features, conditional inference forest, [92] to compare how well the algorithms could detect the informative features from the noise. Their key finding was that features with higher imbalances were ranked lower than features with lower imbalances even when they had the same "built-in" amount of predictive information. A later study [245] extended the work by including continuous features as well as binary features but only examined the case when none of the features contained predictive information. These studies suggest that when modeling data, our results might be measuring spurious properties of the features rather than their predictive information.

In this study, we seek to extend this line of work by considering the data typical of DBER and EDM studies. That is, data that includes a mix of continuous, categorical, and binary features with varying degrees of predictive or explanatory ability, and a binary outcome feature that might be imbalanced. In addition, new techniques for ranking random forest features have been developed, such as the AUC-importance [94], which are designed for imbalanced data sets and hence, might prove fruitful for DBER and EDM research. Finally, we wish to extend the work to regression techniques commonly used for educational data and explore how these biases might manifest in these techniques.

Specifically, we ask three research questions:

1. How might known random forest feature selection biases change when the outcome is imbalanced as is often the case in EDM and DBER studies and does the AUC-permutation importance affect those biases?

2. How might known machine learning biases manifest in traditionally explanatory techniques such as logistic regression?

3. How might penalized regression techniques successfully applied in other disciplines be used in EDM and DBER to combat any discovered biases?

It should be noted that our overarching goal is to compare existing approaches to analyzing data typically found in DBER and EDM research and not to introduce our own new promising method for analyzing such data.

The rest of the paper proceeds as follows. In Sec. 6.2, we provide an overview of the algorithms and approaches we mentioned in the introduction and that we use in the rest of the paper. In Sec. 6.3, we explain how we constructed our simulation data and carried out our neutral comparison simulation study [246]. In Sec. 6.4, we provide the results of our simulation study. In Sec. 6.5, we apply what we learned in the simulation study to a graduate admissions data set from United States universities. In Sec. 6.6, we provide answers to our research questions, compare our findings with similar studies, and consider how our choices might have influenced the results. In Sec. 6.7, we propose future directions for this work, both in terms of the data and algorithms. Finally, in Sec. 6.8, we provide the conclusions from our study and outline a set of recommendations.

6.2 Background

Here, we introduce the two paradigms of statistical modeling and then provide an overview of the algorithms we used in our study.

6.2.1 Paradigms of Statistical Modeling

When discussing modeling data, there are two prominent paradigms, both of which are used in DBER and EDM: prediction and explanation [74, 247]. Shmueli [62] provides an overview of these approaches and we summarize the key points here.

Explanatory modeling or explanation is focused on the causal effect of some set of inputs X on some outcome Y. That is, given some data set, explanation is concerned with which inputs produce a statistically significant effect when modeling the outcome. Traditional logistic or linear regression are examples of explanatory models. Under this approach, models are evaluated based on how well they fit the data using some statistic. In the case of logistic regression or linear regression, common statistics are Pseudo-R^2 and R^2.

In contrast, prediction is focused on generating a model for analyzing new data and determining the outcome and not necessarily the causal effect. Under this paradigm, having two sets of data, one to train the model and one to test the predictive capabilities of the model, is essential as to provide an estimate of the model's predictive ability.

Because prediction is not focused on the causal effects, statistical significance has no role in assessing features in predictive models. Instead, features are assessed based on whether they improve predictions of the model. While a feature with a small effect might be statistically significant, it might not have predictive power because a predictive model might perform just as well without the feature as with it.

As a corollary to this, we should not expect a model with high explanatory power to necessarily have high predictive power or vice versa, and hence, features with high explanatory power might not have high predictive power.

6.2.2 Explanatory Methods

6.2.2.1 Traditional Logistic Regression

When the outcome, Y, is binary, logistic regression is the standard technique for explanatory modeling. Under this approach, the probability, p, of finding the outcome of $Y = 1$ is given by

$$\log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{6.1}$$

when $x_1, x_2, .., x_n$ are the input features and the β are the coefficients. Under this formula, logistic regression has a similar form as linear regression.

We can rearrange the equation to solve for the odds which becomes

$$odds(x_1, x_2, ..., x_n) = \frac{p}{1-p} = b^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)} \tag{6.2}$$

where b is traditionally the natural base, e.

Under this formulation, it makes sense to talk about the odds ratio (OR) or the change in odds as a result of increasing an input feature x_j by 1 unit. More formally,

$$OR_{x_j} = \frac{odds(x_1, x_2, ...x_j + 1, .., x_n)}{odds(x_1, x_2, ...x_j, .., x_n)} = e^{\beta_j} \tag{6.3}$$

which means that the exponentials of the coefficients correspond to the odds ratio for each feature. Notice that the odds ratio is independent of the value of x_j.

Because a β of 0 means no effect, an odds ratio of 1 is equivalent to no effect [63]. Likewise, an odds ratio greater than 1 means an increase in the odds while an odds ratio less than 1 means a decrease in the odds.

An important caveat to this is what a unit increase is and what the odds ratio is in reference to. Often, continuous features are normalized so that the mean is 0 and the variance is 1 or scaled so that an increase of a unit has a tangible meaning. For example, SAT scores are only reported in multiples of 10 so scoring one point higher on the SAT is meaningless. Instead, the researcher would want to adjust the scale of the scores so that an increase of 1 unit corresponded to 10 points better on the test (or another meaningful increment).

For continuous features, what the odds ratio is in reference to is answered by the scale choice. For categorical features, especially unordered categorical features, the answer is nontrivial. An increase of 1 unit might not be meaningful or even possible (e.g., what would an increase of 1 unit of race mean?). In that case, it is customary to use one-hot encoding and create separate, binary features for each label. For example, for race, we could create 6 features: white, Asian, Black, Latinx, Native, Multi-racial. Under this approach with binary features, an increase of a unit

corresponds to changing categories, such as Black compared to non-Black student, which depends on the arbitrary choice of which label is assigned $x_j = 1$ and which is assigned $x_j = 0$. As [63] notes, it is often preferable to invert the odds ratios which are less than 1 to easily compare all odds ratios, which is equivalent to swapping our label for $x_j = 0$ and $x_j = 1$.

6.2.2.2 Penalized Regression

When the data contains issues that might make modeling difficult (i.e., small sample size, correlations, and more features than data points), adding a penalty to logistic regression might be beneficial. This idea is based on the bias-variance trade-off in which we can increase the bias of the coefficient to reduce its variability or vice versa [248]. As a result, penalized regression can be useful for feature selection, which is often an important first step in EDM [239]. Because logistic regression does not have a closed-form solution while linear regression does, we will present the penalized algorithms in the context of linear regression.

For typical least squares linear regression with m features and n cases, we are trying to solve the expression

$$\text{argmin}_\beta(||Y - X^T\beta||^2) \tag{6.4}$$

where Y is a $n \times 1$ vector of the outputs, β is a vector of the $m \times 1$ vector coefficients, and X is a $m \times n$ matrix of the input data.

When we use penalized regression instead, we add a penalty, $\mathbf{P}$, that might depend on the coefficients or data.

$$\text{argmin}_\beta(\ ||Y - X^T\beta||^2 + \mathbf{P}(\beta, X)) \tag{6.5}$$

In this study, we consider two types of penalization for explanatory methods, Firth and Log-F penalization, although many more exist. See Ensoy et al. [249] for an overview of methods often used in cases of separation, where an input feature perfectly predicts the outcome, or rare events.

Under Firth penalization, we try to combat the asymptotic bias of the coefficient estimates, which inversely depend on the sample size to some power. Specifically, the Firth method adds a penalty that removes the asymptotic bias to order $O(n^{-1})$, making it especially useful for small data sets [241]. It does so by penalizing by the Jeffreys invariant prior ([250], which is inversely related to the amount of information in the data. That is, the penalty is larger the less the data allows us to determine the coefficients. For a simple one-feature model, the penalty is equivalent to adding 0.5 to each cell of the 2x2 contingency table of the feature and the outcome [251], making this penalization especially useful in the case of separation. In theory, this penalization should then shrink the confidence intervals of the features with more imbalance because the more uncertainty would have resulted in a higher penalty.

The Jeffreys invariant prior is not without issues, such as being dependent on the data, which are summarized in Greenland and Mansournia [242]. To overcome these, Greenland and Mansournia proposed a log-F(m, m) distributed penalty. The penalty has a tuning parameter, m, that controls the amount of penalization with a higher m providing more accurate estimates of smaller β but less accurate estimates of larger β. When little is known about the data, Greenland and Mansournia recommend taking $m = 1$ to allow for a wider range of possible values. For a single parameter model, the choice $m = 1$ makes the Log-F penalty equivalent to the Firth penalty.

In addition to overcoming issues with the Jeffreys prior, the log-F penalty can be implemented via data augmentation, meaning that any software capable of performing logistic regression can also do Log-F penalization. For a chosen m, the researcher adds m pairs of rows to their data for each feature, where one row has outcome $Y = 1$ and the other has outcome $Y = 0$. In the pair of rows, the researcher then selects one feature to have value 1 and all of the other features to have values of 0, with the choice of feature unique to each pair of rows. The weights for each row are set to be $m/2$ and any intercept feature should be set to 0 in these added rows. An example of this for a 2-feature model with $m = 1$ is shown in Table 6.1.

It should be noted that despite similarity in name, log-F penalized regression has no relation to the recently proposed LogCF framework [252].

Table 6.1: Log-F data augmentation example for a two feature and m=1 example. The last four rows are the augmented data.

Outcome	Feature 1	Feature 2	Intercept	Weight
1	0.748	0.10	1	1
...	...	...	...	...
1	1	0	0	1/2
0	1	0	0	1/2
1	0	1	0	1/2
0	0	1	0	1/2

6.2.3 Predictive Methods

6.2.3.1 Penalized Regression

In addition to using penalized logistic regression as an explanatory method, there are also penalties designed for using regression as a predictive tool. Two of the most common are Ridge and Lasso, which are described in detail in Hastie, Tibshirani, and Friedman [248]. Again, we present the penalties in the context of linear regression.

Ridge penalization adds a penalty to the regression equation proportional to the square of the βs.

$$\text{argmin}_{\beta}(\ ||Y - X^T\beta||^2 + \lambda||\beta||^2) \tag{6.6}$$

Equivalently, it requires the sum of the squared β coefficients to be less than some value.

$$\text{argmin}_{\beta}(\ ||Y - X^T\beta||^2) \tag{6.7}$$

$$\text{subject to} \sum_{j=1}^{m} \beta_j^2 \leq t \tag{6.8}$$

Here, λ, or equivalently t, controls the degree of penalization, with a higher value associated with a stronger penalty.

Ridge penalization is often used in cases of multi-collinearity because it reduces the variability of the coefficients. That is, for two correlated features without penalization, one could be extremely

positive and the other extremely negative to offset each other. With the squaring of the coefficients under Ridge penalization, the coefficients can no longer offset each other and hence, must shrink. Mathematically, Ridge penalization is equivalent to scaling each β by $\frac{1}{1+\lambda}$.

Instead of penalizing based on the squared β, we can penalize based on the absolute value of the β; this is the premise of Lasso penalization. Mathematically, Lasso penalization seeks to solve

$$\text{argmin}_{\beta}(\ ||Y - X^T\beta||^2 + \lambda|\beta|) \tag{6.9}$$

Equivalently, it requires the sum of the absolute value of the β coefficients to be less than some value.

$$\text{argmin}_{\beta}(\ ||Y - X^T\beta||^2) \tag{6.10}$$

$$\text{subject to} \sum_{j=1}^{m} |\beta_j| \leq t \tag{6.11}$$

Again, λ controls the amount of penalization. Here though, the Lasso penalty is designed for feature selection because it shrinks some β to zero while shifting the values of the others.

Lasso is not designed for correlated features, and hence, it can encounter issues in those cases. For example, if two features are correlated, either could be shrunk to zero without reducing the accuracy of the model. Therefore, Lasso can exhibit variability concerns under correlation.

One way around this is to combine the penalties into a single penalty, which is the idea between Elastic net [253]. Mathematically, the Elastic net penalty is

$$\text{argmin}_{\beta}(\ ||Y - X^T\beta||^2 + \lambda(\alpha||\beta||^2 + (1-\alpha)|\beta|)) \tag{6.12}$$

.

where λ controls the overall penalization and α controls the amount of mixing of the Lasso and Ridge penalties, with the special case $\alpha = 0$ reducing to Lasso penalization and $\alpha = 1$ reducing to Ridge regression.

While these algorithms are typically used for prediction, various methods for using these algorithms in an explanatory manner have been developed along with corresponding p-values or other feature selection techniques [254–258]. We will only use these algorithms as predictive tools, but we include references to these approaches here for completeness.

6.2.3.2 Forest Methods

Random forest is an ensemble method of decision trees based on the Classification and Regression Trees (CART) framework [91]. For each decision tree, a subset of features, often noted $mtry$, is randomly selected and used to predict the outcome. To grow the tree, features are split into two groups with the specifics of the splits determined by which ones minimize the Gini Index, a measure of variance, the most. After all trees have been grown, the algorithm uses some method of aggregating results, such as a majority vote of the trees, to determine what the overall prediction is. Because the features are split, categorical features do not need to be one-hot encoded like they would in logistic regression.

To determine which features are relevant to the prediction, the features are often assessed by the mean decrease in the Gini Index across all trees, with a larger value meaning the feature is more predictive of the outcome. However, Strobl et al. showed that the Gini Index is biased toward continuous features and features with many categories [92]. That is, because continuous and features with many categories have many possible split points (infinite in the case of continuous), it more likely the algorithm can find an ideal split than the algorithm could for a binary feature that has only 1 split point. Therefore, these features will be viewed as more important because they appear to better separate the classes.

As a result, alternative measures such as accuracy permutation importance have become popular. To use accuracy permutation importance, each feature is randomly permuted one at a time and the change in predictive accuracy is recorded. The idea is that when a feature that is more predictive of the outcome is permuted, the predictive accuracy will decrease more than when a feature with less predictive information is permuted. As a result, the changes in predictive accuracy can be used to

rank the features in the model qualitatively. More recently, an alternative based on the AUC, which is the probability that the positive case ranks higher than the negative case over all possible pairs of positive and negative cases, has been proposed by Janitza et al. [94]. This AUC-permutation importance is claimed to perform better than the accuracy permutation importance measure when the outcome is imbalanced. It is important to note that both these importances only make sense in the context of the model and relative to each other.

Because the Gini Index is also used to create feature splits, the entire algorithm can be biased when the data contains binary, categorical, and continuous features (as is often the case in DBER and EDM). To correct this problem, Strobl et al. proposed conditional inference forests [92], which are based on the conditional inference framework [259]. Rather than minimize the Gini Index to find ideal splits, conditional inference forests use the conditional inference independence test to determine which feature to split and how to split it. Simulation studies by Strobl et al. have found that using conditional inference forests with subsampling without replacement does in fact, correct the biases shown by traditional random forest [92].

For more details about these algorithms, see the appendix A.

6.3 Methodology

6.3.1 Data Creation

To conduct our simulation study, we first needed to generate our simulated data. To create binary features with varying degrees of imbalance and information, we considered a 2x2 contingency table, Table 6.2. We used labeling conventions similar to those of Olivier, Bell, and Rapallo for the reader's convenience because we reference their formulas here [260].

Table 6.2: 2x2 contingency table of fractions for generic binary feature.

	$x_j = 0$	$x_j = 1$	Total
$Y = 0$	π_{00}	π_{01}	π_{0+}
$Y = 1$	π_{10}	π_{11}	π_{1+}
Total	π_{+0}	π_{+1}	1.0

Table 6.3: Examples of changing only one of the feature imbalance, outcome imbalance, or odds ratio for an N=1000 dataset.

(a) Reference table

	$x_j = 0$	$x_j = 1$	Total
$Y = 0$	300	200	500
$Y = 1$	300	200	500
Total	600	400	1000

(b) Changing only the feature imbalance

	$x_j = 0$	$x_j = 1$	Total
$Y = 0$	400	100	500
$Y = 1$	400	100	500
Total	800	200	1000

(c) Changing only the outcome imbalance

	$x_j = 0$	$x_j = 1$	Total
$Y = 0$	450	300	750
$Y = 1$	150	100	250
Total	600	400	1000

(d) Changing only the odds ratio

	$x_j = 0$	$x_j = 1$	Total
$Y = 0$	360	140	500
$Y = 1$	240	260	500
Total	600	400	1000

For some binary feature x_j, let the fraction of cases with $x_j = 0$ be π_{+0} and the fraction of cases with $x_j = 1$ be π_{+1}. Likewise, for the binary outcome feature Y, let the fraction of cases with $Y = 0$ be π_{0+} and the fraction of cases with $Y = 1$ be π_{1+}. Then the feature imbalance is represented by the ratio $\pi_{+0} : \pi_{+1}$ and the outcome imbalance is represented by $\pi_{0+} : \pi_{1+}$. We pick $x_j = 1$ and $Y = 1$ to be the minority classes, though the choice is arbitrary.

To quantify the amount of information contained in a feature for predicting or explaining the outcome, we will use the odds ratio which is $OR = \frac{\pi_{00}/\pi_{01}}{\pi_{10}/\pi_{11}} = \frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}}$ using the notation in Table 6.2.

By specifying the feature imbalance (in the form of π_{+1}), the outcome imbalance (in the form of π_{1+}) and the odds ratio, we can uniquely express the values in the 2x2 table. Furthermore, any one of the three can be changed while the remaining two can be held constant, allowing us to manipulate the feature imbalance, the outcome imbalance, and the odds ratio systematically. An example with counts for a hypothetical data set with 1000 samples is shown in Table 6.3.

To determine the values in the 2x2 table, we can rearrange the formula for the odds ratio in terms of π_{+1}, π_{1+}, and π_{11} found in the literature to solve for π_{11} [260]. Doing so, we find that

$$\pi_{11} = \frac{1 + (\pi_{+1} + \pi_{1+})(OR - 1) - Q}{2(Q - 1)} \tag{6.13}$$

where

$$Q = \sqrt{(1 + (\pi_{1+} + \pi_{+1})(OR - 1))^2 + 4OR(1 - OR)\pi_{+1}\pi_{1+}} \tag{6.14}$$

In the case that $OR = 1$, that is the feature contains no predictive or explanatory information for the outcome, the expression for π_{11} is indeterminate. In that case, the feature and outcome are independent so $\pi_{11} = \pi_{+1}\pi_{1+}$.

Once we know π_{11}, we can use Table 6.2 to compute the remaining values. That is

$$\pi_{10} = \pi_{1+} - \pi_{11} \tag{6.15}$$

$$\pi_{01} = \pi_{+1} - \pi_{11} \tag{6.16}$$

$$\pi_{00} = 1 + \pi_{11} - \pi_{1+} + \pi_{+1} \tag{6.17}$$

To model continuous features, we assumed the features were normally distributed with a separate distribution for each outcome class. For $Y = 0$, we modeled the feature as $\mathcal{N}(0, 1)$ and for $Y = 1$, we modeled the features as $\mathcal{N}(\mu, 0)$ where μ was a parameter we controlled. By increasing μ, the distributions would have less overlap, and hence, the value of a specific point would provide more information about the outcome.

For our study, we choose the same feature imbalances and odds ratio as found in Boulesteix et al., which correspond to $\pi_{+1} = \{0.5, 0.4, 0.25, 0.1, 0.05\}$ and $OR = \{3, 1.5, 1\}$, creating 15 binary features [244]. We then created five continuous features with $\mu = \{0.75, 0.50, 0, 0, 0\}$, for a total of 20 features. As the number of features in DBER and EDM studies are on the order of 10, we choose to keep the total number of features on the order of 10 rather than on the order of 100 as in the Boulesteix et al. study [244].

We then generated these features for five outcome imbalances, $\pi_{1+} = \{0.5, 0.4, 0.3, 0.2, 0.1\}$, and three sample sizes, $N = \{100,\ 1,000,\ 10,000\}$ for a total of 15 simulated data sets. A visual

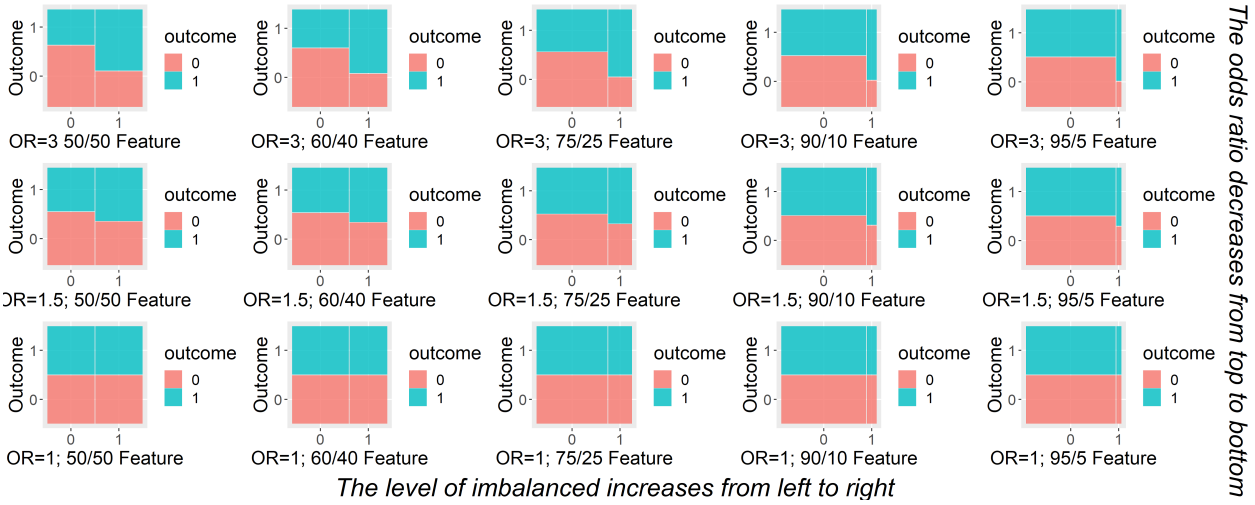

Figure 6.1: Distribution of binary features in the simulated $\pi_{1+} = 0.5$, $N = 1,000$ model.

depiction of the binary features in the $\pi_{1+} = 0.5$ and $N = 1,000$ case is shown in Fig. 6.1 and a visual depiction of the continuous features in that same case are shown in Fig. 6.2.

6.3.2 Procedures

6.3.2.1 Forest Algorithms

To analyze our data set using forest algorithms, we first randomly selected 70% of the cases for the training set and kept the remaining 30% of the data set for the testing set. Our prior work with random forest suggests that the size of the train/test split did not qualitatively affect the conclusions around variable importance and selected features [65]. We then used the `randomForest` function from the `randomForest` package [261] to create random forest models and the `cforest` function from the `party` package [92, 96, 259] to create conditional inference forests in R [98].

For both models, we set the number of trees to 500 as that is the default in the `cforest` algorithm and simulation studies of random forest have found that errors rates level off on the order of a few hundred trees [100]. For the number of features per tree, we picked $\sqrt{p}$ where p is the number of features, which is also aligned with the recommendations of Svetnik et al. [100]. We have called this m previously to distinguish from probability in the logistic model but use p here because it is the common symbol in the random forest literature.

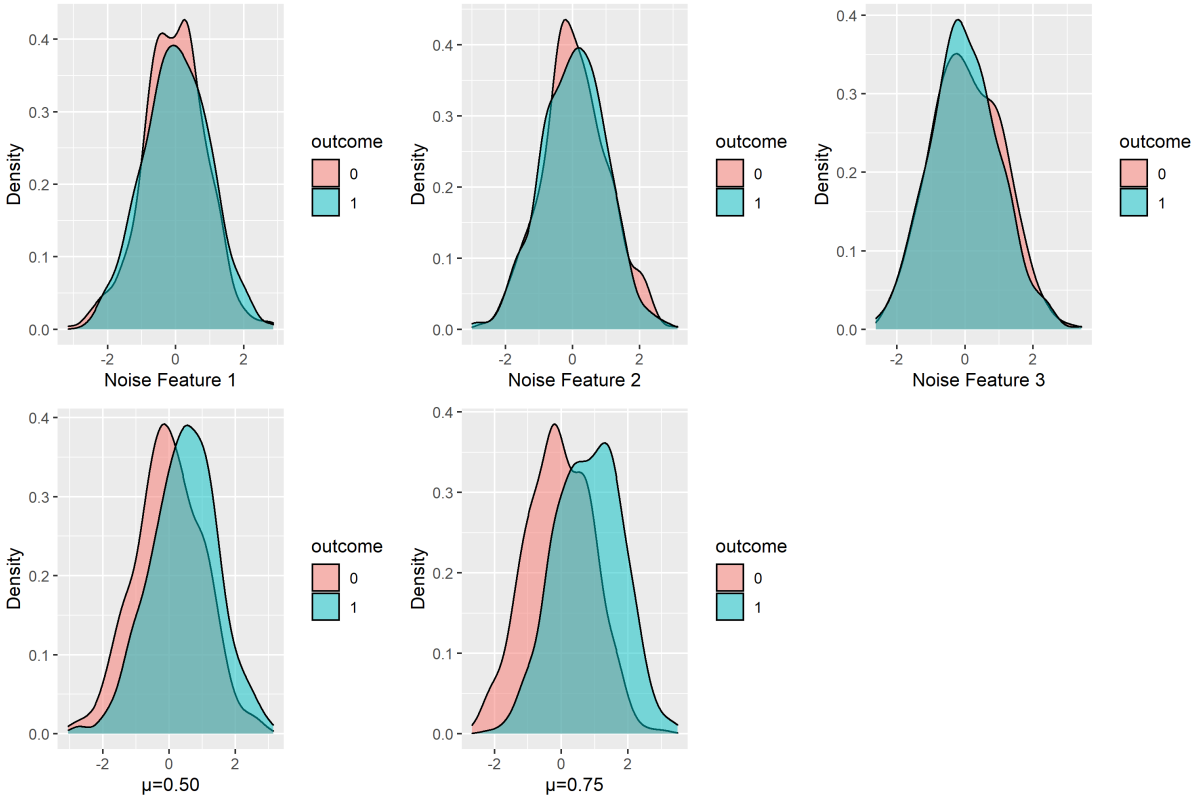

Figure 6.2: Distribution of continuous features in the simulated $\pi_{1+} = 0.5$, $N = 1,000$ model.

For the random forest algorithm, we then computed the Gini importance. For the accuracy permutation importance and the conditional inference algorithm, we computed the AUC permutation importance and accuracy permutation importance. We repeated this procedure of splitting the data, running the model, and calculating the importances 30 times so that the resulting distribution of the importances would be approximately normal according to the central limit theorem [262].

Next, we determined the rank of each feature based on its average value over the 30 runs, where the feature with the largest importance value would have rank 1. This type of approach is often used in screening studies to determine relevant features, which is what we are doing here [95].

To evaluate bias in the Boulesteix et al. paper, they approached bias as the difference from the expected value of zero in the null case and argued that bias when the features have odds ratios different 1 was not well defined [244]. Because we are interested in selecting features, we can create a definition of bias based on the rank of the feature. If a forest algorithm is biased, we would expect to see that features with higher imbalance should have larger rank (i.e. be farther from 1)

than features with identical odds ratios but smaller imbalances.

Using these ranks, we can also define bias in terms of the features detected by the algorithm. Assuming no bias, features with identical odds ratios should be detected at the same rate, regardless of their imbalance.

To determine if a feature was detected, we adopt the convention that detected means different from noise. We define detected as being ranked above the first noise feature, which has $OR = 1$ or $\mu = 0$. We picked this convention so that it is somewhat analogous to the definition statistically significant, which for explanatory models, is that the probability of obtaining a result at least as extreme as the result observed under the assumption of the null hypothesis is less than some threshold, typically 0.05.

6.3.2.2 Regression Algorithms

To use logistic regression in an explanatory manner, we did not use a train/test split as that approach is characteristic of a predictive approach and instead, used all of the data as is customary for explanatory modeling. To create a logistic regression model, we used the `glm` function that is part of base R with the option `family='binomial'` to use logistic instead of linear regression. Because log-F is based on data augmentation, we also used `glm` for that approach. Prior work suggests that a choice of $m = 1$ performed better than a choice of $m = 2$ and $m = 1$ is a good starting choice when nothing is known about the size of the odds ratios [242, 263]. Even though we "know" the true values of the odds ratios because we built them in, we want to approach the problem as if it were real data and we do not have any prior information about the features. We then used the default weights of $m/2$ for the log-F model.

To run the Firth penalization, we used the `brglm` function from the `brglm` package [264, 265]. Per the function's documentation, the choice of `pl` is irrelevant for logistic regression so we left it at its default value.

For all three approaches, we used the `confint` function to compute the confidence intervals. For the Firth penalization, we picked `ci.method` to be `'mean'` as the `brglm` documentation suggests

it is a less conservative approach. We then say that a feature was detected or statistically significant if zero is not in the confidence interval or in the case of odds ratios instead of the raw coefficients, 1 [266].

To get a sense of how the odds ratio varied based on the data, we also ran a bootstrapped simulation. That is, we randomly selected 80% of the cases and ran the standard logistic regression, Firth penalization, and log-F penalization models on that data. We did this 10,000 times.

To create the Lasso, Ridge, and Elastic net models, we again used a train/test split because these algorithms are designed for prediction rather than explanation. To align with the bootstrapping procedure, we used 80% of the cases for the training data and 20% for the testing data. Because Lasso and Ridge have a single tuning parameter, λ that controls the amount of penalization, we used the `cv.glmnet` function to find the optimal value of λ. We then used the `glmnet` function to train the Lasso and Ridge models with their respective best value of λ [267]. We again repeated this process 10,000 times.

Finally, we used the `train` function from the `caret` package to find the best values of α and λ for Elastic net and create the model [268]. We again did this 10,000 times with 80% of the data used as training cases.

To analyze the bootstrapped results and generate confidence intervals for the values of the odds ratios, we used the percentile bootstraps [269]. Under this approach, all of the bootstrap estimates are sorted smallest to highest. For a given α, the bootstrap confidence interval is interval lying between the $100 \times \frac{\alpha}{2}$ and $100 \times (1 - \frac{\alpha}{2})$ percentiles. We chose $\alpha = 0.05$ to form a 95% bootstrapped confidence interval.

6.3.3 Neutral Comparison Study Rationale

Following the call of Boulesteix, Lauer, and Eugster for neutral comparison studies in the computational sciences, we address these three criteria and why we believe we have met their criteria [246].

A. The main focus of the article is the comparison itself. It implies that the primary goal of the article is not to introduce a new promising method. As stated in the introduction, we are not

introducing a method that we have developed and the focus of our paper is on comparing different methods rather than showing the usefulness of a certain method.

B. The authors should be reasonably neutral. We have not developed any of the algorithms or techniques used in this study and hence, we have no stake in which method might perform best. We also have experience using predictive and explanatory methods and have used these techniques in our previous work.

C. The evaluation criteria, methods, and data sets should be chosen in a rational way, Our methods and simulated data are based on a previously published simulated study, so we believe they are rational. We believe our evaluation criteria for detecting is rational because it is intuitive, objective, and based on prior approaches. We acknowledge that other approaches do exist and we address those in the discussion.

6.4 Simulation Results

6.4.1 Forest Algorithm Results

When looking at a subset of the results in Fig. 6.3, we see similar results to Boulesteix et al. [244]. That is, for the Gini importance, represented by plot A, continuous features are ranked higher than binary features regardless of whether they are noise features or not. For the permutation importances, we see that more balanced features tend to have larger importances than less balanced features even when they have the same odds ratios. For example, when looking at Fig. 6.3D, we see that the features *OR=3, 60/40* and *OR=3, 50/50* have much higher importances than the *OR=3, 90/10* and *OR=3, 95/5* features. Similar trends are shown in plots B and C.

To compare across outcome imbalances, we aggregated all 3 sample sizes and 5 outcome imbalances into a single plot for each importance method. The results are shown in Fig. 6.4.

Again, the Gini importance, Fig. 6.4A, shows a preference toward continuous features and against binary features for all sample sizes and outcome imbalances. More specifically, there was not a single sample size or outcome imbalance in which any of the categorical features were detected.

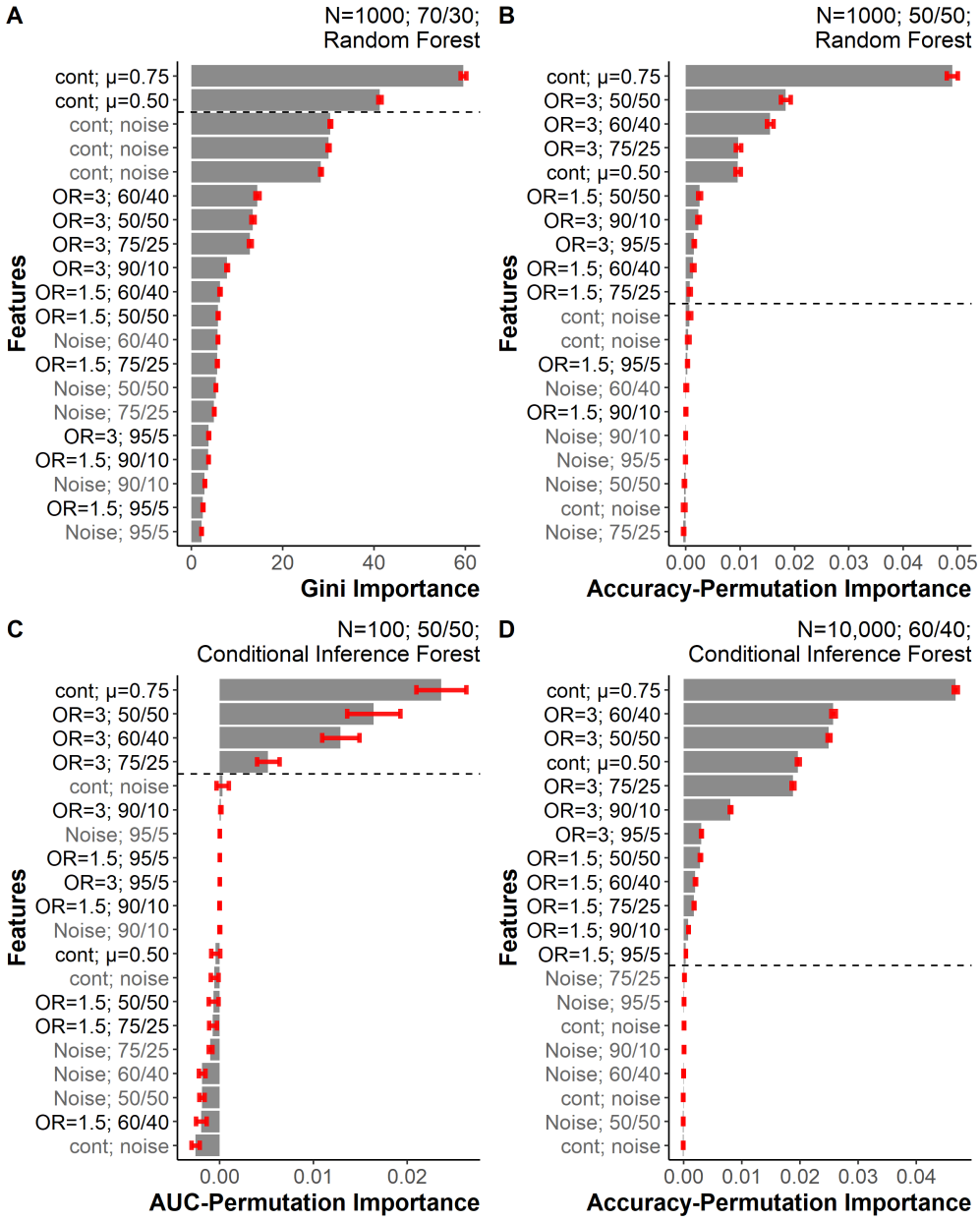

Figure 6.3: Importance values for a subset of the random forest models. Feature names shown in black were constructed to be informative while feature names in grey were constructed to be noise. Plot A shows the N=1000 70/30 outcome imbalance case with the standard random forest algorithm and Gini importance, plot B shows the N=1000 50/50 outcome imbalance case with the standard random forest algorithm and accuracy permutation importance, plot C shows the N=100, 50/50 outcome imbalance case with the conditional inference forest and AUC-permutation importance, and plot D shows the N=10,000 60/40 outcome imbalance case with conditional inference forest and accuracy-permutation importance. For all of the permutation importances, features with less imbalance tend to have larger importances than more imbalanced features for identical odds ratios.

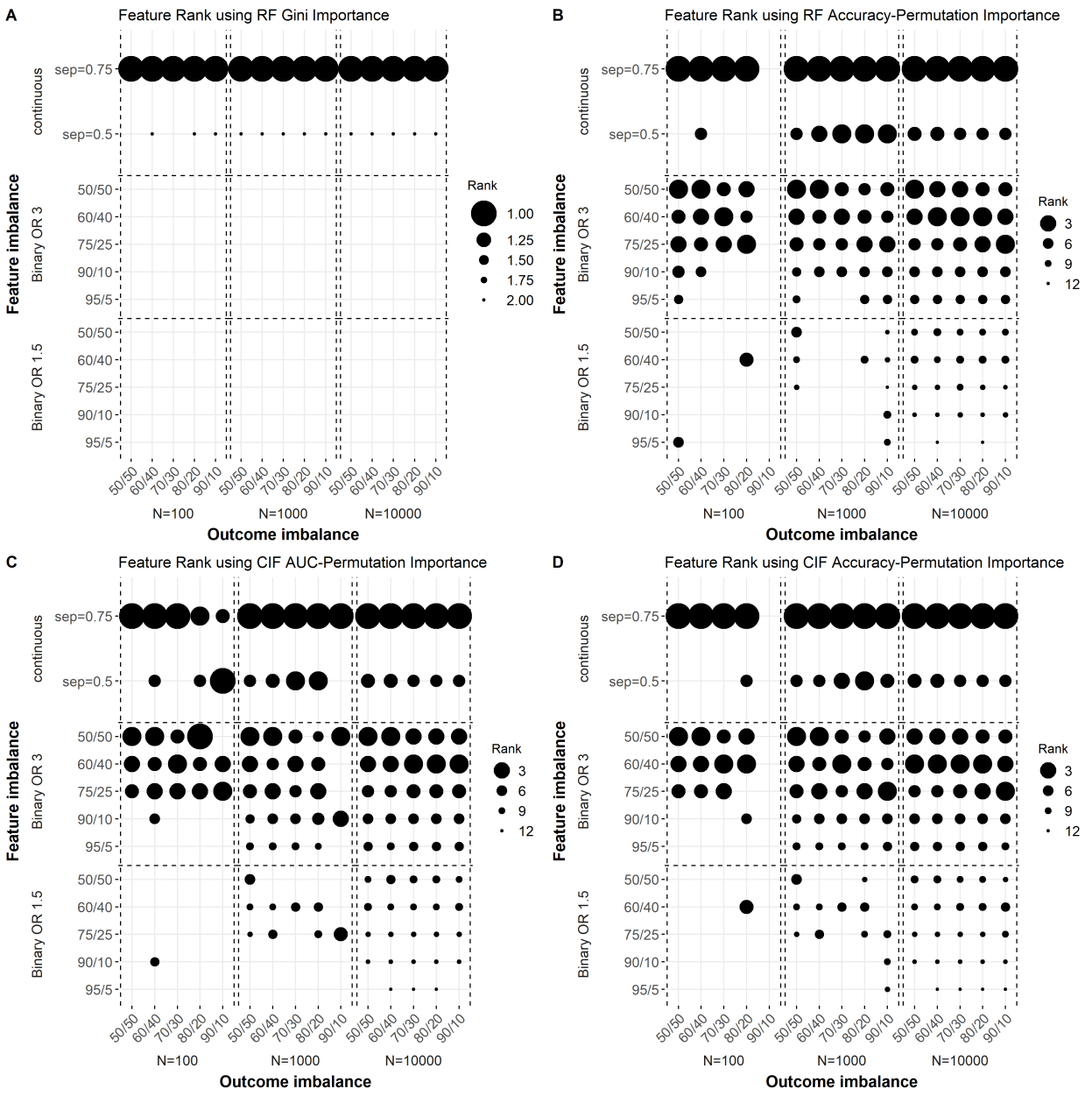

Figure 6.4: The ranks of the informative features for the four importance measures, grouped by the sample size and outcome imbalance. Noise features are not shown and any feature ranked below a noise feature was assigned a rank of 0. Here, a larger circle reflects a higher rank, meaning the feature was more predictive of the outcome. Overall, features with lower imbalance rank higher than features with higher imbalance for a given odds ratio and the result is not affected by the outcome imbalance or the specific permutation importance or forest algorithm used.

For the permutation algorithms, the results are similar regardless of whether the accuracy-based or AUC-based permutation method is used. Regardless of outcome imbalance and sample size, more balanced features with smaller imbalances tend to rank higher than less balanced features with identical odds ratios. This result is reflected in the plots by the decreasing dot size from top to bottom in any of the rectangles formed by the dotted lines. In cases of high outcome imbalance, moderately imbalanced features might rank higher than the balanced feature (e.g. the *Binary OR=3* features for the *N=10,000 90/10* case in Fig. 6.4D), but the most imbalanced features never rank higher than the balanced feature with the odds ratio. In fact, the *OR=1.5 50/50* feature ranks higher than the *OR=3, 95/5* feature for some of the models.

When looking at sections of columns of the plots in Fig. 6.4, we notice that most informative features cannot be detected for $N = 100$, regardless of which algorithm is used. In fact, only the less imbalanced *OR=3* features and the more predictive continuous feature can be detected and even then, that depends on the level of outcome imbalance.

For the $N = 1000$ case, most of the *OR=3* features can be detected. However, only the less imbalanced *OR=1.5* features are detected in most cases. Across the three permutation-based importances, there does not appear to be a consistent pattern for which *OR=1.5* features are detected based on the outcome imbalance.

For the $N = 10,000$ case, nearly all of the features can be detected, with the exception being the highly imbalanced *OR=1.5 95/5* feature. Again, there is not a consistent pattern as to when this feature will not be detected based on the outcome imbalance. While the *OR=1.5, 95/5* feature is never detected in the *50/50* outcome imbalance, it is sometimes detected in the *70/30* and *90/10* outcome imbalance cases, making a pattern difficult to generalize based on the outcome imbalance.

6.4.2 Logistic regression results

In addition to detecting features, logistic regression provides an estimate of the odds ratio, which can give us an idea of how accurately algorithms are modeling the built-in odds ratios. Because the odds ratios and confidence intervals determine detection, we present those first. The odds ratio

results are shown in Fig. 6.5.

From the $N = 100$ case, plot A, we see that the 95% confidence intervals for most features span at least an order of magnitude regardless of the feature imbalance or outcome imbalance. However, the width of the confidence interval tends to increase with both increasing feature imbalance and increasing outcome imbalance. For example, for the *OR=3 90/10* and *OR=3 95/5* features with a *90/10* outcome imbalance, the confidence intervals are too wide to fit on a plot that spans 6 orders of magnitude. In some cases, the width of the confidence interval for a balanced feature with a highly imbalanced outcome can be comparable to a highly imbalanced feature with a balanced outcome such as *OR=1.5 60/40* with a *90/10* outcome imbalance and *OR=3; 95/5* with a *50/50* outcome balance.

Given the width of the confidence intervals, our built-in value of the odds ratio is always contained in the confidence intervals. However, when looking at the actual estimate of the odds ratio, we see varying degrees of accuracy. For some features, like *OR=1.5; 50/50*, the *80/20* outcome imbalance was the most accurate estimate while for *OR=3; 60/40*, the *60/40* outcome imbalance was the most accurate imbalance. In general, there was no specific trend where the discrepancy between the estimated value and the built-in value varied with increasing feature or outcome imbalance. In addition, there was no consistent trend where the estimated odds ratio over- or under-estimated the built-in value.

For the $N = 1,000$ and $N = 10,000$ cases, we notice that the confidence intervals have considerably shrunk and now span on the order of a single magnitude. This is true even for most imbalanced features with the most imbalanced outcome. Nevertheless, the width of the confidence intervals still tend to increase with both increasing feature imbalance and outcome imbalance.

As with the $N = 100$ case, the built-in values are included in the confidence intervals and there is no consistent trend as to whether the estimated odds ratio over- or under-estimates the built-in value. We note that there is an exception to this for *cont; noise2* and *50/50* outcome imbalance on the $N = 10,000$ plot where the noise feature is found to have an odds ratio less than 1.

Next, we can conduct an analysis similar to what we did with the forest algorithms and determine

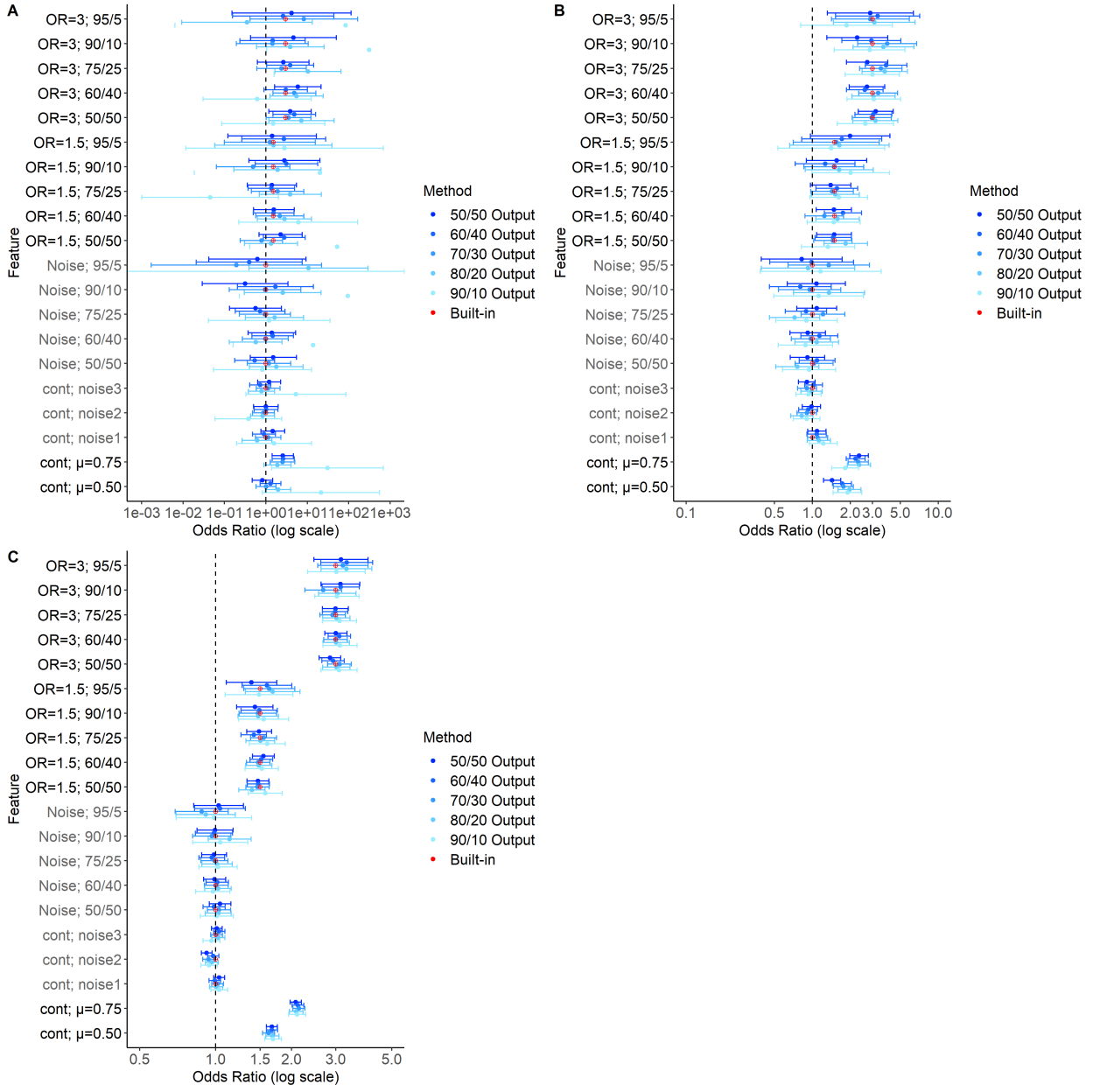

Figure 6.5: Values of the odds ratios and 95% confidence intervals found by logistic regression models compared by outcome imbalance. Our built-in value is represented by the circled plus. Plot A is a sample size of $N = 100$, plot B is a sample size of $N = 1,000$ and plot C is a sample size of $N = 10,000$. Confidence intervals that span beyond the scale are removed from the plot. Note the log scale on the horizontal axis.

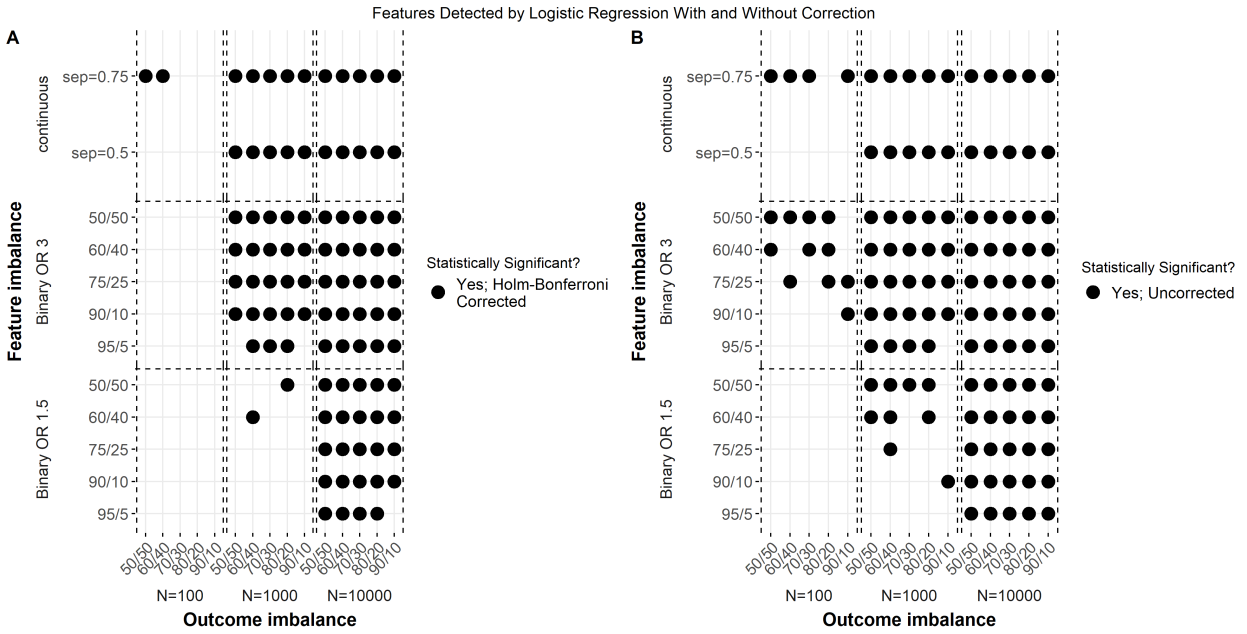

Figure 6.6: Analog of Fig. 6.4 but using logistic regression as the algorithm and statistical significance as the criteria for detection, $\alpha = 0.05$. Plot A uses the Holm-Bonferroni correction to control for multiple tests while plot B uses the uncorrected p-values.

which features are detected by logistic regression. Here, because we are using logistic regression in an explanatory manner, we use the p-value to determine whether a feature is detected, with statistical significance meaning less than a chosen cutoff, α. Because we are conducting multiple tests of statistical significance, we should control for false positives. Therefore, we present the results with and without a Holm-Bonferroni correction [195], which is less conservative than the traditional Bonferroni correction and has been used in DBER work before [270–272]. The correction is applied within each data set because for a study with real data, we would only have one data set. The results are shown in Fig. 6.6

For $N = 100$, when we apply the Holm-Bonferroni correction, the continuous feature with the largest μ is the only one to be detected and even then, only for minor outcome imbalances. If instead we do not apply any corrections, logistic regression is able to detect a few of the *OR=3* features however these tend to be the ones with lower imbalances. That is, even with a generous definition of statistical significance, logistic regression is unable to detect features with moderate odds ratios or features with large odds ratios but higher imbalances.

For $N = 1000$, logistic regression is able to detect both continuous features and most of the *OR=3* features regardless of whether we applied a correction to the p-values or not. Unlike the $N = 100$ case, we are able to detect some of the *OR=1.5* features though only features with lower imbalances and this depends on whether we apply a correction or not. When we apply the correction, we were only able to detect two of the *OR=1.5* features across any of the five outcome imbalances, while if we did not apply the correction, we could detect ten.

Finally, for $N = 10,000$, we were able to detect all of the informative features, regardless of whether we applied a correction or not. However, one of the continuous noise features was marked as statistically significant in the *50/50* outcome imbalance and the *70/30* outcome imbalance cases. One of these disappeared when we applied the p-value correction while one did not, suggesting that with enough data, random variations in the data might appear as signals.

6.4.3 Penalized regression results

Given the result from Sec. 6.4.2 that most features are detected for $N = 10,000$ even without correction, we chose to focus on the $N = 100$ and $N = 1000$ cases as areas where penalized regression might offer a benefit. To get a representative picture of how penalized regression might help, we then applied the algorithms to the *50/50*, *70/30*, *90/10* imbalanced outcome data sets, representing no imbalance, medium imbalance, and high imbalance.

6.4.3.1 Confidence interval approach

Because Firth and Log-F penalized regression are designed for explanatory approaches, we can use them to generate confidence intervals. The results for the $N = 100$ data sets are shown in Fig. 6.7 and the results for the $N = 1000$ data sets are shown in Fig. 6.8. Here, we only present the uncorrected 95% confidence intervals because if we do not find a benefit on the uncorrected confidence intervals, we would not find one on the corrected versions.

For the $N = 100$ case, we notice that the Firth and Log-F penalizations tend to have smaller confidence intervals and in many cases, are closer to the built-in odds ratio than traditional logistic

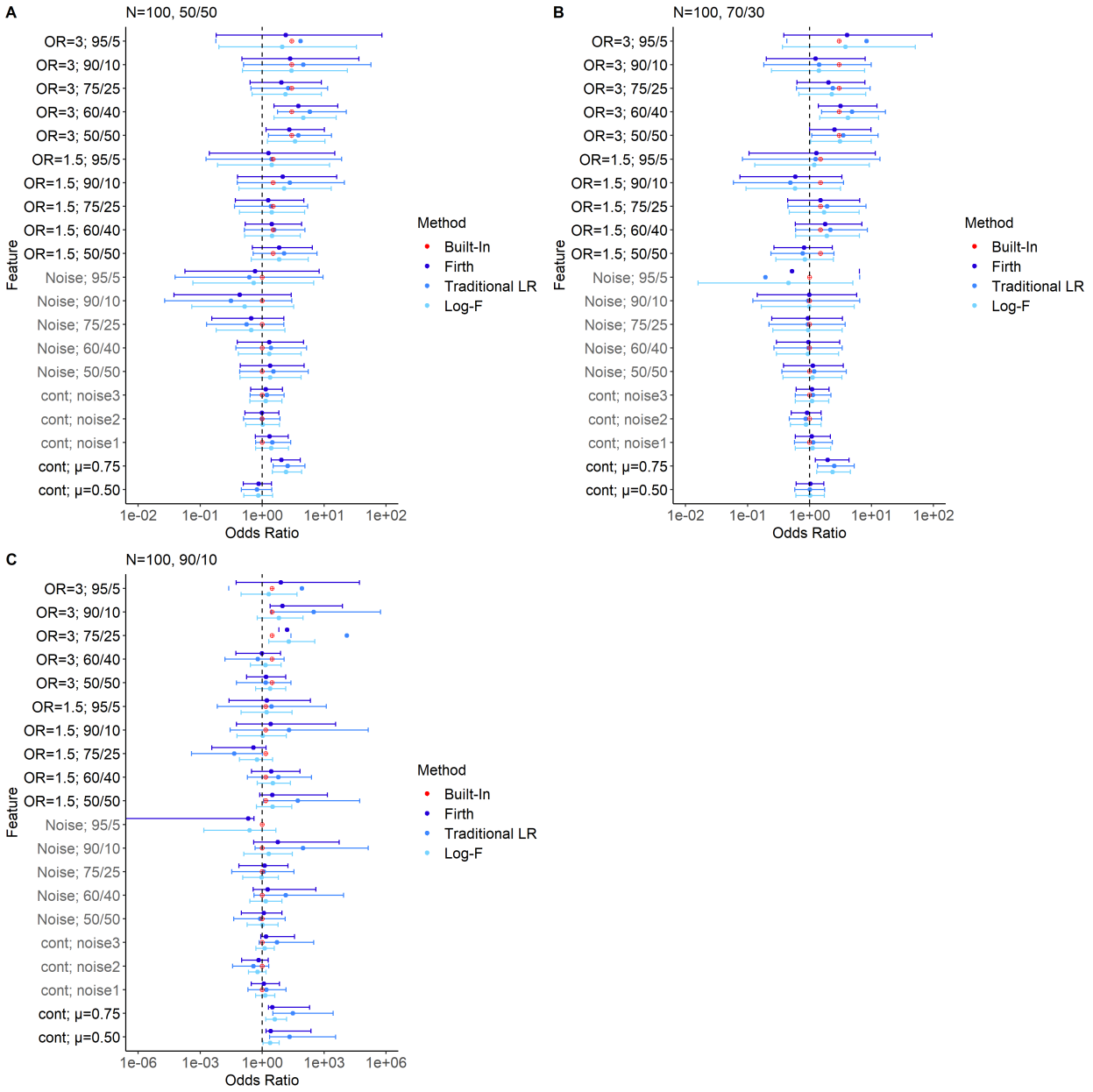

Figure 6.7: 95% confidence intervals for Firth penalized, traditional, and Log-F penalized logistic regression for the $N = 100$ data sets. Plot A shows the *50/50* outcome imbalance, plot B shows the *70/30* outcome imbalance, and plot C shows the *90/10* outcome imbalance. Confidence intervals that span beyond the scale are removed from the plot. For higher outcome imbalance, Firth and Log-F penalizations can considerably shrink the confidence intervals.

regression is.

For the *50/50* case, all three algorithms produce similar confidence intervals for more balances features such as *OR=3; 50/50*. For the highly skewed features such as *OR=3; 95/5*, Firth and Log-F

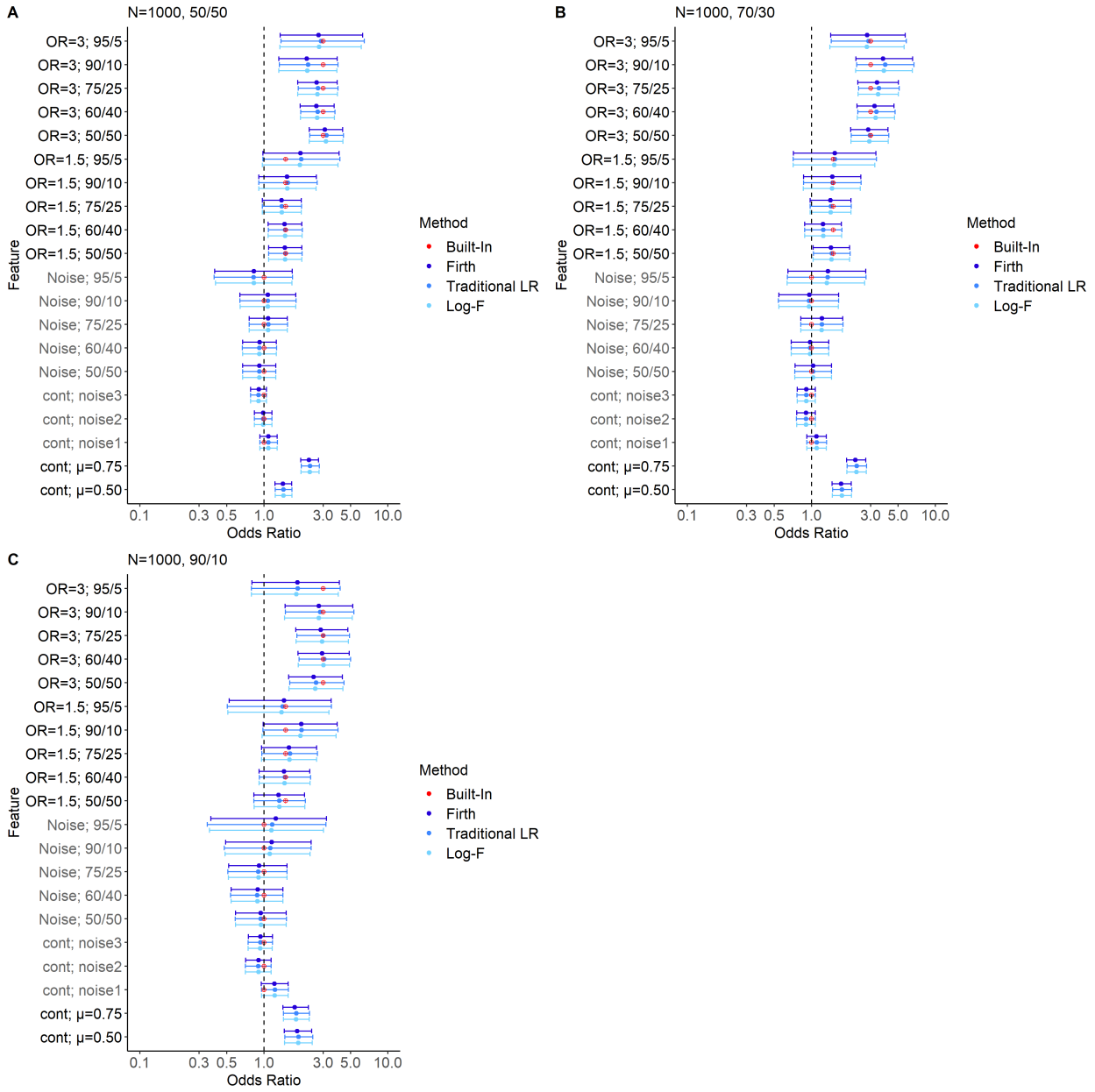

Figure 6.8: 95% confidence intervals for Firth penalized, traditional, and Log-F penalized logistic regression for the $N = 1000$ data sets. Plot A shows the *50/50* outcome imbalance, plot B shows the *70/30* outcome imbalance, and plot C shows the *90/10* outcome imbalance. For higher outcome imbalance, Firth and Log-F penalizations can shrink the confidence intervals.

penalizations do shrink the confidence interval with Log-F appearing to offer a greater benefit. However, none of the shrinking makes a difference as to whether the feature would be statistically significant or not compared to traditional logistic regression.

When we instead look at the moderately imbalanced *70/30* case, we see similar results. That

is, the Firth and Log-F penalizations appear to provide a greater benefit in terms of shrinking the confidence interval for features with greater imbalance, though again, the benefit is not enough to change whether a feature would be detected.

For the highly imbalanced *90/10* case, both penalizations reduce the confidence intervals regardless of the feature's imbalance. The benefits are most clear however for the most imbalanced features. For example, for *OR=3; 90/10*, Log-F penalization reduces the width of the confidence interval by nearly 3 orders of magnitude compared to the traditional logistic regression. As in the *50/50* and *70/30* cases, the penalizations do not affect whether a feature would be statistically significant, but the penalizations still do produce more accurate estimates of the built-in odds ratios than traditional logistic regression does.

Looking at the $N = 1000$ results in Fig. 6.8, we notice that the confidence intervals of the penalized regression methods are similar in length to those of traditional logistic regression. This result is true regardless of the feature imbalance or the outcome imbalance.

When it comes to estimating our built-in odds ratio, the penalized methods do not offer much of an improvement over traditional logistic regression. Indeed, for lower imbalanced features, all three methods tend to provide similar estimates, while for higher imbalanced features, there is no clear trend as to which method will provide an estimate closest to that of the built-in value.

6.4.3.2 Bootstrap approach

In addition to only considering whether the algorithm detects a feature, we can also get a sense of what range the estimated odds ratio will fall in using the five different penalization approaches. The results from the $N = 100$ data sets are shown in Fig. 6.9 and the results from the $N = 1000$ data sets are shown in Fig. 6.10.

From the $N = 100$ plot, we see that spread of the estimated values varies between the different methods. For higher feature imbalances, traditional logistic regression and Firth penalized regression often have the widest distributions. Because Lasso shrinks the coefficients to zero (or equivalently, odds ratios to 1) and Ridge reduces the variance of the estimate, these two methods

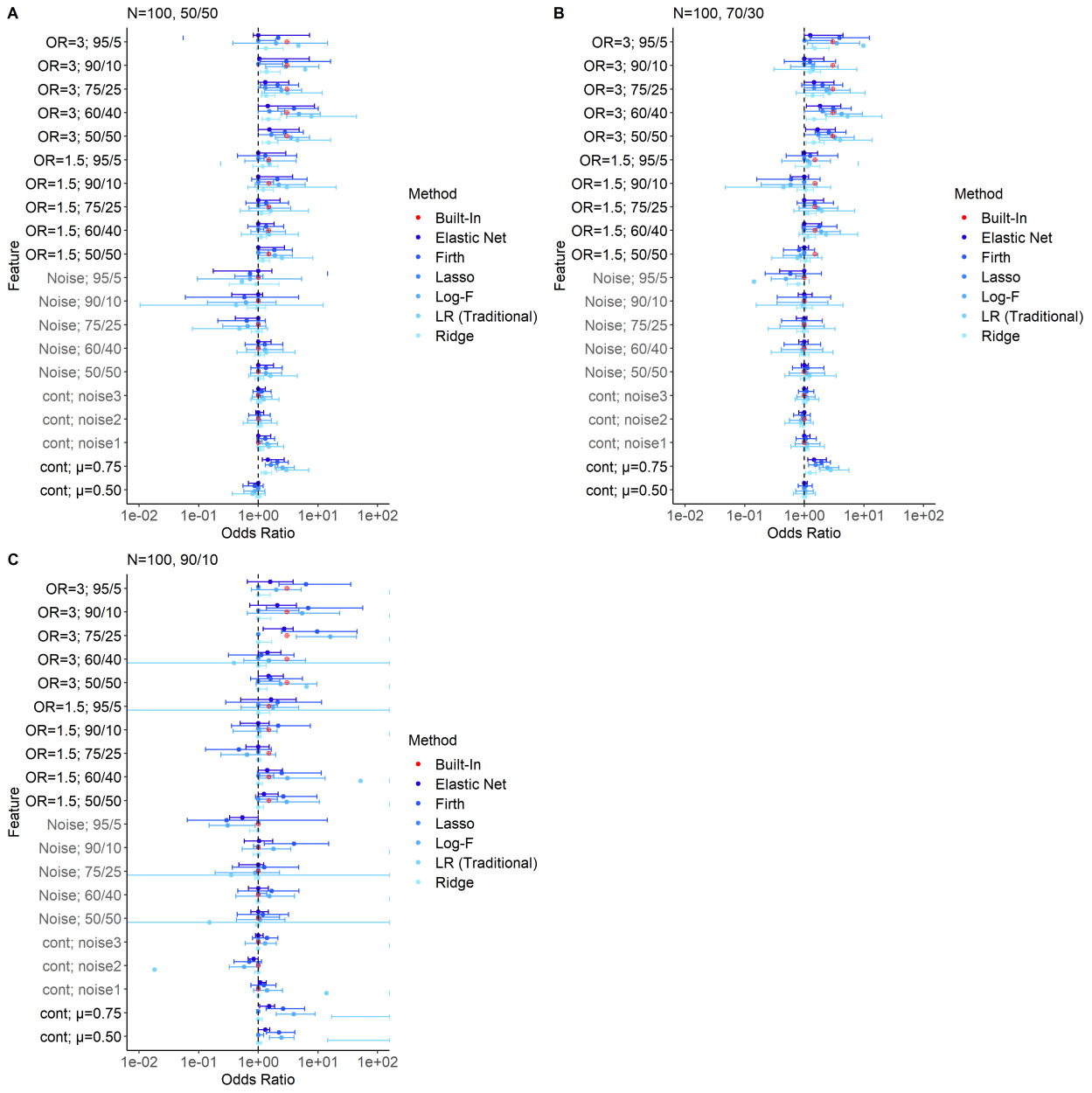

Figure 6.9: 95% percentile bootstraps of the odds ratio for Elastic net, Firth, Lasso, Log-F, no, and Ridge penalizations on the $N = 100$ data. Dots represent the median value. Plot A shows the *50/50* outcome imbalance, plot B shows the *70/30* outcome imbalance, and plot C shows the *90/10* outcome imbalance

often have the most compact distributions.

Likewise, in terms of the median estimate of the odds ratio, we see variation between the methods. Because Lasso shrinks estimates and Ridge scales estimates, these two underestimate the built-in odds ratio. We also find this behavior with Elastic net, which is a middle group between

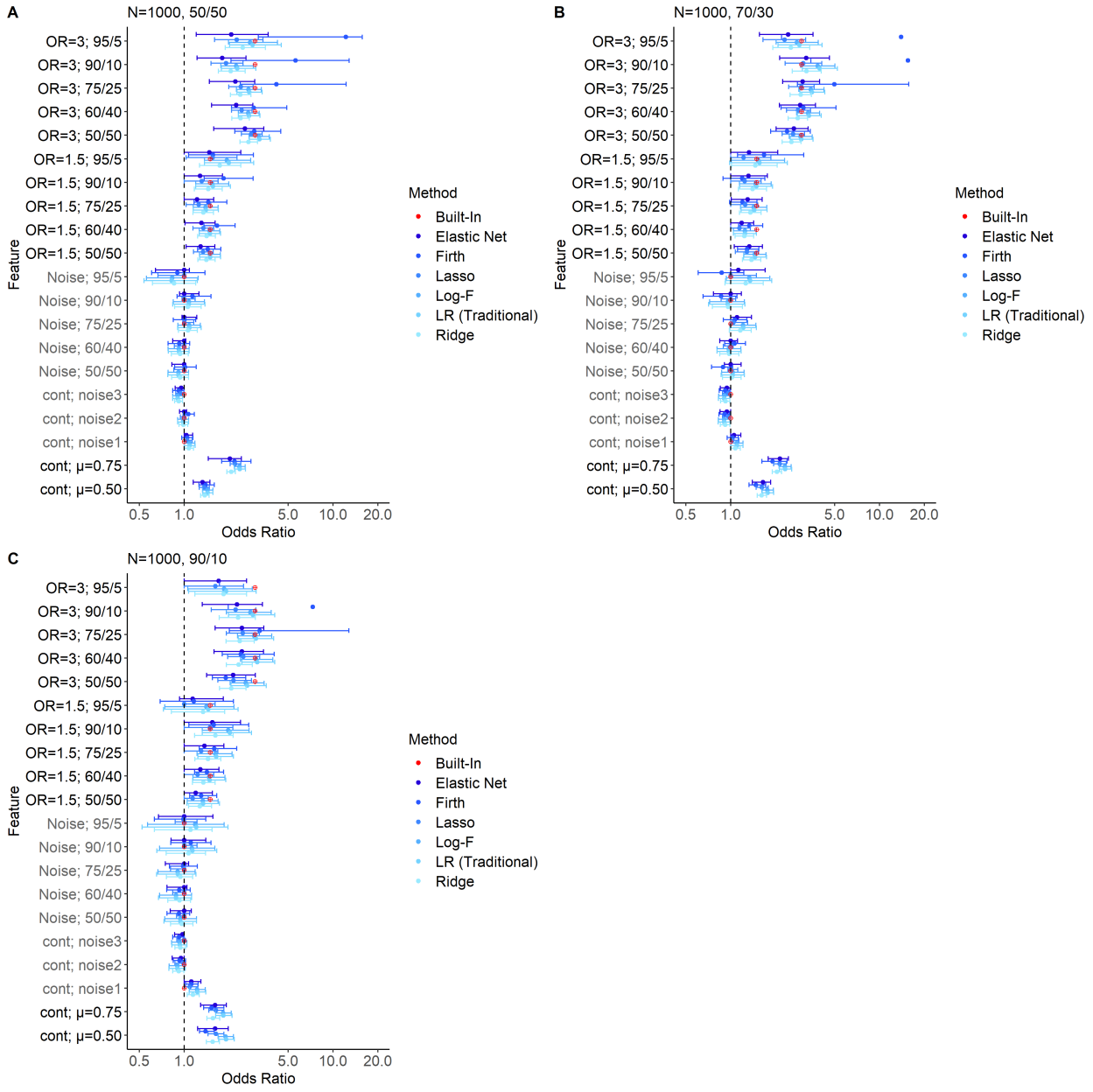

Figure 6.10: 95% percentile bootstraps of the odds ratio for Elastic net, Firth, Lasso, Log-F, no, and Ridge penalizations on the $N = 1,000$ data. Dots represent the median value. Plot A shows the *50/50* outcome imbalance, plot B shows the *70/30* outcome imbalance, and plot C shows the *90/10* outcome imbalance

the two. However, Elastic net often includes the built-in odds ratio within its interval even when Lasso and Ridge do not. This result is especially true for higher feature and outcome imbalances.

Log-F penalization on the other hand often takes a middle ground on both estimates and distribution width. Regardless of the feature or outcome imbalance, Log-F does not consistently

over- or under-estimate the built-in odds ratio and does not have the widest distribution of the estimates.

From the $N = 1000$ results shown in Fig. 6.10, we see that the six methods tend to produce similar results for more balanced features, even at higher outcome imbalances. The exception is the Firth penalization for higher imbalance features (e.g. *OR=3; 90/10*). For these higher imbalance features, the Firth penalization estimates can be nearly an order of magnitude larger than the estimates produced by other methods.

As in the $N = 100$ case, we find that Lasso and Ridge tend to underestimate the built-in odds ratio for the $N = 1000$ case. Unlike the $N = 100$ case, however, the built-in value is included in the bootstrapped confidence interval.

In most cases, Elastic net, Log-F, and logistic regression tend to have similar distribution widths and have the built-in odds ratios within their intervals. While Elastic net under-predicts the built-in value, Log-F penalized and traditional logistic regression do not show a consistent pattern as to whether they over- or under-predict the built-in value.

6.5 Application to Real Data

In this section, we apply the results of our simulation study to a graduate admissions data set.

Our data set comes from the application records of over 5,000 applicants to 6 Big Ten or Midwestern universities over a two-year period. The data includes the applicant's GRE scores, undergraduate GPA, undergraduate university, demographics such as binary gender, race, domestic status, whether the applicant made the shortlist, and whether the applicant was admitted to the program. Details about these features can be found in Posselt et al. [54].

We can then treat each university as a separate case study, which is an approach we have used in our previous work [75]. Doing so allows us to vary the sample size and the outcome imbalance. For the six programs in the data set, the smallest program had $N = 140$ applicants over the two year period while the largest had $N = 1228$. When considering whether the applicant made the shortlist or was admitted, the outcome imbalance ranged from $53/47$ to $83/17$, which means that the sample

Table 6.4: Feature and outcome imbalances for the binary features from actual graduate school admission data

Feature	School			
	School 1 Admit	School 2 Admit	School 3 Shortlist	School 3 Admit
Outcome	59/41	83/17	59/41	76/24
Gender	79/21	81/19	85/15	85/15
Domestic	NA	71/29	50/50	50/50
Year	57/43	55/45	51/49	51/49
Race=Asian	87/13	64/36	52/48	52/48
Race=Black	96/4	99/1	99/1	99/1
Race=Latinx	81/19	91/9	99/1	99/1
Race=Multi	96/4	97/3	93/7	93/7
BinaryNoise1	60/40	60/40	60/40	60/40
BinaryNoise2	75/25	75/25	75/25	75/25
BinaryNoise3	90/10	90/10	90/10	90/10
BinaryNoise4	95/5	95/5	95/5	95/5
N	140	431	1228	1228

size and outcome imbalances are on the same scale as the data we used in our simulation study.

We then selected four of the twelve possible combinations of shortlist or admit and the six programs that represent a small and medium data set with a more balanced and less balanced outcome. Specifically, we modelled school 1's admission (N=140, 59/41), school 2's shortlist (N=431, 78/22), and school 3's shortlist and admission (N=1228, 60/40 and 78/22) respectively. In the initial paper using this data, Posselt et al. analyzed shortlist and admissions separately and hence, we do so here [54].

6.5.1 Methods

To analyze the real data, we used five approaches. First, we use logistic regression and random forest with the Gini importance as they are the "default" methods. Based on the results of the simulation study, we then choose to use Log-F, as it performed either better or no worse than Firth, Elastic net, as it performed better than Lasso or Ridge and retains the benefits of both, and conditional inference forest with the AUC importance, as all of the permutation based importance

Table 6.5: McFadden Pseudo R^2 values for the explanatory models

	School 1	School 2	School 3 shortlist	School 3 admit
Logistic Regression	0.256	0.215	0.199	0.203
Log-F	0.252	0.215	0.199	0.203

measures performed similarly.

To mimic the simulation study and know which features were certainly noise, we added four binary noise features (imbalances of 60/40, 75/25, 90/10, and 95/5, which we refer to as BinaryNoise1, BinaryNoise2, BinaryNoise3, BinaryNoise4) and three continuous noise features. The binary features and their imbalances for the four data sets are shown in Table 6.4.

To run the models, we used the same R packages as in the simulation study. However, for real data, we should be interested in how well the model fits and hence, need to include some measure of that. For the logistic regression based methods, we used the standard McFadden pseudo-R^2 implemented in the `DescTools` package via the `PseudoR2` function [273], where a good value is between 0.2 and 0.4 [274]. While other choices of pseudo-R^2 exist, Menard suggests that there is little reason to prefer one over another, but McFadden's might be preferable because it is intuitive [275].

To connect the forest methods with the logistic regression methods, we also computed the AUC for each model, which follows the recommendation of Aiken et al. [74]. To do so, we used the `AUC` function from the `ModelMetrics` package [276]. We interpreted an AUC of at least 0.7 as a good model [93].

For the predictive methods, Elastic net, random forest, and conditional inference forest, we used the same procedure as in the simulation study except now used a 80/20 train/test split for all methods and calculated the AUC on both the training and testing data sets.

6.5.2 Results

First, we present the metrics used to assess our model, which are shown in Table 6.5 and Table 6.6. We notice that except for school 3 shortlist, all of the pseudo R^2 are within the accepted range.

Table 6.6: AUC values for the various models on the four data sets

	School 1	School 2	School 3 shortlist	School 3 admit
Logistic Regression	0.826	0.806	0.790	0.799
Log-F	0.825	0.805	0.790	0.799
Elastic (Train)	0.830	0.807	0.785	0.799
Elastic (Test)	0.690	0.734	0.771	0.779
Random Forest (Train)	0.547	0.529	0.688	0.613
Random Forest (Test)	0.564	0.521	0.686	0.616
Conditional Inference Forest (Train)	0.749	0.517	0.793	0.674
Conditional Inference Forest (Test)	0.594	0.500	0.681	0.597

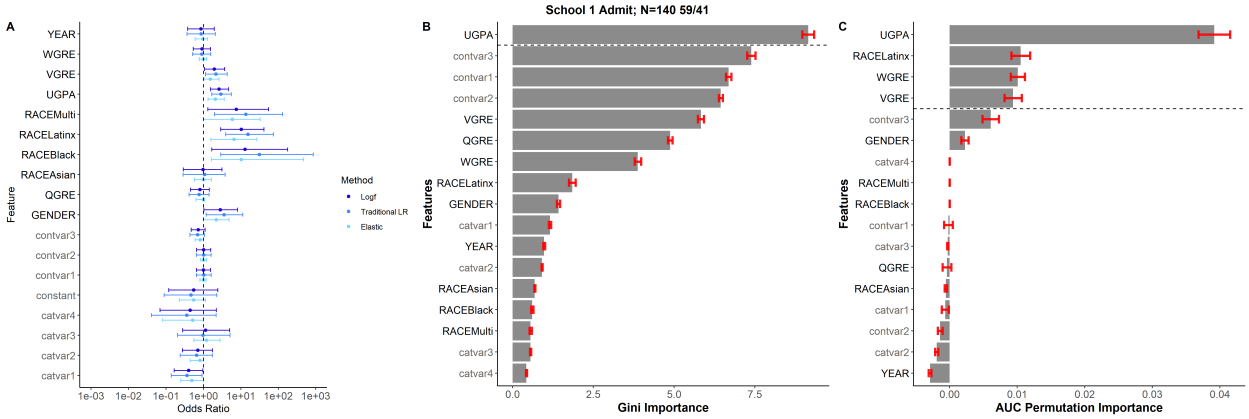

Figure 6.11: Comparison of the odds ratio (A), Gini importance (B), and AUC-permutation importance (C) for the features in school 1. Notice that RaceLatinx has a similar odds ratio as RaceBlack and RaceMulti according to (A) but only RaceLatinx is detectable in (C). RaceLatinx is less imbalanced than RaceBlack and RaceMulti.

When looking at the AUC values, we notice that the regression models outperform the forest models and in most cases, the forest models do not produce an AUC in the acceptable range. A review of physics education research literature found that less than 10% of papers reported out-of-sample metrics, so we cannot say if these results are typical for this type of data [74]. As our goal is not to make the best model but rather to extract features, we did not do any parameter tuning for the forests. We discuss more about these metrics in the discussion.

Because the conclusions from the four data sets are similar, we share only the results of school 1 and provide plots for the other data sets in the appendix for completeness. The results of algorithms applied to the school 1 data set are shown in Fig. 6.11.

When looking at plot A, we notice that Log-F noticeably shrinks the confidence interval for highly skewed features like *RaceBlack*. In exchange though, the estimate of the odds ratio is shrunk closer to $OR = 1$ for nearly all the features.

Even though Elastic net is showing the percentile bootstrapped confidence interval instead of the statistical confidence interval, the results tend to be aligned with the other methods. That is, the median value is on the same order of magnitude of the other estimates and the end points of the confidence interval are also on the same order of the magnitude as the other estimates.

When comparing the different methods, we see that none of the three algorithms would have led to different conclusions about which features are statistically significant or not. From plot A, the statistically significant features would be *VGRE*, *UGPA*, *RaceMulti*, *RaceLatinx*, *RaceBlack*, and *BinaryNoise1*. In the case of *BinaryNoise1*, which is supposed to be a noise feature, we note that due to the random nature of generating the feature, the odds ratio was smaller than 1 and hence, the algorithms appear to have detected that small difference.

When we move to plot B, we note that the continuous features are all ranked above the binary features as expected. As a result, all features except for one rank lower than the first noise feature.

Finally, when we move to plot C, we notice that only four features are detected, which is smaller than the regression approaches. Because prediction and explanation have different goals, we would not expect them to identify the same features. Yet, multiple approaches identifying the same features suggest that these features are in fact, distinct from noise.

One interesting point to note is that we see some ranking issues based on imbalance. For example, using a 2x2 contingency table to calculate the theoretical odds ratios, *RaceMulti* should have an odds ratio of 2.96 while *RaceLatinx* should have an odds ratio of 2.25. However, because *RaceLatinx* has an imbalance of 80/20 while *RaceMulti* has an imbalance of 96/4, *RaceLatinx* is detected by the AUC-permutation importance while *RaceMulti* is not.

6.6 Discussion

Here we address our research questions and consider how our choices and approaches might have impacted the conclusions we can draw from this study. We include a summary of the advantages and disadvantages of each algorithm based on our study and prior work in Table 6.7.

6.6.1 Research Questions

How might known random forest feature selection biases change when the outcome is imbalanced as is often the case in EDM and DBER studies and does the AUC-permutation importance affect those biases? When we vary the outcome imbalance as well as the feature imbalance, we still observe the same general trend as seen in Boulesteix et al. [244]. That is, features with higher imbalance are less likely to be detected compared to features with lower imbalances but the same odds ratio. In fact, the bias might become worse for high outcome imbalances because it is harder to train a "good" model when most of the cases have the same outcome.

In opposition to the claims of Janitza et al, [94], we do not find the AUC permutation importance to outperform the accuracy permutation importance. In fact, we find that the AUC permutation importance and the accuracy permutation importance perform similarly, regardless of the outcome imbalance. Further, we did not find any consistent differences in terms of the features detected by either random forest or conditional inference forest even though conditional inference forest is supposed to be better suited for categorical data [92].

We also see this preference for features with smaller imbalances in the real data. For example, for school 1, we saw that the less imbalanced *RaceLatinx* was detected over the more imbalanced *RaceBlack* and *RaceMulti* even though the theoretical odds ratio of *RaceLatinx* was smaller than that of the other two features.

Across the real data and simulated data, we see the expected bias with the Gini importance in which the continuous features are ranked higher than any of the categorical features. This result is most noticeable in Fig. 6.4 plot A where only continuous features are detected and Fig. 6.11 plot B

Table 6.7: Summary of advantages and disadvantages for each algorithm used in this study

Method	Advantages of algorithm	Disadvantages of algorithm
RF + Gini	-Default choice for many random forest implementations	-Biased in favor of continuous features, regardless of whether they are informative of the outcome or not
RF + accuracy permutation importance, CIF + accuracy permutation importance, CIF+ AUC permutation importance	-Can be used with continuous & categorical features -Categorical features do not need to be binarized -Comparable performance to logistic regression for feature selection without needing to check any assumptions	-Ability to detect features decreases with increasing feature imbalance and outcome imbalance -Questionable performance for small N
Logistic Regression	-standard algorithm for classification, implemented in most software -odds ratios have a "real-world" interpretation	-Width of confidence interval increases for increased outcome and feature imbalance and can become infinite in some cases
Firth penalization	-Able to shrink confidence intervals for imbalanced features in small N situations	-Not widely implemented in software -Advantages compared to logistic regression disappear for larger N
Log-F penalizations	-Able to shrink confidence intervals for imbalanced features in small N situations -Based on data augmentation so no special software needed -Coefficient estimates are similar to those of traditional logistic regression	-Advantages compared to logistic regression disappear for larger N
Lasso	-Shrinks some coefficients to zero which can be useful for feature selection	-Less able to detect less informative features from noise compared to other algorithms in the study
Ridge	-Effective at shrinking the width of the distribution of estimated odds ratios	-All coefficients are scaled by the same amount and are underestimated
Elastic Net	-Combines the benefits of Lasso and Ridge penalizations, often performing better than either approach individually	-Requires hyperparameter tuning to determine the ideal amount of mixing between Lasso and Ridge

where all of the continuous noise features outrank all but one feature.

How might known machine learning biases manifest in traditionally explanatory techniques such as logistic regression? We see similar biases in logistic regression as we see in the random forest for feature selection. For a sample size of $N = 100$ with a multiple comparison correction, we are unable to detect most features and even without correction, we can only detect low imbalance *OR=3* features. The uncorrected results are similar to those of the permutation importances for the forest algorithms.

For $N = 1000$, we can detect most *OR=3* features and without correction, low imbalance *OR=1.5* features. Again, the uncorrected logistic regression results resemble those of the forest algorithms but seem to be more aligned with the conditional inference forest results than the random forest results.

Once we get to a large sample size, $N = 10,000$, we can detect nearly all features, just as we can for the forest algorithms. However, for logistic regression, we also get an occasional false positive. Given the size of the data, it is not unreasonable that the logistic regression model might be picking up on minor differences in the noise features which it treats as a signal.

With explanatory techniques like logistic regression, we could also investigate how well they estimated the built-in odds ratio. We found that while the built-in value is almost always in the confidence interval, this has more to do with the width of the intervals than the ability of the algorithms. In general, confidence interval width increases with feature and outcome imbalance and decreases with sample size. The decrease in width as the sample size increases corresponds to what we would expect based on the conclusions of Nemes et al. [277].

We also observed the same general trend for the real data. Features with higher imbalances tend to have the widest confidence intervals, which can span several orders of magnitude.

How might penalized regression techniques successfully applied in other disciplines be used in EDM and DBER to combat any discovered biases?

While none of the five techniques we tried, Firth penalization, Log-F penalization, Lasso, Ridge, or Elastic net, corrected the bias, they did show promise for use in future EDM and DBER studies.

For explanatory methods, Firth and Log-F were found to shrink the confidence intervals, especially for highly imbalanced features and highly imbalanced outcomes. While Firth can still show wide confidence intervals, the Firth confidence intervals were found to be smaller than those of traditional logistic regression. On the other hand, Log-F provided at worst similar performance to Firth penalization and for higher imbalances, seemed to shrink the width of the confidence interval more than Firth penalization did. We found that both of these methods were most useful for smaller data sets, $N = 100$, while for the medium and larger data sets, their performance was similar.

When it came to the distributions of the estimated odds ratios, Log-F often showed a smaller distribution. While the results were comparable for the small data sets, for medium data sets and features with high imbalance, Firth penalization overestimated the odds ratio and had more variability. Conversely, Log-F produced more accurate and less variable distributions.

For predictive methods, Lasso, Ridge, and Elastic net were only used in a bootstrap, so we cannot cannot discuss the confidence interval width. We can however discuss the distribution of estimated odds ratios.

For Lasso and small data sets, we find that many of the features are shrunk to zero, especially for higher imbalances. For example, even for a small, balanced sample, many of the *OR=1.5* features were shrunk to zero while the other methods did not treat them as consistent with noise. Elastic showed similar results although the effect was not as severe.

For Ridge and small data sets, the distribution of the estimated odds ratio was often the smallest for a given data set. Given that Ridge is designed to shrink the variability of the estimates, this finding is not surprising.

For medium data sets, Lasso, Ridge, and Elastic net performed similarly to the other methods in terms of the distribution of estimated odds ratios.

While our results generally agree with other studies, a true comparison is difficult because each study used its own subset of the algorithms, including ones we used in our study as well as ones we did not. Therefore, which algorithm performed best and under what circumstances depends just as

much on the algorithms it was compared to as the algorithm itself.

In general, other studies have tended to find that Firth penalization does outperform logistic regression in the case of outcome imbalance [251, 278–280] and Log-F penalization shows promise when working with imbalanced data and can outperform Firth-penalization [263, 281].

Likewise, studies like Pavlou et al. have found that Ridge penalization works well except when there are many noise features while Lasso performs better when there are many noise features but limited correlations, which is consistent with our results Pavlou et al. [282]. Their study also found that elastic seemed to perform well in all cases, which generally matches what we found.

In terms of our finding that none of the methods fixed the issues around feature or outcome imbalance, Van Calster reported a similar finding for shrinkage techniques [283]. Specifically, they found that despite working well on average, shrinkage techniques often did not work well on individual data sets, even in cases where the techniques could have provided the most benefit such as in small sample size or low events per variable cases. Even though the techniques did not solve any of the issues in our study, they still showed promise for reducing the scale of the confidence intervals and warrant greater adoption by the DBER and EDM communities.

6.6.2 Limitations and Researcher Choices

In this section, we shift our focus from the results of the research questions and instead consider how our choices around constructing the simulated data, tuning or not tuning our models, defining "detected features," and assessing the models might have impacted the conclusions we can draw from this study.

6.6.2.1 Our data sets

For our simulation study, we used the same levels of information as in the Boulesteix et al. study which we wished to extend [244]. We followed their convention that $OR = 3$ corresponded to a large effect while $OR = 1.5$ corresponded to a moderate effect. However, Olivier noted that what constitutes a large, medium, or small odds ratio depends on the feature imbalance, outcome

imbalance, and correlations [260]. Therefore, even though we are using the same odds ratios for the different imbalances, they might not necessarily contain the same amount of predictive or explanatory power in a “large", “medium", or “small” sense.

One noticeable difference between our study and the Boulesteix et al. study was the number of features [244]. While we argued that DBER and EDM studies usually have the number of features on the order of 10 rather than 100, one could argue that we still had too many features based on our sample size. For example, a rule of thumb is that there should be at least 10 cases of the minority outcome for each feature in the model, referred to as the events per variable [284, 285]. In that case, we would have needed at least a sample size of 400 for the 50/50 outcome imbalance and a sample size of 2,000 for the 90/10 outcome imbalance case.

However, recent work has called into question whether this rule of thumb is supported by evidence [278]. Van Smeden et al found that events per variable did not have a strong relation to predictive performance of models and instead, recommended that a combination of the number of predictors, the total sample size and the events fraction be used to assess sample size criteria [286]. Likewise, Courvoisier et al. found that logistic regression can encounter problems even if the events per variable were greater than 10 and concluded that there is no single rule for guaranteeing an accurate estimate of parameters for logistic regression [287]. Even if the rule of thumb were true for logistic regression, Pavlou et al. claims that penalized regression is effective when the events per variable is less than 10 [288].

6.6.2.2 Hyperparameter tuning

For our simulation study, we did not do extensive hyperparameter tuning for the forest algorithms. We did this because 1) Probst, Wright, and Boulesteix found that random forest is robust against hyperparameter specification, its performance depends less on the hyperparameters than other machine learning methods, and its default choice of hyperparameters are often good enough [289] and 2) Couronné, Probst, and Boulesteix state that for a method to become a standard tool (as random forest is in EDM and is becoming in DBER), it needs to be easy to use by researchers

without computational backgrounds and cannot involve complex human interaction, which is not true of hyperparameter tuning [290] .

In addition, we only do hyperparameter tuning for Lasso, Ridge, and elastic because testing multiple values for λ is built in to the `glmnet` algorithm that we used to run the models. Even then, recent work suggests that optimizing λ for small or sparse data sets results in substantial variability of the coefficients and the found λ might be negatively correlated with the optimal values, meaning that hyperparameter tuning might not have been advisable for our data in the first place [291].

However, for completeness and to minimize computation time needed, we did experiment with multiple choices for the number of trees in the forest, n_{tree}, and the number of features used for each tree, $mtry$. For the conditional inference forests with the AUC importance, a sample size of $N = 1000$ and outcome imbalances of 50/50, 60/40, 70/30, 80/20, and 90/10, we tried $n_{tree} = \{50, 100, 500, 1000, 5000\}$ and $mtry = \{1, p/3, \sqrt{p}, p/2, p\}$ where p is the total number of features in the model. We did not find any meaningful differences in which features were selected and no set of the hyperparameters consistently performed better than the default ($n_{tree} = 500, mtry = \sqrt{p}$). Therefore, we used the default choices for throughout the study.

6.6.2.3 Determining Detected Features

For the forest algorithms, we chose to use the simple and intuitive method of whether the feature ranked above the first noise feature to determine which features were detected. We were able to do this because we knew which features were noise and in the case of the real data, we added features we created to be noise. We could have, however, used a variety of other methods to detect features though each has its own limitations in the context of our study. See Hapfelmeier and Ulm for an overview of different approaches, some comparisons, and their own novel method [292].

In general, the techniques for feature selection in forest algorithms fall into two broad categories. First, there are elimination techniques that pull out a subset of the features based on criteria. For example, Díaz-Uriarte and Alvarez de Andrés used a recursive backward elimination technique that removes a certain fraction of features until only 2 remain [97]. The technique then selects the model

with the fewest features that performs within 1 standard error of the best model using whatever metric the researcher chooses. These type of methods are not appropriate for this study because they can restrict the features too much. That is, by having some cutoff or elimination procedure, features which contain only a small amount of predictive information could be eliminated even though they are in fact predictive.

The second common approach is to use some type of permutation test to generate a p-value. Under this approach, either the outcome or each individual feature is permuted and then run through the model to produce some metric. This is then done a large number of times to get a distribution of the metric. Then the unpermuted data is run through the model to get the actual value of the metric. The p-value is then the fraction of cases where the permuted metric is as extreme as the actual value of the metric [293]. This approach has been used in various random forest studies [294, 295] and has been extended into the PIMP heuristic for correcting the Gini importance bias [296]. While these methods provide an analogous method for comparing with explanatory methods, they can be computationally intensive as they require the distributions to be conducted from scratch for each model.

As a way to reduce the computational complexity, Janitza, Celik, and Boulesteix proposed that the negative importances, which are assumed to be noise features because they are making the predictions worse, could be used to construct a null distribution [297]. Under this approach, the distribution of the negative importances are reflected across the axis to give create the distribution for positive values. The same procedure as above can then be used to calculate the p-values. While this procedure is computationally feasible, DBER and EDM studies often have on the order of 10 features, which means there are a limited number of features which could have negative importances and thus, mirroring the distribution would be of little use.

6.6.2.4 Assessing Our Models

For our real data, we noticed that many of the models did not produce out of sample AUCs in the acceptable range of at least 0.7. Here we try to address that.

First, we acknowledge that one type of model should not always perform better than another; this is the basis of the no free lunch theorems for optimization [298]. Various studies comparing logistic regression and random forest find a similar result where which algorithm performs best depends on the data set [290, 299, 300]. Therefore, the fact that logistic regression models perform better than the forest models is not necessarily a problem. In fact, by comparing multiple models and finding that some work better than others, we can have greater confidence in our results that we are detecting a signal in the data and not just modeling the noise.

Second, we need to acknowledge that overfitting is happening with the Elastic net and conditional inference forests. This overfitting can be detected by looking for differences in the training and testing set AUCs, where a higher training AUC is characteristic of overfitting. The amount of overfitting seems worse for the smaller data sets as shown in Table 6.6. This result is not unexpected because with smaller data sets, there are fewer cases to learn from. The noise in the model might then be seen as a signal and treated as though it contains predictive information.

If we look at the other models and their results in Table 6.6, we notice that the forest models and Elastic net perform best on the school 3 data sets, which correspond to the medium sized data sets in the simulation study and largest of the real data sets. For the forest algorithms specifically, they perform best on the school 3 shortlist data set, which happens to have a smaller outcome imbalance than school 3 admit. This result suggests that to effectively use the predictive approach, the data set should not be too imbalanced and based on the results for school 1 and school 2, the amount of data should be on the order of 1,000 cases.

Additionally, the higher AUC for logistic regression and Log-F might be thought of as their own type of overfitting. Due to the train/test procedure of the predictive paradigm, these two methods are working with the full data set rather than just 80% of the cases, corresponding to a 25% increase in data to work with and hence, learn from. With the "extra" data, these models might be better able to detect trends in the data and separate them out from noise.

While there do exist techniques for detecting overfitting in logistic regression, many of them use some type of testing or validation data set. For example, the Copas test of overfitting recommends

splitting the data in half, using one half of the data to develop the regression model, using that model with the other half of the data to make predictions of the outcome, and then perform a linear regression with the predictions and actual values, testing whether the coefficient is different from 1 [301]. If it were, that would provide evidence of overfitting. However, this approach is nearly equivalent to using logistic regression in a predictive manner rather than the way it is traditionally used in DBER and EDM.

For a technique that aligns the explantory nature of logistic regression, we can examine the residual plots. Because, logistic regression produces discrete residuals, using binned residual plots instead might be helpful [302]. Under this approach, cases are divided into bins and the average value in each bin is plotted against the average residual in that bin. This approach allows the otherwise binary residual to take on any value of the form $\frac{i}{n_{bin}}$ where n_{bin} is the number of cases in the bin and $\{i \in \mathbb{Z} : -n_{bin} \leq i \leq n_{bin}\}$.

When implemented via the `arm` package [303], 95% confidence intervals are generated and we can get an idea of how good the model is by examining what fraction of the binned residuals fall within the intervals. When we do so, we find that the fraction of residuals falling outside of the confidence intervals are between 0.20 for School 3 shortlist and 0.34 for School 3 admit, suggesting the models might in fact, not fit well. There does not appear to be a pattern based on the sample size or outcome imbalance. The plots are shown in Fig. E.4 in appendix E.

6.7 Future Work

While we considered six approaches to logistic regression, two machine learning algorithms, and three importance measures, these are not the only approaches we could have used. Indeed, these are not even the only logistic regression or random forest techniques we could have used but chose these algorithms as a starting point. Future work could then consider how other modifications of logistic regression or random forest might improve upon the problems we have identified here.

For example, for logistic regression algorithms, Puhr et al. proposed two modifications to Firth penalization, a post-hoc adjustment of the intercept and iterative data augmentation, that

showed promise in their simulation study [304]. Based on their results, they recommend using their methods or penalization by Cauchy Priors [305], which we did not include in this study, as better options than Log-F when confidence intervals were of interest. Furthermore, a later study comparing logistic regression, Firth penalization, and the modifications to Firth penalization found that the modifications to Firth's method worked best in terms of parameter estimation bias for rare events and small sample cases [306].

In terms of forest algorithms, there are several variants that might be useful for the data we encounter in DBER and EDM. For example, Balanced Random Forest and Weighted Random Forest have been developed for working with imbalanced outcomes [307] and Oblique Random Forests have been developed to allow for diagonal cuts in the feature space rather than the horizontal or vertical cuts allowed under traditional random forest algorithms [308]. In their study, Menze et al. found that Oblique Random Forests outperform traditional random forest when the data is numerical rather than discrete, which might show promise for our data depending on the ratio of numerical features to categorical or binary features [308].

Alternatively, there are non-CART-based approaches to random forest [309]. Loh and Zhou conducted a simulation study of various approaches to random forest and variable importance [310], finding that forests grown using the GUIDE algorithm, which is implemented for both classification [311] and regression [312], was unbiased while the random forest and conditional inference forest approaches we used here were not. In their study, a method was unbiased "if the expected values of its scores are equal when all variables are independent of the response variable," which would correspond to a case in our study where the odds ratios were 1 for all features. Nevertheless, such an approach might still be worth looking into.

There are also newer importance measures that show promise. In a simulation study, Nembrini, K onig, and Wright proposed a modification of the Gini importance, which they claim removed its bias toward features with more categories and the biases observed here regarding feature imbalance [245]. However, their simulated studies with feature importance only considered null cases in which none of the features were predictive of an outcome. Nevertheless, further study of

this approach might be fruitful.

In contrast to the algorithms used to analyze the data, future work should also explore how changes to the data itself might affect algorithm performance. For example, we could use the risk ratio to encode the level of information in a feature instead of the odds ratio. In theory, risk ratio provides a more intuitive way to quantify the amount of information in a feature because it is based on the ratio of probabilities rather than a ratio of odds. Zhang and Yu proposed a method to convert the odds ratio to a risk ratio [313], though more recent work has called this approach into question and suggests alternatives [314, 315]. Because these two measures are related but not the same, there might be additional insights related to which features are detected based on how we define the amount of "predictiveness" they have.

To better replicate real DBER and EDM data, future work could also explore how the amount of correlation between the features affects the results. In the case of correlated features, new issues with permutations emerge, including the model needing to extrapolate to regions where the model was not trained to calculate a feature importance [95]. Various approaches have been developed for correlated data with forest algorithms [95, 96, 316], which warrant future study, especially for the type of data we see in DBER and EDM studies.

Additionally, future work can extend the data beyond binary features and include categorical features. While logistic regression often requires categorical features to be binarized, doing so can cause a loss of information. For example, treating exam responses as correct or incorrect hides information about the specific incorrect answer the student chose and possible patterns [317] and combining demographics into a single "underrepresented" category can hide the struggles of students of different races and ethnicities [318]. As random forest implementations can often handle categorical features directly, future work can consider how the biases explored in this study might manifest in categorical data and how aggregating or segregating features might introduce its own biases.

Finally, future work can consider how both the data and algorithms affect how features are detected. Recently, Pangastuti et al. found that a combination of bagging, boosting, and SMOTE

improved random forest's classification ability on a large, imbalanced educational data set [319]. We need to be careful with such approaches, however, because our data is not just data but represents actual students. Therefore, we need to be certain that our conclusions are based on the student data and not simulated students "created" to make the data easier to analyze.

6.8 Conclusion and Recommendations

Our work suggests that for both predictive and explanatory models, feature and outcome imbalances can cause algorithms to detect different features despite the same built-in amount of information. We found this to be true for random forest, conditional inference forest, logistic regression, as well as in various penalized regression algorithms. On a practical level, this means that if we are using these algorithms for determining which features might be related to some outcome of interest, we might be introducing false negatives into our results, potentially missing factors that are related to the outcome.

Based on the results of this study, we propose three recommendations for DBER and EDM researchers. First, for smaller data sets with highly imbalanced features, we recommend using a penalized version of logistic regression such as Log-F. Even though Firth penalization was often comparable to Log-F, Firth penalization is not implemented in all statistical software. Log-F however can be used with any statistical software that can perform logistic regression because it is based on data augmentation. If the outcome is also imbalanced, it is even more essential to consider penalized approaches.

Second, for medium or large data sets ($N \geq 1,000$) traditional logistic regression and random forest or conditional inference forest with a permutation importance perform similarly, so either approach works. While the algorithms still do not perform perfectly, none of them provided a consistent advantage over another. We recommend that researchers first consider whether the research questions are best answered using predictive or explanatory techniques and then which affordances of the algorithms are most relevant to the study.

Finally, we call on researchers to include information about their features in their publications,

including the features themselves, their distributions, and in the case of categorical or binary features, their class frequencies as others outside the DBER and EDM communities have done in the past [320]. A simple example of how this might be done is shown in Table 6.4. In addition, we recommend that researchers include data set characteristics or so-called "meta-features" as well. Some examples include sample size, the number of features, the number of numerical features, the number of categorical features, and the percentage of observation of the majority class or outcome balance [290]. Just as there have been calls for increased reporting of demographics in the DBER and EDM communities to understand how results might depend on the sample population or generalize [13, 321], we are calling for the same with the explanatory and predictive models we create, partially addressing some of the questions raised by Knaub, Aiken, and Ding in their analysis of physics education research quantitative work [322]. By doing so, we hope for greater acknowledgement of possible sources of bias or false negatives in feature selection as a result of the data or algorithms used in DBER and EDM studies.

APPENDICES

APPENDIX A

RANDOM FOREST BACKGROUND

The following appendix comes from the supplemental material of Young et al. [65]. The published version includes Grant Allen, John M. Aiken, Rachel Henderson, and Marcos D. Caballero as co-authors. It is reproduced here without changes.

A.1 The Random Forest algorithm

A.1.1 Decision tree learning

Random forests have their roots in the decision tree learning [323]. Decision tree learning uses a set of binary decisions to develop a model for the data set. For our purposes, we focus on classification trees where the object of the model is a class label (i.e., a particular categorical outcome). The decision tree algorithm is provided with the classes and the data that should predict the classes (i.e., input variables). Conceptually, the decision tree algorithm searches the input variables for the one that best segregates the data into separate classes. That choice of "best" can be user specified, but if left to the algorithm, it will be the variable for which a majority of members of a class appear on one side of the decision (termed "branch") and not on the other side. For input variables that are continuous data, the algorithm further decides on the binary decision that best splits the data. For example, if "age $<$ 55" was the binary decision, the algorithm both chose "age" as the input variable and "55" as the cut-off. The algorithm continues to make these decisions, splitting the data into more and more branches until all branches terminate in a single class (termed "leaves") or until a user-specified level. Tracing the path back from any leaf (single class or multiclass) to the starting point shows all the decisions that were made to obtain that leaf. In this sense, decision tree learning is a glass-box algorithm – a researcher can see every step along the path.

Although the researcher can view all parts of the model and how it was constructed, any single decision tree is strongly tied to the data used to construct it. This leads to overfitting of the

data [323, 324]. That is, the decision tree algorithm can produce a model that exactly provides unique classifications for the data it is given. As such, applying that model to predict classes in a new data set will often produce false predictions as the model was so strongly tied to the initial data set on which it was trained. Overfitting is a common problem in machine learning techniques where a single analysis is conducted [325]. To deal with this, some decisions trees are pruned [248] – reduced to a smaller size by removing leaves that predict only small classes. However, as computational time has become less expensive, ensemble methods that develop a series of models from random selections of data are a more common method for combating overfitting [326]. The Random Forest algorithm is one such ensemble method, which grows out of decision tree learning.

A.1.2 Random decision trees

A Random Forest is grown from a set of random decision trees and stems from a technique known as "tree bagging." Tree bagging, or simply bagging, refers to running the decision tree algorithm a specified number of times on a random selection of data with replacement [327]. Data selected to be used in the algorithm is "bagged." Each time the algorithm runs, it produces a decision tree from a randomly selected set of data. The input variables are scored based on how well they classify the data. Those that continually classify well earn higher scores, while those that do not are given lower scores. The result is a set of input variables that have been tested on a variety of data sets so that those input variables that are most important for classification can be found. Random forests use the bagging technique, but also randomly select a subset of input variables [91]. That is, data to develop the Random Forest model are randomly selected and only a subset of input variables (again, randomly chosen) are used in the classification. The same scoring procedure for these input variables is used. Both the bagging and Random Forest algorithms combat overfitting by leveraging random selection as opposed to pruning. Conceptually, when randomly sampling data, the trained model will sometimes produce a good model and at other times it will not. By checking the quality of each decision tree model after each run, the algorithms ensure that consistently-appearing predictors from good models are carried into the complete bagging or Random Forest

model and others are not. Due to randomly selecting a subset of the input variables in addition to a bootstrapped sample of the data, Random Forests have been shown to have significant advantage over bagging alone [328].

A.1.3 Tuning Random Forest Parameters

Random forests have two tuning parameters that can be adjusted to obtain a reliable model, n_{in} – the number of input variables selected at random, and n_{trees} – the number of trees in the forest. A review of how n_{in} can affect the predictions of a Random Forest has been conducted by Svetnik et al. [100]. For a variety of choices of n_{in}, Svetnik et al. compared error rates (fraction of false positives and false negatives) predictions and found that for most choices of n_{in} error rates were well maintained between 20-25%. For lower n_{in}, the model does not develop enough robust comparisons between different input variables to be reliable leading to slightly elevated error rates. While for higher n_{in}, the model overfits the training data leading to higher scores for less important variables, which again produce slightly higher error rates. For a given number of input variables, N, Svetnik et al. suggested n_{in} include $N/2$, $N/4$, and $\sqrt{N}$ where each is rounded up or down to the nearest integer. They tested these suggested choices and found that all choices produce similar error rates even as the number of input variables is varied between 3 and 100. After 100 variables, the choice $n_{in} = \sqrt{N}$ performs slightly better than the other choices, but only marginally so (2% difference in error rates).

The number of trees in the forest, n_{trees}, describes the number of times that the algorithm randomly selects data and input variables to perform the classification task. Here, more is better, but only up to a point, after which adding addition trees does not improve the classification. There is no penalty for running the algorithm many times besides wasting computational resources. An estimate of the model's performance for a given number of trees is the Out-Of-Bag error (OOB error). When the algorithm selects data to train, it leaves some data "out of the bag." For any given tree in the forest, we can predict the classifications for the data left out of the bag. The error rate associated with that prediction is an estimate of the Out-Of-Bag error for that tree. The estimate

		Known Classes	
		Yes	No
Model	Yes	N_{TP}	N_{FP}
Predicts	No	N_{FN}	N_{TN}

Figure A.1: The confusion matrix counts the number of each predicted classification by the model and compares that to the what the data indicates. In this case, a two class system with binary classifications leads to a 2 x 2 matrix. For M classes, the matrix continues to be square and grows to be M x M.

for the total OOB error is the average across all the trees. In their work, Svetnik et al., found that OOB error stabilized when $n_{trees} > 10^2$. For 3 orders of magnitude beyond that the OOB error remained flat at 0.2, that they found a ~20% average misclassification of data left out of the bag for any number of trees from $10^2 - 10^5$.

It is common to determine the "best" parameters by using a grid search [329]. Here, a number of random forests are constructed with different combinations of n_{in} and n_{trees} to find the combination with the strongest validation scores. Typically, one performs a coarse-grain search allowing the tunable parameters to vary greatly. This is much like searching for the appropriate order of magnitude for each parameter. Once a reasonable range is found for each parameter, a finer-grain search is performed within the bounds determined from the coarse grained search. This search can continue *ad infinitum*, but it is typically only done until reasonable estimates of the "best" parameters are found given the expected error and available computational time.

A.2 Validating the Random Forest model

A Random Forest will develop a model of data, but that does not always mean that model is meaningful. Moreover, because that model is developed from a random selection of data and input variables, individual trees in the model might be terrible predictors. By abiding by suggested parameter choices [100], one can be somewhat confident in the model. However, additional validation of the model can be conducted to be able to provide evidence of that confidence. These metrics and the associated curves are developed from how well the model predicts classifications of the test data – that is, the data that was not used to train the model.

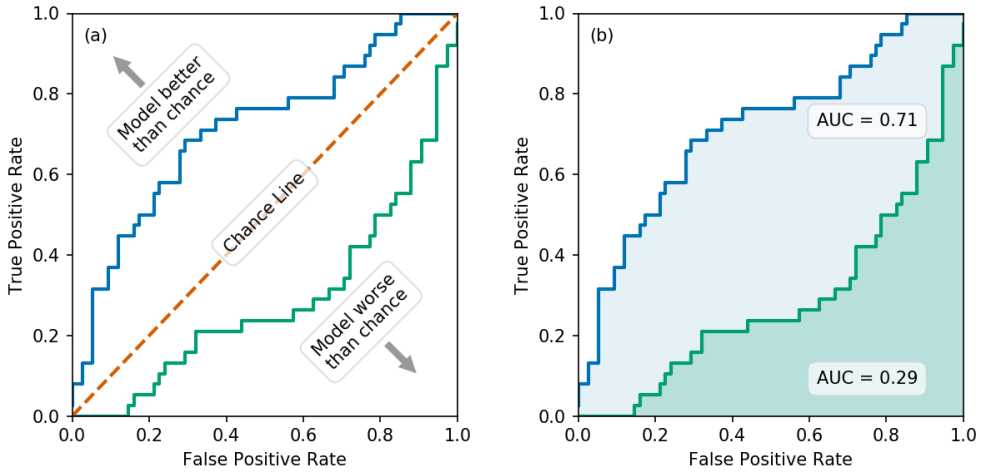

Figure A.2: (a) Sample receiver operating characteristic (ROC) curves that demonstrate two models: one that is better than chance (blue) and one that is worse than chance (green). These ROC curves are plotted along with the chance line (orange dotted). Models that are demonstrably better than chance have ROC curves that tend towards the upper-left corner of the space as the arrow indicates. Models that are worse than chance tend towards the bottom-right corner. (b) For both models, the area under the ROC curves (AUC) are shown (blue and green shading) and computed. AUC provides a measure of the quality of the model. It is indicative of the probability of accurately classifying a random sample from the data.

A.2.1 Confusion Matrix

The simplest tool for understanding how well the Random Forest model predicts classifications in the new data set is the confusion matrix. Most other measures associated with validity of the model are derived from the confusion matrix. Conceptually, the confusion matrix keeps track of true positives (N_{TP}), true negatives (N_{TN}), false positives (N_{FP}), and false negatives (N_{FP}). The sum of all these measures is the total number of observations in the data set that is being tested (N_{test}).

$$N_{test} = N_{TP} + N_{TN} + N_{FP} + N_{FN} \tag{A.1}$$

For a two class system (e.g., Yes/No), these values can be organized into the 2 x 2 matrix where the columns describe the known classes in the data set and the rows describe the classes predicted by the model (Fig. A.1). The confusion matrix provides a quick check of the predictions of the Random Forest model. Essentially a good model will have strong diagonal elements, that is, high

numbers of true predictions, and small off-diagonal elements, low numbers of false predictions.

A.2.2 Associated measures

From the confusion matrix, a number of associated measures may be derived. Here we provide those that are common to report in the Random Forest literature. Additional measures exist, but are not reported here [1]. The accuracy of the model (ACC) is the fraction of true predictions compared to the total number of observations in the test data set,

$$ACC = \frac{N_{TP} + N_{TN}}{N_{test}}. \tag{A.2}$$

This accuracy of the model can vary between 0 and 1, with 0.5 being equal to chance predictions. A model that predicts worse than chance will have $ACC < 0.5$. The sensitivity, or the true positive rate (TPR) compares the number of predicted true positives to the total number of actual positives appearing in the test data,

$$TPR = \frac{N_{TP}}{N_{TP} + N_{FN}}. \tag{A.3}$$

This rate varies between 0 and 1. For a good model of the data, we expect this number to be closer to 1. The fall-out, or false positive rate (FPR) compares the number of predicted false positives negatives to the total number of actual negatives in the data,

$$FPR = \frac{N_{FP}}{N_{FP} + N_{TN}}. \tag{A.4}$$

This rate varies between 0 and 1. For a good model of the data, we expect this number to be closer to 0. Pure guessing (i.e., chance) would yield 0.5 for both of these values. Taken together, these values are plotted together for a range of discrimination thresholds in a receiver operating characteristic curve, which indicate how much better (or worse) the model is than chance.

[1] In some publications, certain likelihood ratios ($LR+ = TPR/FPR$ and $LR- = FNR/TNR$), odds ratios ($LR+/LR-$), and the F_1 score are reported, but ACC, TPR, and FPR are the most commonly reported metrics.

A.2.3 Receiver operating characteristic curve

The receiver operating characteristic curve (ROC curve) provides a visualization of the quality of a binary model [330]. In it, TPR is plotted against FPR for a variety of discrimination thresholds. These thresholds vary from 0 to 1 and describe the probability above which an observation is placed into one class compared to another. Conceptually, it determines, at a given probability of classification, the expected rates of true positives compared to false positives.

A sample ROC for mock data is plotted in Fig. A.2(a). The chance line, in which TPR and FPR are equal for all thresholds, is plotted in orange. The blue curve is the ROC curve for a model that is better than chance. The humped-shaped of curve is common for good models as this shape are indicative of TPR values above chance for all thresholds. On the other hand, the green curve is the ROC curve for a model that is worse than chance. This shape is characteristic as well as it is indicative of TPR below chance for all thresholds.

A quantitative measure of the quality of the model is the area under the ROC curve (AUC). This measure is visually represented in Fig. A.2(b) where AUC for the better than chance model is indicated by the blue and the green shading taken together. The AUC for the worse than chance model is indicated by the green shading alone. AUC is indicative of the probability that model will rank a randomly chosen positive instance higher than a randomly chosen negative one. From Fig. A.2, the probability for the first model is approximately 0.71 while for the second it 0.29. AUCs above 0.7 are typically considered reasonable models with 0.8 and above considered to be good models [93]. A perfect classifier will have an area under the curve of 1 while the chance curve will have an area under the curve of 0.5.

A.3 Feature selection

One of the more useful aspects of machine learning, and of the Random Forest algorithm in particular, is the ability to determine which features are more important than other features in the data. For classification tasks, that means finding the input variables that consistently separate the results into classes. There is an analogy to regression analysis where the important input variables

in a Random Forest classifier act as statistically significant correlates with the outcome variable. However, because Random Forests are not rooted in traditional statistical analysis, the important features do not arise from correlation – linear or otherwise.

Feature selection makes use of these important features to reduce the overall number of input variables needed to classify the data. This is similar to using regression models of increasing complexity to find the minimal model that explains the outcomes sufficiently. These important features can be used as the sole input variables and the resulting model can be validated using the techniques described in Sec. A.2. A good reduced model will maintain high accuracy, produce an ROC curve that is still above the chance line for all thresholds, and have an AUC that is similarly well above chance while using the minimum amount of features.

To determine the importance of an input variable (termed "feature importance"), the standard Random Forest algorithm (CART) continuously compares how well each input variable in a single decision tree separates the data set into classes. For the CART algorithm, the measure of how well this occurs is either the Gini impurity or the information gain, depending on user selection and choice of tool. For the simplest implementation of the Random Forest classifier, the feature importance is related to the Gini impurity (I_G) [91], which is the total decrease in node impurity. I_G is computed for each input variable (node) and is then averaged for each input variable over all the trees in the forest. Conceptually, I_G for an input variable is probability of the input variable showing up in a given class multiplied by the probability of a misclassification within that factor summed over all classes [331]. Thus the higher I_G for an input variable in a given tree, the less favorable choice it is for splitting the tree. For example, a high I_G input variable would not be selected as the input variable for the first branch in a given tree. For this implementation, important features are those which consistently produce the best splits for a large proportion of trees in the forest. The feature importances are often distributed normally around some mean and are reported with error that is inversely proportional to the square root number of trees in which the input variable was randomly selected for use.

A.4 Bias and improvements

While the CART algorithm and the associated Gini-based feature selection are commonly used in Random Forest classification, both are subject to biases that for certain kinds of data (including those analyzed in this paper) can lead to inaccurate models.

First, the Gini-based feature selection described above is not reliable when the input variables vary in scale of measurement or in the number of categories (possible responses) [92]. This is because variables with more categories can be split into two groups in more ways than variables with less categories can and hence, it is more likely a favorable split could be found. Furthermore, Gini-based feature importances can be biased if the variables are correlated [332]. In such cases, accuracy-based permutation variable importances can be used [243, 244].

Accuracy-based permutation variable importances are based on the idea that if a input variable is associated with the outcome variable, then permuting the input variable should break that association, and therefore, the accuracy (ACC) should decrease. The input variables that change the accuracy the most are then said to have the largest variable importance.

However, these accuracy based variable importances are biased when the sample is unbalanced, that is, the categories in the predicted variable do not occur in equal frequencies [94]. Janitza et al. suggest modifying the accuracy-based permutation variable importances to be based on the AUC instead of accuracy because accuracy is biased toward the majority class while the AUC is not. When applying this modification, they found that the AUC-based permutation feature importances are better able to discriminate between variables that are good predictors and variables that are poor predictors than the accuracy-based permutation feature importances are when the sample is unbalance. When the sample is balanced, they reported no significant different between the two methods.

Because the Random Forest algorithm is based off classification and regression trees (CART) [309] and CART uses the Gini impurity to determine the split points, the Random Forest algorithm itself is biased for the same reasons the Gini-based feature importance is. This bias can be corrected using conditional inference forests based on the framework proposed by Hothorn et al [92, 259].

The conditional inference forest algorithm breaks the determining the split points into two steps, unlike the Random Forest algorithm, which selects the input variable and its split point in a single step. First, algorithm tests the global null hypothesis that there is no association between any of the input variables and predicted variable at some predetermined level of significance. If this test fails, the algorithm terminates. If the null hypothesis can be rejected, the algorithm selects the input variable with the strongest association to the predicted variable as measured by its P value. The algorithm then splits the input variable into two groups that maximizes a chosen test statistic. Finally, the algorithm returns to the first step and retests the global null hypothesis and either terminates or continues with the next input variable with the highest association to the predicted variable. This revised version of the Random Forest algorithm has been found to be unbiased even when the variables vary in scale of measurement or in the number of categories provided that boostrapping is not used (subsampling must occur without replacement or the original biases are still present) [92].

A.4.1 Determining “significant" features

While we have detailed various improvements to variable selection, none of these methods can by themselves determine whether the variables are actually important, only how important they are with respect to each other. Various approaches to determine the actual important variables have been proposed such as selecting the top 10% of the variables ordered by an importance measure [333], selecting all variables above the absolute value of the most negative importance values [328], using recursive backward elimination to select the fewest number of variables that result in an OOB rate within 1 standard error of the best OOB rate [97], generating a null distribution of each variable and then assigning a *p*-value based on the fraction of null importances greater than the actual importance value [292], and mirroring the distribution of negative importances to generate an overall null distribution for the importances and again assigning a *p*-value based on the fraction of null importances greater than the actual importance value [297]. Each of these approaches has its own benefits and problems such related to ease of implementation and computation time required

and to our knowledge, there is no standard choice of procedure. As we had a limited number of negative importances and limited computational power, we used recursive backward elimination in this study. Since these “significant" factors are not determined by tests of statistical significant but rather by how much the model changes when they are removed, we refer to the selected factors as meaningful factors.

APPENDIX B

CHAPTER 3 ANALYSIS OF FEATURES

In this appendix, we describe the data used to answer research questions 2 and 3 to give the reader a better idea of the distributions of physics GRE scores and GPA in the data set. Because the data are skewed left and exhibit ceiling effects (many applicants have 4.0 GPAs or 990 physics GRE scores), quartiles are used to describe the various features. To maximize the amount of information shown about the data, we use raincloud plots [334, 335], which show the distribution, the density plot, and traditional box plot. Kolmogorov-Smirnov tests suggest the distributions are not significantly different whether we include applicants who may have applied to multiple schools in our data set, so we include possible duplicates in our analysis.

Fig. B.1 shows the physics GRE scores and undergraduate GPAs of each applicant based on whether they attended a large undergraduate physics program (top 25% nationally in yearly

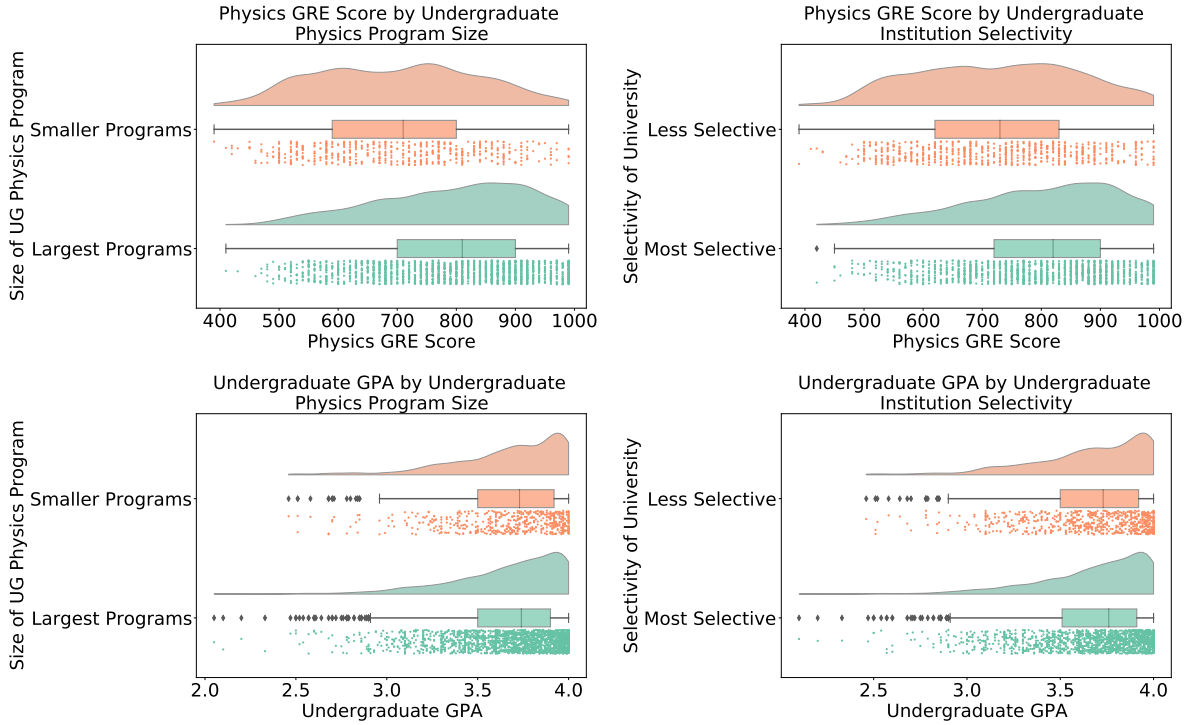

Figure B.1: Distribution of physics GRE scores & undergraduate GPAs by the size of the undergraduate physics program & institutional selectivity for each applicant.

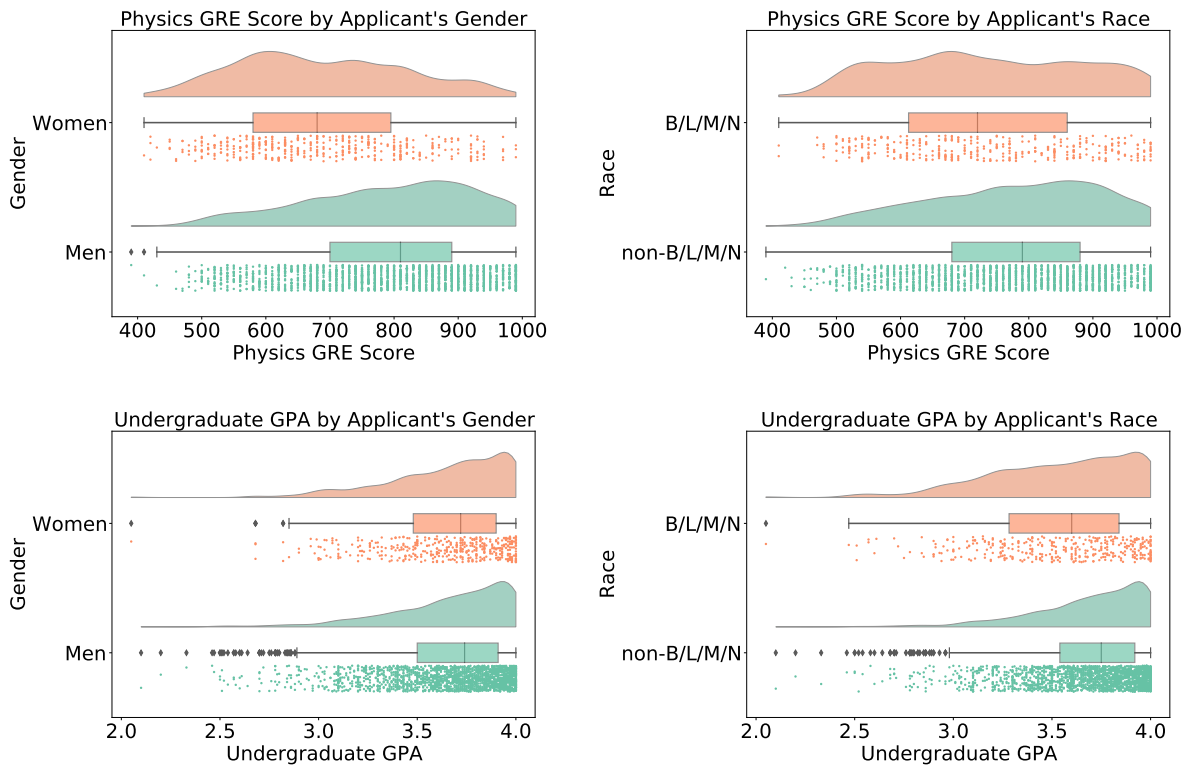

Figure B.2: Distribution of physics GRE scores and undergraduate GPAs by gender and whether the applicant identified as a member of racial or ethnic group currently underrepresented in physics.

physics bachelor‘s degrees) or attended a selective university (categorized as most competitive or highly competitive based on Barron‘s Selectivity Index). We notice that the physics GRE score distributions are shifted to the right for applicants from large physics departments or selective institutions, signifying higher scores. Indeed, the median physics GRE scores of applicants from large programs or selective institutions are nearly 100 points higher than those of applicants from smaller or less selective institutions. However, in terms of GPA, the median GPA is approximately the same, regardless of whether the applicant graduated from a larger or smaller physics department or attended a more or less selective institution.

Fig. B.2 shows the physics GRE and undergraduate GPAs by gender and race. As expected, men score higher on the physics GRE than women do and Asian and white applicants score higher than Black, Latinx, Multiracial, or Native applicants, though the gaps appear larger than those reported in [52].

When comparing GPAs, we find that men and women have similar GPAs, as recently reported

in [156] when comparing men and women's STEM GPAs. Likewise, our data also shows a racial GPA gap with non-B/L/M/N applicants having a median GPA higher than that of B/L/M/N applicants by 0.15.

When looking across both figures, we notice that the physics GRE score distributions of smaller and less selective programs resemble the physics GRE distributions of women and B/L/M/N applicants while the physics GRE score distributions of the largest and most selective programs resemble the physics GRE score distributions of men and non-B/L/M/N applicants. To see if gender and race are confounding variables in our analysis, we examined the fraction of women and B/L/M/N applicants in each group. If this were the case, the smaller and less selective programs should have a greater fraction of women and B/L/M/N applicants than the larger and more selective programs.

However, we did not find this to be the case. Applicants from more selective institutions were 16% women while applicants from less selective institutions were 18% women (15% and 14% respectively for B/L/M/N applicants). For institution size, applicants from larger institutions were 16% women compared to 21% women from smaller institutions (14% and 17% for B/L/M/N applicants respectively). Thus, it does not appear that differences in who attends (in terms of gender and race) larger or more selective institutions are responsible for the observed differences in scores.

APPENDIX C

CHAPTER 3 SUPPLEMENTAL FIGURES

In this appendix, we include plots that show institutional effects by gender and race. As we note in the discussion, the programs included in this study were actively trying to increase the diversity of their graduate programs. Thus, we are unable to determine from our data whether women were admitted at higher rates than men were and B/L/M/N applicants were admitted at higher rates than non- B/L/M/N applicants were because they stood out or if the admissions committees highlighted these applicants from the start. Therefore, we do not comment on any possible interactions. For completeness, the plots are shown here.

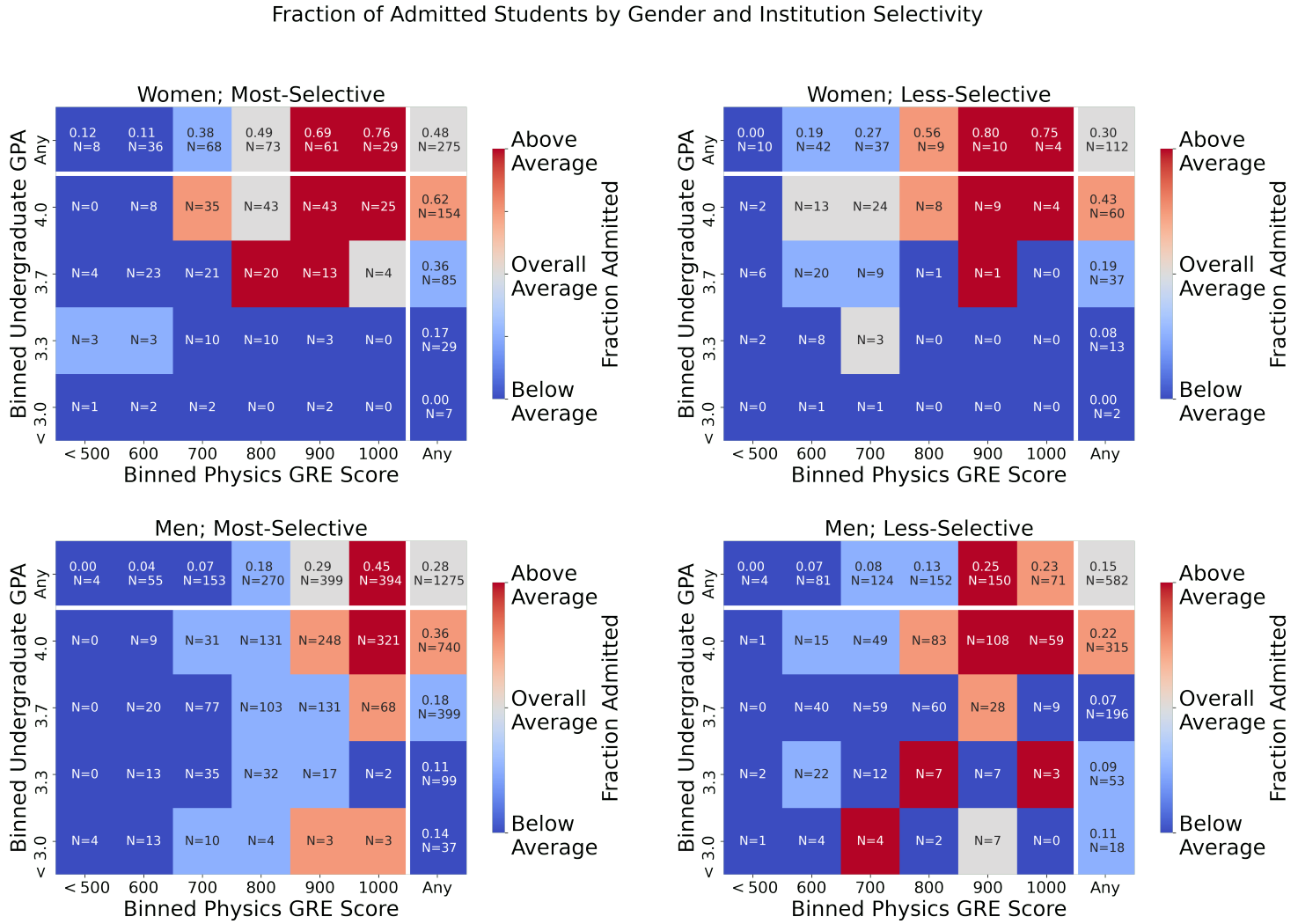

Figure C.1: Admission fractions of applicants split by their gender and the selectivity of their undergraduate institutions.

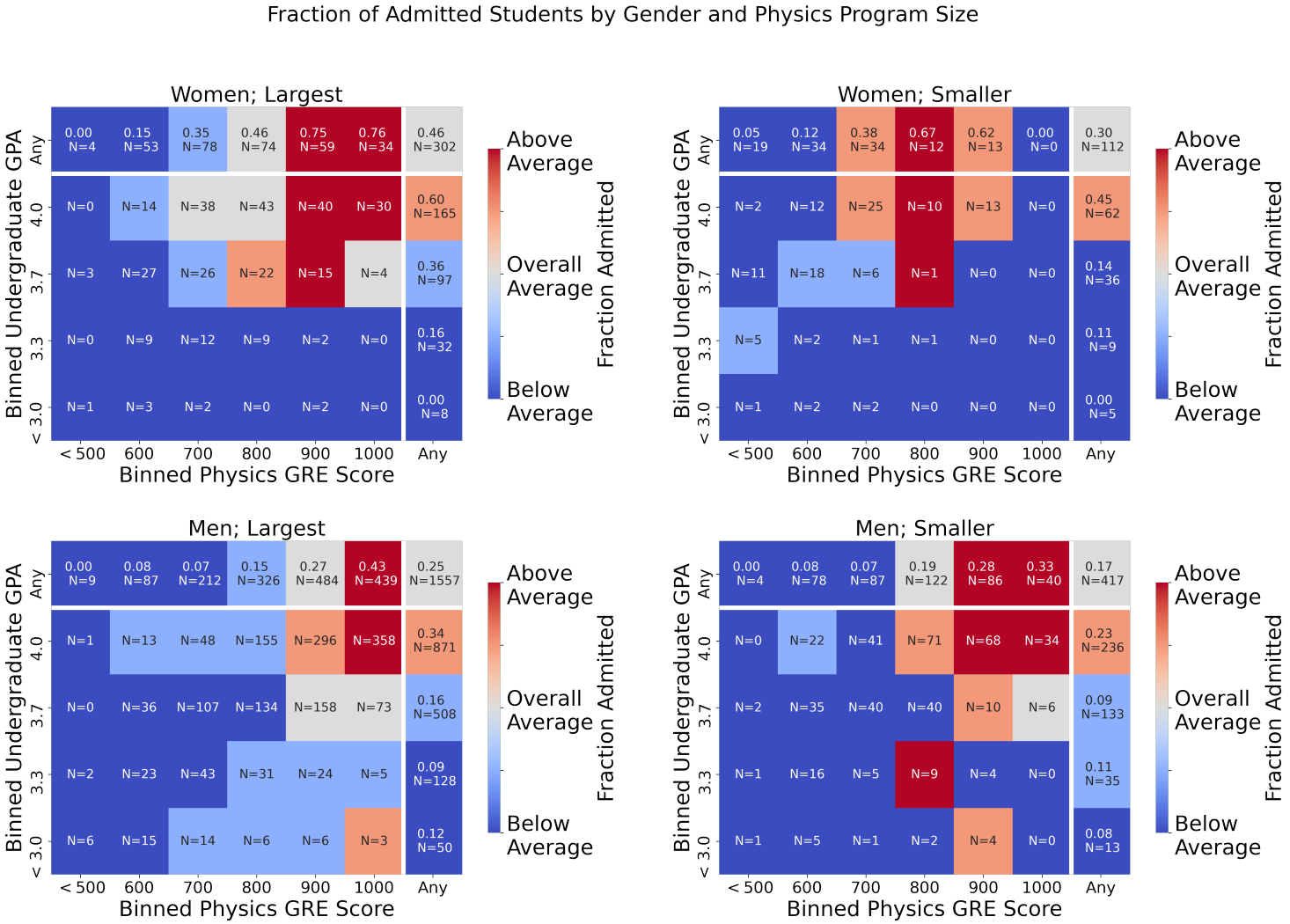

Figure C.2: Admission fractions of applicants split by their gender and the size of their undergraduate institutions.

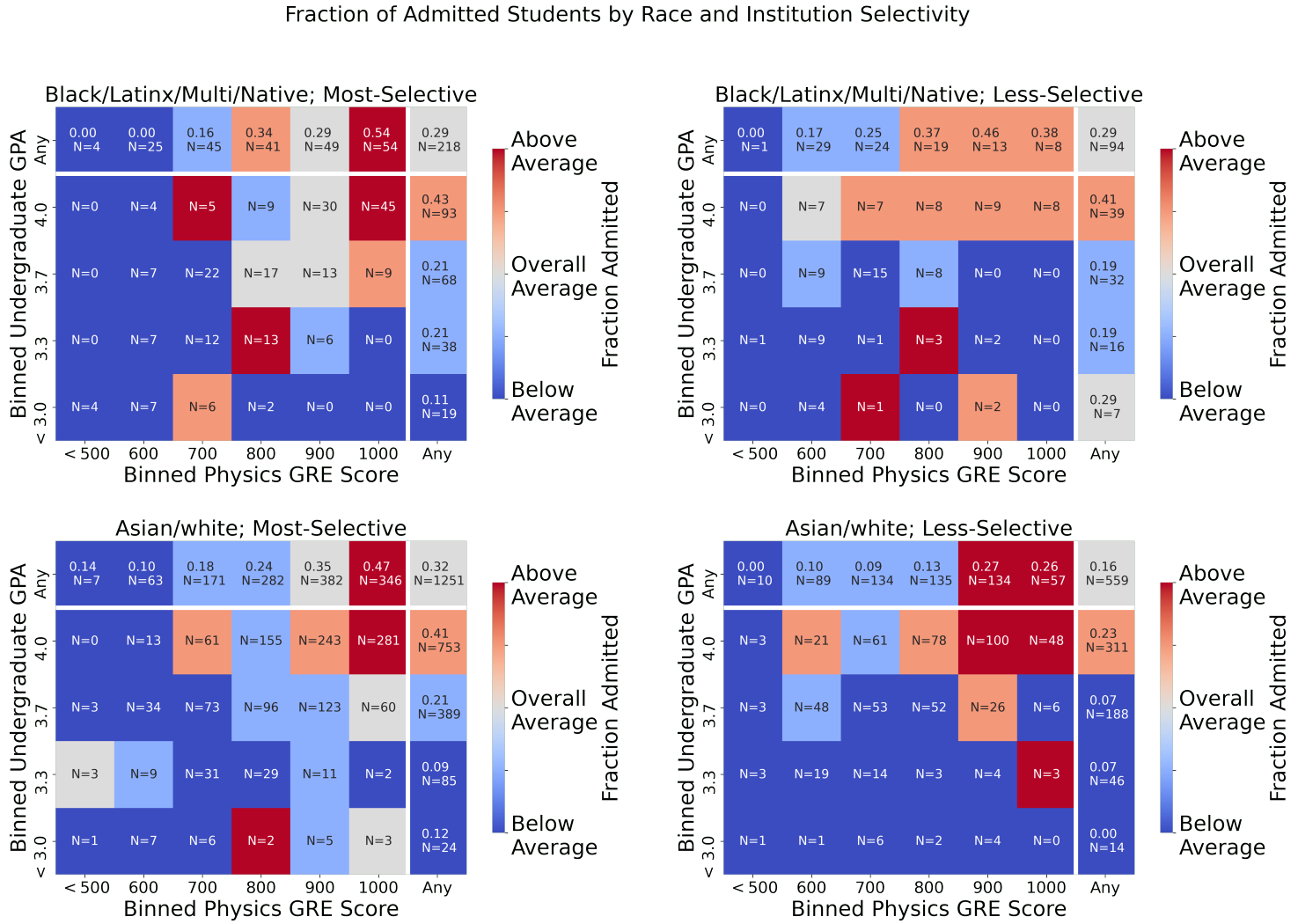

Figure C.3: Admission fractions of applicants split by their race and the selectivity of their undergraduate institutions.

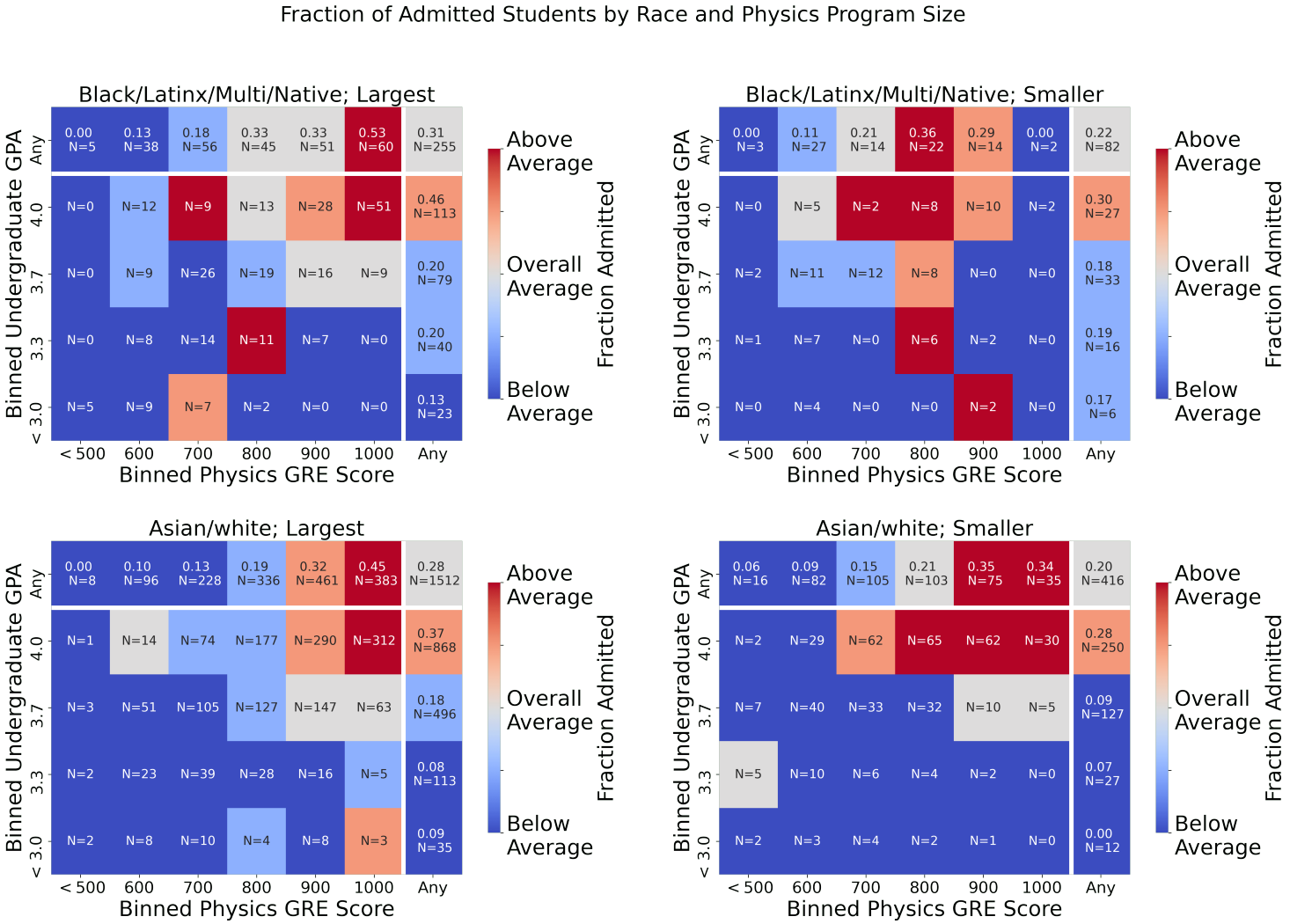

Figure C.4: Admission fractions of applicants split by their race and the size of their undergraduate institutions.

APPENDIX D

CHAPTER 4 SUPPLEMENTAL FIGURES

Here we present figures showing the results split by admission status and gender, undergraduate institution selectivity, and undergraduate physics program size. Given the relatively small sample sizes, we did not conduct tests of statistical significance.

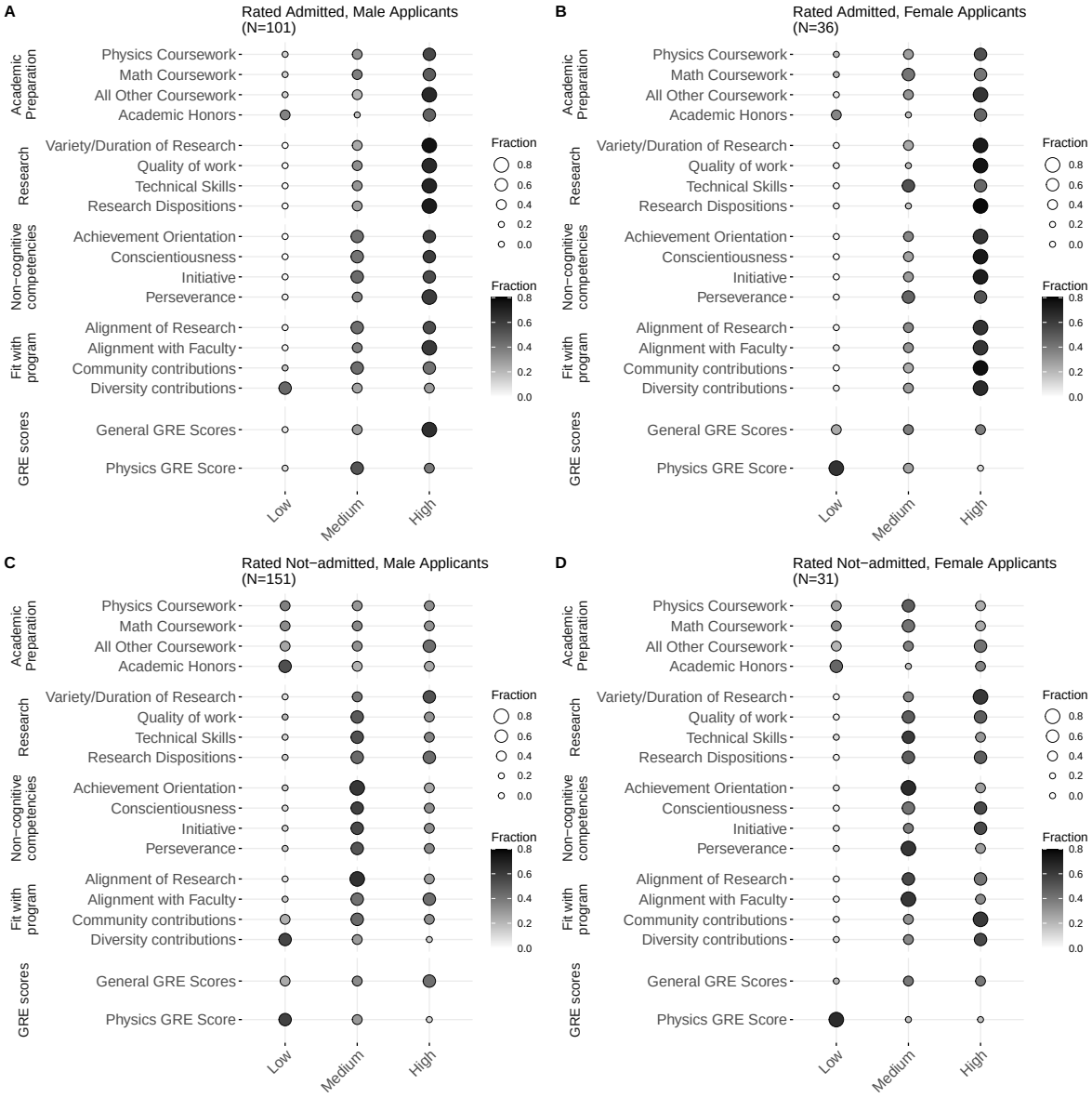

Figure D.1: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant was male or female and whether they were admitted or not.

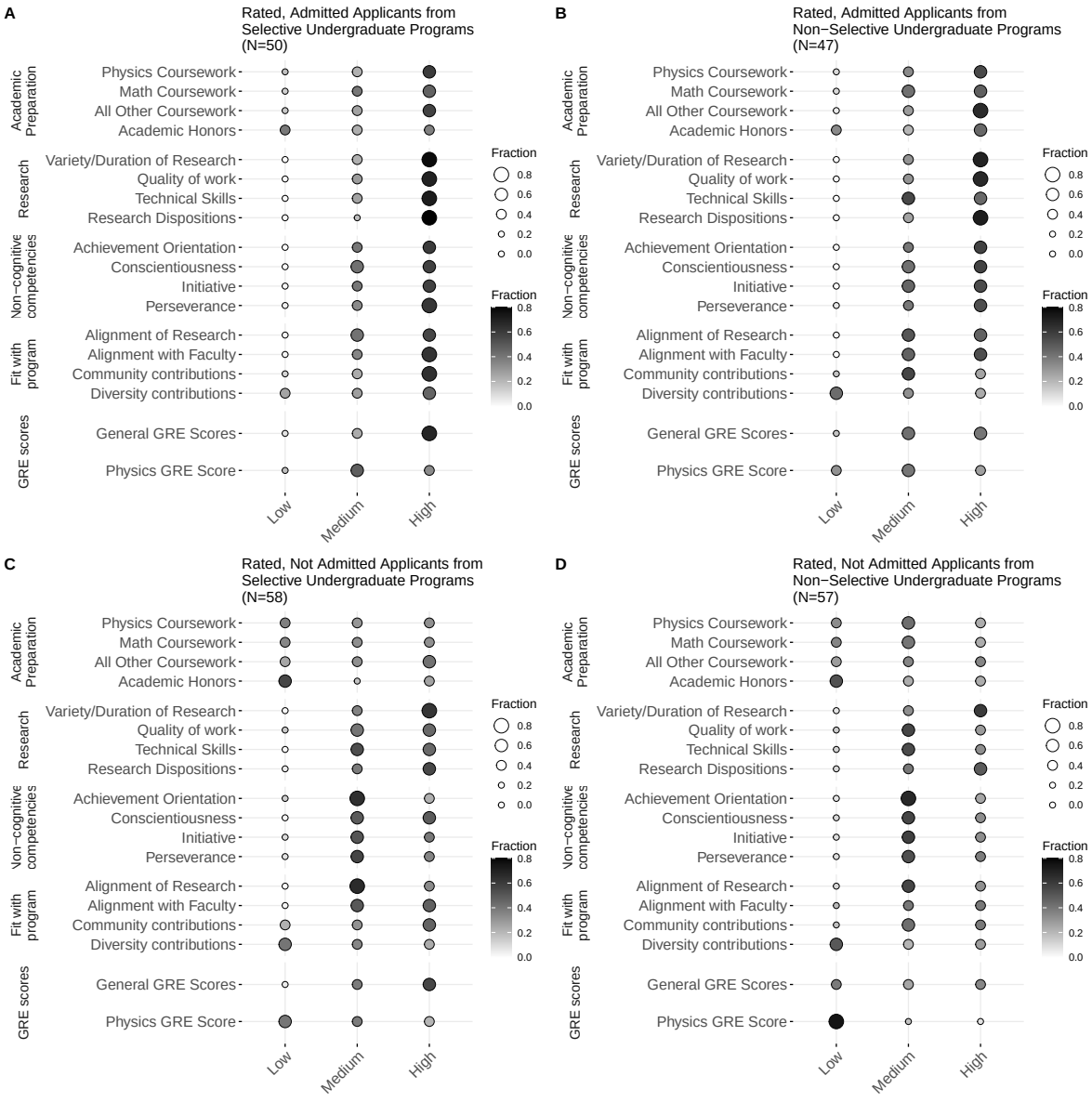

Figure D.2: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant attended a more selective or less selective undergraduate university and whether they were admitted or not.

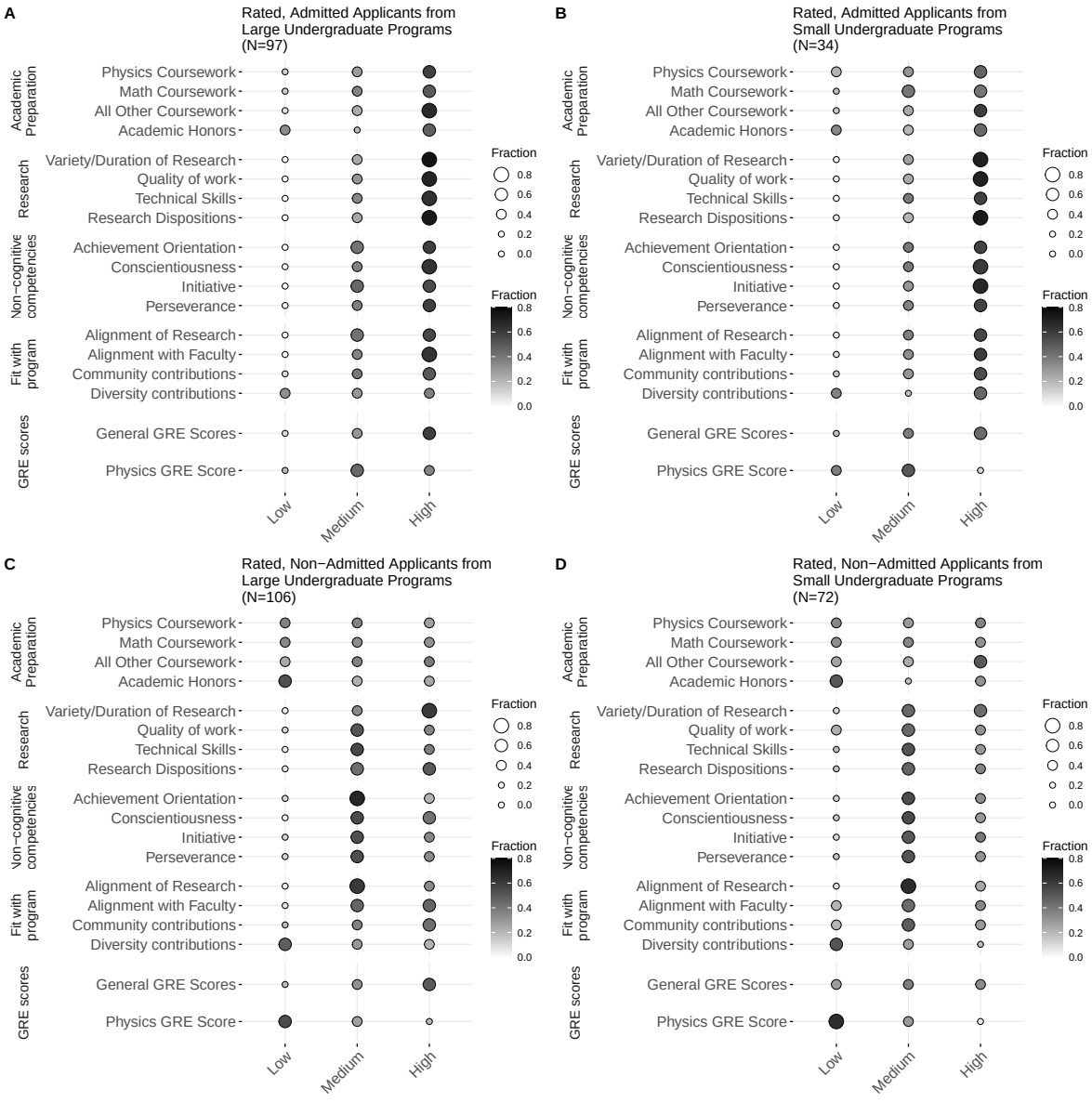

Figure D.3: Faculty ratings of domestic applicants on 18 constructs split by whether the applicant attended a university with a larger or smaller physics program and whether they were admitted or not.

APPENDIX E

CHAPTER 6 SUPPLEMENTAL FIGURES

Here we provide the plots for the additional three data sets from Sec. 6.5 for completeness. The plots show the same general results as discussed in the main manuscript. We also include the residual plots from the discussion.

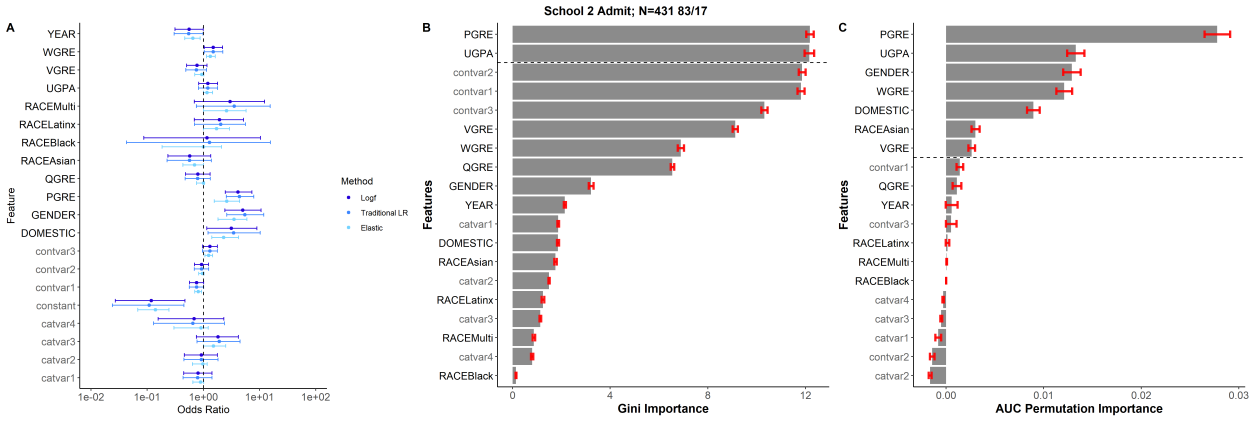

Figure E.1: Comparison of the odds ratio, Gini importance, and AUC-permutation importance for the features in the school 2 admit data set.

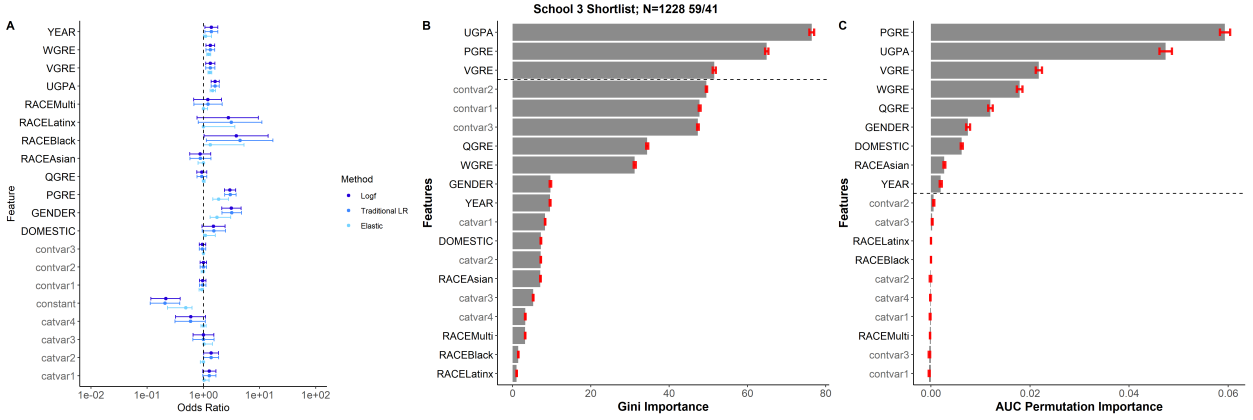

Figure E.2: Comparison of the odds ratio, Gini importance, and AUC-permutation importance for the features in the school 3 shortlist data set.

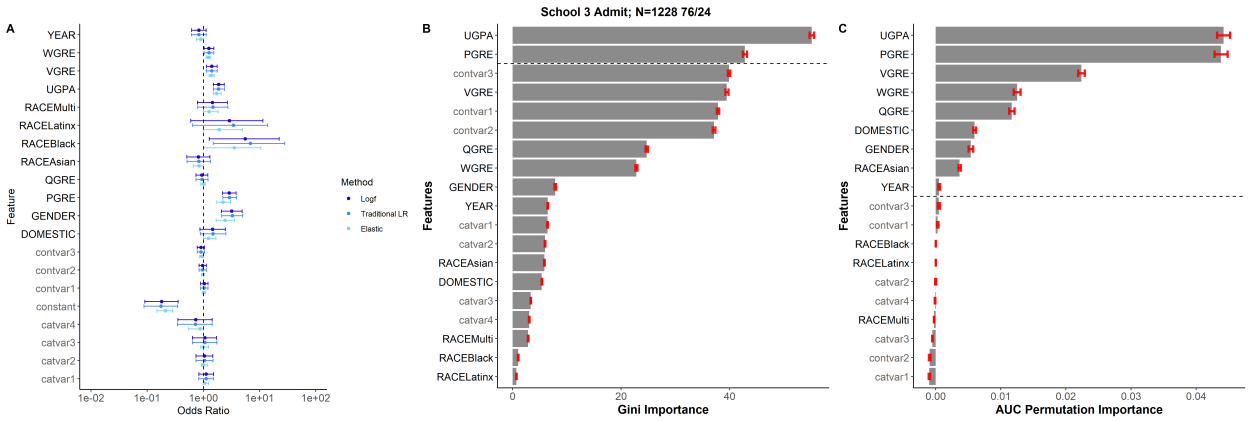

Figure E.3: Comparison of the odds ratio, Gini importance, and AUC-permutation importance for the features in the school 3 admit data set.

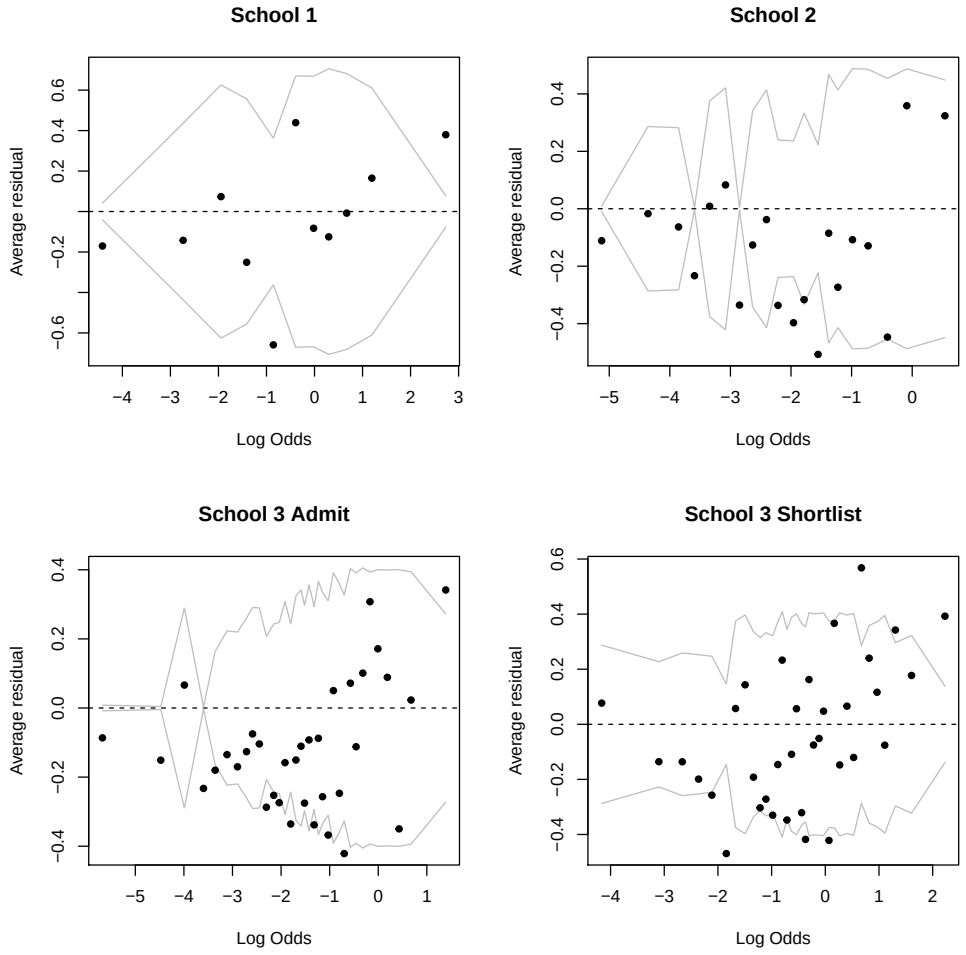

Figure E.4: Plots of the Log-odds vs the average residual in each bin for the four schools. Across all plots, between 20% and 34% of the points fall outside of the confidence intervals, suggesting the logistic regression models might not be fitting the data especially well.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] GRE Subject Test Interpretative Data.

[2] Rosemary S. Russ and Tor Ole B. Odden. Physics Education Research as a Multidimensional Space: Current Work and Expanding Horizons. In Charles Henderson and Kathleen A. Harper, editors, *Getting Started in PER*. American Association of Physics Teachers, College Park, MD, 2018.

[3] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2019. Technical report, September 2020.

[4] Physicists and Astronomers : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics.

[5] Employment and Careers in Physics, April 2020.

[6] U.S. Census Bureau QuickFacts: United States.

[7] Patrick J. Mulvey and Starr Nicholson. Physics Bachelor's Degrees: 2018, August 2020.

[8] Lillian C. McDermott and Edward F. Redish. Resource Letter: PER-1: Physics Education Research. *American Journal of Physics*, 67(9):755–767, September 1999.

[9] Jennifer L Docktor and José P Mestre. Synthesis of discipline-based education research in physics. *Physical Review Special Topics-Physics Education Research*, 10(2):020119, 2014.

[10] Tor Ole B. Odden, Alessandro Marin, and Marcos D. Caballero. Thematic analysis of 18 years of physics education research conference proceedings using natural language processing. *Physical Review Physics Education Research*, 16(1):010142, June 2020.

[11] National Research Council. *Adapting to a Changing World–Challenges and Opportunities in Undergraduate Physics Education*. National Academies Press, Washington, D.C., June 2013.

[12] Lin Ding. Theoretical perspectives of quantitative physics education research. *Physical Review Physics Education Research*, 15(2):020101, July 2019.

[13] Stephen Kanim and Ximena C. Cid. Demographics of physics education research. *Physical Review Physics Education Research*, 16(2):020106, July 2020.

[14] Marcos D Caballero, Bethany R Wilcox, Leanne Doughty, and Steven J Pollock. Unpacking students' use of mathematics in upper-division physics: where do we go from here? *European Journal of Physics*, 36(6):065004, 2015.

[15] Brian Farlow, Marlene Vega, Michael E. Loverude, and Warren M. Christensen. Mapping activation of resources among upper division physics students in non-Cartesian coordinate systems: A case study. *Physical Review Physics Education Research*, 15(2):020125, September 2019.

[16] Mary Bridget Kustusch, Corinne Manogue, and Edward Price. Design tactics in curriculum development: Examples from the Paradigms in Physics ring cycle. *Physical Review Physics Education Research*, 16(2):020145, December 2020.

[17] Qing X. Ryan and Benjamin P. Schermerhorn. Students' use of symbolic forms when constructing equations of boundary conditions. *Physical Review Physics Education Research*, 16(1):010122, April 2020.

[18] Tao Tu, Chuan-Feng Li, Zong-Quan Zhou, and Guang-Can Guo. Students' difficulties with partial differential equations in quantum mechanics. *Physical Review Physics Education Research*, 16(2):020163, December 2020.

[19] Bethany R. Wilcox and Giaco Corsiglia. Cross-context look at upper-division student difficulties with integration. *Physical Review Physics Education Research*, 15(2):020136, October 2019.

[20] J Christopher Moore and Louis J Rubbo. Scientific reasoning abilities of nonscience majors in physics-based courses. *Physical Review Special Topics-Physics Education Research*, 8(1):010106, 2012.

[21] Jessica Watkins, Janet E Coffey, Edward F Redish, and Todd J Cooke. Disciplinary authenticity: Enriching the reforms of introductory physics courses for life-science students. *Physical Review Special Topics-Physics Education Research*, 8(1):010112, 2012.

[22] Ramón S Barthelemy, Melinda McCormick, and Charles Henderson. Gender discrimination in physics and astronomy: Graduate student experiences of sexism and gender microaggressions. *Physical Review Physics Education Research*, 12(2):020119, 2016.

[23] Elizabeth Gire and Edward Price. Arrows as anchors: An analysis of the material features of electric field vector arrows. *Physical Review Special Topics-Physics Education Research*, 10(2):020112, 2014.

[24] CD Porter and AF Heckler. Effectiveness of guided group work in graduate level quantum mechanics. *Physical Review Physics Education Research*, 16(2):020127, 2020.

[25] Rachel E Scherr, Mike A Lopez, and Marialis Rosario-Franco. Isolation and connectedness among Black and Latinx physics graduate students. *Physical Review Physics Education Research*, 16(2):020132, 2020.

[26] Ben Van Dusen, Ramón S Barthelemy, and Charles Henderson. Educational trajectories of graduate students in physics education research. *Physical Review Special Topics-Physics Education Research*, 10(2):020106, 2014.

[27] Ying Cao and Bárbara M Brizuela. High school students' representations and understandings of electric fields. *Physical Review Physics Education Research*, 12(2):020102, 2016.

[28] Emily A Dare and Gillian H Roehrig. "If I had to do it, then I would": Understanding early middle school students' perceptions of physics and physics-related careers by gender. *Physical Review Physics Education Research*, 12(2):020117, 2016.

[29] Jessie Durk, Ally Davies, Robin Hughes, and Lisa Jardine-Wright. Impact of an active learning physics workshop on secondary school students' self-efficacy and ability. *Physical Review Physics Education Research*, 16(2):020126, October 2020.

[30] Lisa M Goodhew and Amy D Robertson. Exploring the role of content knowledge in responsive teaching. *Physical Review Physics Education Research*, 13(1):010106, 2017.

[31] Ericka Lawton, Carrie Obenland, Christopher Barr, Matthew Cushing, and Carolyn Nichol. Improving high school physics outcomes for young women. *Physical Review Physics Education Research*, 17(1):010111, March 2021.

[32] Amy D Robertson and Abigail R Daane. Energy Project professional development: Promoting positive attitudes about science among K-12 teachers. *Physical Review Physics Education Research*, 13(2):020102, 2017.

[33] Xiaoming Zhai, Barbara Schneider, and Joseph Krajcik. Motivating preservice physics teachers to low-socioeconomic status schools. *Physical Review Physics Education Research*, 16(2):023102, October 2020. Publisher: American Physical Society.

[34] Jessica L. Alzen, Laurie S. Langdon, and Valerie K. Otero. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*, 5(1):56, December 2018.

[35] Stephanie V. Chasteen and Rajendra Chattergoon. Insights from the Physics and Astronomy New Faculty Workshop: How do new physics faculty teach? *Physical Review Physics Education Research*, 16(2):020164, December 2020.

[36] Dimitri R Dounas-Frazer and HJ Lewandowski. Electronics lab instructors' approaches to troubleshooting instruction. *Physical Review Physics Education Research*, 13(1):010102, 2017.

[37] Alice Olmstead and Chandra Turpen. Assessing the interactivity and prescriptiveness of faculty professional development workshops: The real-time professional development observation tool. *Physical Review Physics Education Research*, 12(2):020136, 2016.

[38] Alanna Pawlak, Paul W. Irving, and Marcos D. Caballero. Learning assistant approaches to teaching computational physics problems in a problem-based learning course. *Physical Review Physics Education Research*, 16(1):010139, June 2020.

[39] Danny Doucette, Russell Clark, and Chandralekha Singh. Professional development combining cognitive apprenticeship and expectancy-value theories improves lab teaching assistants' instructional views and practices. *Physical Review Physics Education Research*, 16(2):020102, July 2020.

[40] Melanie Good, Emily Marshman, Edit Yerushalmi, and Chandralekha Singh. Graduate teaching assistants' views of broken-into-parts physics problems: Preference for guidance overshadows development of self-reliance in problem solving. *Physical Review Physics Education Research*, 16(1):010128, May 2020.

[41] Tong Wan, Constance M Doty, Ashley A Geraets, Christopher A Nix, Erin K H Saitta, and Jacquelyn J Chini. Evaluating the impact of a classroom simulator training on graduate teaching assistants' instructional practices and undergraduate student learning. *Physical Review Physics Education Research*, 17(1):010146, June 2021.

[42] Ben Van Dusen and Jayson Nissen. Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation. *Physical Review Physics Education Research*, 16(1):010117, April 2020.

[43] Geraldine L Cochran, Andrea G Van Duzor, Mel S Sabella, and Brian Geiss. Engaging in self-study to support collaboration between two-year colleges and universities. In *2016 Physics Education Research Conference Proceedings*, pages 76–79, 2016.

[44] Renee Michelle Goertzen, Eric Brewe, Laird H Kramer, Leanne Wells, and David Jones. Moving toward change: Institutionalizing reform through implementation of the Learning Assistant model and Open Source Tutorials. *Physical Review Special Topics-Physics Education Research*, 7(2):020105, 2011.

[45] Jacqueline Doyle and Geoff Potvin. In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis. pages 107–110, December 2015. ISSN: 2377-2379.

[46] Julie R Posselt. *Inside graduate admissions*. Harvard University Press, 2016.

[47] Geoff Potvin, Deepa Chari, and Theodore Hodapp. Investigating approaches to diversity in a national survey of physics doctoral degree programs: The graduate admissions landscape. *Physical Review Physics Education Research*, 13(2):020142, December 2017.

[48] Rachel E. Scherr, Monica Plisch, Kara E. Gray, Geoff Potvin, and Theodore Hodapp. Fixed and growth mindsets in physics graduate admissions. *Physical Review Physics Education Research*, 13(2):020133, November 2017.

[49] Geraldine L. Cochran, Theodore Hodapp, and Erika E. Alexander Brown. Identifying barriers to ethnic/racial minority students‘ participation in graduate physics. In *Physics Education Research Conference Proceedings*, PER Conference, pages 92–95, Cincinnati, OH, March 2018.

[50] Deepa Chari and Geoff Potvin. Admissions practices in terminal master's degree-granting physics departments: A comparative analysis. *Physical Review Physics Education Research*, 15(1):010104, January 2019.

[51] Deepa Chari and Geoff Potvin. Understanding the importance of graduate admissions criteria according to prospective graduate students. *Physical Review Physics Education Research*, 15(2), September 2019.

[52] Casey W. Miller, Benjamin M. Zwickl, Julie R. Posselt, Rachel T. Silvestrini, and Theodore Hodapp. Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Science Advances*, 5(1):eaat7550, January 2019.

[53] Lindsay Owens, Benjamin M. Zwickl, Scott V. Franklin, and Casey W. Miller. Identifying qualities of physics graduate students valued by faculty. In *Physics Education Research Conference Proceedings*, 2019.

[54] Julie Posselt, Theresa Hernandez, Geraldine Cochran, and Casey Miller. Metrics First, Diversity Later? Making the Shortlist and Getting Admitted to Physics PhD Programs. *Journal of Women and Minorities in Science and Engineering*, 25(4), 2019.

[55] Lindsay M. Owens, Benjamin M. Zwickl, Scott V. Franklin, and Casey W. Miller. Physics GRE Requirements Create Uneven Playing Field for Graduate Applicants. In *2020 Physics Education Research Conference Proceedings*, pages 382–387, September 2020.

[56] Mike Verostek, Ben Zwickl, and Casey Miller. Do admissions metrics predict PhD completion directly, or indirectly through graduate GPA? *arXiv:2104.08591 [physics]*, April 2021. arXiv: 2104.08591.

[57] Urie Bronfenbrenner. *The ecology of human development: Experiments by nature and design*. Harvard university press, 1979.

[58] Abhilash Nair. *THE RELEVANCE OF PHYSICS*. PhD thesis, Michigan State University, 2018.

[59] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, January 2015.

[60] Cristobal Romero and Sebastian Ventura. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1075.

[61] Bill Cope and Mary Kalantzis. Big Data Comes to School: Implications for Learning, Assessment, and Research. *AERA Open*, 2(2):2332858416641907, April 2016.

[62] Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289–310, August 2010.

[63] Elli J. Theobald, Melissa Aikens, Sarah Eddy, and Hannah Jordt. Beyond linear regression: A reference for analyzing common data types in discipline based education research. *Physical Review Physics Education Research*, 15(2):020110, July 2019.

[64] John M. Aiken, Rachel Henderson, and Marcos D. Caballero. Modeling student pathways in a physics bachelor's degree program. *Physical Review Physics Education Research*, 15(1):010128, May 2019.

[65] Nicholas T. Young, Grant Allen, John M. Aiken, Rachel Henderson, and Marcos D. Caballero. Identifying features predictive of faculty integrating computation into physics courses. *Physical Review Physics Education Research*, 15(1):010114, February 2019.

[66] Cabot Zabriskie, Jie Yang, Seth DeVore, and John Stewart. Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research*, 15(2):020120, August 2019.

[67] Jie Yang, Seth DeVore, Dona Hewagallage, Paul Miller, Qing X. Ryan, and John Stewart. Using machine learning to identify the most at-risk students in physics classes. *Physical Review Physics Education Research*, 16(2):020130, October 2020.

[68] Seth DeVore, Jie Yang, and John Stewart. Extending Machine Learning to Predict Unbalanced Physics Course Outcomes. *arXiv:2002.01964 [physics]*, February 2020.

[69] Nils J. Mikkelsen, Nicholas T. Young, and Marcos D. Caballero. Investigating institutional influence on graduate program admissions by modeling physics Graduate Record Examination cutoff scores. *Physical Review Physics Education Research*, 17(1):010109, February 2021.

[70] Lei Bao. Theoretical comparisons of average normalized gain calculations. *American Journal of Physics*, 74(10):917–922, October 2006.

[71] Jayson Nissen, Robin Donatello, and Ben Van Dusen. Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*, 15(2), July 2019.

[72] Cole Walsh, Martin M. Stein, Ryan Tapping, Emily M. Smith, and N.G. Holmes. Exploring the effects of omitted variable bias in physics education research. *Physical Review Physics Education Research*, 17(1):010119, March 2021.

[73] Computational Mathematics, Science and Engineering.

[74] John M. Aiken, Riccardo De Bin, H. J. Lewandowski, and Marcos D. Caballero. A Framework for Evaluating Statistical Models in Physics Education Research. *arXiv:2106.11038 [physics]*, June 2021.

[75] Nicholas T. Young and Marcos D. Caballero. Using machine learning to understand physics graduate school admissions. In *2019 Physics Education Research Conference Proceedings*, Provo, UT, January 2020. American Association of Physics Teachers.

[76] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, 2015.

[77] Rachel Ivie. Beyond Representation: Data to Improve the Situation of Women and Minorities in Physics and Astronomy, March 2018.

[78] N. Gupta, A. Sawhney, and D. Roth. Will I Get in? Modeling the Graduate Admission Process for American Universities. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 631–638, December 2016.

[79] Austin Waters and Risto Miikkulainen. GRADE: Machine Learning Support for Graduate Admissions. *AI Magazine*, 35(1):64–64, March 2014.

[80] Thomas Lux, Randall Pittman, Maya Shende, and Anil Shende. Applications of Supervised Learning Techniques on Undergraduate Admissions Data. In *Proceedings of the ACM International Conference on Computing Frontiers*, CF '16, pages 412–417, New York, NY, USA, 2016. ACM. event-place: Como, Italy.

[81] Kanadpriya Basu, Treena Basu, Ron Buckmire, and Nishu Lal. Predictive Models of Student College Commitment Decisions Using Machine Learning. *Data*, 4(2):65, June 2019.

[82] James S Moore. An expert system approach to graduate school admission decisions and academic performance prediction. *Omega*, 26(5):659–670, October 1998.

[83] Julie R. Posselt. Disciplinary Logics in Doctoral Admissions: Understanding Patterns of Faculty Evaluation. *The Journal of Higher Education*, 86(6):807–833, November 2015.

[84] Constitution of the State of Michigan - Article I § 26.

[85] Indiana University Center for Postsecondary Research. *The Carnegie Classification of Institutions of Higher Education, 2018 edition*. Bloomington, IN.

[86] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2013. Technical report, American Institute of Physics, August 2014.

[87] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2014. Technical report, American Institute of Physics, September 2015.

[88] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2015. Technical report, American Institute of Physics, September 2016.

[89] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2016. Technical report, American Institute of Physics, September 2017.

[90] Pamela Paxton and Kenneth A. Bollen. Perceived Quality and Methodology in Graduate Department Ratings: Sociology, Political Science, and Economics. *Sociology of Education*, 76(1):71–88, 2003.

[91] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[92] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, January 2007.

[93] Miguel B. Araújo, Richard G. Pearson, Wilfried Thuiller, and Markus Erhard. Validation of species–climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513, September 2005.

[94] Silke Janitza, Carolin Strobl, and Anne-Laure Boulesteix. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14:119, April 2013.

[95] Giles Hooker and Lucas Mentch. Please Stop Permuting Features: An Explanation and Alternatives. *arXiv:1905.03151 [cs, stat]*, May 2019. arXiv: 1905.03151.

[96] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, July 2008.

[97] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, January 2006.

[98] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[99] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, July 2006.

[100] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, November 2003.

[101] Alexander Hapfelmeier, Torsten Hothorn, Kurt Ulm, and Carolin Strobl. A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1):21–34, January 2014.

[102] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

[103] Donald Rubin. Basic Ideas of Multiple Imputation for Nonresponse. *Survey Methodology*, 12(1):37–47, 1986.

[104] Gregory Attiyeh and Richard Attiyeh. Testing for bias in graduate school admissions. *The Journal of Human Resources; Madison*, 32(3):524–548, 1997.

[105] Casey Miller and Keivan Stassun. A test that fails. *Nature*, 510(7504):303–304, June 2014.

[106] Statement on the Use of the GRE in Admissions to Graduate Physics Programs. *American Association of Physics Teacher*, July 2019.

[107] AAS Statement on Limiting the Use of GRE Scores in Graduate Admissions in the Astronomical Sciences, October 2018.

[108] Nicholas T. Young and Marcos D. Caballero. Physics Graduate Record Exam does not help applicants "stand out". *Physical Review Physics Education Research*, 17(1):010144, June 2021.

[109] Raeshanda Wilson. Predicting Graduate School Success: A Critical Race Analysis of the Graduate Record Examination. *Doctor of Education in Secondary Education Dissertations*, May 2020.

[110] Emily M. Levesque, Rachel Bezanson, and Grant R. Tremblay. Physics GRE Scores of Prize Postdoctoral Fellows in Astronomy. *arXiv:1512.03709 [astro-ph, physics:physics]*, December 2015.

[111] Laura A Lopez. Demographic Effects of Removing the Physics GRE Requirement in Graduate Admissions, October 2019.

[112] Katie Langin. A wave of graduate programs drops the GRE application requirement. *Science*, May 2019.

[113] About the GRE Subject Tests (For Test Takers).

[114] GRE Guide to the Use of Scores, 2019.

[115] Nancy D. Morrison, William V. Dixon, Casey W. Miller, and The Women in Astronomy IV Graduate School Admissions White Paper Group. Women in Astronomy IV White Paper: Graduate Admissions in a Post-GRE World. *Bulletin of the AAS*, 51(4), 2020.

[116] Emily M. Levesque, Rachel Bezanson, and Grant R. Tremblay. Why astronomy programs are moving on from the physics GRE. *Physics Today*, March 2017.

[117] Julie R. Posselt. Trust Networks: A New Perspective on Pedigree and the Ambiguities of Admissions. *The Review of Higher Education*, 41(4):497–521, June 2018.

[118] Alexander R. Small. Range restriction, admissions criteria, and correlation studies of standardized tests. *arXiv:1709.02895 [physics]*, September 2017. arXiv: 1709.02895.

[119] Reuben M Baron and David A Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

[120] Judith J. M. Rijnhart, Jos W. R. Twisk, Iris Eekhout, and Martijn W. Heymans. Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. *BMC Medical Research Methodology*, 19(1):19, January 2019.

[121] Andrew F. Hayes and Michael Scharkow. The Relative Trustworthiness of Inferential Tests of the Indirect Effect in Statistical Mediation Analysis: Does Method Really Matter? *Psychological Science*, 24(10):1918–1927, October 2013.

[122] Nathaniel Amos and Andrew F. Heckler. Mediating relationship of differential products in understanding integration in introductory physics. *Physical Review Physics Education Research*, 14(1):010105, January 2018.

[123] Susanne Ditlevsen, Ulla Christensen, John Lynch, Mogens Trab Damsgaard, and Niels Keiding. The Mediation Proportion: A Structural Equation Approach for Estimating the Proportion of Exposure Effect on Outcome Explained by an Intermediate Variable. *Epidemiology*, 16(1):114–120, January 2005.

[124] Laurence S. Freedman. Confidence intervals and statistical power of the 'Validation' ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, 96(1):143–153, June 2001.

[125] Andrew F. Hayes. PROCESS: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling. Technical report, 2012.

[126] Kristopher J. Preacher, Derek D. Rucker, and Andrew F. Hayes. Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivariate Behavioral Research*, 42(1):185–227, June 2007.

[127] M. B. Weissman. Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller et al. *Science Advances*, 6(23):eaax3787, June 2020.

[128] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2017. Technical report, American Institute of Physics, October 2018.

[129] Starr Nicholson and Patrick J. Mulvey. Roster of Physics Departments with Enrollment and Degree Data, 2018. Technical report, American Institute of Physics, October 2019.

[130] National Center for Education Statistics. NCES-Barron's Admissions Competitiveness Index Data Files: 1972, 1982, 1992, 2004, , 2008, 2014, January 2017.

[131] Raj Chetty, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. Mobility Report Cards: The Role of Colleges in Intergenerational Mobility. Technical Report w23618, National Bureau of Economic Research, Cambridge, MA, July 2017.

[132] Adrienne L. Traxler, Ximena C. Cid, Jennifer Blue, and Ramón Barthelemy. Enriching gender in physics education research: A binary past and a complex future. *Physical Review Physics Education Research*, 12(2):020114, August 2016.

[133] Tiffani L. Williams. 'Underrepresented Minority' Considered Harmful, Racist Language, June 2020.

[134] Robert T Teranishi. Race, ethnicity, and higher education policy: The use of critical quantitative research. *New Directions for Institutional Research*, 2007(133):37–49, 2007.

[135] Frank J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, March 1951.

[136] Casey Miller and Julie Posselt. Broadening Participation in Graduate Education through Holistic Review, November 2018. Presented at the Bridge Program and National Mentoring Community Conference.

[137] Julia D Kent and Maureen Terese McCarthy. Holistic review in graduate admissions: A Report from the Council of Graduate Schools. *Washington DC: Council of Graduate Students*, 2016.

[138] Alexander Rudolph, Gibor Basri, Marcel Agüeros, Ed Bertschinger, Kim Coble, Meghan Donahue, Jackie Monkiewicz, Angela Speck, Keivan Stassun, Rachel Ivie, Christine Pfund, and Julie Posselt. Final Report of the 2018 AAS Task Force on Diversity and Inclusion in Astronomy Graduate Education. *Bulletin of the AAS*, 51(1), 2020.

[139] Kris Dunlap. Journey to a more holistic admissions review process by implementing an evaluation rubric, July 2018.

[140] Marenda A. Wilson, Max A. Odem, Taylor Walters, Anthony L. DePass, and Andrew J. Bean. A Model for Holistic Review in Graduate Admissions That Decouples the GRE from Race, Ethnicity, and Gender. *CBE Life Sciences Education*, 18(1), 2019.

[141] John R. Searle. How to Derive "Ought" From "Is". *The Philosophical Review*, 73(1):43–58, 1964.

[142] Patrick J. Mulvey, Starr Nicholson, and Jack Pold. Trends in Physics PhDs. Technical report, February 2021.

[143] Theodore Hodapp and Erika Brown. Making physics more inclusive. *Nature*, 557(7707):629–632, May 2018. Number: 7707 Publisher: Nature Publishing Group.

[144] The AIP National Task Force to Elevate African American Representation in Undergraduate Physics & Astronomy. The Time is Now: Findings from TEAM-UP Report to Increase the Number of African Americans with Bachelor's Degree in Physics and Astronomy. Technical report, American Institute of Physics, 2020. Publisher: APS.

[145] Brian J. Rybarczyk, Leslie Lerea, Dawayne Whittington, and Linda Dykstra. Analysis of Postdoctoral Training Outcomes That Broaden Participation in Science Careers. *CBE—Life Sciences Education*, 15(3):ar33, September 2016.

[146] Özlem Sensoy and Robin DiAngelo. "We Are All for Diversity, but . . .": How Faculty Hiring Committees Reproduce Whiteness and Practical Suggestions for How They Can Change. *Harvard Educational Review*, 87(4):557–580, December 2017.

[147] Arri Eisen and Douglas C. Eaton. A Model for Postdoctoral Education That Promotes Minority and Majority Success in the Biomedical Sciences. *CBE Life Sciences Education*, 16(4), 2017.

[148] Needhi Bhalla. Strategies to improve equity in faculty hiring. *Molecular Biology of the Cell*, 30(22):2744–2749, October 2019.

[149] Michelle I. Cardel, Emily Dhurandhar, Ceren Yarar-Fisher, Monica Foster, Bertha Hidalgo, Leslie A. McClure, Sherry Pagoto, Nathanial Brown, Dori Pekmezi, Noha Sharafeldin, Amanda L. Willig, and Christine Angelini. Turning Chutes into Ladders for Women Faculty: A Review and Roadmap for Equity in Academia. *Journal of Women's Health*, 29(5):721–733, February 2020.

[150] Julie R Posselt. *Equity in Science*. Standford University Press, 2020.

[151] KerryAnn O'Meara, Dawn Culpepper, and Lindsey L. Templeton. Nudging Toward Diversity: Applying Behavioral Design to Faculty Hiring. *Review of Educational Research*, 90(3):311–348, June 2020. Publisher: American Educational Research Association.

[152] Julie R. Posselt. Toward Inclusive Excellence in Graduate Education: Constructing Merit and Diversity in PhD Admissions. *American Journal of Education*, 120(4):481–514, August 2014. Publisher: The University of Chicago Press.

[153] Inclusive Graduate Education Network.

[154] Casey Miller and Julie Posselt. Equitable Admissions in the Time of COVID-19. *Physics*, 13, December 2020.

[155] GRE Subject Tests Fees (For Test Takers).

[156] Kyle M. Whitcomb and Chandralekha Singh. Not all disadvantages are equal: Racial/ethnic minority students have largest disadvantage of all demographic groups in both STEM and non-STEM GPA. *arXiv:2003.04376 [physics]*, March 2020.

[157] Stuart Rojstaczer and Christopher Healy. Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record*, 114(7):1–23, 2012. Place: US Publisher: Teachers College, Columbia University.

[158] Sang Eun Woo, James LeBreton, Melissa Keith, and Louis Tay. Bias, Fairness, and Validity in Graduate Admissions: A Psychometric Perspective. August 2020.

[159] Sandra L. Petersen, Evelyn S. Erenrich, Dovev L. Levine, Jim Vigoreaux, and Krista Gile. Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark. *PLOS ONE*, 13(10):e0206570, October 2018.

[160] Joshua D. Hall, Anna B. O'Connell, and Jeanette G. Cook. Predictors of Student Productivity in Biomedical Graduate School Applications. *PLOS ONE*, 12(1):e0169121, January 2017. Publisher: Public Library of Science.

[161] Linda Sealy, Christina Saunders, Jeffrey Blume, and Roger Chalkley. The GRE over the entire range of scores lacks predictive ability for PhD outcomes in the biomedical sciences. *PLOS ONE*, 14(3):e0201634, March 2019. Publisher: Public Library of Science.

[162] Carrie Hawkins. The Impact of a Holistic Admissions Review Process in a Doctor of Physical Therapy Program. *Graduate Theses, Dissertations, and Capstones*, August 2020.

[163] Annie M. Francis, L. B. Klein, Sharon Holmes Thomas, Kirsten Kainz, and Amy Blank Wilson. Holistic Admissions and Racial/Ethnic Diversity: A Systematic Review and Implications for Social Work Doctoral Education. *Journal of Social Work Education*, 0(0):1–18, April 2021. Publisher: Routledge _eprint: https://doi.org/10.1080/10437797.2021.1895927.

[164] Wendy I. Pacheco, Richard J. Noel, James T. Porter, Caroline B. Appleyard, and Hannah Sevian. Beyond the GRE: Using a Composite Score to Predict the Success of Puerto Rican Students in a Biomedical PhD Program. *CBE—Life Sciences Education*, 14(2):ar13, June 2015.

[165] Blaire Lauren Moody Rideout. *A Study of the Inter-Rater Reliability of University Application Readers in a Holistic Admissions Review Process*. PhD thesis, Bowling Green State University, 2017.

[166] Tim Kautz, James J Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans. Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *National Bureau of Economic Research*, 2014. Publisher:.

[167] Brent W. Roberts. Back to the Future: Personality and Assessment and Personality Development. *Journal of research in personality*, 43(2):137–145, April 2009.

[168] Mathilde Almlund, Angela Lee Duckworth, James Heckman, and Tim Kautz. Personality psychology and economics. In *Handbook of the Economics of Education*, volume 4, pages 1–181. Elsevier, 2011.

[169] W. E. Sedlacek. Noncognitive Measures for Higher Education Admissions. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, pages 845–849. Elsevier, Oxford, January 2010.

[170] Terence J. Tracey and William E. Sedlacek. Noncognitive Variables in Predicting Academic Success by Race. *Measurement and Evaluation in Guidance*, 16(4):171–178, 1984. Publisher: Routledge _eprint: https://doi.org/10.1080/00256307.1984.12022352.

[171] Terence J. Tracey and William E. Sedlacek. Prediction of College Graduation Using Noncognitive Variables by Race. *Measurement and Evaluation in Counseling and Development*, 19(4):177–184, January 1987. Publisher: Routledge _eprint: https://doi.org/10.1080/07481756.1987.12022838.

[172] Niki Medrinos. *BEYOND THE SAT/ACT: AN EXAMINATION OF NON-COGNITIVE FACTORS THAT CONTRIBUTE TO STUDENTS' COLLEGE SUCCESS*. PhD thesis, Temple University, 2014.

[173] Stephen Carp, Kyle Fry, Brittany Gumerman, Kevin Pressley, and Alyssa Whitman. Relationship Between Grit Scale Score and Academic Performance in a Doctor of Physical Therapy Program: A Case Study. *Journal of Allied Health*, 49(1):29–36, February 2020.

[174] Scott K. Stolte, Stephanie B. Scheer, and Evan T. Robinson. The Reliability of Non-Cognitive Admissions Measures in Predicting Non-traditional Doctor of Pharmacy Student Performance Outcomes. *American Journal of Pharmaceutical Education*, 67(1):18, September 2003.

[175] Kristin Zakariasen Victoroff and Richard E. Boyatzis. What Is the Relationship Between Emotional Intelligence and Dental Student Clinical Performance? *Journal of Dental Education*, 77(4):416–426, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.0022-0337.2013.77.4.tb05487.x.

[176] Christopher Peskun, Allan Detsky, and Maureen Shandling. Effectiveness of medical school admissions criteria in predicting residency ranking four years later. *Medical Education*, 41(1):57–64, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2929.2006.02647.x.

[177] Jay Burmeister, Erin McSpadden, Joseph Rakowski, Adrian Nalichowski, Mark Yudelev, and Michael Snyder. Correlation of admissions statistics to graduate student success in medical physics. *Journal of Applied Clinical Medical Physics*, 15(1):375–385, 2014.

[178] Chan Kulatunga Moruzi and Geoffrey R. Norman. Validity of Admissions Measures in Predicting Performance Outcomes: The Contribution of Cognitive and Non-Cognitive Dimensions. *Teaching and Learning in Medicine*, 14(1):34–42, 2002. Publisher: Routledge _eprint: https://doi.org/10.1207/S15328015TLM1401_9.

[179] Melissa C. O'Connor and Sampo V. Paunonen. Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, 43(5):971–990, October 2007.

[180] Patrick Kyllonen, Alyssa M. Walters, and James C. Kaufman. Noncognitive Constructs and Their Assessment in Graduate Education: A Review. *Educational Assessment*, 10(3):153–184, September 2005. Publisher: Routledge _eprint: https://doi.org/10.1207/s15326977ea1003_2.

[181] Robert E. Ployhart and Brian C. Holtz. The Diversity–Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection. *Personnel Psychology*, 61(1):153–172, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.2008.00109.x.

[182] William E. Sedlacek. Why we should use noncognitive variables with graduate and professional students. *The Advisor: The Journal of the National Association of Advisors for the Health Professions*, 24(2):32–39, 2004.

[183] Casey W Miller. Using Non-Cognitive Assessments in Graduate Admissions to Select Better Students and Increase Diversity. page 10, 2015.

[184] Yuanyuan Chen, Shuaizhang Feng, James J. Heckman, and Tim Kautz. Sensitivity of self-reported noncognitive skills to survey administration conditions. *Proceedings of the National Academy of Sciences*, 117(2):931–935, January 2020.

[185] Equity in Graduate Education - Non-Cognitive Assessment, 2021.

[186] Penny Salvatori. Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. *Advances in Health Sciences Education*, 6(2):159–175, May 2001.

[187] Jacqueline M. Zeeman, Jacqueline E. McLaughlin, and Wendy C. Cox. Validity and reliability of an application review process using dedicated reviewers in one stage of a multi-stage admissions model. *Currents in Pharmacy Teaching and Learning*, 9(6):972–979, 2017.

[188] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, October 2012. Publisher: National Academy of Sciences Section: Social Sciences.

[189] Asia A. Eaton, Jessica F. Saunders, Ryan K. Jacobson, and Keon West. How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors' Biased Evaluations of Physics and Biology Post-Doctoral Candidates. *Sex Roles*, 82(3):127–141, February 2020.

[190] Nicolás E. Barceló, Sonya Shadravan, Christine R. Wells, Nichole Goodsmith, Brittany Tarrant, Trevor Shaddox, Yvonne Yang, Eraka Bath, and Katrina DeBonis. Reimagining Merit and Representation: Promoting Equity and Reducing Bias in GME Through Holistic Review. *Academic Psychiatry: The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, 45(1):34–42, February 2021.

[191] David M. Quinn. Experimental Evidence on Teachers' Racial Bias in Student Evaluation: The Role of Grading Scales. *Educational Evaluation and Policy Analysis*, 42(3):375–392, September 2020. Publisher: American Educational Research Association.

[192] Saleem Razack, Torsten Risør, Brian Hodges, and Yvonne Steinert. Beyond the cultural myth of medical meritocracy. *Medical Education*, 54(1):46–53, January 2020.

[193] Keivan G. Stassun, Susan Sturm, Kelly Holley-Bockelmann, Arnold Burger, David J. Ernst, and Donna Webb. The Fisk-Vanderbilt Master's-to-Ph.D. Bridge Program: Recognizing, enlisting, and cultivating unrealized or unrecognized potential in underrepresented minority students. *American Journal of Physics*, 79(4):374–379, March 2011.

[194] Sonia F. Roberts, Elana Pyfrom, Jacob A. Hoffman, Christopher Pai, Erin K. Reagan, and Alysson E. Light. Review of Racially Equitable Admissions Practices in STEM Doctoral Programs. *Education Sciences*, 11(6):270, June 2021. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[195] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

[196] Cassandra M. Guarino and Victor M. H. Borden. Faculty Service Loads and Gender: Are Women Taking Care of the Academic Family? *Research in Higher Education*, 58(6):672–694, September 2017.

[197] Nicholas T. Young and Marcos D. Caballero. The Physics GRE does not help applicants "stand out". *arXiv:2008.10712 [physics]*, August 2020. arXiv: 2008.10712.

[198] Charles D Brown III. Commentary: Disentangling anti-Blackness from physics. *Physics Today*, July 2020.

[199] Katemari Rosa and Felicia Moore Mensah. Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life. *Physical Review Physics Education Research*, 12(2):020113, August 2016.

[200] Paul H. Barber, Tyrone B. Hayes, Tracy L. Johnson, and Leticia Márquez-Magaña. Systemic racism in higher education. *Science*, 369(6510):1440–1441, September 2020.

[201] Danielle Dickens, Maria Jones, and Naomi Hall. Being a Token Black Female Faculty Member in Physics: Exploring Research on Gendered Racism, Identity Shifting as a Coping Strategy, and Inclusivity in Physics. *The Physics Teacher*, 58(5):335–337, May 2020.

[202] Chanda Prescod-Weinstein. Making Black Women Scientists under White Empiricism: The Racialization of Epistemology in Physics. *Signs: Journal of Women in Culture and Society*, 45(2):421–447, January 2020. Publisher: The University of Chicago Press.

[203] Kelly Ochs Rosinger, Karly Sarita Ford, and Junghee Choi. The Role of Selective College Admissions Criteria in Interrupting or Reproducing Racial and Economic Inequities. *The Journal of Higher Education*, pages 1–25, September 2020.

[204] David Kalsbeek, Michele Sandlin, and William Sedlacek. Employing Noncognitive Variables to Improve Admissions, and Increase Student Diversity and Retention. *Strategic Enrollment Management Quarterly*, 1(2):132–150, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sem3.20016.

[205] La'Tonia Stiner-Jones and Wolfgang Windl. 2019 Best Diversity Paper: Work in Progress: Aligning What We Want With What We Seek: Increasing Comprehensive Review in the Graduate Admissions Process. June 2020.

[206] ETS. Curated Approaches.

[207] Quinn Capers, Leon McDougle, and Daniel M. Clinchot. Strategies for Achieving Diversity through Medical School Admissions. *Journal of Health Care for the Poor and Underserved*, 29(1):9–18, 2018.

[208] Jayson M. Nissen, Manher Jariwala, Eleanor W. Close, and Ben Van Dusen. Participation and performance on paper- and computer-based low-stakes assessments. *International Journal of STEM Education*, 5(1):21, December 2018.

[209] Ivan Tomek. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, November 1976.

[210] T. Ryan Hoens and Nitesh V. Chawla. Imbalanced Datasets: From Sampling to Classifiers. In *Imbalanced Learning*, pages 43–59. John Wiley & Sons, Ltd, 2013. Section: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118646106.ch3.

[211] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pages 225–228, May 2016.

[212] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, June 2004.

[213] Siriporn Sawangarreerak and Putthiporn Thanathamathee. Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression. *Information*, 11(11):519, November 2020. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[214] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 358–364, New York, NY, USA, February 2020. Association for Computing Machinery.

[215] Paula Branco, Rita P. Ribeiro, and Luis Torgo. UBL: an R Package for Utility-Based Learning. *CoRR*, abs/1604.08079, 2016.

[216] Chuck Powell. CGPfunctions, December 2020.

[217] Liangyuan Hu, Jung-Yi Joyce Lin, and Jiayi Ji. Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *arXiv:2104.02769 [stat]*, April 2021. arXiv: 2104.02769.

[218] Robert C. MacCallum, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1):19–40, 2002.

[219] Julie R. Irwin and Gary H. McClelland. Negative Consequences of Dichotomizing Continuous Predictor Variables. *Journal of Marketing Research*, 40(3):366–371, August 2003. Publisher: SAGE Publications Inc.

[220] M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, M. K. Eagan, J. M. Esson, J. K. Knight, F. A. Laski, M. Levis-Fitzgerald, C. J. Lee, S. M. Lo, L. M. McDonnell, T. A. McKay, N. Michelotti, A. Musgrove, M. S. Palmer, K. M. Plank, T. M. Rodela, E. R. Sanders, N. G. Schimpf, P. M. Schulte, M. K. Smith, M. Stetzer, B. Van Valkenburgh, E. Vinson, L. K. Weir, P. J. Wendel, L. B. Wheeler, and A. M. Young. Anatomy of STEM teaching in North American universities. *Science*, 359(6383):1468–1470, March 2018. Publisher: American Association for the Advancement of Science Section: Education Forum.

[221] Kelley Commeford, Eric Brewe, and Adrienne Traxler. Characterizing active learning environments in physics using latent profile analysis. *arXiv:2105.02897 [physics]*, May 2021.

[222] Hannah Müggenburg. Beyond the limits of memory? The reliability of retrospective data in travel research. *Transportation Research Part A: Policy and Practice*, 145:302–318, March 2021.

[223] R Behrens and R Del Mistro. Analysing changing personal travel behaviour over time: methodological lessons from the application of retrospective surveys in Cape Town. In *8th international conference on survey methods in transport: harmonisation and data quality, Annecy*, 2008.

[224] Allen L. Edwards. *The social desirability variable in personality assessment and research*. The social desirability variable in personality assessment and research. Dryden Press, Ft Worth, TX, US, 1957. Pages: viii, 108.

[225] Nitesh V Chawla. Data Mining for Imbalanced Datasets: An Overview. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer, Boston, MA, 2009.

[226] Elsa Vazquez Arreola and Jeffrey R. Wilson. Bayesian multiple membership multiple classification logistic regression model on student performance with random effects in university instructors and majors. *PLOS ONE*, 15(1):e0227343, January 2020.

[227] Mei-Shiu Chiu. Gender differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis. *JEDM | Journal of Educational Data Mining*, 12(2):48–77, August 2020.

[228] Katherine P. Dabney and Robert H. Tai. Comparative analysis of female physicists in the physical sciences: Motivation and background variables. *Physical Review Special Topics - Physics Education Research*, 10(1):010104, February 2014.

[229] Charles Henderson, Melissa Dancy, and Magdalena Niewiadomska-Bugaj. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Physical Review Special Topics - Physics Education Research*, 8(2):020104, July 2012.

[230] Rachel Ivie, Susan White, and Raymond Y. Chu. Women's and men's career choices in astronomy and astrophysics. *Physical Review Physics Education Research*, 12(2):020109, August 2016.

[231] Silvija Maslov Kruzicevic, Katarina Josipa Barisic, Adriana Banozic, Carlos David Esteban, Damir Sapunar, and Livia Puljak. Predictors of Attrition and Academic Success of Medical Students: A 30-Year Retrospective Study. *PLOS ONE*, 7(6):e39144, June 2012.

[232] Cathryn A. Manduca, Ellen R. Iverson, Michael Luxenberg, R. Heather Macdonald, David A. McConnell, David W. Mogk, and Barbara J. Tewksbury. Improving undergraduate STEM education: The efficacy of discipline-based professional development. *Science Advances*, 3(2):e1600193, February 2017.

[233] Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero, and Sebastián Ventura. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, 38(3):315–330, April 2013.

[234] Kenneth I. Maton, Tiffany S. Beason, Surbhi Godsay, Mariano R. Sto. Domingo, TaShara C. Bailey, Shuyan Sun, and Freeman A. Hrabowski. Outcomes and Processes in the Meyerhoff Scholars Program: STEM PhD Completion, Sense of Community, Perceived Program Benefit, Science Identity, and Research Self-Efficacy. *CBE—Life Sciences Education*, 15(3):ar48, September 2016.

[235] Sergi Rovira, Eloi Puertas, and Laura Igual. Data-driven system to predict academic grades and dropout. *PLOS ONE*, 12(2):e0171207, February 2017.

[236] Linda J. Sax, Kathleen J. Lehman, Ramón S. Barthelemy, and Gloria Lim. Women in physics: A comparison to science, technology, engineering, and math education over four decades. *Physical Review Physics Education Research*, 12(2):020108, August 2016.

[237] Kelly Spoon, Joshua Beemer, John C. Whitmer, Juanjuan Fan, James P. Frazee, Jeanne Stronach, Andrew J. Bohonak, and Richard A. Levine. Random Forests for Evaluating Pedagogy and Informing Personalized Learning. *JEDM | Journal of Educational Data Mining*, 8(2):20–50, December 2016. Number: 2.

[238] Robert H. Tai, Xiaoqing Kong, Claire E. Mitchell, Katherine P. Dabney, Daniel M. Read, Donna B. Jeffe, Dorothy A. Andriole, and Heather D. Wathington. Examining Summer Laboratory Research Apprenticeships for High School Students as a Factor in Entry to MD/PhD Programs at Matriculation. *CBE—Life Sciences Education*, 16(2):ar37, June 2017.

[239] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4, Part 1):1432–1462, March 2014.

[240] Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, 9:137–163, 2001.

[241] David Firth. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, 80(1):27–38, 1993. Publisher: [Oxford University Press, Biometrika Trust].

[242] Sander Greenland and Mohammad Ali Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34(23):3133–3143, 2015.

[243] Kristin K. Nicodemus. Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, 12(4):369–373, July 2011.

[244] Anne-Laure Boulesteix, Andreas Bender, Justo Lorenzo Bermejo, and Carolin Strobl. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics*, 13(3):292–304, May 2012.

[245] Stefano Nembrini, Inke R. König, and Marvin N. Wright. The revival of the Gini importance? *Bioinformatics*, 34(21):3711–3718, November 2018.

[246] Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. A Plea for Neutral Comparison Studies in Computational Sciences. *PLOS ONE*, 8(4):e61562, April 2013.

[247] Cristobal Romero and Sebastian Ventura. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355, 2020.

[248] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2nd edition edition, 2009.

[249] C. Ensoy, T. W. Rakhmawati, C. Faes, and M. Aerts. Separation Issues and Possible Solutions: Part I – Systematic Literature Review on Logistic Models - Part II – Comparison of different methods for separation in logistic regression. *EFSA Supporting Publications*, 12(9):869E, 2015.

[250] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, September 1946.

[251] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

[252] Fu Chen and Ying Cui. LogCF: Deep Collaborative Filtering with Process Data for Enhanced Learning Outcome Modeling. *Journal of Educational Data Mining*, 12(4):66–99, December 2020.

[253] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[254] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, September 2011.

[255] Benjamin Hofner, Luigi Boccuto, and Markus Göker. Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics*, 16(1):144, May 2015.

[256] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, September 2010.

[257] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, June 2016.

[258] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2):413–468, April 2014.

[259] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, September 2006.

[260] Jake Olivier and Melanie L. Bell. Effect Sizes for 2×2 Contingency Tables. *PLoS ONE*, 8(3):e58777, March 2013.

[261] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[262] C. C. Heyde. Central Limit Theorem. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat04559.

[263] M. Shafiqur Rahman and Mahbuba Sultana. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC medical research methodology*, 17(1):33, 2017.

[264] Ioannis Kosmidis. *brglm: Bias Reduction in Binary-Response Generalized Linear Models*. 2020.

[265] Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometirka*, 2020.

[266] Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Confidence Interval or P-Value? *Deutsches Ärzteblatt International*, 106(19):335–339, May 2009.

[267] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, February 2010. Number: 1.

[268] Max Kuhn. *caret: Classification and Regression Training*. 2020.

[269] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, January 1982.

[270] Bethany R Wilcox and Heather J Lewandowski. Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics. *Physical Review Physics Education Research*, 12(1):010123, 2016.

[271] Bethany R Wilcox and Heather J Lewandowski. Students' views about the nature of experimental physics. *Physical Review Physics Education Research*, 13(2):020110, 2017.

[272] Zhongzhou Chen, Kyle M. Whitcomb, Matthew W. Guthrie, and Chandralekha Singh. Evaluating the effectiveness of two methods to improve students' problem solving performance after studying an online tutorial. In *2019 Physics Education Research Conference Proceedings*, pages 99–104, December 2019.

[273] Andri Signorell. *DescTools: Tools for Descriptive Statistics*. 2020.

[274] Daniel McFadden. Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. Technical Report 474, Cowles Foundation for Research in Economics, Yale University, 1977.

[275] Scott Menard. Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, 54(1):17–24, February 2000.

[276] Tyler Hunt. *ModelMetrics: Rapid Calculation of Model Metrics*. 2020.

[277] Szilard Nemes, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*, 9:56, July 2009.

[278] Maarten van Smeden, Joris A. H. de Groot, Karel G. M. Moons, Gary S. Collins, Douglas G. Altman, Marinus J. C. Eijkemans, and Johannes B. Reitsma. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1), December 2016.

[279] Hyungwoo Kim, Taeseok Ko, No-Wook Park, and Woojoo Lee. Comparison of Bias Correction Methods for the Rare Event Logistic Regression. *Korean Journal of Applied Statistics*, 27(2):277–290, April 2014.

[280] Sam Doerken, Marta Avalos, Emmanuel Lagarde, and Martin Schumacher. Penalized logistic regression with low prevalence exposures beyond high dimensional settings. *PLOS ONE*, 14(5):e0217057, May 2019. Publisher: Public Library of Science.

[281] Emmanuel O. Ogundimu. Prediction of default probability by using statistical models for rare events. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1143–1162, 2019.

[282] Menelaos Pavlou, Gareth Ambler, Shaun Seaman, Maria De Iorio, and Rumana Z. Omar. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 35(7):1159–1177, 2016.

[283] Ben Van Calster, Maarten van Smeden, Bavo De Cock, and Ewout W Steyerberg. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, 29(11):3166–3178, November 2020.

[284] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996.

[285] Peter C Austin and Ewout W Steyerberg. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*, 26(2):796–808, April 2017.

[286] Maarten van Smeden, Karel GM Moons, Joris AH de Groot, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8):2455–2474, August 2019.

[287] Delphine S. Courvoisier, Christophe Combescure, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V. Perneger. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 64(9):993–1000, September 2011.

[288] Menelaos Pavlou, Gareth Ambler, Shaun R. Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z. Omar. How to develop a more accurate risk prediction model when there are few events. *BMJ*, 351:h3868, August 2015.

[289] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

[290] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):270, July 2018.

[291] Hana Šinkovec, Georg Heinze, Rok Blagus, and Angelika Geroldinger. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *arXiv:2101.11230 [stat]*, January 2021.

[292] A. Hapfelmeier and K. Ulm. A new variable selection approach using Random Forests. *Computational Statistics & Data Analysis*, 60:50–69, April 2013.

[293] Markus Ojala and Gemma C. Garriga. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11:1833–1863, June 2010.

[294] Xiang Chen, Ching-Ti Liu, Meizhuo Zhang, and Heping Zhang. A forest-based approach to identifying gene and gene–gene interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19199–19203, 2007.

[295] Minghui Wang, Xiang Chen, and Heping Zhang. Maximal conditional chi-square importance in random forests. *Bioinformatics*, 26(6):831–837, March 2010.

[296] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, May 2010.

[297] Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 12(4):885–915, November 2016.

[298] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997. Conference Name: IEEE Transactions on Evolutionary Computation.

[299] Kaitlin Kirasich. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, 1(3):25, 2018.

[300] Andreas Wålinder. *Evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis*. PhD thesis, Linnaeus University: Sweden, 2014.

[301] J. B. Copas. Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

[302] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[303] Andrew Gelman and Yu-Sung Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2020.

[304] Rainer Puhr, Georg Heinze, Mariana Nold, Lara Lusa, and Angelika Geroldinger. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, 36(14):2302–2317, June 2017.

[305] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, December 2008.

[306] Hülya Olmuş, Ezgi Nazman, and Semra Erbaş. Comparison of penalized logistic regression models for rare event case. *Communications in Statistics - Simulation and Computation*, 0(0):1–13, October 2019.

[307] Chao Chen, Andy Liaw, and Leo Breiman. Using Random Forest to Learn Imbalanced Data. Technical Report 66, University of California, Berkley, July 2004.

[308] Bjoern H. Menze, B. Michael Kelm, Daniel N. Splitthoff, Ullrich Koethe, and Fred A. Hamprecht. On Oblique Random Forests. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 453–469, Berlin, Heidelberg, 2011. Springer.

[309] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[310] Wei-Yin Loh and Peigen Zhou. Variable importance scores. *arXiv:2102.07765 [cs]*, February 2021.

[311] Wei-Yin Loh. Improving the precision of classification trees. *The Annals of Applied Statistics*, 3(4):1710–1737, December 2009.

[312] Wei-Yin Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica sinica*, pages 361–386, 2002.

[313] J. Zhang and K. F. Yu. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, 280(19):1690–1691, November 1998.

[314] Louise-Anne McNutt, Chuntao Wu, Xiaonan Xue, and Jean Paul Hafner. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*, 157(10):940–943, May 2003.

[315] I. Karp. Re: "Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes". *American Journal of Epidemiology*, 179(8):1034–1035, April 2014.

[316] Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach. *arXiv:2006.04628 [cs, stat]*, June 2020.

[317] R. Padraic Springuel, Michael C. Wittmann, and John R. Thompson. Reconsidering the encoding of data in physics education research. *Physical Review Physics Education Research*, 15(2), July 2019.

[318] Devyn Shafer, Maggie S. Mahmood, and Tim Stelzer. Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students. *Physical Review Physics Education Research*, 17(1):010113, March 2021.

[319] Sinta Septi Pangastuti, Kartika Fithriasari, Nur Iriawan, and Wahyuni Suryaningtyas. Data Mining Approach for Educational Decision Support. *EKSAKTA: Journal of Sciences and Data Analysis*, 2(1):33–44, February 2021. Number: 1.

[320] Sander Greenland, Mohammad Ali Mansournia, and Douglas G. Altman. Sparse data bias: a problem hiding in plain sight. *British Medical Journal*, 352, April 2016.

[321] Luc Paquette, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, and Ryan Baker. Who's Learning? Using Demographics in EDM Research. *JEDM | Journal of Educational Data Mining*, 12(3):1–30, October 2020. Number: 3.

[322] Alexis V. Knaub, John M. Aiken, and Lin Ding. Two-phase study examining perspectives and use of quantitative methods in physics education research. *Physical Review Physics Education Research*, 15(2):020102, July 2019.

[323] Lior Rokach and Oded Maimon. *Data mining with decision trees: theory and applications*. World scientific, 2014.

[324] Max Bramer. *Principles of data mining*, volume 180. Springer, 2007.

[325] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. Publisher: ACM.

[326] Thomas G Dietterich and others. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000.

[327] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[328] Carolin Strobl, James Malley, and Gerhard Tutz. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological methods*, 14(4):323–348, December 2009.

[329] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[330] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006. Publisher: Elsevier.

[331] Matthew Kirk. *Thoughtful machine learning: A test-driven approach*. " O'Reilly Media, Inc.", 2014.

[332] Kristin K. Nicodemus and James D. Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics (Oxford, England)*, 25(15):1884–1890, August 2009.

[333] D. M. Reif, A. A. Motsinger-Reif, B. A. McKinney, M. T. Rock, J. E. Crowe, and J. H. Moore. Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination. *Genes and Immunity*, 10(2):112–119, March 2009.

[334] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A. Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, 4:63, April 2019.

[335] Kirstie Whitaker, Tom Rhys Marshall, Tim Van Mourik, Paula Andrea Martinez, Davide Poggiali, Hao Ye, and Marius Klug. RainCloudPlots/RainCloudPlots: WellcomeOpenResearch, August 2019.