Editors' Suggestion | Featured in Physics

Analyzing admissions metrics as predictors of graduate GPA and whether graduate GPA mediates Ph.D. completion

Mike Verostek*

Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627, USA and Department of Physics and Astronomy, Rochester Institute of Technology, Rochester, New York 14623, USA

Casey W. Miller and Benjamin Zwickl

School of Physics and Astronomy, Rochester Institute of Technology, Rochester, New York 14623, USA

(Received 19 May 2021; accepted 9 July 2021; published 7 September 2021)

An analysis of 1955 physics graduate students from 19 Ph.D. programs shows that undergraduate grade point average predicts graduate grades and Ph.D. completion more effectively than GRE scores. Students' undergraduate GPA (UGPA) and GRE Physics (GRE-P) scores are small but statistically significant predictors of graduate course grades, while GRE quantitative and GRE verbal scores are not. We also find that males and females score equally well in their graduate coursework despite a statistically significant 18 percentile point gap in median GRE-P scores between genders. A counterfactual mediation analysis demonstrates that among admission metrics tested only UGPA is a significant predictor of overall Ph.D. completion, and that UGPA predicts Ph.D. completion indirectly through graduate grades. Thus UGPA measures traits linked to graduate course grades, which in turn predict graduate completion. Although GRE-P scores are not significantly associated with Ph.D. completion, our results suggest that any predictive effect they may have is also linked indirectly through graduate GPA. Overall our results indicate that among commonly used quantitative admissions metrics, UGPA offers the most insight into two important measures of graduate school success, while posing fewer concerns for equitable admissions practices.

DOI: 10.1103/PhysRevPhysEducRes.17.020115

I. INTRODUCTION

As physics graduate admission committees across the country consider eliminating GRE scores from consideration when evaluating applicants [1,2], it is important to continue examining the GRE's ability to predict success in graduate school in order for programs to make informed policy choices. Although GRE scores are among the numeric metrics that best predict admission into U.S. graduate programs [3,4], there are significant disparities in typical GRE performance between students of different demographic backgrounds [5]. Combined with the fact that physics remains one of the least diverse of all the science, technology, engineering, and mathematics (STEM) fields [6], the prospect that GRE tests limit the ability of certain students to enter graduate school has led researchers to begin questioning the utility of GRE exam scores in the graduate admissions process in comparison to other quantitative metrics such as undergraduate GPA (UGPA) [1,7,8]. Among some of the findings in this body of work are indications that earning high marks on the GRE Physics (GRE-P) test fails to help students "stand out" to admissions committees who would have overlooked them due to an otherwise weak application [8], and that typical physics Ph.D. admissions criteria such as the GRE-P exam fail to predict Ph.D. completion despite limiting access to graduate school for underrepresented groups [1].

Yet overall Ph.D. completion is only one measure of "success" in graduate school. Graduate faculty often cite high grades, graduation in a reasonable amount of time, and finding a job after graduation as indications of successful graduate students [9]. It is therefore crucial for admissions committees to understand how these other measures of success are related to common quantitative admissions metrics as well. In particular, studying the role of graduate grade point average (GGPA) is important for both historical and practical reasons. Among physics graduate students, positive relationships between GRE-P scores, first-year graduate grades, and cumulative graduate grades have traditionally been touted as evidence for the exam's utility in evaluating applicants [10,11]. Several other studies [12–15] suggest that a number of common admissions metrics are correlated with GGPA as well. At a practical

*mveroste@ur.rochester.edu

level, gaining a better understanding of which factors best predict graduate grades is valuable due to the fact that performance in graduate classes can influence whether students will ultimately complete a Ph.D. For instance, programs may institute GPA requirements that prevent students from continuing study if their course grades do not meet certain criteria.

Predictive validity analyses of GRE scores across all STEM disciplines consistently find that scores on the GRE quantitative (GRE-Q) and verbal (GRE-V) tests are more effective predictors of graduate grades than Ph.D. completion [12,13]. For instance, recent studies on Ph.D. admissions in the biomedical field found that students' GRE-Q and GRE-V scores are poor predictors of Ph.D. completion, but are more associated with first-semester and cumulative graduate school grades [14,15]. In contrast, studies cited by the Educational Testing Service (ETS) such as the meta-analysis of GRE predictive validity by Kuncel *et al.* [11] show a positive correlation between GRE subject scores, graduate grades, and Ph.D. completion. Kuncel *et al.* find that GRE subject tests show larger correlations with GGPA than GRE-Q, GRE-V, or UGPA, which they attribute to the subject-specific knowledge that the GRE subject tests are purported to measure. Still, GRE-Q and GRE-V scores, which the authors presume to be broad measures of cognitive ability, are shown to only moderately correlate with GGPA but do not significantly correlate with Ph.D. completion. Kuncel *et al.* also find undergraduate grade point average (UGPA) correlates with GGPA but not completion.

Despite voluminous research on the efficacy of quantitative admissions metrics in predicting graduate success, there remains a dearth of studies specifically examining these metrics in the context of physics graduate education. No current study elucidates the relationships between undergraduate grades, GRE scores, and physics graduate grades. Moreover, studies such as Ref. [1] do not incorporate graduate grades into models of Ph.D. completion despite its theoretical and structural importance on the road to graduate success. This paper aims to fill these gaps in the current literature.

The primary goal of this paper is to extend the analysis of Miller *et al.* [1], using the same dataset to examine the correlations of common quantitative admissions statistics with graduate physics GPA, as well as the role that graduate GPA plays in predicting whether a student completes their Ph.D. program. Whereas Ref. [1] did not utilize information on student graduate course performance, this paper incorporates graduate GPA into several models in order to determine whether commonly used admissions metrics predict Ph.D. completion of U.S. students directly, or indirectly via graduate GPA; a discussion of the theoretical motivation for why graduate grades may mediate the relationship between admissions metrics and Ph.D. completion is offered in Sec. II. Hence, while the analysis presented in Miller *et al.* [1] was primarily focused on simply identifying the measures that best correlated with Ph.D. completion, this analysis explores questions regarding both how and why those correlations occurred.

Exploring whether graduate GPA mediates the relationship between common admissions metrics and Ph.D. completion affords us the opportunity to employ statistical methods from the literature on causal inference [16–23]. In doing so we lay out methods of calculating the direct and indirect effects of common admissions metrics on Ph.D. completion, as well as the assumptions needed for those effects to have a causal interpretation. This approach allows us to gain useful information from the present analysis, while careful examination of the assumptions required for causal interpretation will help guide future studies.

Use of statistical methods developed in the causal inference literature allows us to build on the findings in Ref. [24] by incorporating the ranking of a student's Ph.D. program along a mediating pathway to completion rather than as a covariate in regression analysis. We also present models with various combinations of GRE-P and GRE-Q scores to show that variance inflation due to collinearity is minimal, and is therefore not a concern. These analyses are included in the Supplemental Material [25].

We seek to answer two primary research questions in this paper:

1. How do commonly used admissions metrics and demographic factors relate to physics graduate GPA?
2. What role does graduate GPA play in predicting Ph.D. completion, and do quantitative admissions metrics predict Ph.D. completion indirectly through graduate GPA?

To answer these questions, we begin by exploring the relationships between variables using bivariate correlations. We then examine the unique predictive effects of different admission metrics on graduate GPA using a multiple linear regression model. These results lay the groundwork for a mediation analysis, which is used to examine the role that graduate GPA plays in Ph.D. completion by breaking down effects into direct and indirect components. All of the primary analyses are performed using data on U.S. physics graduate students, with a review of equivalent analyses for international students included in the Supplemental Material [25].

II. BACKGROUND AND MOTIVATION

Before outlining the quantitative methods employed in this analysis, we briefly describe the student performance metrics used in this study and the broad individual student characteristics they help to measure. We discuss the underlying constructs hypothesized by the GRE quantitative, verbal, and subject tests, as well as undergraduate and graduate grades, and several external factors that influence

these scores. The GRE analytical writing test is not included since it is not used enough in physics graduate admissions to warrant investigation. This section serves as a theoretical motivation for the models of Ph.D. completion analyzed in this study.

The GRE is a series of standardized tests designed to help admissions committees predict future academic success of students coming from different backgrounds [26,27]. While the GRE-Q assesses basic concepts of arithmetic, algebra, geometry, and data analysis, the GRE-V assesses reading comprehension skills and verbal and analytical reasoning skills. These tests are specifically constructed to measure "basic developed abilities relevant to performance in graduate studies" [28]. In their meta-analysis of GRE predictive validity, Kuncel *et al.* frame the GRE-Q and GRE-V as most related to declarative and procedural knowledge and suggest that they are best described as measures of general cognitive ability [11]. In contrast, the GRE subject tests "assess acquired knowledge specific to a field of study" [26], indicating that the GRE subject tests are ostensibly a direct measure of a student's knowledge of a particular area of study. Indeed, admissions committees often interpret high GRE subject scores as strong evidence of a student's discipline-specific knowledge [9]. Other research suggests that higher scores on standardized subject tests could also reflect greater student interest in that subject area [29].

The individual characteristics measured by a student's undergraduate grades include both academic knowledge and a collection of noncognitive factors [30]. Much research exists on the meaning and value of grades, particularly at the K–12 level, and a review [31] of the past century of grading research finds that grades assess a multidimensional construct comprising academic knowledge, engagement (including motivation and interest), and persistence. Consistently over the past 100 years only about 25% of variance in grades is attributable to academic knowledge as measured by standardized tests [32], with recent research suggesting that much of the unexplained variance is represented by a students ability to negotiate the "social processes" of school [33]. We therefore regard UGPA as broadly measuring student academic achievement across a wide range of subjects in addition to several aspects of noncognitive traits such as motivation, interest, and work habits. However, we also recognize the limitations inherent in compressing students' college academic performance into a single number, including the loss of information pertaining to student growth over time and time to degree completion.

We conceptualize graduate GPA similarly, treating it as a measure of subject-specific academic knowledge as well as other nonacademic characteristics. In addition to the broad research on grades described above, research specifically addressing the factors leading to graduate success supports this interpretation of GGPA. Interviews conducted with over 100 graduate school faculty reveal that graduate success, which they define as a student's ability to earn high graduate grades and eventually complete their degree in a timely manner, is largely dependent on noncognitive characteristics [9]. Interviewees deemed motivation, work ethic, maturity, and organizational skills as crucial to student success in graduate school. In a separate review of noncognitive predictors pertaining to graduate success, graduate GPA is specifically linked to a variety of personality (e.g., extroversion and conscientiousness) and attitudinal factors (e.g., motivation, self-efficacy, and interests) [34]. Indeed, the authors of the review characterize graduate grades as a complex composite of many of the cognitive and noncognitive factors related to graduate school success.

These conceptions of grades and GRE scores compel us to hypothesize that graduate GPA mediates the relationship between common quantitative admissions metrics and Ph.D. completion. As a construct measuring subject-specific knowledge and several noncognitive characteristics, we expect UGPA and GRE-P to most strongly link to GGPA. Despite the drawbacks of cumulative UGPA (such as grade inflation and masking of individual growth), we expect UGPA to be associated with graduate course performance since it captures aspects of both academic and some nonacademic characteristics. We also expect GRE-P scores to be related to graduate grades due to their requirement of specific physics knowledge. Finally, while we expect GRE-Q may have a small predictive effect on GGPA as a general cognitive measure, we do not necessarily expect a similar relationship for GRE-V as its content is generally disparate from physics curricula.

On both a theoretical and structural level we expect graduate grades to predict Ph.D. completion. Graduate GPA may offer insight into a student's mastery of advanced physics concepts as well as their personality and attitudes. All of these contribute to successful physics Ph.D. completion, but likely vary in importance depending on choice of research area [34]. Structurally, a satisfactory performance in graduate courses is implicit on the path to completing a Ph.D. For example, GGPA requirements can act as thresholds for being allowed to continue studying in a Ph.D. program. Poor course performance may negatively influence personal factors (e.g., self-efficacy, identity), limit access to research opportunities (e.g., repeating classes, ease of finding a research lab), or may indicate a lack of preparation for research, all of which could hinder Ph.D. completion. It is also more temporally proximal to Ph.D. completion than the other metrics included in the study.

Lastly, we note that although this discussion has focused on students' individual traits that may predict success in graduate school, there are undoubtedly a number of external factors that can influence student attrition. Socioeconomic factors, mental health, family responsibilities, work duties, external job prospects, and departmental

culture are all variables that would play a role in a comprehensive model of graduate school persistence [9,35–39]. These uncollected pieces of information may act as "confounding" variables that can bias results, and we discuss their influence on the present study in Sec. V B.

III. METHOD

A. Data

Student level data for both this study and [1] were requested from physics departments that awarded more than 10 Ph.D.'s per year for students who matriculated between 2000 and 2010, including information on undergraduate GPA (UGPA), GRE-Q, GRE-V, GRE-P, and graduate GPA (GGPA). Data collected also included the final disposition of students (Ph.D. earned or not), start and finish years, and demographic information. GPA data are analyzed on a 4.0 scale while GRE scores are on the percentile scale.

We received data from 27 programs (approximately a 42% response rate), which spanned a broad range of National Research Council (NRC) rankings. The sample used in Ref. [1] consisted of all students in 21 programs for which start year was available. Given that the median time to degree across physics Ph.D. programs is 6 years, some students who started before 2010 were still active at the time of data collection in 2016. The probability of not completing the physics Ph.D. has an exponential time dependence with a time constant of 1.8 yr. Thus, students who have been in their programs for three time constants have only a 5% chance of not completing. These students were thus categorized as completers in this study.

These data covered 3962 students (see Table 1). Of this subset, two programs did not report GGPA data for their students. Hence, the sample for this study excludes these students, thereby reducing the sample size to 3406 students across 19 programs. This corresponds to approximately 11% of matriculants to all U.S. physics Ph.D. programs during the years studied.

Among the sample of U.S. students, 16% are women ($N = 317$). Although the authors generally advocate for a nuanced treatment of gender in physics education research and recognize the deficits associated with treating gender as a fixed binary variable [40], the present dataset spans the years 2000 to 2010 during which the data collected by programs only allowed for the binary option of male or female. Hence, we must treat gender as a dichotomous variable in this analysis.

TABLE I. Demographic breakdown of the data used in this analysis. To focus on issues of diversity and inclusion most strongly associated with U.S. applicants to physics graduate programs, we analyze only the data from U.S. graduate students.

	U.S.	Non-U.S.	Total
Male	1638	1164	2802
Female	317	287	604
Total	1955	1451	3406

The racial composition of the dataset is 61.6% White, 1.3% Black, 2.1% Hispanic, 0.2% Native American, 3.5% Asian, 1.0% multiple or other races, and 30.2% undisclosed. Excluding the cases for which race was unavailable, the sample is thus roughly representative of annual Ph.D. production in U.S. physics for gender, race, and citizenship [41]. We include race as a covariate in each analysis presented; however, small N, particularly for Black, Hispanic, Native American, and Asian students, often precludes useful interpretation of the results pertaining to these subsets.

In order to focus on issues of diversity and inclusion associated most strongly with U.S. applicants, we use only the subset of data from domestic graduate students. This decision is further motivated by research suggesting that it is difficult for admission committees to directly compare scores earned by U.S. and international students, indicating that separate analyses are appropriate [9]. Using the subset of students who are from the U.S. reduces the total sample size for the study to $N = 1955$. A cursory visualization of the variables in the data set, as shown in Fig. 1, shows that the distributions of scores for U.S. and Non-U.S. students are markedly different, which further justifies separate analyses of these two student populations.

Examining the distributions of scores in Fig. 1, the presence of non-normality is evident in nearly all of the variables. Each of the continuous variables in the dataset fail the Shapiro-Wilk test of normality at the α level of 0.05. However, these tests are often of limited usefulness; in general distributions with skewness $|\hat{\gamma}_1| > 3$ or kurtosis $|\hat{\gamma}_2| > 10$ likely indicate that they violate any assumption of normality [45]. For this dataset, the GGPA distribution skewness $\hat{\gamma}_1 = -3.31$ and kurtosis $\hat{\gamma}_2 = 20.43$, indicating severe non-normality. The GRE-Q distribution also falls into the problematic range ($\hat{\gamma}_1 = -2.11$ and $\hat{\gamma}_2 = -9.33$). Ceiling effects are also present, since many students earned 4.0 grade point averages or earned the maximum score on the GRE examinations.

The data collection process was limited to gathering only cumulative graduate GPA rather than first-year graduate grades, which were not recorded by some programs. Thus, depending on whether a student persisted in a program, their graduate GPA may be based on many courses while others are based on only a few courses. The data set is also necessarily subject to range restriction, since data on student performance in graduate school is automatically limited to include only students who were accepted to undertake graduate study. We cannot know how students who were not accepted into graduate school would have performed had they been accepted. Range restriction may act to attenuate the strength of observed effects in subsequent analyses [46].

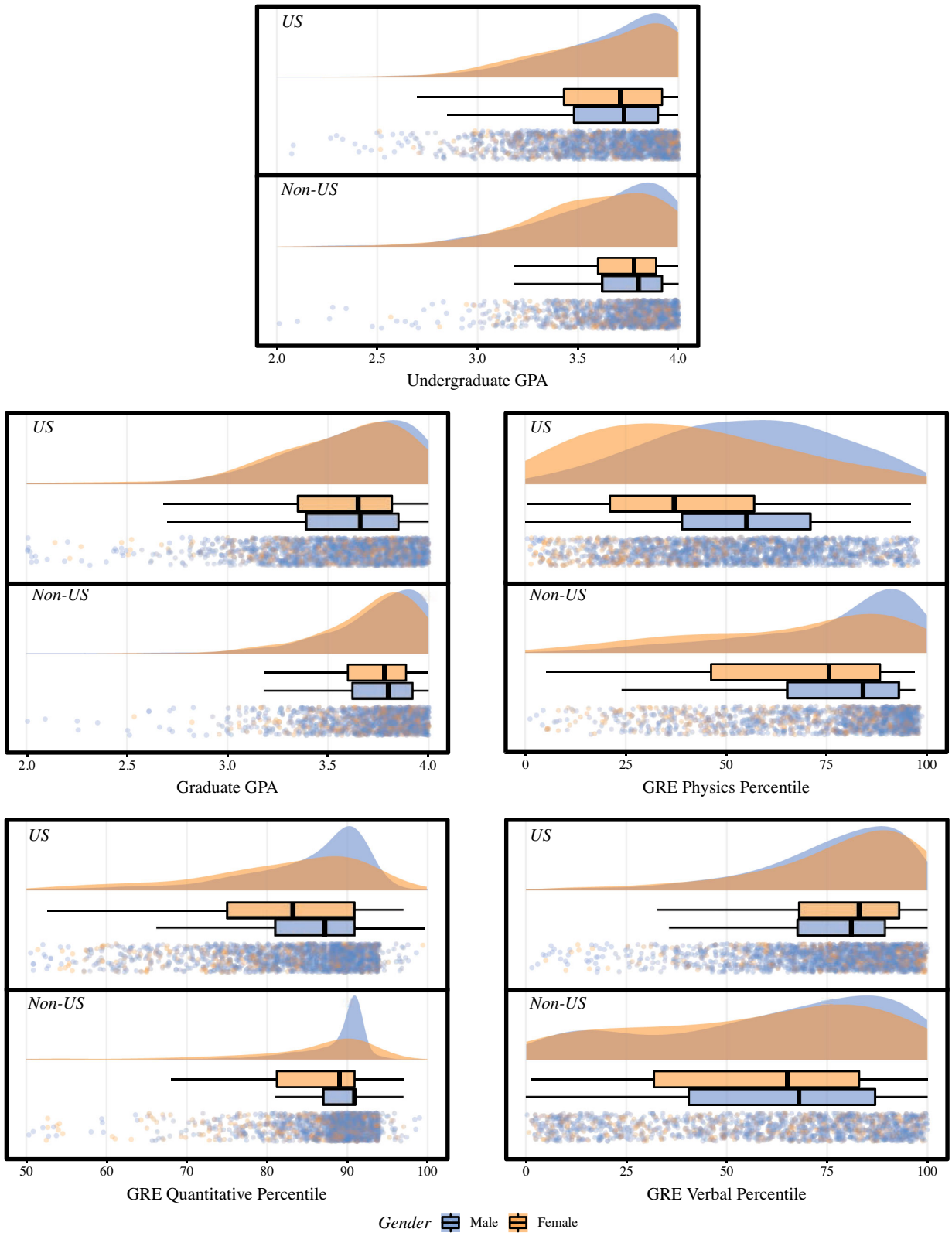

FIG. 1. Distributions of the quantitative metrics included in the data. "Raincloud plots" show density plots, boxplots, and scatterplots of the data. We see that despite significant score gaps between U.S. male and female GRE-P test takers, no such gap exists in subsequent GGPA performance. UGPA distributions for male and female applicants are also similar. Code for generating raincloud plots courtesy of Ref. [42]. All figures generated with the R package `ggplot2` [43]. Figure themes adapted from Ref. [44].

Although not used in a majority of this study, we briefly explore the role of the doctoral programs NRC ranking in Ph.D. completion [47]. Since the NRC only gives confidence intervals for program rank, we created a ranking for this study by averaging the 5% and 95% confidence bounds for the NRC regression-based ranking (NRC-R) and rounded this up to the nearest five to protect the confidentiality of participating programs. This led to a ranking range of 5 to 105. We divided the programs into terciles of approximately equal number of records, and categorized as tier 1 (highest ranked, NRC-R ≤ 20), tier 2 ($25 \leq$ NRC-R ≤ 55), and tier 3 (NRC-R > 55).

Multiple imputation (predictive mean matching) is used to impute missing UGPA and GRE-P scores. Predictive mean matching is used due to the non-normality of the data. 160 students do not have data for either UGPA or GRE-P, while 400 are missing only UGPA and 263 are missing only GRE-P. All multiple imputation is conducted using the `mice` package in R [48]. 20 imputed datasets are used for each analysis. For consistency, incomplete variables are imputed using the same imputation model used in Ref. [1], in which the imputation model utilizes all other variables in the dataset aside from graduate GPA and Ph.D. completion. The model utilizes GRE-Q, GRE-V, program tier, gender, and race, as well as complete cases of UGPA and GRE-P. Although the imputation approach presented here is theoretically sound, we also present a comparison of several different models of data imputation in the Supplemental Material [25].

B. Methods to explore the role of graduate grades

The goal of this section is twofold. First, we seek to gain a cursory look at how graduate GPA is related to common admissions metrics. In doing this, we also wish to determine whether it is reasonable that admissions metrics could indirectly predict completion through graduate GPA. This section presents a series of analyses meant to elucidate the relationships between standard admissions metrics, students' GGPA, and students' final disposition. To make our analysis maximally accessible to readers of different statistical backgrounds, we describe here in detail the methods used in this section.

Bivariate correlation coefficients provide information about the level of association between two variables, and are therefore a useful starting point for analysis. We construct a correlation matrix (see Table II) for all variables in the sample using Pearson correlation coefficients, which are equivalent to the standardized slope coefficients for a linear model predicting y from x. These are given by

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{1}$$

for any two continuous variables x and y. Calculating r_{xy} gives us a first glance at the relationships between the continuous variables UGPA, GGPA, GRE-P, GRE-Q, and GRE-V.

When the x variable is treated as dichotomous (e.g., gender in this dataset), Eq. (1) reduces to the point-biserial correlation coefficient r_{pb},

$$r_{pb} = \frac{\bar{y_1} - \bar{y_0}}{\sigma_y}\sqrt{pq}, \tag{2}$$

where $\bar{y_1}$ and $\bar{y_0}$ are the means of the continuous y variable for the two x groups 1 and 0, q and p are the proportions of data belonging to these two groups, and σ_y is the standard deviation for the y variable. Like the Pearson coefficient, the quantity r_{pb} ranges from -1 to 1 and indicates the strength of association between two variables. Conveniently, a significance test for the point-biserial correlation is identical to performing an independent t test on the data [49]. Thus, the point-biserial correlation coefficient yields information about whether two group means are statistically different. For instance, the point-biserial correlation tests whether the GGPA of male students are statistically different from those of female students (we find that GGPAs are not significantly different by gender, see Table II).

When x and y are both dichotomous, the Pearson coefficient reduces to the phi coefficient,

TABLE II. A matrix showing bivariate correlations between continuous and dichotomous variables used in subsequent analyses. Correlations are shown in the lower diagonal while confidence intervals for those correlations are shown in the upper diagonal. For example, the correlation between GGPA and GRE-P is 0.22 and a 95% CI of (0.18, 0.27), indicating a weak correlation. Means and standard deviations are also presented in the first column. GPAs are on a 4.0 scale while GRE scores are in terms of percentiles. Correlations are calculated for U.S. students only.

Measure (M + SD)	UGPA	GRE-Q	GRE-V	GRE-P	GGPA	Final Disp.	Gender
UGPA (3.6 ± 0.3)	⋯	(0.25, 0.35)	(0.12, 0.22)	(0.26, 0.37)	(0.24, 0.33)	(0.10, 0.20)	(−0.11, 0.01)
GRE-Q (83.3 ± 10.4)	0.30	⋯	(0.30, 0.37)	(0.47, 0.54)	(0.13, 0.22)	(0.10, 0.18)	(−0.16, −0.07)
GRE-V (76.3 ± 18.7)	0.17	0.33	⋯	(0.23, 0.32)	(0.06, 0.15)	(0.02, 0.10)	(−0.02, 0.07)
GRE-P (52.9 ± 23.2)	0.31	0.51	0.28	⋯	(0.18, 0.27)	(0.10, 0.19)	(−0.33, −0.24)
GGPA (3.5 ± 0.5)	0.29	0.18	0.10	0.22	⋯	(0.39, 0.46)	(−0.06, 0.03)
Final Disp.	0.15	0.15	0.06	0.14	0.43	⋯	(−0.09, −0.01)
Gender	−0.05	−0.11	0.03	−0.30	−0.01	−0.05	⋯

$$\phi = \sqrt{\frac{\chi^2}{n}}, \tag{3}$$

where χ^2 is the chi-squared statistic for a 2×2 contingency table and n is the total number of observations in the data. The phi coefficient also ranges from -1 to 1 and indicates the strength of association between two binary variables. This quantity allows us to examine whether final disposition is significantly associated with gender. We find that the association between gender and final disposition just meets the threshold for statistical significance ($\phi = -0.05 \pm 0.04$, $p = 0.04$), likely due to the large sample size of our data, but the very small phi coefficient indicates that the practical strength of this relationship is negligible [50].

To characterize how GGPA and other numerical predictors vary across program tier, we conduct several one-way analysis of variance (ANOVA) tests using program tier as the independent variable. ANOVA tests allow us to determine whether there are significant differences between different groups, such as students in different program tiers. These tests produce an F statistic, which is interpreted as the ratio of between-group variability to within-group variability. Thus, higher values of F indicate that between-group variability is large compared to within-group variability, which is unlikely if the group means all have a similar value.

Lastly we present the results of a multiple regression analysis in which we regress GGPA on common admissions metrics and demographic factors. Regression allows us to examine the unique predictive effects of these predictors.

The classical linear regression model is written mathematically for an outcome variable Y as

$$Y_i = \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_k X_{ik} + \epsilon_i, \tag{4}$$

where $i = 1, \ldots, n$, the number of observations in the data, and k represents the number of predictors in the model. Error terms ϵ_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ. $\hat{\alpha}$ is the vector of regression coefficients that minimizes the sum of squared errors

$$\Sigma_{i=1}^{n} = (Y_i - \hat{\alpha} X_i)^2 \tag{5}$$

for the given data. The regression coefficients can be interpreted as the difference in the outcome variable Y, on average, when comparing two groups of units that differ by 1 in one predictor X while keeping all the other predictors the same.

We report both unstandardized and standardized versions of the regression coefficients. Unstandardized coefficients are the result of regression analysis using the original, unscaled variables. Thus, the unstandardized regression coefficients represent the predicted average change in the outcome variable Y when the corresponding predictor X is changed by one unit. This allows for a straightforward interpretation since the variables are not scaled, but does not yield insight into the relative predictive strengths of the independent variables since they are scaled differently. Standardized regression coefficients result from regression analyses using continuous variables that have been mean-centered and divided by their standard deviation, resulting in variables with variances equal to 1. Thus standardized regression coefficients represent the average number of standard deviations changed in the outcome variable when a predictor variable is increased by 1 standard deviation. By calculating the standardized coefficients, we exchange a simple interpretation of score change for an interpretation of which variables have the greatest effect on the dependent variable.

C. Mediation analysis methods

Using mediation analysis we seek to answer the question of whether graduate GPA mediates the predictive ability of common admissions metrics on Ph.D. completion. Whereas analyses such as logistic regression [1] yield information about whether independent variables such as UGPA and GRE-P affect final disposition of a graduate student, they do not offer insight into the explanation of why and how UGPA, GRE-P, and other admissions metrics affect completion. Mediation analysis is one technique that allows us to probe the underlying process by which some variables influence others [19–23].

Figure 2 graphically depicts a prototypical mediation model, where X, Y, and M represent the model's independent, dependent, and mediating variables. C represents a covariate. As a hypothetical example, let's say that previous research has shown a positive relationship between the use of active learning activities in physics class and student

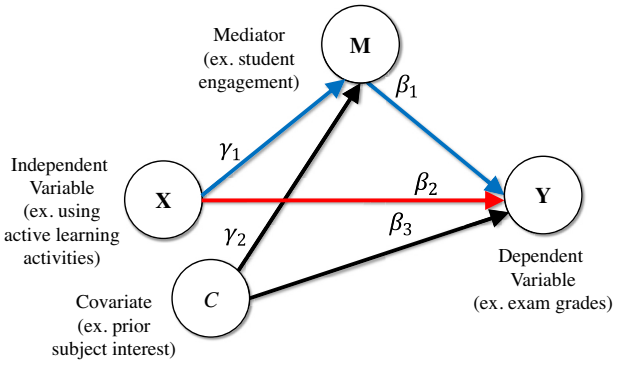

FIG. 2. A qualitative graphical depiction of a mediated relationship between two variables. X, Y, and M represent the model's independent, dependent, and mediating variables while C represents a covariate. A researcher who observes a positive relationship between using active learning activities in the classroom and the exam grades of students might posit that a third variable, student engagement, is actually responsible for causing the observed relationship.

exam grades. Researchers might hypothesize that this relationship is actually due to a third mediating variable, student engagement. Using active learning activities in class may cause students to become more engaged with the material, making their subsequent exam grades increase. Engagement is a mediating variable in this case. Meanwhile, since students who are already interested in physics could be predisposed to being more engaged and performing better on exams, the researcher might take students' prior physics interest into account by including it as a covariate in their analyses.

In this section, we wish to discern whether graduate GPA mediates the relationship between common admissions metrics and students' likelihood of completing graduate school. In practice, mediation analysis is done by simultaneously estimating a set of regression equations [51]. The goal is to partition the total effect of the independent variable X on the dependent variable Y into two parts: the direct effect of X on Y and the indirect effect of X on Y through the mediating variable M. For the simple example given above, the set of regression equations to be solved are

$$Y_i = \beta_0 + \beta_1 M_i + \beta_2 X_i + \beta_3 C_i + \epsilon_{yi,} \tag{6}$$

$$M_i = \gamma_0 + \gamma_1 X_i + \gamma_2 C_i + \epsilon_{mi}. \tag{7}$$

Traditional mediation literature [52,53] defines the direct effect of X on Y as the coefficient β_2 and the indirect effect of X on Y as the product of coefficients $\gamma_1\beta_1$, corresponding to the products of the path coefficients along the mediated path shown in Fig. 2. Statistically significant values of $\gamma_1\beta_1$ indicate that the relationship between X and Y is mediated by M. This method demonstrates the general intuitive ideas underlying mediation analysis, but is subject to several important limitations. Foremost among the limitations associated with traditional mediation analysis is that its applicability to model categorical variables (e.g., binary outcomes) and nonlinearities is not well defined, as these situations preclude the use of sums and products of coefficients [19,54,55]. The difficulties associated with binary outcomes are therefore problematic for a model predicting final disposition, a binary outcome. Furthermore, traditional mediation models leave the causal interpretation of their results ambiguous [56].

Recent work in the field of causal inference [16–18] has formalized and generalized mediation analysis to resolve these limitations, allowing for categorical outcomes while also clarifying that under certain conditions the results may be interpreted causally. In this framework, often called the "potential outcomes" framework, the traditional product-of-coefficients mediation analysis is a special case for which the mediator and outcome variables are both continuous, while the functional forms of the direct and indirect effects for other situations become more complicated [51].

For the primary analysis of this paper, we calculate the direct and indirect effects defined by the potential outcomes framework for the case of a continuous independent variable (UGPA, GRE-P, and GRE-Q), a continuous mediator (GGPA), and a dichotomous outcome (final disposition). The simultaneous regression equations to be calculated are still Eqs. (6) and (7), except the binary Y is replaced with Y^*, a continuous unobserved latent variable which represents the observed binary variable. Once estimated, the direct and indirect effects reduce to simple differences in probability of completing a Ph.D. between students across different values of the independent variables. Mathematically, the effects for a change in the independent variable from a value x_0 to x_1 at a particular value of the control c are given by

$$\mathrm{IE} = \Phi[\mathrm{probit}(x_0, x_1)] - \Phi[\mathrm{probit}(x_0, x_0)], \tag{8}$$

$$\mathrm{DE} = \Phi[\mathrm{probit}(x_1, x_1)] - \Phi[\mathrm{probit}(x_0, x_1)], \tag{9}$$

where Φ represents the normal cumulative distribution function and $\mathrm{probit}(x_a, x_b)$ is given by

$$\mathrm{probit}(x_a, x_b) = [\beta_0 + \beta_2 x_a + \beta_3 c + \beta_1(\gamma_0 + \gamma_1 x_b + \gamma_2 c)]/\sqrt{v(x_a)}, \tag{10}$$

and $v(x_a)$ is

$$v(x_a) = \beta_1^2 \sigma_m^2 + 1. \tag{11}$$

Note that these expressions are all still simply combinations of the coefficients from the regression equations (6) and (7). Thus, the potential outcomes framework allows us to calculate the total predicted change in probability of completing a Ph.D. due to an independent variable and decompose it into that variable's indirect effect on final disposition through GGPA as well as its direct effect (see Fig. 4).

Using this mediation framework can help to give powerful insights into nuanced relationships between the variables in an observational study [57]. However, giving a truly causal interpretation to the results of this mediation analysis requires a set of strong assumptions to be met, and in practice it can be difficult for any observational study fully meet these conditions [58]. Hence, we try to avoid making explicitly causal claims in our discussion of the results. We discuss the assumptions needed for a causal interpretation as well as the robustness of the current study to violations of those assumptions in Sec. V B.

Mediation analyses were conducted using the `mediation` package in R [59]. Checks for the consistency of results across different computational approaches were done by performing duplicate mediation analyses in the R package `medflex` [60] as well as the statistical software Mplus [61].

IV. RESULTS

A. Results of exploring the role of graduate grades

Correlations.—An initial question related to predicting a student's final disposition is whether UGPA, GRE scores, and GGPA are reliably correlated with a student's final outcome. We are also interested in the strength of association between GGPA, UGPA, and GRE scores, as this information yields insight into whether GGPA could serve as a mediating variable in predicting final disposition. Table II contains the bivariate correlations (Pearson, point-biserial, and phi) between each pair of measures for the sample in the lower diagonal. The 95% confidence intervals are reported in the upper diagonal. Confidence intervals that do not include a value of zero indicate that the correlation is statistically significant. The means and standard deviations of the continuous variables are presented in the table's first column (GPA data is analyzed on a 4.0 scale while GRE scores are on the percentile scale).

Inspection of Table II reveals that GGPA is the predictor most strongly correlated with final disposition ($r_{pb} = 0.43$). This value is statistically significant ($p < .001$) and positive, meaning that students with higher GGPA are more likely to finish their Ph.D. program successfully. This trend is visually apparent in Fig. 3(a), which shows boxplots of GGPA grouped by Ph.D. completers and non-completers. We also observe that UGPA ($r_{xy} = 0.29,\ p < .001$), GRE-Q ($r_{xy} = 0.18,\ p < .001$), and GRE-P ($r_{xy} = 0.22,\ p < .001$) are all positively correlated with GGPA, albeit weakly, meaning that students with higher scores in these metrics tend to earn higher GGPAs. Taken together, the observation that higher UGPA and GRE scores positively correlate to GGPA, which in turn correlate with a student's likelihood of completion, implies that GGPA might play an important role in mediating the influence of these admissions metrics on Ph.D. completion.

The lack of a statistically significant correlation between gender and GGPA indicates that the disparity in scores on the GRE-P between males and females does not manifest itself in subsequent GGPA performance. Indeed, there is no statistical difference between average GGPA for males and females ($r_{pb} = -0.01,\ p = 0.57$, equivalent to a non-significant independent t-test), as demonstrated in Fig. 3(b). Yet there exists a statistically significant difference between males and females in GRE-P performance in our data ($r_{pb} = -0.30,\ p < .001$, equivalent to a statistically significant independent t-test). Thus GGPA does not differ between genders despite the known performance gap between males and females on the GRE-P exam. Furthermore, the phi coefficient measuring the association between gender and Ph.D. completion is negligible despite barely meeting the threshold of statistical significance ($\phi = -0.05,\ p = 0.04$).

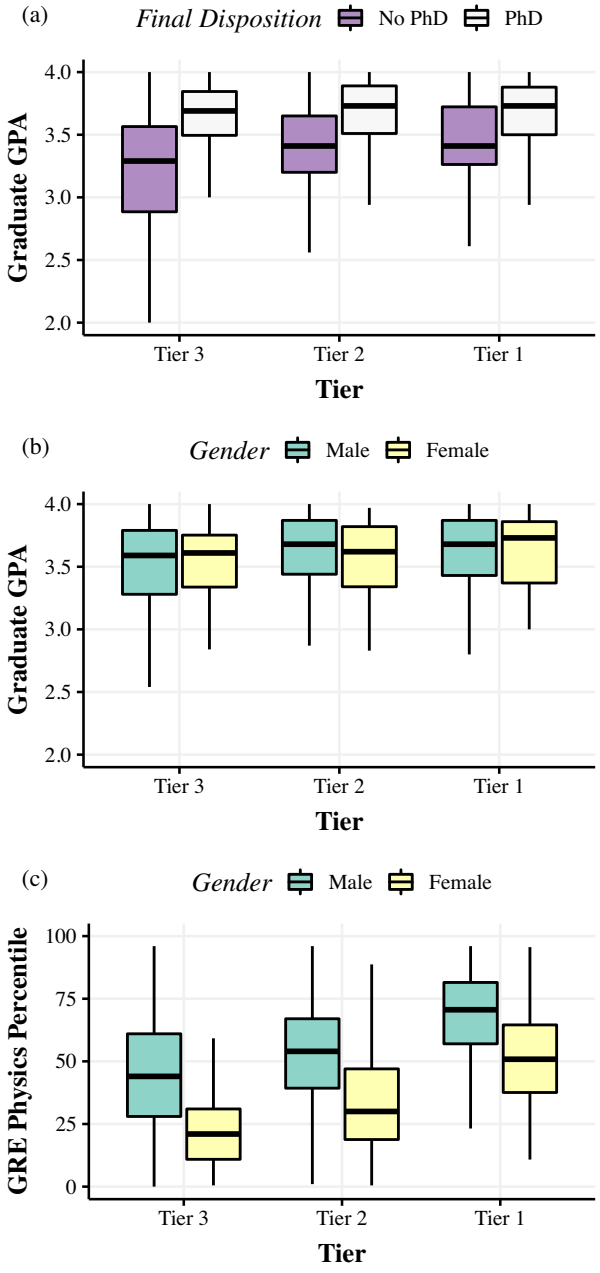

FIG. 3. (a) Graduate GPA by program tier and final disposition. Students who do not complete a Ph.D. earn lower graduate grades than students who complete their programs ($r_{pb} = 0.43$, $p < 0.001$). Tier 3 students tend to earn slightly lower graduate course grades than tier 1 or 2 students. (b) Graduate GPA by program tier and gender. Male and female students earn similar graduate grades ($r_{pb} = -0.01,\ p = 0.57$, equivalent to a non-significant independent t test), and this trend holds across program tier. (c) GRE Physics by program tier and gender. Across all program tiers there is a significant gap in scores between male and female GRE-P test takers ($r_{pb} = -0.30$, $p < 0.001$, equivalent to a statistically significant independent t test). Score distributions trend upward for higher tier programs.

Still, the bivariate correlations shown in Table II do not control for possible relationships between the variables of interest. For instance, the moderate correlation between GRE-Q and GRE-P ($r_{xy} = 0.51$, $p < .001$) indicates that there may be a spurious relationship between one of these variables and GGPA. In addition, we observe low but statistically significant correlations between UGPA and GRE-P ($r_{xy} = 0.31$, $p < .001$), GRE-Q ($r_{xy} = 0.30$, $p < .001$), and GRE-V ($r_{xy} = 0.17$, $p < .001$). This is expected, as UGPA likely contains some information regarding the specific aspects of students' aptitudes tested by GRE exams. These results motivate the use of multiple regression analysis later in this section in order to disentangle the unique effects of each independent variable on GGPA. That analysis reveals that when we isolate the unique predictive effects of each variable in the regression model, UGPA and GRE-P remain significant but weak predictors of GGPA, while GRE-Q does not retain statistical significance.

Similarly, although UGPA ($r_{pb} = 0.15$, $p < .001$), GRE-P ($r_{pb} = 0.14$, $p < .001$) and GRE-Q ($r_{pb} = 0.15$, $p < .001$) are positively correlated with Ph.D. completion, the magnitudes of these correlations are very weak and do not account for other parameters that may be associated with completion. Multivariate approaches allows us to isolate how individual metrics relate to Ph.D. completion, which we explore in the mediation analysis presented in Section IV B. Consistent with previous studies of Ph.D. completion using multivariate approaches [1], we find that when accounting for other parameters, only UGPA remains a statistically significant predictor of completion.

Results of one-way independent ANOVA tests show that the main effect of program tier on GGPA is significant, $F(2, 1949) = 26.31$, $p < .001$, which reflects the upward trend in GGPA from Tier 3 to Tier 1 and 2 programs. A Tukey post hoc test reveals that the GGPA was significantly higher for students at Tier 2 ($M = 3.59$, $SD = 0.40$, $p < .001$) and Tier 1 ($M = 3.60$, $SD = 0.42$, $p < .001$) institutions than those at Tier 3 institutions ($M = 3.41$, $SD = 0.59$). There was no statistically significant difference between the Tier 1 and Tier 2 groups ($p = 0.97$).

Multiple regression.—To disentangle the unique effects of each predictor on GGPA we conduct a multiple linear regression analysis. Multiple linear regression allows us to simultaneously fit many independent variables to measure each of their relative effects on a single dependent variable, GGPA. Analyzing the raw coefficients fitted by the regression analysis yields insight into the predicted change in GGPA due to changes in one variable while holding all others constant. Standardized coefficients allow for a comparison of the relative effect sizes of the independent variables.

Our model includes all available GRE scores (GRE-P, GRE-Q, and GRE-V) as well as UGPA in order to examine the unique predictive effects of each measure. We considered the possibility that including both GRE-P and GRE-Q in the same model would raise collinearity concerns, but find these concerns unfounded. The bivariate correlation ($r_{xy} = 0.51$) between the two is not high enough to warrant genuine concern [62,63]; furthermore, the variance inflation factor (VIF) for every imputed dataset's regression model was below 1.75, well below the commonly cited threshold of 10. Hence, we deem the model posed in the study as most appropriate to answer the research questions raised in this study (further discussion regarding collinearity concerns in the data is available in the Supplemental Material [25]).

The results obtained in our analysis are summarized in Table III. Significant predictive effects at the 95% threshold were found for the numerical metrics UGPA ($\beta = 0.24$, $t = 8.88$, $p < 0.01$) and GRE-P ($\beta = 0.15$, $t = 5.02$, $p < 0.01$). Students with higher UGPA and GRE-P scores therefore tend to receive higher GGPAs. Among statistically significant predictors, the highest standardized coefficient is UGPA, which is larger than the GRE-P coefficient by approximately 50%.

The regression model predicts that for a 0.10 score increase in UGPA, GGPA is expected to increase on average by 0.035 points, holding all other predictors fixed. Meanwhile, a 10 percentile increase in GRE-P score is associated with a 0.031 point increase in GGPA on average, again holding other predictors fixed.

A significant predictive effects was found for gender ($\beta = 0.13$, $t = 2.11$, $p = 0.04$). The positive β coefficient

TABLE III. Coefficients of a multiple regression analysis modeling graduate GPA as a function of common quantitative admissions metrics. Reference categories are White for race and male for gender.

Multiple regression results (*$p < 0.05$, **$p < 0.01$).		
	Dependent variable—GGPA	
Independent variable	Coefficient (standard error)	Standardized coefficient (β)
Intercept	1.92** (0.15)	−0.06
UGPA	0.35** (0.04)	0.24**
GRE-P	31×10^{-4}** (6×10^{-4})	0.15**
GRE-Q	16×10^{-4} (12×10^{-4})	0.03
GRE-V	3×10^{-4} (6×10^{-4})	0.01
Black	−0.11 (0.09)	−0.23
Hispanic	−0.01 (0.07)	−0.02
Native Am.	0.01 (0.23)	0.03
Asian	−0.02 (0.06)	−0.05
Other	−0.24* (0.11)	−0.51*
Undisclosed	0.07** (0.02)	0.15**
Gender	0.06* (0.03)	0.13*
N	1955	
Adjusted R-Squared	0.11	

TABLE IV. Results of mediation analyses that show the predicted change in probability of Ph.D. completion due to changes in admissions metrics from their 25th percentile values to their 75th percentile values among students in the study. The predicted total effect of UGPA is to increase a student's overall probability of Ph.D. completion by 6.0% ($p < 0.01$). That effect is entirely attributable to the indirect effect of UGPA on completion through graduate GPA. The total effects associated with GRE-P and GRE-Q are not statistically significant.

Mediation analysis results for score changes from 25th to 75th percentiles			
Independent variable	Direct effect	Indirect effect	Total effect
UGPA	−0.001	0.060	0.060
	(−0.033, 0.034)	(0.037, 0.084)	(0.030, 0.090)
	$p = 0.97$	$p < 0.01$	$p < 0.01$
GRE-P	−0.006	0.042	0.037
	(−0.047, 0.035)	(0.023, 0.063)	(−0.002, 0.075)
	$p = 0.78$	$p < 0.01$	$p = 0.08$
GRE-Q	0.024	0.009	0.034
	(−0.009, 0.060)	(−0.007, 0.023)	(−0.004, 0.068)
	$p = 0.13$	$p = 0.26$	$p = 0.07$

indicates that if a female student and male student have the same admissions scores, the female student would earn a higher grades in her graduate classes. We include race as a parameter in the analysis as a categorical variable, though all groups except White had small N (the most students in the dataset identified as White, $N = 1205$, while the fewest identified as Native American, $N = 4$). None of the race categories were statistically significant parameters except "other."

Notably the GRE-Q coefficient is not statistically significant, suggesting that it has little predictive effect on graduate course performance. Yet in Ref. [1] GRE-Q was seen to link to overall completion. This observation, along with those made above regarding the predictive effects of UGPA and GRE-P on GGPA, motivate the mediation analysis in the following section. Some predictors, namely, GRE-Q, may link more directly to Ph.D. completion, while others such as UGPA and GRE-P may indirectly predict completion through their effect on GGPA, which itself links to completion.

B. Mediation analysis results

Having demonstrated that graduate GPA (GGPA) is correlated with several quantitative admissions metrics of interest as well as final Ph.D. completion, we now turn to the results of a mediation analysis to determine the extent to which GGPA mediates the relationship between these variables. As described previously, these results yield insight into whether better performance in undergraduate coursework and on GRE examinations increase a student's likelihood of completing a Ph.D. program directly or indirectly via GGPA. Over the course of performing these analyses we explored numerous mediation models using different combinations of covariates (e.g., gender and race) to explore their effect on the results. For instance, we probed the effects of including these variables as moderators in the analysis to account for varying predictive effects among different demographics. However, results were consistent regardless of how these covariates were included in the model.

We perform a separate mediation analysis for each predictor variable (UGPA, GRE-P, and GRE-Q). For each analysis, we begin by calculating the direct and indirect effects of each metric on Ph.D. completion using Eqs. (8) and (9). We also calculate the total effect, which is the sum of the direct and indirect effects. The total effect is comparable to the result one would obtain by using logistic regression to predict changes in probability of Ph.D. completion as was done in Miller *et al.* [1].

This procedure requires us to choose both a control and treatment value for the admissions metrics, since the output of the mediation analysis is the predicted difference in probability of Ph.D. completion if a student who earned the control score had earned the treatment score instead.

Table IV shows the results of a mediation analysis in which we calculate the predicted change in probability of Ph.D. completion if a student had earned a 3.90 undergraduate GPA rather than a 3.46 undergraduate GPA. This change corresponds to a shift from the 25th to 75th percentile of undergraduate GPAs in our data. We observe that the predicted total effect of this change is to increase the student's overall probability of Ph.D. completion by 6.0% ($p < 0.01$). That effect is entirely attributable to the indirect effect of UGPA on Ph.D. completion through graduate GPA, since the direct effect is estimated to be −0.1% and is not statistically significant ($p = 0.97$), while the indirect effect is estimated to be 6.0% and is statistically significant ($p < 0.01$). Meanwhile, mediation analysis using GRE-P as the predictor variable estimates that due to a change from the 25th to 75th percentile of GRE-P scores among students in this study, the probability of Ph.D. completion increases by 3.7% (this change corresponds to a shift in GRE-P percentile ranking from 35 to 71 among the overall physics test-taking population). This

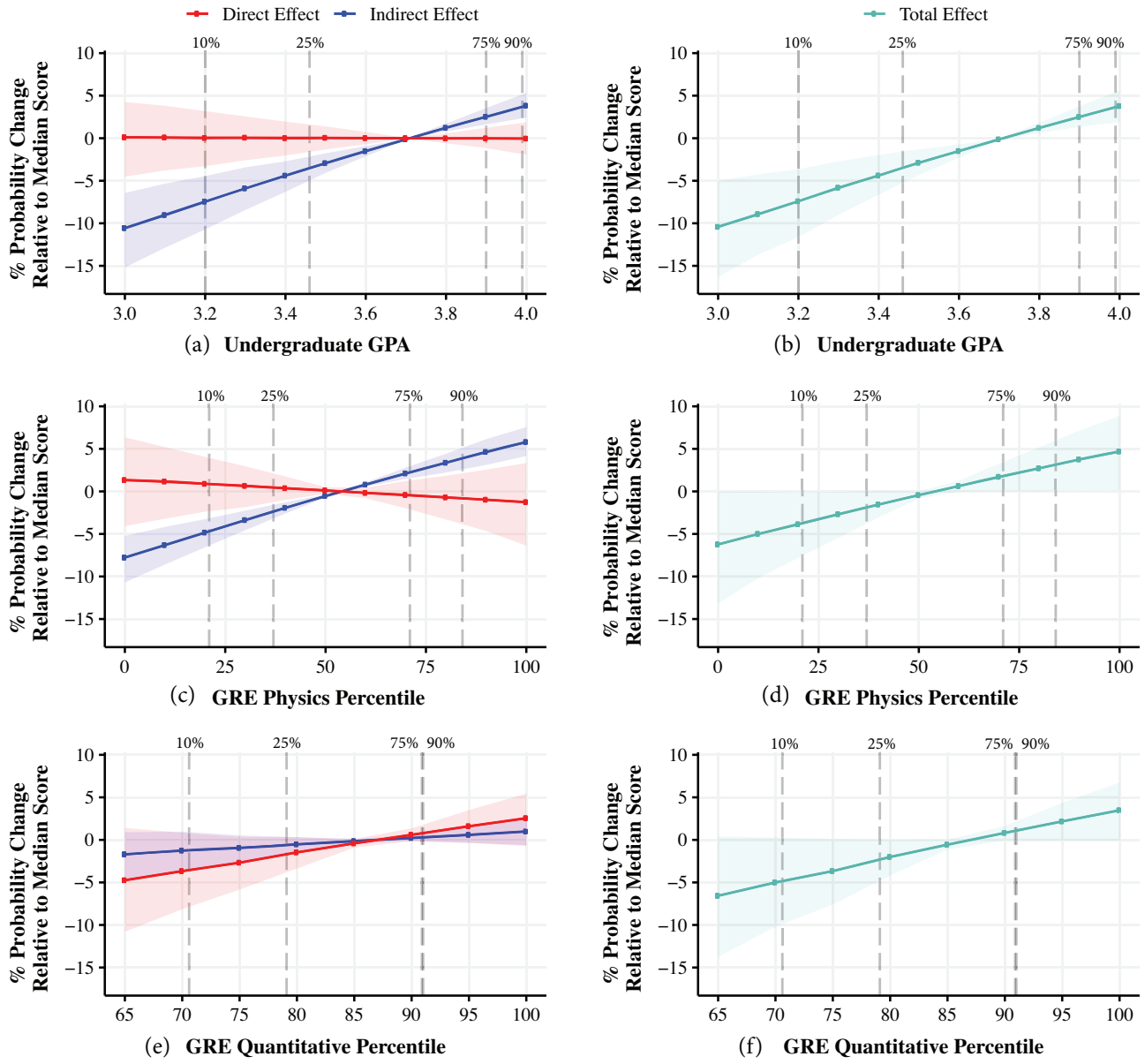

FIG. 4. Mediation analysis results predicting a student's change in probability of completion, split into direct (red) and indirect via graduate GPA (blue) effects. Points on each plot represent individual calculations of probability change relative to each admission metric's median value. 95% confidence intervals are indicated by the shaded region around each line; hence, if the shaded region contains the line $y = 0$, the effect is not statistically significant at the $\alpha = 0.05$ level. The median value of the x-axis variable is clearly shown on the plot as the point where the direct and indirect effect lines intersect, corresponding to a total probability change of 0%. The total effect of UGPA as well as the indirect effects associated with UGPA and GRE-P are statistically significant across all magnitudes of score change.

result just misses the threshold of statistical significance ($p = 0.08$). However, any effect that may be associated with a higher GRE-P score is attributable to the indirect effect of GRE-P on GGPA, which is estimated to be 4.2% and is statistically significant ($p < 0.01$). For a change in GRE-Q score from the 25th to 75th percentile of scores among students in this study (a change in GRE-Q percentile ranking from 79 to 91 among the overall physics test-taking population), the direct (2.4%, $p = 0.13$) and indirect (0.9%, $p = 0.26$) predictive effects on Ph.D. completion are not statistically significant. Their sum, the total effect, estimates an increase in Ph.D. completion probability of 3.4%, and like the total effect of GRE-P also just misses the threshold for statistical significance ($p = 0.07$).

To examine how the predicted probability of Ph.D. completion changes over a broad range of UGPA and GRE scores, we repeat this single mediation analysis for a range of treatment values, as suggested in Ref. [22]. We opt

to choose our data's median values as the baseline control value of each metric, then compute the direct and indirect effects for a variety of treatment values with respect to this baseline. The result is a plot of predicted Ph.D. completion probability change as each score is varied. Figure 4 graphically summarizes the results of the mediation analysis. Figures 4(a), 4(c), and 4(e) display the results of the three separate analyses predicting Ph.D. completion probability changes relative to the median value of each independent variable. Points on each plot represent individual calculations of probability change relative to each admission metric's median value. Hence, the median value of the x-axis variable is clearly shown on the plot as the point where the direct (red) and indirect effect (blue) lines intersect, corresponding to a total probability change of 0%.

Percent probability changes due to direct effects of each admissions metric on Ph.D. completion are shown in red, while indirect effects on Ph.D. completion transmitted through GGPA are shown in blue. Plots of the total effect of each variable on Ph.D. completion probability, which is the sum of the direct and indirect effects, are shown in Figs. 4(b), 4(d), and 4(f). The shaded ribbons around the lines representing the best estimates of probability change show the 95% confidence interval. Hence, if this shaded region contains the line $y = 0$, the effect is not statistically significant at the $\alpha = .05$ level.

In agreement with the results of the example analysis shown in Table IV, the total predictive effect of UGPA on Ph.D. completion is statistically significant while the total effects of GRE-P and GRE-Q do not reach the threshold for statistical significance, indicated by the error bands that encompass the $y = 0$ line [Figs. 4(d) and 4(f)]. Still, as suggested by the result in the Total Effect column of Table IV for GRE-P ($p = 0.08$) and GRE-Q ($p = 0.07$), these effects are close to reaching statistical significance; for reference, each point in Fig. 4(d) the GRE-P confidence ribbon surpasses the line $y = 0$ by less than 0.5%. Thus these results provide some evidence that scoring more highly on the GRE-P and GRE-Q are positively associated with higher rates of completion, although the effects are not statistically significant.

Also consistent with the results of the example analysis shown in Table IV the results of the three mediation analyses indicate that predictive effects of UGPA and GRE-P on a student's Ph.D. completion are entirely mediated by GGPA, defined by the fact that the indirect effect of these admissions metrics are statistically significant across all magnitudes of score change while their direct effects are not. These indirect effects are shown in Fig. 4 by the blue lines in the UGPA and GRE-P plots, whose error ribbons do not contain the line $y = 0$. The interpretation of this result is that UGPA effectively predicts a student's GGPA, which in turn predicts Ph.D. completion. Similarly, any increase in Ph.D. completion probability associated with increases in GRE-P scores are a result of the indirect effect through graduate GPA, although the total effect is not statistically significant.

With regard to GRE-Q, given the weak relationship between GRE-Q and GGPA revealed by the multiple regression analysis in Sec. IV A, it is unsurprising that the indirect effect shown in blue on Fig. 4 is nearly zero. Indeed, as indicated by the earlier results in Table IV any predictive effect from GRE-Q on completion appears to stem from the direct effect of GRE-Q on Ph.D. completion, although the direct effect does not achieve statistical significance at the $\alpha = 0.05$ level.

Lastly, despite using different statistical methods and omitting program tier as a predictor in our model, we observe that the results obtained in this study are qualitatively consistent with those reported by Miller *et al.* [1]. Table V shows a comparison of the total effects predicted by the mediation model presented here with the results of the logistic regression model presented in Ref. [1], again using as an example the predicted changes in probability of Ph.D. completion due to shifts from the 25th to 75th percentile in score for the different admissions metrics. In both analyses, the predictive effects associated with changes in UGPA are statistically significant and are the largest in magnitude among the tested metrics. Effects associated with changes in GRE-P scores are not statistically significant at the $\alpha = 0.05$ level in either model but are close ($p = 0.08$ in this study and $p = 0.09$ in Ref. [1]). The only minor inconsistency between the two

TABLE V. Comparison of results from the current study to results presented by Miller *et al.* [1] in which the authors predict Ph.D. completion using a logistic regression model. Despite using different statistical methods and omitting program tier as a predictor in our model, we observe that the results obtained in this study are qualitatively consistent with those reported previously. In both cases, UGPA is the strongest predictor of Ph.D. completion among admissions metrics tested.

	Current study		Previous study (Ref. [1])	
Independent variable	Total effect value (Probability scale)	Predicted change in completion probability (25th–75th percentile)	Ref. [1] coefficient (Log odds scale)	Predicted change in completion probability (25th–75th percentile)
UGPA	0.060 ($p < 0.01$)	6.0%	0.60 ($p < 0.01$)	3.6%
GRE-P	0.037 ($p = 0.08$)	3.7%	5×10^{-3} ($p = 0.09$)	3.1%
GRE-Q	0.034 ($p = 0.07$)	3.4%	10×10^{-3} ($p = 0.04$)	2.1%

analyses is revealed in the results of the GRE-Q models. Whereas in Ref. [1] the predictive effect of GRE-Q on completion was barely significant ($p = 0.04$), it is not statistically significant here ($p = 0.07$). However, as demonstrated in Table V, effects associated with GRE-Q are not strong in either model and both are very close to the $\alpha = 0.05$ threshold for statistical significance, indicating that the two results are still approximately consistent.

V. DISCUSSION

A. Interpretation of results

The results presented in Sec. IV B give new insight into how admissions committees may contextualize the use of quantitative admissions metrics. Clearly, no existing metric provides unassailable evidence that a student will complete their graduate program. However, among the imperfect quantitative admissions metrics commonly used by admissions committees, the consistent message from this work and others is that undergraduate GPA offers the most promising insight into whether physics graduate students will earn a Ph.D.. Moreover, there is no significant difference between male and female applicants' UGPAs as there is for the GRE-P, meaning that its use in ranking applicants is less likely to skew diversity of admitted students.

As demonstrated by the multiple regression analyses in Sec. IV A, UGPA is most strongly associated with graduate course performance among the variables tested. In some ways this is an expected result: UGPA, the metric that directly measures a student's in-class performance, is most effective at predicting future in-class performance in graduate school. Still, UGPA would seem to vary greatly depending on the student's particular undergraduate institution while a standardized exam like the GRE-P is consistent across all students. It is possible that UGPA may also be signaling socio-emotional skills such as achievement orientation and conscientiousness, which are known to predict high levels of performance both in and out of the classroom [64–67]. These sorts of socio-emotional skills are also shown in the relevant research literature to lack the race, gender, and culture of origin gaps that are found on many standardized tests [68–70].

While previous work showed that UGPA was an effective predictor of Ph.D. completion, mediation analysis demonstrates that relationship is entirely transmitted through UGPA's ability to predict GGPA. Thus, although UGPA is the best predictor of a student's final disposition, our analysis indicates that it is not a direct measure of Ph.D. completion. Rather, the observed relationship between undergraduate grades and completion is explained by the intervening variable GGPA. Regarding the magnitude of the observed effects, we see that changes in score from the 25th to 75th percentile in UGPA (3.46 to 3.90) are associated with an 6% increase in completion probability. Changes across a broader range of UGPAs from the 10th to 90th percentile (3.2 to 3.98), the mediation model predicts an 11% increase in Ph.D. completion probability.

Multiple regression and mediation analyses also yield improved insight into the information provided by GRE-P. Consistent with prior published work by ETS [10,11], regression analysis reveals that GRE-P is an effective predictor of graduate grades. However, the effect associated with UGPA is approximately 50% larger than the effect associated with GRE-P. Regarding the relationship between GRE-P scores and Ph.D. completion, any association between GRE-P performance and Ph.D. completion is entirely mediated by graduate course performance, similar to UGPA. This particular indirect effect indicates that a student who scores more highly on the GRE-P is more likely to perform better in graduate school courses, which may slightly improve their probability of graduation (although the total effect of GRE-P is not statistically significant). However, this indirect effect is still smaller than the indirect effect associated with UGPA, as illustrated in Fig. 4.

Notably, despite the existence of a large gender gap in GRE-P scores (the median GRE-P percentile for females is 35 and 57 for males), male and female graduate students earn nearly indistinguishable graduate grades (Fig. 3). Moreover, there is no practical relationship between gender and Ph.D. completion (Table II). We have also done a preliminary analysis of this same data examining the relation between gender and time to Ph.D. completion, and found that no statistically significant difference exists in the time it takes for male and female physics graduate students to complete doctoral degrees. The disparity in GRE-P scores between male and female test takers is therefore anomalous, as it does not appear to be related to differences in ability or level of preparation and is not reflected in subsequent graduate performance.

Results of multiple regression and mediation analyses that show GRE-Q is not strongly associated with increased graduate course performance are unsurprising given the GRE-Q's task relevance is lower than subject tests and undergraduate grades. Indeed, in its *Guide to the Use of Scores* [26], ETS describes the GRE-Q as testing "high school mathematics and statistics at a level that is generally no higher than a second course in algebra; it does not include trigonometry, calculus or other higher-level mathematics." As shown in Table IV, the association between GRE-Q scores and Ph.D. completion just misses the $\alpha = 0.05$ threshold of statistical significance ($p = 0.07$). Combined with previous studies in which this weak association was statistically significant [1], evidence suggests a weak relationship between GRE-Q scores and completion.

Both direct and indirect effects of GRE-Q on completion were not statistically significant and therefore we cannot discern with certainty which small effect is more important. Considering the case of a possible direct effect between

GRE-Q and completion, the low task relevance makes it unlikely that GRE-Q is a measure of research competence or perseverance. One hypothesis is that socioeconomic status (SES) could be confounding the relationship between GRE-Q, an exam consisting of high school level mathematics questions, and Ph.D. completion. Students with lower SES may have fewer opportunities academically and may perform worse on a standardized exam like the GRE-Q. Indeed, it is estimated that roughly 20% of variance in standardized test scores can be explained by SES [71] In general, previous research [72] indicates that SES impacts whether students possess the proper resources to support them should financial, health or other external circumstances make it difficult to complete. Moreover, doctoral students from lower social classes are more likely to experience a lower sense of belonging in graduate school, often due to the residual financial burdens that are not mitigated by graduate stipends [73]. Reduced sense of belonging in graduate school drives lower interest in pursuing advanced careers in the field, and ultimately a lower likelihood of completing a Ph.D.

B. Limitations and future research

1. Assumptions in causal mediation analysis

The causal effect framework laid out in Sec. III C requires several assumptions to be satisfied in order for effect estimates to be properly identified. All of the assumptions underlying causal mediation analysis refer to "confounding variables," which are variables that influence two other variables simultaneously, thereby causing a spurious association between them. In this section we discuss whether it is plausible that our study has satisfied these assumptions, as well as suggestions for future researchers seeking to perform similar analyses.

Although the assumptions prescribed by causal mediation analysis may be stated in multiple ways [74,75], we describe them in the manner presented by Ref. [19], who condenses them into four requirements. These assumptions, displayed graphically in Fig. 5, are that there exists (i) no unmeasured treatment-outcome confounders, (ii) no unmeasured mediator-outcome confounders, (iii) no unmeasured treatment-mediator confounders, and (iv) no mediator-outcome confounder affected by the treatment.

The final assumption is equivalent to assuming that there is not an alternative mediating variable for which we have not accounted [74]. Concerns about this assumption are handled using methods for multiple mediators, discussed in detail in the Supplemental Material [25].

Assumptions 1–3 require us to be sure we have included all variables that could be confounding the relationships between X (UGPA, GRE-P or GRE-Q), Y (Ph.D. completion), and M (GGPA) in our regression models. These are shown graphically by variables C_1, C_2, and C_3 in Fig. 5. Some of the C variables may influence more than one of X, Y, and M, or may influence each other, so categorization

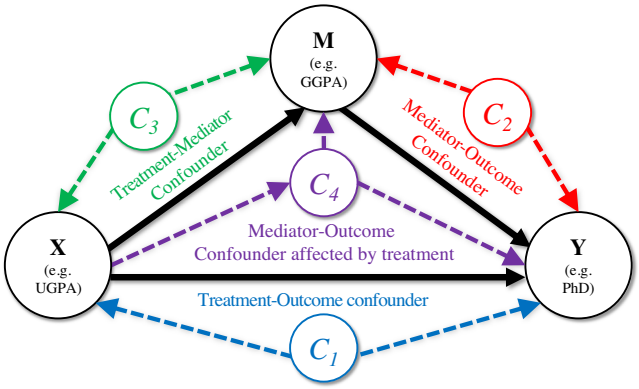

FIG. 5. Mediation diagram showing the relationship of confounding variables C_{1-4} to the variables of interest X, Y, and M. To satisfy the underlying assumptions of causal mediation analysis, researchers must account for these confounding variables in their regression models.

into C_1, C_2, and C_3 is not exclusive. Accounting for variables C_1 and C_3 correspond to assumptions normally made in observational studies to calculate total effects [76], while accounting for variables labeled C_2 is important specifically for the estimation of direct and indirect effects.

2. Sensitivity analysis for violation of assumptions

Unfortunately, as is the case in any regression analysis, whether these assumptions are met is not testable; however, there exist several ways to probe the robustness of our findings under certain violations. In particular, sensitivity analyses have been developed [19,22] to determine how strong a confounding effect between the mediator and the outcome would have to be in order to make statistically significant effects become no longer significant.

We perform such sensitivity analyses on each of the imputed datasets used in the single mediator models presented in Sec. IV B to assess their robustness. We then report the results of the sensitivity analyses averaged across the imputed datasets.

Let R_M^2 and R_Y^2 represent the proportions of original variances explained by the unobserved confounder for the mediator GGPA and the outcome Ph.D. completion, respectively. The result of a sensitivity analysis is a single value representing the product $R_M^2 \times R_Y^2$ that identifies the amount of original variance in the mediator and outcome that the confounder would have to explain in order to make the observed effect vanish. Hence, the sensitivity analysis results in a family of solutions for which the equation $R_M^2 \times R_Y^2 = \text{const}$ is satisfied.

Our sensitivity analysis reveals that for a confounder to explain enough variance to make the indirect effect of UGPA on GGPA vanish, the product $R_M^2 \times R_Y^2$ would have to be, on average, 0.072 (standard deviation = 0.006). For context, the variables included in the mediation models presented here, as a group, are able to explain 11% of

variance in the GGPA outcome ($R^2 = 0.11$); for Ph.D. completion, $R^2 = 0.32$.

Comparing these values to those required to satisfy the equation $R_M^2 \times R_Y^2 = 0.072$, we see that a mediator-outcome confounder would have to be a better predictor of both GGPA and Ph.D. completion than any quantity included in the model so far. For instance, suppose a mediator-outcome confounder satisfied the equation $R_M^2 \times R_Y^2 = 0.15 \times 0.45 = 0.072$, sufficient to nullify the indirect effect of UGPA on completion. The relationship between the confounder and GGPA ($R_M^2 = 0.15$) would be larger than all other variables combined in the model. The link between the confounder and Ph.D. completion ($R_Y^2 = 0.45$) would be larger than all other variables combined in the model as well.

Since the values of R_M^2 and R_Y^2 must be so high relative to the rest of the variables in our model to nullify the indirect effect of UGPA on Ph.D. completion, this result appears to be fairly robust.

Results of sensitivity analyses on the indirect effect of GRE-P on Ph.D. completion were similarly robust: for a confounder to explain enough variance to make the indirect of GRE-P on GGPA vanish, the product $R_M^2 \times R_Y^2$ on average would have to be 0.069 (standard deviation = 0.001). This is essentially the same as the analyses on the robustness of the UGPA result, so its interpretation is the same as above.

More details regarding the sensitivity analyses performed, including contour plots of $R_M^2 \times R_Y^2$ products for which indirect effects of UGPA and GRE-P vanish, are available in the Supplemental Material [25].

3. Conceptualizing possible unmeasured confounders

Thoroughly considering variables that could conceivably act as confounders in Fig. 5 would benefit future researchers studying graduate admissions. Nonquantitative aspects of a student's admission credentials such as letters of recommendation and prior research experience may be associated with Ph.D. completion, and could represent several C variables in the diagram. For example, prior research experience could act as a mediator-outcome confounder, labeled C_2 in Fig. 5. A student with more research experiences prior to entering graduate school may be more likely to complete a Ph.D. than a student with fewer research experiences, since they have already become familiar with the expectations associated with scientific research and have already demonstrated the motivation to pursue it independently. More research experience may also translate to better graduate course performance, particularly if graduate courses are well aligned with the goal of preparing students for future research.

As discussed earlier, information on student socioeconomic status could be useful as well. We also see SES as potentially representing several confounding relationships. For instance, it could act as a treatment-outcome confounder, influencing both GRE scores and Ph.D. completion. It is estimated that roughly 20% of variance in standardized test scores can be explained by SES [71], so SES could be influencing performance on tests such as the GRE-Q and GRE-P. Students with lower SES may have fewer resources to support them should financial, external circumstances arise that make it difficult for them to complete their Ph.D. [36,72]. Including data on these and other possible confounders would bolster causal claims in future analyses. Furthermore, our current models only explain a small amount of the overall variance in the outcome variable, and many other factors are surely at play.

Lastly, although graduate course performance and Ph.D. completion represent some aspects of graduate school success (as evidenced by their inclusion in GRE validation studies), they are certainly crude metrics. Future work should explore other outcomes as well, including success measures such as research productivity, job attainment, or graduate student satisfaction.

VI. CONCLUSIONS

Using data visualization, regression analyses, and mediation analyses, we investigated the role that graduate GPA plays on a physics graduate student's path to Ph.D. completion. We aimed to answer two primary research questions: (i) How do commonly used admissions metrics and demographic factors relate to physics graduate GPA?, and (ii) What role does graduate GPA play in predicting Ph.D. completion, and does it mediate the influence of these other predictor variables on Ph.D. completion? Broadly, we find that across the dynamic range of scores in the data, undergraduate GPA was a better predictor of both graduate GPA and final disposition than GRE scores.

Regarding the first research question of how various admissions metrics and demographic factors relate to physics graduate GPA, we see that significant but weak predictive effects at the 95% threshold were found for the numerical metrics undergraduate GPA ($\beta = 0.24$, $t = 8.88$, $p < .01$) and GRE Physics ($\beta = 0.15$, $t = 5.02$, $p < .01$); GRE Quantitative and Verbal scores are not significantly associated with graduate GPA. The regression model predicts that for a 0.10 score increase in undergraduate GPA, a student's graduate GPA is expected to increase on average by 0.035 points, holding all other predictors fixed. Meanwhile, a 10 percentile increase in GRE Physics score is associated with a 0.031 point increase in graduate GPA on average, again holding other predictors fixed. For comparison, a change in UGPA from the 25th to 75th percentile of scores in our data predicts a 0.15 point increase in graduate GPA, whereas a change in GRE-P from the 25th to 75th percentile of scores in our data predicts a 0.11 increase in graduate GPA.

We also observe that the graduate GPAs are not statistically different between males and females. Hence, the statistically significant gap in performance by gender on the GRE Physics exam (within our data the median GRE

Physics percentile for females is 35 and 57 for males) does not carry over to subsequent graduate course performance. The large difference in performance is unexplained, yet is potentially problematic for promoting diversity in physics graduate school. Multiple regression analysis did not reveal race to be a statistically significant predictor of graduate GPA, but unfortunately it is difficult to properly interpret relationships between race and graduate grades due to a small N. Small sample size precludes useful interpretation of the results pertaining to Black, Hispanic, Native American, and Asian students.

As to the second research question of whether graduate GPA mediates the influence of these other predictor variables on Ph.D. completion, we find that UGPA predicts Ph.D. completion indirectly through graduate grades. Only UGPA is a statistically significant predictor of overall Ph.D. completion (a change in UGPA from the 25th to 75th percentile of scores in our data predicts a 6% increase in Ph.D. completion probability, $p < 0.01$), and that effect is entirely attributable to the indirect effect of UGPA on Ph.D. completion through graduate GPA. The indirect effect associated with UGPA on Ph.D. completion was statistically significant across all magnitudes of score changes, while the direct effect was not (see Fig. 4). Thus UGPA effectively predicts graduate course performance, which is then associated with degree completion. The association between GRE-P scores and Ph.D. completion is not statistically significant (a change in GRE-P from the 25th to 75th percentile of scores in our data predicts a 3.7% increase in Ph.D. completion probability, $p = 0.08$). However, like UGPA, the indirect effect associated with increases in GRE-P score was also statistically significant across all magnitudes of score change, meaning that any predictive effect that GRE-P score may have is therefore also linked indirectly through graduate GPA.

Although these models explain some of the variance in student outcomes [the variables included in the mediation models, as a group, explain 11% of variance in graduate GPA ($R^2 = 0.11$); for Ph.D. completion, $R^2 = 0.32$], much of the variation lies in factors outside the models, in both unmeasured student characteristics prior to admission and unmeasured aspects of the graduate student experience.

No standardized test measures the research and project management skills it takes to successfully complete a multi-year research project, yet those are the skills that are so highly valued in Ph.D. graduates. The GRE-P utilizes two-minute theoretical physics problems to ascertain aspects of students physics knowledge, but it neglects the broad range of computational and experimental skills used in contemporary physics research. Because undergraduate GPA reflects a mix of courses that include theory, experiment, computation, and in some cases research projects, it could be a more useful measure of research-relevant skills. However, our result that UGPA only indirectly predicts Ph.D. completion seems to indicate those research-relevant skills are not a major part of the overall UGPA. Identifying a broader set of applicant characteristics that predict graduate student outcomes is essential.

By better understanding and improving graduate education, we have the opportunity to meet societal goals of a highly skilled advanced STEM workforce that reflects the diversity of our society. While adjusting admissions practices may offer some improvements by adjusting which students are allowed to undertake graduate study, such efforts do nothing to improve the graduate student experience and train graduate students more effectively for STEM careers. The potential for innovation and improvement within graduate education is large and is the area deserving substantial increased attention for education research and programmatic implementation. While our study gives admissions committees greater insight into how and why various quantitative scores link to completion, our discussion of the limitations also points to areas where future researchers can build. We encourage the continued study of not only the physics graduate admissions process, but also the ongoing experience of students in Ph.D. programs, how they are taught, mentored, and supported through their growth as individuals within a larger scientific community.

The authors are pleased to acknowledge valuable discussions with Nicholas Young, Julie Posselt, and Rachel Silvestrini. This work was supported by NSF Grants No. 1633275 and No. 1834516.

[1] C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, and T. Hodapp, Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion, Sci. Adv. **5**, 7550 (2019).

[2] GRE requirements and admissions fees for U. S./Canadian Astronomy & physics programs (2021), available at https://docs.google.com/spreadsheets/d/19UhYToXOPZkZ3CM469ru3Uwk4584CmzZyAVVwQJJcyc/edit#gid=0.

[3] G. Attiyeh and R. Attiyeh, Testing for bias in graduate school admissions, J. Hum. Resour. **32**, 524 (1997).

[4] J. R. Posselt, T. E. Hernandez, G. L. Cochran, and C. W. Miller, Metrics first, diversity later? making the short list and getting admitted to physics Ph.D. programs, J. Women Minorities Sci. Engineering **25**, 283 (2019).

[5] C. Miller and K. Stassun, A test that fails, Nature (London) **510**, 303 (2014).

[6] NSF-Science and Engineering Doctorates: (SED Interactive), available at https://www.nsf.gov/statistics/2016/nsf16300/digest/ (2021).
[7] E. M. Levesque, R. Bezanson, and G. R. Tremblay, Physics GRE scores of prize postdoctoral fellows in astronomy (2015), arXiv:1512.03709.
[8] N. T. Young and M. D. Caballero, Physics Graduate Record Exam does not help applicants stand out, Phys. Rev. Phys. Educ. Res. **17**, 010144 (2021).
[9] M. Walpole, N. W. Burton, K. Kanyi, and A. Jackenthal, Selecting successful graduate students: In-depth interviews with GRE users, Educational Testing Service Reports No. GREB-99-11R, No. RR-02-08, 2002.
[10] L. M. Schneider and J. B. Briel, Validity of the GRE: 1988-1989 summary report, Educational Testing Service Technical Report No. GREB-90-01VSS, 1990, https://www.ets.org/research/policy_research_reports/publications/report/1990/awmu.
[11] N. R. Kuncel, S. A. Hezlett, and D. S. Ones, A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance, Psychol. Bull. **127**, 162 (2001).
[12] S. L. Petersen, E. S. Erenrich, D. L. Levine, J. Vigoreaux, and K. Gile, Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark, PLoS One **13**, e0206570 (2018).
[13] N. R. Kuncel, S. Wee, L. Serafin, and S. A. Hezlett, The validity of the Graduate Record Examination for masters and doctoral programs: A meta-analytic investigation, Educ. Psychol. Meas. **70**, 340 (2010).
[14] J. D. Hall, A. B. O'Connell, and J. G. Cook, Predictors of student productivity in biomedical graduate school applications, PLoS One **12**, e0169121 (2017).
[15] L. Moneta-Koehler, A. M. Brown, K. A. Petrie, B. J. Evans, and R. Chalkley, The limitations of the GRE in predicting success in biomedical graduate school, PLoS One **12**, e0166742 (2017).
[16] J. M. Robins and S. Greenland, Identifiability and exchangeability for direct and indirect effects, Epidemiology **3**, 143 (1992).
[17] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, England, 2000).
[18] J. Pearl, Direct and indirect effects, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI '01* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001), pp. 411–420.
[19] T. VanderWeele, *Explanation in Causal Inference: Methods for Mediation and Interaction* (Oxford University Press, New York, 2015).
[20] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Guilford Press, New York, 2013).
[21] J. Pearl, The causal mediation formula—a guide to the assessment of pathways and mechanisms, Prevention Sci. **13**, 426 (2012).
[22] K. Imai, L. Keele, and D. Tingley, A general approach to causal mediation analysis, Psychol. Methods **15**, 309 (2010).
[23] L. Valeri and T. J. Vanderweele, Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros, Psychol. Methods **18**, 137 (2013).
[24] C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, and T. Hodapp, Response to comment on, "Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion", Sci. Adv. **6**, eaba4647 (2020).
[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.020115 for results of analyses using Non-U.S. student data and alternate imputation methods.
[26] GRE Guide to the Use of Scores, available at https://www.ets.org/gre/guide, 2021.
[27] C. Wendler and B. Bridgeman, *The Research Foundation for the GRE Revised General Test: A Compendium of Studies* (Educational Testing Service, Princeton, NJ, 2014).
[28] J. Briel, K. O'Neill, and J. Scheuneman, *GRE Technical Manual* (Educational Testing Service, Princeton, NJ, 1993).
[29] W. W. Willingham, J. M. Pollack, and C. Lewis, Grades and test scores: Accounting for observed differences, J. Educ. Measure. **39**, 1 (2002).
[30] A. J. Bowers, What do Teacher Assigned Grades Measure? A one page research summary (2016).
[31] S. M. Brookhart, T. R. Guskey, A. J. Bowers, J. H. McMillan, J. K. Smith, L. F. Smith, M. T. Stevens, and M. E. Welsh, A century of grading research: Meaning and value in the most common educational measure, Rev. Educ. Res. **86**, 803 (2016).
[32] A. J. Bowers, What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school, Educ. Res. Eval. **17**, 141 (2011).
[33] A. J. Bowers, Reconsidering grades as data for decision making: more than just academic knowledge, J. Educ. Adm. **47**, 609 (2009).
[34] P. C. Kyllonen, A. M. Walters, and J. C. Kaufman, The role of noncognitive constructs and other background variables in graduate education, ETS Res. Report Series **2011**, i (2011).
[35] L. M. Owens, K. Shar, B. M. Zwickl, and C. W. Miller, Student decits vs. Sense of belonging: Exploring faculty and student perspectives on retention in physics graduate programs (2020).
[36] L. M. Owens, B. Zwickl, S. Franklin, and C. Miller, Misaligned visions for improving graduate diversity: Student characteristics vs. systemic/cultural factors, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).
[37] S. K. Gardner, Fitting the mold of graduate school: A qualitative study of socialization in doctoral education, Innovative Higher Educ. **33**, 125 (2008).
[38] B. E. Lovitts, *Leaving the ivory tower: The causes and consequences of departure from doctoral study* (Rowman & Littlefield Publishers, Lanham, MD, 2002).
[39] National Academies of Sciences, Engineering, and Medicine and others, *Graduate STEM education for the 21st century* (National Academies Press, Washington, DC, 2018).
[40] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary

past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[41] U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Integrated Postsecondary System (IPEDS), Available at https://nces.ed.gov/ipeds/ (2021).
[42] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, and R. A. Kievit, Raincloud plots: A multiplatform tool for robust data visualization, Wellcome Open Res. **4**, 63 (2019).
[43] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).
[44] K. Desiraju, RPubs—ggplot theme for publication ready figures, Available at https://rpubs.com/Koundy/71792.
[45] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 4th ed. (Guilford Publications, New York, 2015).
[46] A. R. Small, Range restriction, admissions criteria, and correlation studies of standardized tests, arXiv:1709.02895.
[47] National Research Council, *A data-based assessment of research doctorate programs in the United States* (National Research Council, 2011).
[48] S. van Buuren and K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, J. Stat. Softw. **45**, 1 (2011).
[49] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*, 1st ed. (SAGE Publications Ltd., Thousand Oaks, CA, 2012).
[50] J. Scott Jones, Learn to use the phi coefficient measure and test in R with data from the Welsh health survey (Teaching Dataset), https://methods.sagepub.com/dataset/phi-coefficient-whs-2009-r.
[51] B. O. Muthén, L. K. Muthén, and T. Asparouhov, *Regression and Mediation Analysis Using Mplus* (Muthn & Muthn, Los Angeles, CA, 2016).
[52] R. M. Baron and D. A. Kenny, The moderatormediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, J. Personality Social Psychol. **51**, 1173 (1986).
[53] D. P. MacKinnon, *Introduction to Statistical Mediation Analysis* (Routledge, London, 2008).
[54] D. P. MacKinnon, C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman, The intermediate endpoint effect in logistic and probit regression, Clinical Trials (London, England) **4**, 499 (2007).
[55] R. H. Hoyle, *Handbook of Structural Equation Modeling* (Guilford Press, New York, 2012).
[56] M. E. Sobel, Identification of causal parameters in randomized studies with mediating variables, J. Educ. Behav. Stat. **33**, 230 (2008).
[57] K. Imai, L. Keele, D. Tingley, and T. Yamamoto, Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies, Am. Political Sci. Rev. **105**, 765 (2011).
[58] B. Muthén, Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus (2011).
[59] D. Tingley, T. Yamamoto, K. Hirose, L. Keele, and K. Imai, mediation: R package for Causal Mediation Analysis, J. Stat. Softw. **59**, 1 (2014).
[60] J. Steen, T. Loeys, B. Moerkerke, and S. Vansteelandt, medflex: An R package for flexible mediation analysis using natural effect models, J. Stat. Softw. **76**, 1 (2017).
[61] L. Muthén and B. Muthén, Mplus, statistical analysis with latent variables. User's Guide 8.4 (2009).
[62] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carr, J. R. G. Marquz, B. Gruber, B. Lafourcade, P. J. Leito, T. Mnkemller, C. McClean, P. E. Osborne, B. Reineking, B. Schrder, A. K. Skidmore, D. Zurell, and S. Lautenbach, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography **36**, 27 (2013).
[63] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, Multicollinearity in regression analyses conducted in epidemiologic studies, Epidemiology (Sunnyvale) **6**, 227 (2016).
[64] F. L. Schmidt and J. E. Hunter, The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings, Psychol. Bull. **124**, 262 (1998).
[65] M. M. Shultz and S. Zedeck, Admission to law school: New measures, Educ. Psychol. **47**, 51 (2012).
[66] F. Lievens and P. R. Sackett, The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance, J. Appl. Psych. **97**, 460 (2012).
[67] K. Z. Victoroff and R. E. Boyatzis, What is the relationship between emotional intelligence and dental student clinical performance?, J. Dental Educ. **77**, 416 (2013).
[68] F. L. Oswald and L. M. Hough, *Personality and its Assessment in Organizations: Theoretical and Empirical Developments* (American Psychological Association, Washington, DC, 2011).
[69] A. Feingold, Gender differences in personality: A meta-analysis, Psychol. Bull. **116**, 429 (1994).
[70] R. Emmerling, R. E. Boyatzis, and R. J. Emmerling, Emotional and social intelligence competencies: Cross cultural implications, Cross Cultural Manage **19**, 4 (2012).
[71] P. R. Sackett, N. R. Kuncel, A. S. Beatty, J. L. Rigdon, W. Shen, and T. B. Kiger, The role of socioeconomic status in SAT-grade relationships and in college admissions decisions, Psychol. Sci. **23**, 1000 (2012).
[72] L. Owens, B. Zwickl, S. Franklin, and C. Miller, Physics GRE requirements create uneven playing field for graduate applicants, in *Proceedings of the 2020 Physics Education Research Conference*, virtual conference (AIP, New York, 2020), pp. 382–387.
[73] J. M. Ostrove, A. J. Stewart, and N. L. Curtin, Social class and belonging: Implications for graduate students' career aspirations, J. Higher Educ. **82**, 748 (2011).
[74] K. Imai and T. Yamamoto, Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments, Political Anal. **21**, 141 (2013).
[75] J. Pearl, Interpretation and identification of causal mediation, Psychol. Methods **19**, 459 (2014).
[76] T. J. VanderWeele, Mediation analysis: A practitioner's guide, Annu. Rev. Public Health **37**, 17 (2016).