SCIENTIFIC COMMUNITY

Response to comment on "Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion"

Casey W. Miller[1]*, Benjamin M. Zwickl[2], Julie R. Posselt[3], Rachel T. Silvestrini[4], Theodore Hodapp[5]

We provide statistical measures and additional analyses showing that our original analyses were sound. We use a generalized linear mixed model to account for program-to-program differences with program as a random effect without stratifying with tier and found the GRE-P (Graduate Record Examination physics test) effect is not different from our previous findings, thereby alleviating concern of collider bias. Variance inflation factors for each variable were low, showing that multicollinearity was not a concern. We show that range restriction is not an issue for GRE-P or GRE-V (GRE verbal), and only a minor issue for GRE-Q (GRE quantitative). Last, we use statistical measures of model quality to show that our published models are better than or equivalent to several alternates.

INTRODUCTION

Here, we present a deeper rationale for our work (*1*) and additional analyses that support the assumptions and interpretations of our original analysis. Our response defines terms to ensure shared language, and we intentionally connect statistical and conceptual rationales for our decisions. This combination of perspectives provides a broader context on the rigorous use of statistical data for informing educational and other public policy decisions. As more scientists turn to research evidence to inform their policy and practice, e.g., in structuring their graduate education programs, statistical evidence should be one resource among several that collectively enable sound and principled decisions.

Our work used historical data to measure relationships between typical admissions criteria and the probability of Ph.D. completion in physics. This analysis was a retrospective, observational study, which is subject to standard limitations. It was not a designed experiment, which means that it may not identify precise causal relationships. However, causal inference techniques can be used to attempt to distinguish causality from spurious correlation. We used causal inference techniques when studying these data. We assume that the use of grades and test scores as a part of the admissions process means that admissions committees believe these variables are useful in identifying students that will be successful in their programs. We also assume that a "successful" student is one that completes the Ph.D. program. In a directed acyclic graph representation of the causal relationships we study, we treat the covariates within our models as "exposures" that could (given the admissions process) influence completion. The data we have do not support the use of Graduate Record Examination (GRE) scores as a reliable measure of whether a student will successfully complete the Ph.D. in physics. In what follows, we provide additional support for this claim and our published findings using statistical methods to show that (i) collider bias is minimal by reproducing our published findings with a model that excludes tier as a categorical variable, (ii) variance inflation and range restriction are not problematic, and (iii) our model choices were just as good as alternates.

INCLUSION OF PROGRAM TIER

The goal of our analysis was to understand how GRE scores and undergraduate grade point average (UGPA) associate with Ph.D. completion in physics. Data are clustered by program; we therefore included a "tier" variable for each Ph.D. program based on its National Research Council ranking. Doing so enables more precise estimates of the relationships between the input variables and completion by grouping programs that are similar. Highly ranked programs, for example, may select, fund, mentor, and educate students in different ways than lower-ranked programs; our analysis of departments' published admissions criteria documents this difference explicitly with respect to admissions. Ignoring these meaningful differences in our analysis may lead to an overestimation of the magnitude of relationships that other variables have with the outcome. Poststratification (e.g., the use of tier to create clusters of programs that are similar) is a standard practice to mitigate the effects of such selection bias or omitted variable bias (*2*).

Poststratification by ranking tier also allows us to prospectively address the interests of different groups of readers. When higher education leaders turn to research to guide their policy decisions, they are typically interested in data about institutions or programs like their own. For example, faculty in elite physics Ph.D. programs may question the relevance of findings generated on students enrolled in less selective programs and vice versa. By stratifying the sample, different readers can understand how the results may apply to their specific interests.

While it is true that stratification may introduce bias in coefficient estimates in some cases, whether a variable is a collider is not necessarily obvious a priori. We investigated this possibility and found little evidence that rank is a collider here. Two alternate analyses, as detailed in the section "Model without tier," produce similar conclusions to those in our published results, indicating that rank is unlikely to act as a collider. A first alternate approach including an

[1]School of Chemistry and Materials Science, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA. [2]School of Physics and Astronomy, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA. [3]Rossier School of Education, University of Southern California, 3470 Trousdale Parkway, Los Angeles, CA 90089, USA. [4]Industrial and Systems Engineering Department, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA. [5]American Physical Society, One Physics Ellipse, College Park, MD 20740, USA.
*Corresponding author. Email: cmilleratphysics@gmail.com

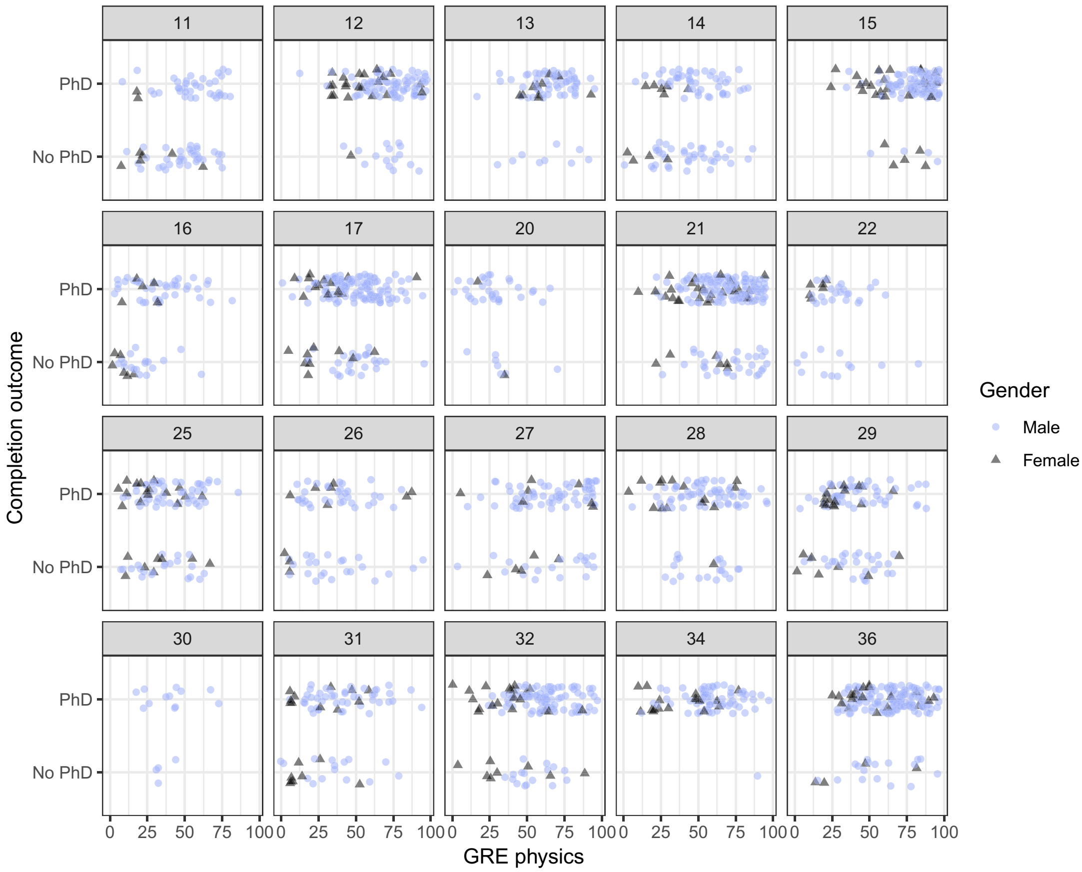

Fig. 1. Ph.D. completion for U.S. men and women by program in our dataset as a function of GRE physics percentile score.

interaction term between tier and the GRE tests was run: No interaction term was associated with Ph.D. completion even at the 0.10 level of significance. A second alternate analysis using a generalized linear mixed model with random effects to account for program-level variations (but excluding tier as a parameter) was consistent with our published results: physics GRE (GRE-P) was not a statistically significant predictor at the 0.05 level in a model that excluded tier and all other admissions metrics [quantitative GRE (GRE-Q), verbal GRE (GRE-V), and undergraduate GPA (ug.GPA)] but included gender and random effects for intercept and slope for GRE-P. This analysis yielded a fixed-effect slope estimate for GRE-P of 0.0058 ± 0.0031 ($P = 0.06$), which is similar to our reported findings (0.005 ± 0.003). Although not detailed in this document for brevity, the GRE-P has an even weaker relationship with Ph.D. completion, and reduced statistical significance, when GRE-Q and ug.GPA are included. These findings demonstrate that including tier as a categorical variable did not confound our results.

MULTICOLLINEARITY AND VARIANCE INFLATION

When explaining complex outcomes like who does and does not finish a Ph.D., a multivariate approach is desirable to enable more precise estimates of individual variables' associations with an outcome. In our case, such an approach was also important given recent evidence that admissions decision makers in most physics Ph.D. programs rely on a combination of undergraduate grades (ug.GPA), and verbal, quantitative, and physics subject GRE scores (i.e., GRE-V, GRE-Q, and GRE-P) (*3*). While multicollinearity and variance inflation can be introduced by using multivariate models, these effects were negligible in our analyses.

In addition to these conceptual rationales for multivariate regression, we offer here statistical rationales for our analysis. One of our initial steps in assessing the data collected for this work was to perform a principal components analysis (PCA) to estimate the potential impact of any collinearity among the four admissions input metrics (ug.GPA, GRE-V, GRE-Q, and GRE-P). The results of the PCA indicated that the correlations between the variables of interest were not large enough to cause concern for including them as independent variables in a multiparameter regression; this point is substantiated in detail below. Given this, we used uncorrelated errors in determining the confidence intervals (CIs) in Fig. 2 of the original publication. Recalculating the CI using correlated errors reduces the CI by about a factor of 3. The magnitude of the CI was the same for male and female in the original Fig. 2 because gender was a categorical variable in the depicted model, meaning it adds only an offset to the model result and does not affect the CI's size.

The indications derived from the PCA are further supported by the variance inflation factors (VIFs) and correlation matrices, both of which support our use of multivariate regression analysis with all four continuous variables.

Multicollinearity can be measured by a VIF. When two variables are independent (i.e., orthogonal and zero correlation), the VIF is 1.

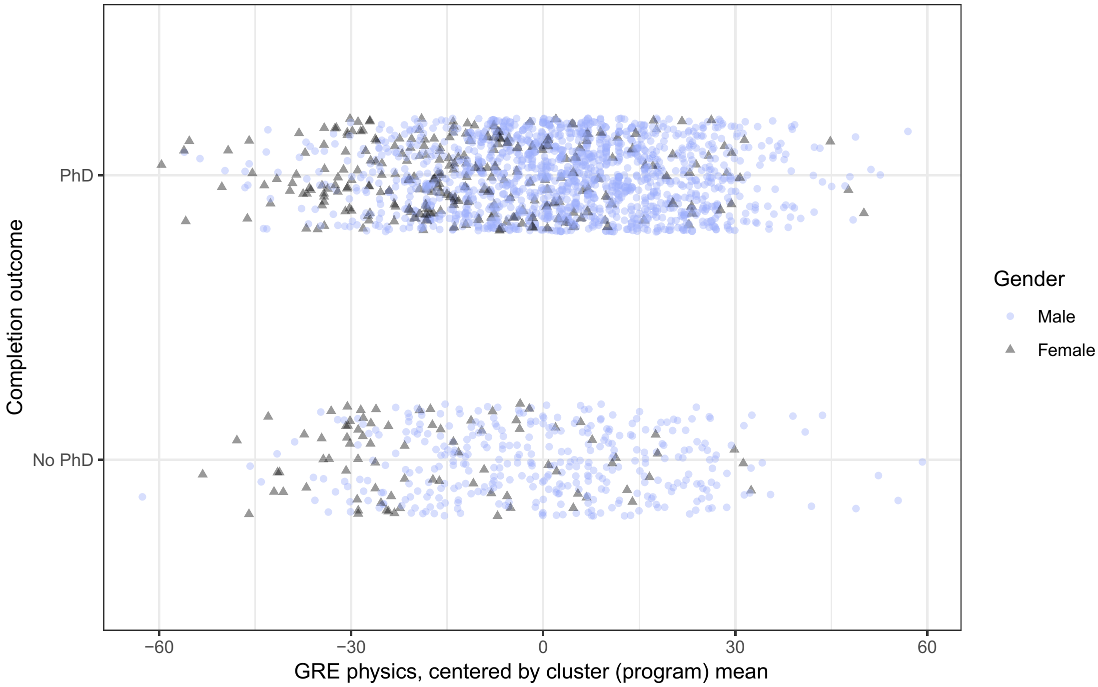

Fig. 2. Ph.D. completion for U.S. men and women in our dataset as a function of mean-centered GRE physics percentile score.

Table 1. Fit parameters from the generalized linear mixed model excluding tier as a parameter and using random effects to account for program-level variations.

Parameter	Fit coefficient	SE	*P* value
Intercept	1.36	0.17	3.00×10^{-15}
GRE-P	0.0058	0.0031	0.06
Female	−0.32	0.12	0.007

As the correlation among variables increases, so does the VIF value. Statisticians agree that when VIF values are between 1 and 5, it is reasonable to include variables together in a multivariate model; when the VIF values exceed 10, multicollinearity is an issue that may lead to errors in interpretation. For more information, see Silvestrini and Burke (*4*) and O'Brien (*5*). As reported in the section "Variance inflation factors," the VIF calculated for each analytic sample was 2 or below. Variance inflation was, thus, not an issue in our analysis.

The bivariate Spearman's correlations among the four continuous variables in each of the four analytic samples ("All students," "U.S. only," "U.S. female," and "U.S. male"), and the correlations among the resultant fit parameters are detailed in the section "Correlation matrices." All correlations are weak to moderate. Some possible reasons for the limited correlation between the GRE-Q and GRE-P include the following:

1) The GRE-Q has limited relation to physics test performance because, according to the test maker, it tests "...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" (*6*). As a result, a high GRE-Q score will not imply a high GRE-P score.

2) The timing of the GRE-P may mask individuals' ability to perform in graduate school because many U.S. students take the test without completing some of the advanced physics courses whose topics appear on the GRE-P. It is noteworthy that the timing of the GRE-P poses a particularly difficult barrier for students at liberal arts and other small colleges where a full physics curriculum is not offered annually; roughly 40% of all physics bachelor's degrees in the United States are granted by such institutions.

3) Undergraduate physics majors' GRE-Q scores are nearly all within just a few standard errors (SEs) of a perfect score. This strong range restriction necessarily limits the strength of any correlation between GRE-Q and any other variable, including GRE-P.

For these reasons, the data do not indicate a problematic correlation between GRE-Q and GRE-P.

RESTRICTION OF RANGE

Range restriction (i.e., a sample whose dependent variable's range is much narrower than that of the population) is a problem that could pose a threat to interpretation in a study using observational data like ours. This potential bias would have been applicable if our study aimed to use a sample of physics students to validate the GRE's utility across disciplines. Our focus on the subset of test takers that were physicists mitigates this possibility. As described above, physicists' GRE-Q scores are restricted relative to the overall test taker population: The 10th percentile U.S. physicist scores at about the median of all test takers, and the median U.S. physicist scores at about the 80th percentile of all test takers. Therefore, the subset of physicist test takers (i.e., those whose intended graduate major was physics, which is a search parameter in the ETS database) is the appropriate group to compare with regard to restriction of range, not the tests' available range of scores.

Table 2. Generalized variance inflation factors for each of the models in Table 2 of our *Science Advances* article.

Model	Parameter	GVIF	Df	GVIF^(1/(2*Df))
All students	ug.GPA	1.14	1	1.07
	GRE.Q	1.56	1	1.25
	GRE-V	1.30	1	1.14
	GRE-P	2.02	1	1.42
	Gender	1.07	1	1.03
	Tier	1.30	2	1.07
	Race	1.93	7	1.05
	Start year	1.11	10	1.01
U.S. only	ug.GPA	1.16	1	1.08
	GRE-Q	1.56	1	1.25
	GRE-V	1.20	1	1.09
	GRE-P	1.71	1	1.31
	Gender	1.09	1	1.04
	Tier	1.55	2	1.12
	Race	1.38	6	1.03
	Start year	1.12	10	1.01
U.S. male	ug.GPA	1.17	1	1.08
	GRE-Q	1.51	1	1.23
	GRE-V	1.18	1	1.09
	GRE-P	1.65	1	1.28
	Tier	1.60	2	1.12
	Race	1.40	6	1.03
	Start year	1.13	10	1.01
U.S. female	ug.GPA	1.21	1	1.10
	GRE-Q	1.81	1	1.35
	GRE-V	1.31	1	1.15
	GRE-P	1.71	1	1.31
	Tier	1.55	2	1.12
	Race	1.58	6	1.04
	Start year	1.37	10	1.02

As the section "Restriction of range" shows, the interquartile ranges (IQRs) of our analytic samples are comparable to ETS-reported ranges of physicist test takers for GRE-P and GRE-V; restriction of range is not a concern for those results. The IQR of GRE-Q in our sample is about two-thirds that for all physicist test takers. Adjusting for this fact may yield somewhat stronger correlations, but the difference is likely to be modest in practice because the distribution's skew means that well above half of physicists' scores are in the range studied. Even so, this pattern is not an issue for our analysis because we are not attempting to validate the examination for the entire population of test takers.

MODEL QUALITY

Our published work conducted a number of robustness and sensitivity analyses for a variety of models and reported on four models

Table 3. Correlation coefficients (Spearman's rank order correlation) among the four quantitative admissions metrics within each of the four analytic samples, averaged across 40 imputations of missing data.

Sample		ug.GPA	GRE-Q	GRE-V	GRE-P
All students	ug.GPA	1	0.18	0.18	0.18
	GRE-Q		1	0.23	0.55
	GRE-V			1	0.14
	GRE-P				1
U.S. only	ug.GPA	1	0.26	0.18	0.28
	GRE-Q		1	0.35	0.55
	GRE-V			1	0.33
	GRE-P				1
U.S. male	ug.GPA	1	0.24	0.16	0.36
	GRE-Q		1	0.34	0.54
	GRE-V			1	0.34
	GRE-P				1
U.S. female	ug.GPA	1	0.26	0.22	0.24
	GRE-Q		1	0.45	0.61
	GRE-V			1	0.42
	GRE-P				1

that included the variables available to and dominantly used by admissions committees in physics, and were representative of findings of other models. Our analysis was not intended to identify the best predictive model with the minimum number of parameters. Here, we provide additional analyses using the full data to provide further evidence that our published models are equivalent to or superior to several alternate models.

In addition to our published models, a few alternate models were explored: (i) excluding GRE-Q from the four original independent parameters, (ii) excluding GRE-P from the four original parameters, (iii) using only GRE-Q, (iv) using only GRE-P, (v) using the average of GRE-Q and GRE-P along with ug.GPA and GRE-V, and (vi) using only ug.GPA. We compare the quality of these models using the Akaike information criteria (corrected for sample size; AICc) (*7*), which provides a standard control for different model complexities. When comparing models, AICc differences of two or less indicate that models are of equivalent statistical quality. In the section "Relative quality of models," we show that the models reported in our paper either have the minimum AICc or are within two of the minimum for all the models noted above, with one exception: Using only ug.GPA for the U.S. women analytic sample is better than the model we published, and better than all other alternate models.

CONCLUSIONS

Our work shows the GRE-Q and GRE-P have limited reliability in identifying Ph.D. completers among applicants in our sample of physics Ph.D. programs, yet these scores can be efficiently used to eliminate women, underrepresented minorities, and U.S. citizens from the discipline. In light of this, should programs rely so much on test scores in pursuit of "the best" when the data show that scores cannot reliably differentiate between finishers and nonfinishers,

Table 4. Correlation coefficients (Spearman's rank order correlation) among the fit parameters within each of the four analytic samples, averaged across 40 imputations of missing data. Demographic factors (race, citizenship, and gender), tier, and start year fixed effect were also calculated but are not shown here.

Model		(Intercept)	ug.GPA	GRE-Q	GRE-V	GRE-P
All students	(Intercept)	1	−0.71	−0.49	−0.03	0.22
	ug.GPA		1	−0.14	−0.08	−0.06
	GRE-Q			1	−0.16	−0.42
	GRE-V				1	−0.17
	GRE-P					1
U.S. only	(Intercept)	1	−0.74	−0.37	−0.10	0.18
	ug.GPA		1	−0.18	−0.06	−0.04
	GRE-Q			1	−0.20	−0.42
	GRE-V				1	−0.13
	GRE-P					1
U.S. male	(Intercept)	1	−0.74	−0.38	−0.12	0.20
	ug.GPA		1	−0.17	−0.05	−0.06
	GRE-Q			1	−0.17	−0.41
	GRE-V				1	−0.15
	GRE-P					1
U.S. female	(Intercept)	1	−0.75	−0.35	−0.04	0.15
	ug.GPA		1	−0.19	−0.08	0.04
	GRE-Q			1	−0.27	−0.43
	GRE-V				1	−0.06
	GRE-P					1

Table 5. AICc analysis for the four analytic samples analyzed through a variety of models. The rows for the four analytic samples indicate the difference between the AICc for each model relative to the minimum among these models. Models with AICc differences of two or less are considered equivalent.

Model	Continuous variables included	Continuous variables excluded	All students	U.S. only	U.S. male	U.S. female
As published	ug.GPA, GRE-Q, GRE-V, and GRE-P		0.0	1.5	1.7	4.5
a	ug.GPA, GRE-V, and GRE-P	GRE-Q	4.1	1.7	1.1	3.0
b	ug.GPA, GRE-Q, and GRE-V	GRE-P	0.3	2.1	1.7	2.9
c	GRE-Q	ug.GPA, GRE-V, and GRE-P	2.6	9.8	4.5	3.8
d	GRE-P	ug.GPA, GRE-Q, and GRE-V	7.9	11.4	4.9	5.2
e	ug.GPA, (GRE-P + GRE-Q)/2, and GRE-V		0.4	0.0	0.0	2.6
f	ug.GPA	GRE-Q, GRE-V, and GRE-P	7.6	3.8	1.6	0.0

let alone who is "the best"? Might a more rational approach be one that acknowledges the limitations of selection due to the uncertainties and biases of both the metrics and gatekeepers and then cultivates the potential of admitted students that matriculate? Ph.D. student outcomes are strongly affected by the quality of the mentoring, research infrastructure, and other resources available to students. Substantial research shows that most students leave doctoral programs for nonacademic reasons (e.g., unwelcoming climates, mentor-mentee conflicts). Addressing these known issues would be a better use of human resources and time than conducting a randomized

Table 6. Data showing the interquartile range (IQR) for different groups of test takers for the indicated tests.

Test	Group	25th	50th	75th	Range 75th–25th
GRE-P	Physicists	26	51	75	49
	Our sample: physicists	42	65	85	43
	U.S. physicists	28	40	63	35
	Our sample: U.S. physicists	35	55	71	36
	U.S. male physicists	21	42	67	46
	Our sample: U.S. male physicists	39	57	73	34
	U.S. female physicists	13	26	47	34
	Our sample: U.S. female physicists	21	37	57	36
GRE-Q	Physicists	69	81	91	22
	Our sample: physicists	81	89	91	10
	United States	20	38	59	
	U.S. physicists	62	78	87	25
	Our sample: U.S. physicists	79	87	91	12
	U.S. male	27	51	69	
	U.S. male physicists	66	78	89	23
	Our sample: U.S. male physicists	79	87	91	12
	U.S. female	17	30	51	
	U.S. female physicists	59	73	84	25
	Our sample: U.S. female physicists	75	83	91	16
GRE-V	Physicists	35	65	86	51
	Our sample: physicists	57	77	89	32
	United States	39	61	80	
	U.S. physicists	65	83	93	28
	Our sample: U.S. physicists	68	81	91	23
	U.S. male	43	69	86	
	U.S. male physicists	65	80	93	28
	Our sample: U.S. male physicists	68	81	91	23
	U.S. female	35	56	76	
	U.S. female physicists	69	83	93	24
	Our sample: U.S. female physicists	70	83	93	23

control trial of GRE-based admissions, which is likely to admit nontrivial biases, including collider effects, due to programs' voluntary participation. As a practical step in this direction, practitioners should take into account both the limited utility of test scores and the disparate impact that can accompany programs' overreliance on scores.

METHODS

Model without tier

To remove tier but not ignore the variability between programs, a series of generalized linear mixed models were computed using the lme4 package in R. Multiple imputation was performed on missing data in the same way as in the published *Science Advances* models. The analyses indicate that collider bias did not affect our published findings.

Our approach began with mean centering the GRE-Q and GRE-P by program. Mean centering simply shifts the distribution of test scores for each program relative to its own mean score: Scores above the program average are positive, while scores below the average are negative. These shifted scores are then used in a mixed-effect logistic regression, which has the advantage of allowing us to fit fixed-effect coefficients for the whole population, while understanding the

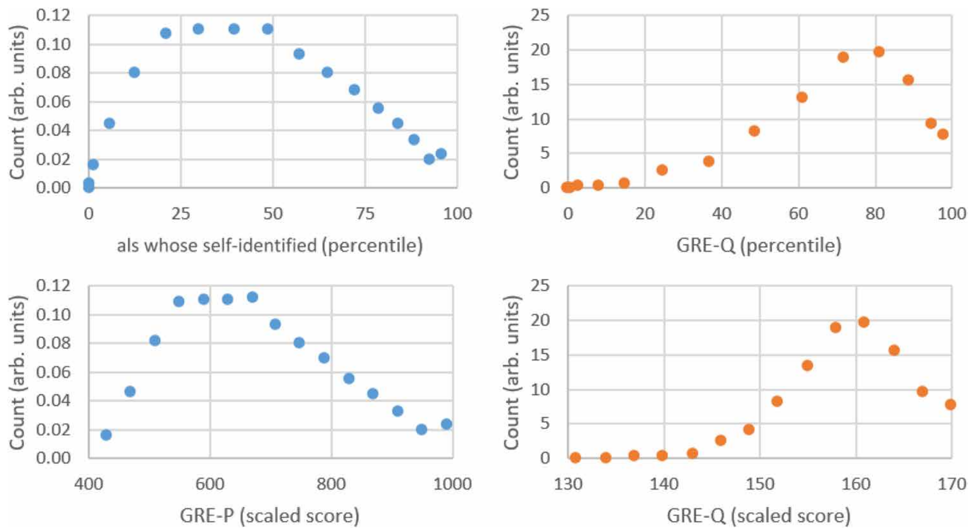

Fig. 3. GRE physics (blue) and GRE quantitative (orange) score distributions for U.S. physicists, plotted as percentile or scaled score. The symbols are bins of width 1 SE error of measure for score differences for each test, decreasing from a perfect score.

program-level variation using random effects. This approach allowed us to account for variation between programs without including a parameter related to tier/rank.

A visualization of the GRE-P scores for completers ("Ph.D.") and noncompleters ("no Ph.D.") from the programs in our dataset that reported GRE-P scores is presented in Fig. 1 by program, while Fig. 2 shows the same data after mean centering.

The most notable feature here is the lack of an obvious difference between the score means and ranges for completers and noncompleters: Students with above-average scores do not graduate, and students with below-average scores do graduate. This pattern alone indicates the limited association between the GRE-P and completion. Were a strong correlation to exist, the right panel of the figure (and the individual graphs in the left panel) would have a strong diagonal character because the "Ph.D." cluster would be centered at a more positive value than the "no Ph.D." cluster. As stated in the main text, this yielded a fixed-effect slope estimate for GRE-P of 0.0058 ± 0.0031 ($P = 0.06$), which is similar to our reported findings (0.005 ± 0.003). We also reiterate that the relationship between GRE-P and Ph.D. completion weakens when GRE-Q and ug.GPA are included, and the result has reduced statistical significance.

The full model specification in the lme4 package was

final.disp ~ GRE.P_cmc + gender + (GRE.P_cmc | program).

The variable final.disp refers to the outcome (completing a Ph.D. or not), GRE.P_cmc refers to the GRE-P score that has been centered by the program's mean score. The term (GRE.P_cmc | program) allows for correlated random effects between the intercept and the GRE-P slope. A summary of the fit parameters from the logistic regression is shown in Table 1. The random variance of intercept between programs (as an SD) was 0.86, the random variance of GRE-P slope between programs (as a SD) was 0.008, and the covariance between intercept and GRE-P slope was −0.005.

Variance inflation factors

The VIF was computed using the vif() function within the car package in R version 3.5.1. See more on the vif() function elsewhere (*8*). The standard VIF is computed when each variable represents just one degree of freedom (Df). However, the generalized variance inflation factor (GVIF) is automatically used when models include some categorical variables that have multiple discrete values (e.g., tier). The VIF results shown in Table 2 are similar between imputations, and we, thus, reported just one imputation.

Correlation matrices

Table 3 shows the correlations between the four quantitative admissions metrics within the dataset. Correlation coefficients are Spearman's rank order correlation averaged across 40 imputations of missing data. Table 4 shows the correlations among the fit parameters resulting from the indicated models were averaged over 40 imputations.

Relative quality of models

We used the AICc to understand the relative quality of our model and various alternate models. AICc differences of two or less indicate that two models are of equivalent quality; one model is better than another if its metric is two to six lower than the other's metric, and a model is substantially better than another if its metric is lower by six or more. Table 5 reports the AICc differences, i.e., the AICc for one model minus that of the minimum AICc.

It is worth additional note that we used the conventional *P* value of 0.05 to indicate statistical significance, with a note indicating that issues with relying on *P* value are known [e.g., risk of observing type I errors (i.e., false positives) increases as sample size increases]. The nebulous nature of the *P* value is, in part, why we commented throughout the article on the practical significance of specific parameters (i.e., reporting both a model's *P* value and probabilities predicted by the model for a range of inputs). The likelihood of finding statistical significance at a specific level is greater with larger sample sizes, and our sample sizes are large relative to those used in analyses conducted by ETS to validate the GRE tests (*9*): A recent validity study used 508 students in CIP Code 40 to validate the GRE for doctoral students in the physical sciences. Three of our analytic samples contain four to eight times as much data, increasing our likelihood of finding statistically significant results, while one contains a comparable amount. Our study only contains physics students, not a heterogeneous group of students from the disciplines comprising CIP Code 40 (astronomy and astrophysics, atmospheric sciences and meteorology, chemistry, geological and earth sciences/geosciences, physics, and materials sciences).

Restriction of range

Table 6 reports the IQRs for groups of GRE test takers and our analytic sample. In addition, included is the range 75th to 25th; comparison of these ranges between the physicist test taker distributions and our sample for each group gives an idea of range restriction. The following nomenclature is used: "physicists" = individuals whose self-identified intended graduate major is physics, as indicated in ETS database; "U.S." = individuals identifying as U.S. citizens, as indicated in the ETS database; "male" = individuals identifying as male, as indicated in the ETS database; and "female" = individuals identifying as female, as indicated in ETS database (*10*). Rows without reference to physicists indicate data for the overall test taker population. The entries show the overall percentile rank that corresponds to the 25th, 50th, and 75th percentiles within each group. For example, a GRE-P reported percentile rank of 63 is the 75th percentile for U.S. physicists.

Percentile versus scaled score

The GRE-P and GRE-Q score distributions for physicist test takers are plotted in Fig. 3 to show the differences between distributions when using percentile and scaled score. While there are differences, they are not substantial enough to cause concern about using one over the other. This may not be the case with the overall test taker population.

REFERENCES AND NOTES

1. C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, T. Hodapp, Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Sci. Adv.* **5**, eaat7550 (2019).
2. P. E. Rossi, R. E. McCulloch, G. M. Allenby, The value of purchase history data in target marketing. *Mark. Sci.* **15**, 301–394 (1996).
3. G. Potvin, D. Chari, T. Hodapp, Investigating approaches to diversity in a national survey of physics doctoral degree programs: The graduate admissions landscape. *Phys. Rev. Phys. Educ. Res.* **13**, 020142 (2017).
4. R. T. Silvestrini, S. E. Burke, in *Linear Regression Analysis with JMP and R* (ASQ Quality Press, 2018).
5. R. M. O'Brien, A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **41**, 673–690 (2007).
6. Educational Testing Service, Overview of the Quantitative Reasoning Measure, (2020); https://ets.org/gre/revised_general/prepare/quantitative_reasoning/.
7. K. P. Burnham, D. R. Anderson, Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304 (2004).
8. J. Fox, Variance inflation factors, rdocumentation.org (2020); https://rdocumentation.org/packages/car/versions/3.0-2/topics/vif.
9. D. M. Klieger, F. A. Cline, S. L. Holtzman, J. L. Minsky, F. Lorenz, New perspectives on the validity of the *GRE* general test for predicting graduate school grades. *ETS Res. Rep. Series* **2014**, 1–62 (2014).
10. Educational testing service, ETS Portal, (2020); portal.ets.org.

Acknowledgments

Funding: C.W.M. was supported by NSF 1633275. B.M.Z. was supported by NSF 1633275. J.R.P. was supported by NSF-INCLUDES 1649297. T.H. was supported by NSF 1143070. **Author contributions:** All authors contributed equally to the writing of this response. B.M.Z. performed statistical analyses. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and in the references cited. Additional data related to this paper may be requested from the authors.

Submitted 6 December 2019
Accepted 1 April 2020
Published 5 June 2020
10.1126/sciadv.aba4647

Citation: C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, 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).

Science Advances

Response to comment on "Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion"

Casey W. Miller, Benjamin M. Zwickl, Julie R. Posselt, Rachel T. Silvestrini and Theodore Hodapp

Sci Adv **6** (23), eaba4647.
DOI: 10.1126/sciadv.aba4647

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.