SCIENTIFIC COMMUNITY

Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller *et al.*

M. B. Weissman*

A recent paper in *Science Advances* by Miller *et al.* concludes that Graduate Record Examinations (GREs) do not help predict whether physics graduate students will get Ph.D.'s. Here, I argue that the presented analyses reflect collider-like stratification bias, variance inflation by collinearity and range restriction, omission of parts of a needed correlation matrix, a peculiar choice of null hypothesis on subsamples, blurring the distinction between failure to reject a null and accepting a null, and an unusual procedure that inflates the confidence intervals in a figure. Release of results of a model that leaves out stratification by the rank of the graduate program would fix many of the problems.

INTRODUCTION

A recent paper by Miller *et al.* (*1*) argues, primarily with regard to GRE scores, that "Typical Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion." They interpret their findings as indicating that lower-than-average scores on admissions examinations do not imply a lower-than-average probability of earning a physics Ph.D. and that GREs are "metrics that do not predict Ph.D. completion." These conclusions, however, rely on standard significance criteria (*2*), which show [see Table 2 of (*1*)] only one predictor for doctoral completion that can be used by admissions committees to select students and is "highly significant" ($P < 0.01$) in the overall sample studied: the Graduate Record Examination quantitative test (GRE-Q). This technical comment describes problems in the analysis. It is not intended to take a position on the complicated nontechnical issue of desirable admissions criteria.

Before evaluating the statistical validity of the analysis, we first clarify the question it aims to answer. That question appears to be how much predictive power for physics Ph.D. completion would be lost by de-emphasizing or dropping the GRE components of physics admissions criteria to diversify student cohorts.

Directly evaluating how well students admitted by different criteria would have done requires either a randomized trial in which similar programs would be randomly assigned to do GRE-aware or GRE-blind admissions (not feasible at the time of the study) or a comparison of nonrandomly assigned programs using modern causal inference methods (*3*) to attempt to reduce systematic errors. There seem to have not been enough GRE-blind programs to allow such an observational study (*1*) . Instead, the strategy used was to create an implicit model of what causes program completion, from which one can estimate what the effect would have been for dropping GREs. Such an analysis would be reasonable if implemented properly.

The authors model Ph.D. completion, a convenient but crude dichotomous proxy for broader goals such as scientific productivity, by a standard logistic regression, with the logit given by a multivariate linear regression on several predictors. The multivariate form is justified as a way to give a better "basis for policy decisions" by avoiding "confounding" (*1*). Since confounding is a purely causal concept, this explanation indicates that the results are intended to tell us what the causal effects of policy choices would be. The model coefficients for the predictors, combined with the ranges of the predictors, may be interpreted as indicating how much incremental predictive power would be lost by dropping each predictor, i.e., what effect that policy change would have on completion rate. The predictors include percentile ranks of GRE scores (quantitative GRE-Q, verbal GRE-V, and physics GRE-P), undergraduate GPA, gender, ethnicity/race, U.S. versus non-U.S. citizenship, year of matriculation, and the rank stratum of the graduate program in which the student ultimately enrolled (*1*), which is one predictor that an admissions committee cannot use as an attribute to distinguish between applicants. Setting aside for now the rank stratum, some such procedure, with the usual caveats, would provide a conventional start to estimating which effects could be excluded from admissions decisions without causing major reductions in degree completion rates.

Several features of the analysis, however, contribute to overestimation of the statistical uncertainty in estimates of the incremental predictive value of GREs, i.e., to the "variance inflation" problem (*4*). Inclusion of the rank stratum can also create systematic underestimation of the predictive power. The inclusion of these features obscures the statistical reliability of the conclusion that those tests help to predict graduation.

RESULTS

Variance inflation from collinearity

The main issue being addressed by the paper is not how well one can distinguish the separate predictive coefficients of GRE-Q and GRE-P, but rather, since they show similar disparities among demographic groups (*1*, *5*), what weight if any should be placed on such tests altogether (*1*). (GRE-V turns out to have essentially no incremental predictive value.) The model presented includes both GRE-P and GRE-Q as separate variables, dividing up their net predictive power into two smaller pieces and inflating the SEs in the estimates of their predictive coefficients via collinearity (*4*, *6*).

A subsequent addendum by Miller *et al.* (*7*) provides the correlation coefficients (in the population studied) both between the GRE-Q and GRE-P percentiles (0.55) and between their estimated predictive coefficients (−0.42). We can recover the net GRE effect size and its nominal statistical significance by combining the two percentiles, giving them equal weight by dividing each by its range in the sample. From Figure 2 of (*1*) , we see that the GRE-P range in the U.S.

Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801-3080, USA.
*Corresponding author. Email: mbw@illinois.edu

is about 1.5 times as large as the GRE-Q range. Adding the Q coefficient to 1.5 times the GRE-P coefficient [from Table 2 of (*1*)], we find that the predictive coefficient of the equal-weight sum is the same to within a 1% range in the "All Students" total sample and in each of the three subsamples described: U.S., U.S. female, and U.S. male. Calculating the SEs of the net coefficients from the reported coefficient SEs (*1*) with their reported correlations (*7*), we find the net GRE predictive effect is 4.5 SE in All Students, far more than the conventional 1.96 SE significance cutoff value for such problems. In the subsamples (U.S., U.S. male, and U.S. female), it is 3.4 SE, 3.0 SE, and 1.5 SE, respectively. Although the U.S. female result does not reach the conventional threshold for significance, due to the small sample, the point estimate is virtually identical to those in the larger groups. Even simply dropping GRE-P, this effect of increasing the coefficient and reducing the SE for GRE-Q would give significance well above the standard threshold except for that smallest subsample.

The net logit change between the 10th and 90th percentiles on that combined score would be reduced from the sum of the separate effects of the two scores [~0.46 and ~0.36 in the United States for Q and P, respectively, estimated from Figure 2 of (*1*)] by a factor $(1.55/2)^{1/2}$ since their correlation is 0.55, giving a net logit effect of ~0.72. (Here, I assume that the percentile range scales approximately with the SD and that slightly changing GRE weightings does not induce a large change in the coefficients of other variables.) This effect somewhat exceeds the corresponding GPA effect (~0.6) in the U.S. subsample and no doubt greatly exceeds the GPA effect for All Students, for which the GPA predictive coefficient falls off sharply (*1*). (Inclusion of different weights for GRE-P and GRE-Q, i.e., inclusion of their equal-weight difference as a predictor, adds very little to the predictive power.) Thus, even before getting to the interesting and important modeling questions, we see that according to the data of the paper (*1*) and addendum (*7*), overall GREs do better than GPA for predicting graduation within the context of the linear logit model.

Stratification: Variance inflation and collider-like bias

The model chosen includes the rank of the graduate program in which the student enrolled, via two adjustable terms for three rank strata (*1*). Admissions committees cannot use the result of future enrollment to decide among competing applicants. This fact raises a question whether such a variable belongs in a model estimating the predictive value of other metrics.

An immediate issue with stratification is that it creates another variance inflation by restricting the range of the predictors. This problem of restricted range in predictive modeling is well known, especially in the context of educational and employment decisions [e.g., (*8*, *9*)], and even in the specific context of physics GREs (*10*). Correlation between outcomes and predictors is suppressed in narrow strata. In one experimental comparison, correlation between scores on a two-component Swedish driving test fell by more than a factor of 2 when restricted to those who passed the first test (*9*). In a 1993 study (*8*), the GRE validity in predicting performance of psychology students in classes on statistics, assessments, and research methods was found to be high (0.55 to 0.70) in a program with little range restriction, in contrast to much lower validity in a range-restricted subset or to typical low validity for predicting grades in more range-restricted programs. The authors' conclusion was "These results support the conventional argument that uncorrected GRE validity estimates based on range-restricted samples are strongly biased toward zero" (*8*).

It is not unreasonable that the Miller *et al.* analysis contains a restricted range, since a school or employer typically does not have performance data on those who either were not offered a position in their institution or did not choose to take it. However, this particular range restriction comes mainly from the choice to stratify students by program rank (*1*). Miller *et al.* (*1*) state that one of the strengths of their study is that it includes a wide range for the predictive variables because it includes schools of very different ranks, but they do not use that range to narrow the statistical uncertainties in the parameter estimates.

A critical question is whether loss of precision is justified by the need to avoid systematic errors. Miller *et al.* say they "…include covariates to render more precise estimates", but including covariates can either remove or add systematic bias depending on which covariates are included and on what one wishes to estimate (*3*, *11*, *12*). In causal inference studies stratifying on a "collider," a downstream variable affected both by the suspected cause and by unmeasured other causes adds a systematic error called collider stratification selection bias to the causal estimand (*3*, *11*, *12*). For example, inadvertent collider conditioning produces a paradoxical effect that maternal smoking appears to protect low–birth weight newborns from mortality, because within the low birth weight stratum, smoking is negatively correlated with even more ominous predictors (*13*).

Miller *et al.* (*1*) find that even after taking into account GPA, GREs, etc., students in the higher-ranked programs have a higher likelihood of completion. The use of their stratified model to evaluate the incremental predictive power of GREs implicitly assumes that this boost is caused entirely by factors that would not change if students with lower scores were admitted to those programs. There are two main possible causes of this boost mentioned in the original paper.

One possibility mentioned (*1*) would be that a typical student has a systematically easier time graduating from higher-rank programs than from lower-rank programs, so the boost would persist even if admissions procedures changed and students who would currently enroll in lower-rank programs were switched to high-rank programs. If this were the main explanation, then rank would be a simple confounder and should be removed by stratification or other methods to improve the estimate of the incremental predictive power of the GREs. No evidence is given to support this possibility, and the actual sign of effect is not obvious.

The other possibility is that the high-ranked programs are getting students with a higher propensity to graduate than predicted by the in-model GREs and GPA because they use a variety of other predictors as well, as documented in (*14*), which shares a co-author with Miller *et al.* (*1*). These predictors include prior research experience, letters of recommendation, etc. (*14*). Unless those predictors are irrelevant to degree completion, they will have some positive predictive value, which will be reflected in the coefficient of the rank variable, with which they will be positively correlated (*1*). If the out-of-model predictors are positively correlated with an in-model predictor, they will increase the coefficient that the model assigns to that predictor beyond what would actually be lost by dropping the predictor, but if they are negatively correlated, they will decrease the coefficient. As a result, the model estimate will depend on stratification because the correlation between the in-model and out-of-model predictors changes as a function of stratification (*11*).

Students with low GREs and GPAs who nonetheless are accepted into high-rank schools are likely to have especially good prior research experience, letters of recommendation, etc., creating a

negative correlation within each stratum between those stratum-correlated out-of-model predictors and the predictors used in the model (*10*). A similar effect occurs in a different context: Although performances on long jumps and 110-m races are likely to be positively correlated in the general population, in the stratum of olympic decathletes, they have a strongly negative correlation (*15*).

The reported data include indications that the odds boost for students in high-ranked programs is likely to be due to the out-of-model predictors used in admissions rather than to any direct student-independent effects (of unknown sign) of differently ranked programs. If some randomly chosen students were boosted in enrolled program rank, their graduation probability would increase from the hypothetical direct effect but not change for the out-of-model selection effect. In the selection case, but not the direct effect case, the stratified model would then assign this random group a negative logit equal to the positive logit assigned to the rank boost. In a causal diagram, the random group assignment would collide with effects of out-of-model selection traits on program rank, and the random group assignment would pick up a logit via collider bias despite having no causal effect on graduation. Something approximately similar to that randomized trial would happen if the boosted students were picked nonrandomly, but based on traits with little direct relevance to graduation probability. Given the almost universal attempt to boost representation of underrepresented minorities, we may see such statistical artifacts in the large negative logits the model assigns to them [seen in Table 2 of (*1*)], which are statistically significant in the overall sample and close in magnitude to the positive logit assigned to the difference between the first and third rank tier. That pattern is more consistent with collider bias in the model than with the more selective programs being easier to complete, although without further information on other possible factors, one cannot precisely sort out such systematic effects. I predict that these negative demographic logits will shrink substantially in a less-stratified (and, as I will argue, probably more accurate) model omitting program rank and could easily fall to zero or turn positive if a fully unstratified model or one including all important predictors were possible.

Since the out-of-model predictors are themselves likely to be positively correlated with in-model predictors, they would be confounders in a model completely lacking range restriction, causing some positive overestimate of the incremental predictive power of the in-model predictors. For the real data, however, the unavoidable limitation to students who have been accepted means that the population under study is systematically restricted compared with the one of interest—all the applicants plus some others who might apply if GREs were dropped (*10*). That unavoidable range restriction effect is not small. For example, if both in and out contributions are independent normally distributed and given equal weight, mere selection of applicants with an overall above-average score gives a correlation coefficient of $-1/(\pi - 1) = -0.47$. Even if the in and out predictors are positively correlated (coefficient r_{OI}) in the entire applicant population, their correlation in the enrolled upper half is $((\pi - 1)r_{OI} - 1)/(\pi - 1 - r_{OI})$. Even without rank strata, the model would underestimate the in-model coefficients if $r_{OI} < 1/(\pi - 1) = 0.47$, which is larger than one would ordinarily expect the correlation to be between disparate predictors such as test scores and research experience. Since the coefficients of the tiers do not show especially large variance inflation (*7*), they cannot be very strongly correlated with the other predictors. (It would be easier to reason accurately about this possibility if the covariances between program tier and other variables were available.) Thus, to the extent that the positive logits for high-ranked programs are caused by their selection of students, even a model omitting rank strata would be likely to underestimate the incremental predictive power of including GREs, or at any rate not overestimate it by very much.

The more finely rank is stratified, the more negative these correlations become (*10*). In the ideal limit of narrow rank stratification and admissions criteria successfully aimed to maximize a particular goal, all power for predicting that goal using any variables other than rank becomes zero regardless of how predictive they are in the unstratified population, since no variation is left within each stratum. That remains true regardless of how much range remains for any individual predictor, how complete the overall range of the data is, and how large the sample size is. That program rank should be a relatively good predictor in the stratified model, thus, tells us little other than that physics admissions committees are making use of the out-of-model predictors that they say they use (*14*) in a way that correlates with program rank.

Null hypotheses for subsamples, anomalous confidence intervals, and dynamic range compression

The Miller *et al.* paper reasonably avoids making a strong prior assumption that each predictor will work equally well in each subsample. As we have seen, however, the point estimate for the net GRE predictive coefficient based on their data is virtually identical in each subsample, providing no evidence that net GRE weighting should differ among them. The paper replaces the conventional null hypothesis of equal effects in different subsamples with null hypotheses of no effect in each subsample. This choice may produce anomalous interpretations. For example, although the point estimate given in Table 2 of (*1*) for the coefficient of the logit for GRE-Q in All Students (0.013 per percentile rank) is statistically significant, and the point estimate among U.S. females (0.017) is larger, the latter fact is described as "we see no differences in Ph.D. completion probability…" in females (*1*). Here, the paper interprets this result as being insufficiently precise to confidently reject a null hypothesis. Such an interpretation can be problematic. For instance, in typical medical trials, when a treatment appears to work better in a subsample than in the overall group, but with larger uncertainty due to the small sample, it would be highly unconventional to conclude that the treatment does not work in the smaller group, even though that possibility cannot be statistically ruled out.

Figure 2 (*1*) illustrates the predictive slopes of the U.S. subsample for GPA and the GREs applied separately to the 10th, 50th, and 90th percentile scores for U.S. females and males. It shows very large "95% confidence intervals associated with Ph.D. completion probability," (p) leaving the visual impression that predictive effects are small compared with uncertainty. Converting to logits, these intervals are roughly ±1.1 for each estimate at the low, middle, and high parts of the distributions for both U.S. males and females. The near equality at the middle and edges of the distribution indicates that these intervals cannot primarily reflect the uncertainty of interest, i.e., uncertainty in the slopes of the logit dependence on the model variables, because that would not show up much in the middle points. For large N in the middle of the parameter range, the 95% confidence intervals for the logit should be $\pm 1.96^*/(Np(1-p))^{1/2}$. For the full U.S. sample with $N = 2315$ and $p = \sim 0.7$, that would be ±0.09, not ±1.1. The confidence intervals shown appear to be based on the number of students (~23) within each integer percentile group rather

than the actual group size from which the probability estimates are calculated, which would inflate them by approximately one order of magnitude.

Rather than directly use the GRE scores themselves in the linear model, the paper uses percentile rankings (*1*), a convenient way to stitch together scores from before and after the GRE scale changed. It is not required, however, since score conversion tables are available. The percentile method has the effect of greatly compressing the dynamic range in the higher scores in the tail of the distribution and magnifying small differences in the middle of the distribution, where most accepted applicants are found. It is possible that this highly nonlinear map from test scores to the predictors used in the linear model reduces the predictive power.

The bottom line

Based even on the incomplete data presented, the statistical uncertainty in estimating how much predictive strength would be lost by dropping or de-emphasizing GREs is not particularly important (*1*). We have seen that in the U.S. subsample, a simple equal-weight sum of the two relevant GREs provides a logit difference of ~0.72, i.e., an odds ratio of ~2.1, even before making any upward correction for a systematic stratification bias or for possible improvement from using test scores rather than percentiles.

Extending those results to the non-U.S. 40% of the sample requires guesswork, because the range data and correlation coefficients for that subsample have not been provided. In the published results, there is no indication that GREs would be a weaker predictor in that group than in the U.S. (*1*). In contrast, the predictive coefficient for GPA is only about half as large in All Students as in the U.S. (*1*). Thus, although no predictors of graduation are especially good, the net equal-weight GRE-P and GRE-Q combination looks better than GPA overall. Results in the addendum (*7*) for formal model evaluation criteria, which include a likelihood measure and a penalty for adding parameters, look consistent with this conclusion, although a simple model including GPA and the net GRE-P and GRE-Q but omitting the irrelevant GRE-V (*1*) is not included. Extending the results to lower scores, particularly relevant for GRE-Q whose range is strongly restricted in the sample, not just the strata (*7*), is uncertain, but past indications are that such dependencies do not become any weaker in the low end (*8*).

DISCUSSION

Miller *et al.* deserve credit for collecting a substantial amount of useful data on a question of wide interest. Since their dataset includes a relatively large predictor range, it opens up the possibility of making more reliable estimates than ones based on data or anecdotes from individual programs. Some of their results are already useful, e.g., showing that the GRE-V adds essentially nothing to the statistical predictive power in this sample (*1*). The correlation coefficients (*7*) let us see that an equal-weight GRE-P and GRE-Q combination is only a slightly stronger predictor than plain GRE-Q, a result that may have little import for admissions decisions but might matter in evaluating costs of admissions requirements. There are anecdotal reports that the direct financial cost of an added test is a barrier for many students, especially from underrepresented groups. There is also a slight statistical hint, about a 1 SE effect, suggesting that different weights for GRE-Q and GRE-P might be useful for U.S. females.

Only a few straightforward further steps would be needed to make much fuller use of these data. Most important by far would be to publish the results for the model without the biasing stratification by program rank. Another very useful step would be to publish results (ranges of predictors, coefficients, and correlations) on the 40% of the sample that are not U.S. citizens, just like on the U.S. subsamples. Publishing ranges of variables for All Students would also be useful. Publishing a comparison of a model that drops GRE-P with one that uses an equal-weight P and Q sum would help in cost-benefit decisions about requiring GRE-P. A small amount of additional calculation would allow an analysis using scores rather than percentiles.

Even a more transparent analysis of these data, however, will potentially be subject to systematic errors of unknown sign, as with any observational study. Since there are now fairly many physics departments unsure of what GRE policy to adopt, a randomized controlled trial now might directly address the causal policy question: What effects do different admissions policies have on student outcomes? Different volunteer departments could be randomly assigned to different GRE policies for a year and then switched for the next year. The resulting cohorts could then be followed not only for degree completion but also for other outcomes. Regardless of the results, it would set an example of scientists trying to use objective scientific methods to help make policy choices.

Even if more reliable outcome estimates become available, interesting arguments (far beyond the scope of this technical comment) will continue over how to balance different goals. The effects of changing criteria may not even be dominated by the individual-level effects discussed here, but by much harder to predict changes in institutional traits and motivational signaling effects. For example, if GRE-P were not used in graduate admissions decisions, many institutions might change undergraduate physics curricula and grading standards, for better or worse or both.

The emerging interest in "preregistration" in the social sciences, in which analysis methods are peer reviewed before data are collected, is a promising tool for statistical studies of educational inputs and outcomes intended to inform policy decisions. Preregistration is not a panacea, but it can help anticipate and mitigate methodological concerns that might arise after publication, as in this instance.

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. V. Amrhein, S. Greenland, B. McShane, Scientists rise up against retire statistical significance. *Nature* **567**, 305–307 (2019).
3. J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, New York, 2018).
4. R. A. Stine, Graphical Interpretation of Variance Inflation Factors. *Am. Stat.* **49**, 53–56 (1995).
5. C. Miller, K. Stassun, A test that fails. *Nature* **510**, 303–304 (2014).
6. D. E. Farrar, R. R. Glauber, Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **49**, 92–107 (1967).
7. C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, T. Hodapp, Typical physics PhD admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. arXiv [physics.ed-ph], 1906.11618.pdf (2019).
8. B. E. Huitema, C. R. Stein, Validity of the GRE without restriction of range. *Psychol. Rep.* **72**, 123–127 (1993).
9. M. Wiberg, A. Sundström, A comparison of two approaches to correction of restriction of range in correlation analysis. *Pract. Assess. Res. Eval.* **14**, 1–9 (2009).

10. A. Small, Range restriction, admissions criteria, and correlation studies of standardized tests. arXiv [physics.ed-ph] 1709.02895.pdf (2017).
11. S. Greenland, Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14**, 300–306 (2003).
12. M. A. Hernán, S. Hernández-Díaz, J. M. Robins, A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
13. T. J. VanderWeele, Commentary: Resolutions of the birthweight paradox: Competing explanations and analytical insights. *Int. J. Epidemiol.* **43**, 1368–1373 (2014).
14. 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).
15. J. Park, V. M. Zatsiorsky, Multivariate statistical analysis of decathlon performance results in olympic athletes (1988-2008). *Int. J. Sport and Health Sciences* **5**, 1128–1131 (2011).

Acknowledgments: I thank E. Fireman, J. Robins, and A. Small for conversations and many colleagues for comments on an arXiv paper. **Funding:** No funding was involved in this work. **Author contribution:** M.B.W. is solely responsible for the content of this paper. **Competing interests:** The author declares that he has 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 author.

Submitted 18 March 2019
Accepted 27 March 2020
Published 5 June 2020
10.1126/sciadv.aax3787

Citation: M. B. Weissman, Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller *et al. Sci. Adv.* **6**, eaax3787 (2020).