Submitted: 7th of December 2023
Published: 5th of March 2024
DOI: 10.26775/OP.2024.03.05
ISSN: 2597-324X

Diversity in STEM: Merit or Discrimination via Inaccurate Stereotype?

Joseph Bronski* Emil O.W. Kirkegaard†

Abstract

Leslie et al. (2015) advocated a model where a stereotype that a given field requires brilliance to succeed scares women away from the field, thus resulting in a self-fulfilling prophecy similar to stereotype threat. Leslie however ignored decades of findings in stereotype accuracy research, where stereotypes are generally known to accurately track real existing differences. As such, a simpler explanation for the data is that the brilliance stereotype results from real existing differences in academic ability between fields of study, which is also the variable that explains the different distribution of demographic groups in these fields due to differences in academic abilities. Chiefly, men's superior mathematical ability explains why they are overrepresented in fields that require strong mathematical talent to succeed (e.g. physics). We present an analysis which suggests that the proportion of a field that is female is better predicted by that field's average math GRE score ($r = -0.79$) than Leslie et al.'s Brilliance stereotype ($r = -0.65$), and the proportion of a field that is Black is predicted equally well by both that field's average GRE score ($r = -0.49$) and Leslie et al.'s Brilliance stereotype ($r = -0.53$). We show that a field's Brilliance stereotype is furthermore closely associated with its average GRE score ($r = 0.58$). Additionally, we show that a field's scientificiness stereotype score is predicted by its GRE math tilt ($r = 0.36$) while a field's conservativeness stereotype score is associated with the actual percent of registered Republicans in that field ($r = 0.55$). We conclude that Leslie et al.'s uncritical reliance on inaccurate stereotype to explain disparities in racial and gender diversity by academic field is deeply flawed. Finally, their results failed to replicate among the doctorate holding public; GRE Math was a better predictor of the percent of a field that is female than brilliance stereotype among doctorate holders ($r = -0.79$ vs. $r = -0.39$).

Keywords: diversity, intelligence, female IQ, black IQ, stereotype

1 Background

The cause of racial and gender disparities among academic fields has long been debated. In 2002, Templer & Tomeo correlated the percent of males in a field with that field's mean GRE score and found $r = 0.76$, a very strong correlation among those typically found in social science (Templer & Tomeo, 2002). As of writing, this finding has been largely ignored, only having been cited 10 times according to Google Scholar.

In contrast, in 2015, Leslie et al. ran a survey asking people how much they thought "brilliance" was required for success in various academic fields, and found a correlation of $r = -0.60$ between this metric and the percent of females in each field (Leslie et al., 2015). They hypothesized that "across the academic spectrum, women are underrepresented in fields whose practitioners believe raw, innate talent is the main requirement for success, *because women are stereotyped as not possessing such talent*" (emphasis added). At no point did they consider the association between each field's mean GRE math score and each field's proportion of women – instead, they chose to only consider the mean composite GREs of *applicants* to each subject, and only for a very small number of subjects. In our study, we have collected data on GRE math and verbal scores for applicants to a much larger spread of fields. A clear picture emerges: female representation is mostly applicant GRE math score by field.

*Independent Researcher, Email: josephbronski7@protonmail.com
†Ulster Institute for Social Research, United Kingdom, Email: emil@emilkirkegaard.dk

For ignored math scores, they received 1395 citations at the time of writing, and numerous popular science news articles covering their paper. This liberal political bias is a systemic issue in social science (Honeycutt & Jussim, 2022). There are other examples of massive citation discrepancies between complementary papers, one favoring a liberal worldview. One instance is "National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track" (Williams & Ceci, 2015) having 434 citations according to Google Scholar as of writing while "Science faculty's subtle gender biases favor male students" (Moss-Racusin et al., 2012) has 3422 citations. Another are two studies, both from the 1990s, one releasing a scale to measure sexism against women, and one to measure sexism against men. The woman-sexism scale (Glick & Fiske, 2018) has 6404 citations while the man-sexism scale (**?**) has only 618 citations.

There is reason to believe that Leslie et al.'s reasoning is flawed, beyond the findings of Templer & Tomeo. Sex differences in systemization capacity and interest appear in infancy and persist to adulthood (Baron-Cohen, 2005), indicating that women should be less likely to earn high math GRE scores and STEM graduate degrees in the absence of any discrimination. More than 99 % of geniuses who have ever lived have been male (Murray, 2003). Only one woman, Judit Polgár, has ever beaten a world champion in chess, and all world champions have been men. Women systematically under-perform men in chess, requiring their own leagues, while "male" leagues are open to women. Chess achievement titles like "grand-master" are also set at 200 ELO points lower for women than for men. This is likely related to the fact that women have lower IQs then men on the order of 3 to 5 points (Nyborg, 2015). Male IQ also has larger variance, males having a standard deviation of 15 vs. females having a standard deviation of about 13.5, leading to more male individuals on the right tail of performance, which is where graduate students are recruited from (Shields, 1982).

Given white female IQ is distributed on $N(97, 13.5)$, and white male IQ is distributed on $N(100, 15)$, we would expect over 80 % of white people over at an IQ of 140 to be men and over 70 % of white people over an IQ of 130 to be men. Assuming the mean high achiever has an IQ in this range, with no personality differences whatsoever, men would account for about 75 % of high achievers despite being only half the population. This is somewhat less men than actually observed, but personality differences likely account for the rest.

While stereotype threat has long been advocated as an explanation for why women perform worse in math, (Sackett et al., 2004) this has recently failed large-scale replication and a meta-analysis indicated substantial publication bias (Flore & Wicherts, 2015; Flore et al., 2018). In 2014, the average male received a 154.3 on the math GRE while the average female received only a 149.4. The standard deviations were 8.6 and 8.1 respectively (ETS, 2014). Assuming the scores are normally distributed, 19 % of males and only 6 % of females meet the average GRE score required to be admitted to a mathematics PhD program (162). This means we should expect about 29 % of math PhD earners to be women, which is nearly what is observed. If there is bias, it is likely mediated through women having lower math GRE scores, not by PhD supervisors directly thinking of women as less brilliant for no rational reason.

In addition, it is a robust finding that stereotypes tend to be based on accurate underlying observations ($0.94 \leq r \leq 0.98$ for occupation stereotypes, for instance) (Jussim et al., 2009). Gender stereotypes in particular have long been found to be closely related to verifiable observations (Hall & Carter, 1999; Kirkegaard & Gerritsen, 2021). Gender ratios in occupations have recently been found to correlate with occupation gender stereotype with $r = 0.50$. GPT-2 accurately learns the proportion of women in occupations through its language training set (Kirk et al., 2021) while teachers accurately track the development the math skill gap between the sexes through elementary and middle school (Robinson & Lubienski, 2011). Also recently, it has been found that the more gender-equal a culture is, the more likely are gender gaps in STEM fields, contradicting the expectation that women are absent from some fields due to cultural and legal barriers. (Stoet & Geary, 2018). It has been suggested that this is because freer women are more likely to choose to pursue careers where they excel relative to men, such as childcare and nursing, as women tend to be better empathizers and care-takers (Baron-Cohen, 2005). It is also possible that sexual dimorphism varies by race, with northern European populations having larger dimorphism across many metrics.

Based on these broad, robust findings in the literature, we predict that upon further examination, Leslie et al.'s hypothesis will falter. In this study, we wish to test the accuracy of stereotypes regarding different academic fields. We also wish to investigate the link between racial and gender diversity, merit by field as measured by GRE scores, and stereotypes of fields. We expect to find that GRE is the best predictor of demographics and that GRE scores and associated demographics drive accurate stereotypes.

2 Methods and Hypotheses

We formulated the following hypotheses:

1. GRE scores are superior predictors of the proportion of women in a field relative to stereotypes.
2. Stereotypes of fields are generally accurate, being significantly associated with metrics in the expected directions. These include stereotypes sex representation, scientificness, and politics.

To test these hypotheses, we conducted an original survey of PhD holders on the surveying site Prolific, asking each participant to rate each field in our list by perceived scientificness and political conservativeness ($n = 507$). Participants were simply given a list of fields and asked to rate them in these domains with likert scales. In addition, we compiled various data from other numerous other sources. From Leslie et al. (2015), we collected "brilliance" stereotype ratings (Leslie et al., 2015). From the National Center for Science and Engineering Statistics we retrieved field demographics, both racial and sexual, and income (*Survey of Earned Doctorates*, 2024). From the ETS we retrieved mean GRE (ETS, 2014), both verbal and mathematical, by field. We also retrieved Republican by field based on voter registration data from (Langbert, 2018b) (also see author correction (Langbert, 2018a)).

Data was arranged by field and correlations between the metrics were computed. We expect to see that 1) GRE scores correlate more with percent female than brilliance; 2) percent female correlates negatively with percent Republican; and 3) the best predictor of the conservativeness stereotype is the proportion of Republicans in a field.

Additionally, we attempted to replicate Leslie et al. (2015)'s brilliance stereotype. Whereas Leslie et al. asked academics four questions about their own disciplines ("Being a top scholar of [discipline] requires a special aptitude that just can't be taught.", "If you want to succeed in [discipline], hard work alone just won't cut it; you need to have an innate gift or talent.", "With the right amount of effort and dedication, anyone can become a top scholar in [discipline].", "When it comes to [discipline], the most important factors for success are motivation and sustained effort; raw ability is secondary."), we asked 98 doctorate holders to rate the following question for 55 disciplines: "talent is needed for success is [discipline]." Responses were either strongly disagree, somewhat disagree, neutral, somewhat agree, or strongly agree.

3 Distributions of Metrics by Field

The percent of women in each field varied between just under 20 % in Physics to more than 80 % in nursing. Subjects with that are widely considered to feature more math tended to have less women.

Not all data was available for every field. Still, it is interesting to see that white people are more common in humanities fields like history, philosophy, and English literature. Asians tend to decrease the white proportion in quantitative fields, and blacks in social sciences.

"Brilliance" indicates the data from Leslie et al. Philosophy was oddly considered the most brilliant, with English literature not far behind.

"Our brilliance" is our attempted replication of Leslie et al.'s metric, with a slightly changed question and sampling methodology. Our results did not significantly correlate with Leslie et al.'s (Figure 9). Our brilliance score was more significantly associated with GRE Math ($r = 0.49$, $p < 0.01$ vs. $r = 0.48$, $p > 0.01$), and the association with percent female was weaker ($r = -0.39$ vs $r = -0.65$). The association with percent Black was about the same, but our result was significant while Leslie et al.'s was not ($r = -0.51$). Furthermore, our inter-rater reliability was very high, with ICC1k $= 0.92, p < 0.001$.

The difference between our results and those of Leslie et al. might be due to differences between the brilliance perceptions of the doctorate holding public, and academic researchers of their own disciplines.

If the difference is simply that between the brilliance perceptions of a field's own practitioners, and those of the highly educated general public, then fields that are below the dotted line in Figure 5 rate themselves higher than the doctorate holding public does. Philosophy, for instance, believe they are the most brilliant of any field, but the doctorate holding public thinks that they are only as brilliant as finance or education. Meanwhile, physicists are almost perfectly self-aware. Neuroscience is humble and rates itself lower than the highly educated general public.

Corporate subjects and religion were seen as the most conservative (Figure 6). The quantitative subjects were the most conservative among those with data available (Figure 7). Relatively non-quantitative subjects, chemistry,

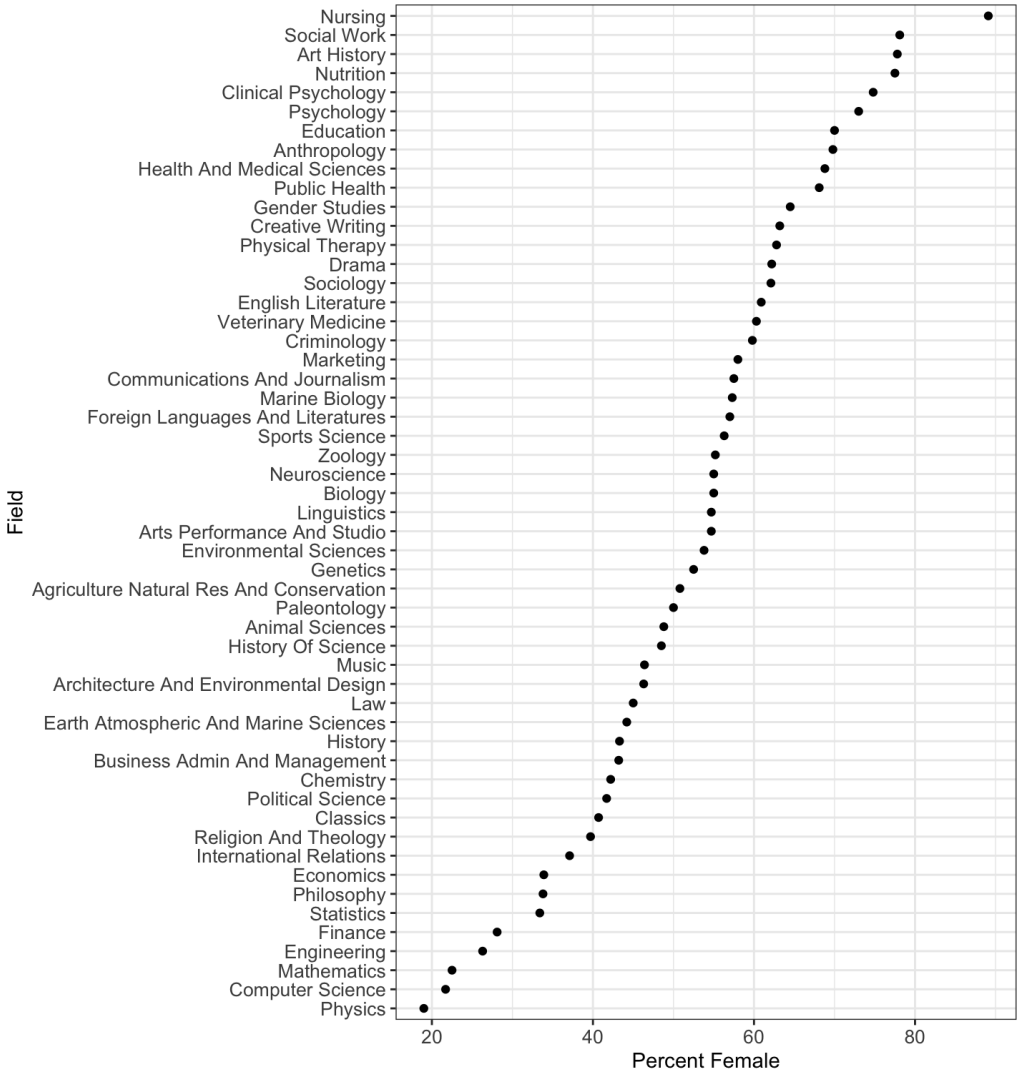

Figure 1: Percent female per field.

biology, and neuroscience, topped the list for scientificness (Figure 8). Finally, the quantitative subjects, physics, statistics, mathematics, and economics had the highest GRE scores (Figure 9).

4 Results

4.1 Women, GRE, and Brilliance

Each hypothesis was vindicated by our analysis. Both mean GRE score and mean GRE math score were better predictors of the proportion of women in a field than "brilliance" stereotype. In particular, GRE math score was the best predictor of the proportion of women in a field by far, with $r^2_{GRE-M} = 62.4$ % and $r^2_{Brilliance} = 42.2$ % ($p = 0.0477$), per Figure 10. Additionally, mean field GRE and GRE math scores significantly predicted field brilliance stereotypes.

As the table shows, GRE Math is a much stronger predictor of percent female by field than the brilliance stereotype is. Consequently, we take this as evidence that female representation in different fields is largely driven by female performance on the GRE math test, which is predicted to be less by studies of female IQ and systematizing abilities, as discussed in the introduction.

A model that predicts percent female by mean math GRE predicts 20 % more of the variance than a model that predicts percent female by brilliance stereotype. When our mean math GRE metric is added to Leslie et al's model, r^2 goes up by 33 %; when Leslie et al's brilliance stereotype metric is added to our mean math GRE

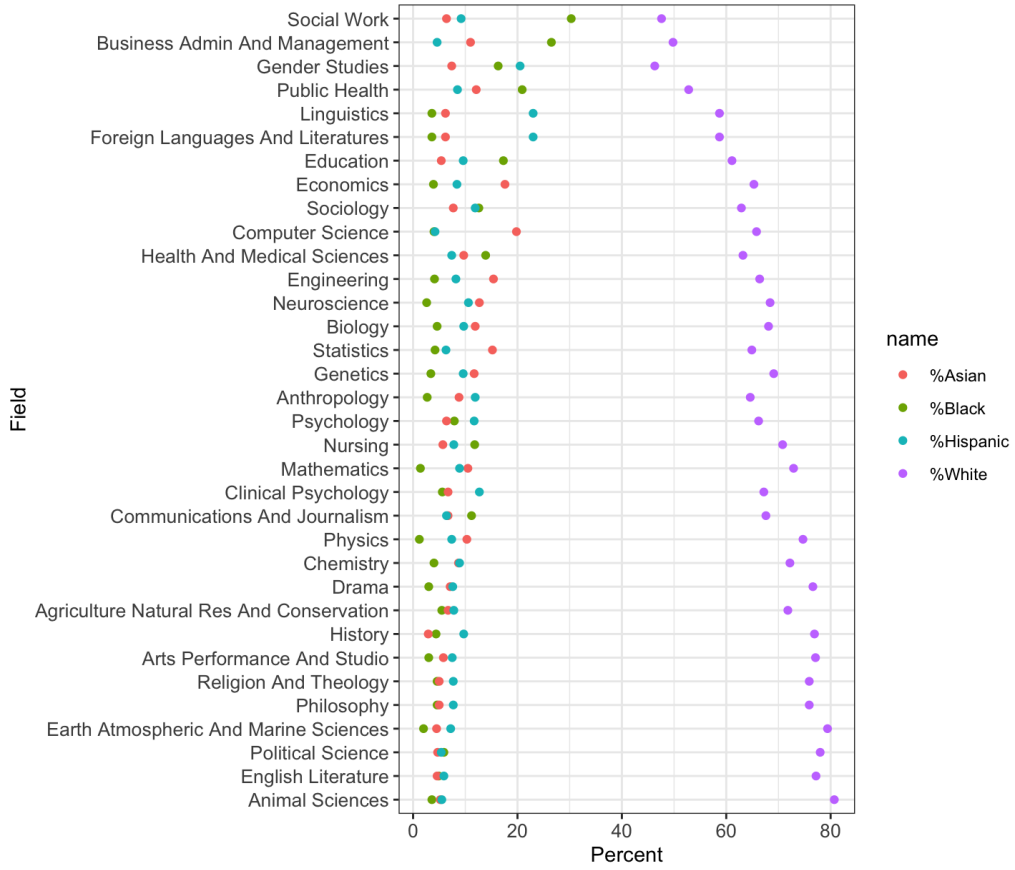

Figure 2: Percent of each race per field.

Table 1: Regression Models predicting Percent Female. Format is $\beta(SE, p)$. All variables are standardized.

Term	Model 1	Model 2	Model 3	Model 4
(Intercept)	-0.077 (0.114, 0.506)	0.009 (0.085, 0.919)	-0.410 (0.175, 0.028*)	-0.082 (0.138, 0.557)
GRE Math	-0.658 (0.120, < 0.001***)	-0.797 (0.091, < 0.001***)		-0.655 (0.130, < 0.001***)
GRE Verbal		-0.103 (0.091, 0.264)	0.253 (0.193, 0.203)	0.009 (0.143, 0.946)
Brilliance	-0.353 (0.124, 0.009*)		-0.735 (0.165, 0.009*)	-0.356 (0.137, 0.016*)
Model Fit	$R^2 = 0.75$	$R^2 = 0.63$	$R^2 = 0.46$	$R^2 = 0.75$

metric, r^2 only goes up 12 %. Another way of thinking about this is that in the joint GRE math - brilliance model, the math GRE metric uniquely captures 33 % of the variance of the field's proportion of females, while the brilliance stereotype uniquely captures only 12 % of the variance.

Table 2: Sequential ANOVA of Brilliance and GRE Math.

	p
Brilliance	< 0.001***
GRE Math	< 0.001***

Additionally, sequential ANOVA (Table 2) reveals that the hypothesis that the effect of ability (math GRE) was entirely mediated by perceived ability (Brilliance) can be rejected at $p < 0.0001$.

Meanwhile, the hypothesis that perceived ability (brilliance) is entirely mediated by mathematical ability (math GRE) and temperamental abilities associated with white people that mediate genius cannot be rejected ($p = 7.42$ %), although the p-value is borderline, and this test was done ad-hoc, meaning this evidence should be considered tentatively (Table 3). Still, one fact that defends this test is that historical, more than 95 % of

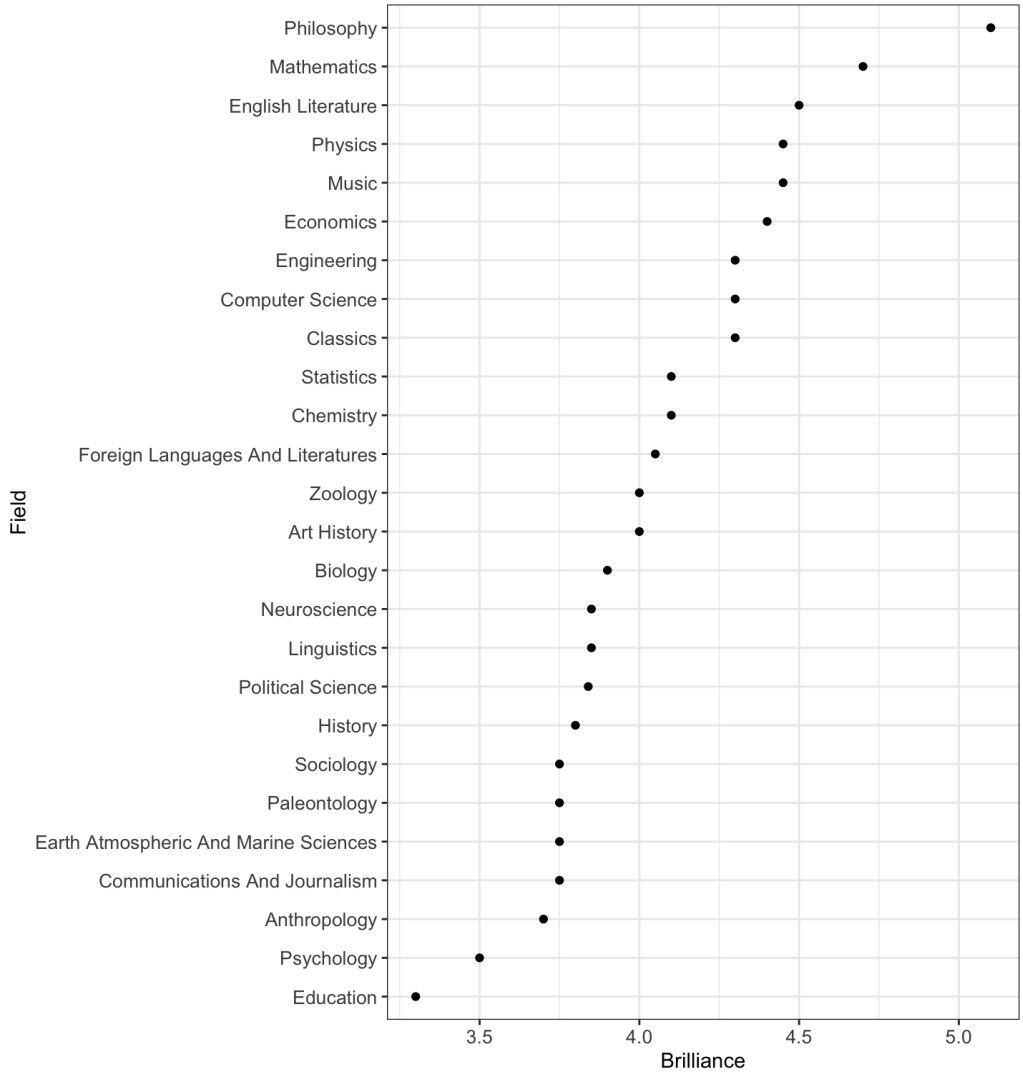

Figure 3: Brilliance stereotype by field. Score represents averaged Likert scale.

Table 3: Sequential ANOVA of GRE Math, Percent White and Brilliance.

	p
GRE Math	$< 0.001^{***}$
%White	0.002^{**}
Brilliance	0.074

geniuses have been white (Murray, 2003). For discussion on temperamental traits associated with white people and their potential role in explaining the achievement gap between European and Asian countries, in the context of high Asian IQ, see Murray (2003).

These data show that 1) a model that predicts the proportion of women in a field with mathematical talent as measured by the math GRE is quantitatively superior to a model that uses the brilliance stereotype by a large, significant margin; 2) the marginal improvement in loss derived from adding the talent metric to the stereotype model is significantly larger than the marginal improvement in loss derived from adding the brilliance metric to the talent model; 3) the idea that "across the academic spectrum, women are underrepresented in fields whose practitioners believe raw, innate that talent is the main requirement for success, *because women are stereotyped as not possessing such talent*" has been falsified by sequential ANOVA. Based on these data, a math GRE discrepancy must be part of the picture, a larger part than stereotyping at that; and 4) a model where brilliance stereotype is entirely mediated by talents associated with the GRE Math test and the presence of

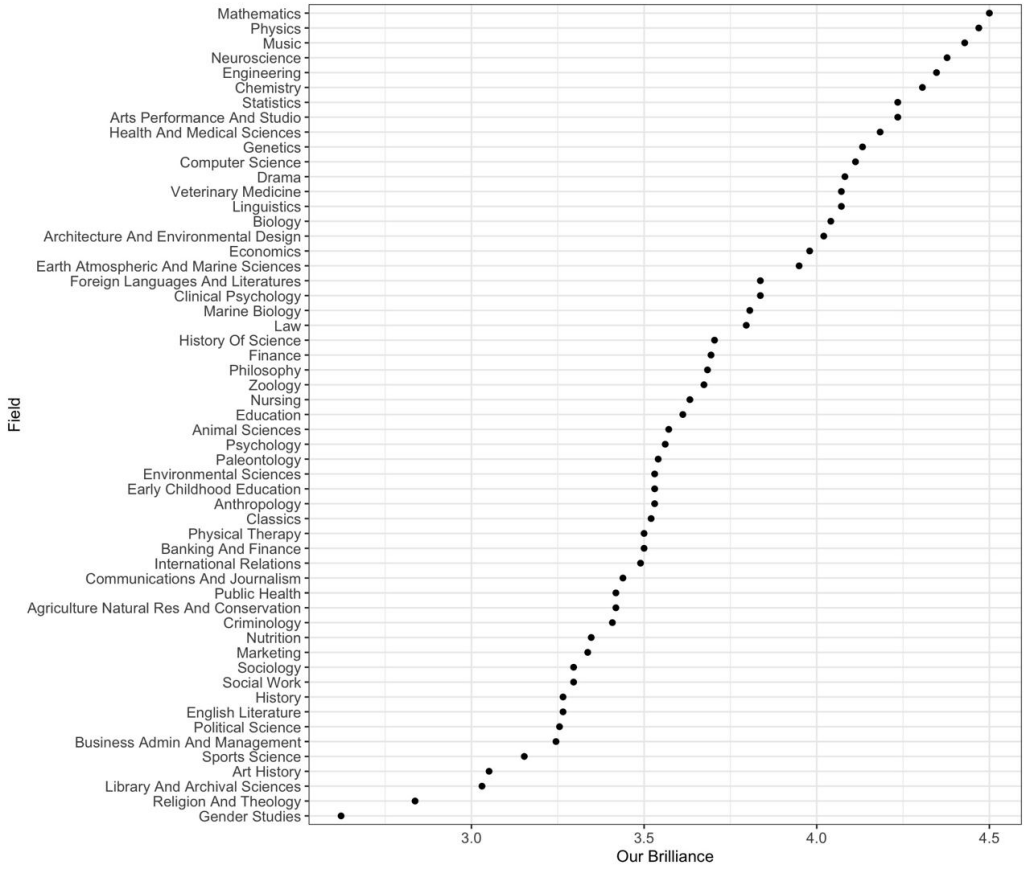

Figure 4: Our brilliance stereotype by field. Score represents averaged Likert scale.

white people in a field cannot be rejected. In such a model, the brilliance stereotype is a basically accurate representation of the amount of mathematical and temperamental talent that is in a field. Women are lower in this talent, and so both measuring the talents directly or using the stereotype predict the number of women in a field, and therefore when these talents have been accounted for by math GRE and the proportion of white people, the stereotype fails to add to the model.

4.2 Blacks, GRE, and Brilliance

Leslie et al. make the same claim regarding the proportion of Black people in academic fields: "We hypothesize that, across the academic spectrum, women are underrepresented in fields whose practitioners believe raw, innate talent is the main requirement for success, because women are stereotyped as not possessing such talent. This hypothesis extends to African Americans' underrepresentation as well, as this group is subject to similar stereotypes." Since Black people may face discrimination even in a world where women do not, we will perform the same analyses to see if Leslie et al.'s hypothesis is any sturdier regarding race.

Social work and Business and Management are large outliers, potentially reflecting differences in interests between demographics (Schmitt et al., 2011).

The first thing to notice is that for percent Black, $r^2_{Brilliance}$ = 28 % and r^2_{GRE} = 24 %. The difference is not statistically significant ($p = 0.413$).

A lot of the variation that brilliance accounts for lines up with the variance that field GRE accounts for (Table 4).

GRE-Math and Brilliance are similar predictors of the percent of Black people in a field (Table 6).

Neither the hypothesis that brilliance is totally mediated by talent as measured by the GRE nor the hypothesis that GRE is totally mediated by brilliance stereotype can be rejected based on sequential ANOVA (Table 5). While these data are less revealing than those regarding female representation, the Black-White IQ gap is a well documented, robust finding, Rushton (2012) showing the presence of a roughly 1 SD gap in cognitive

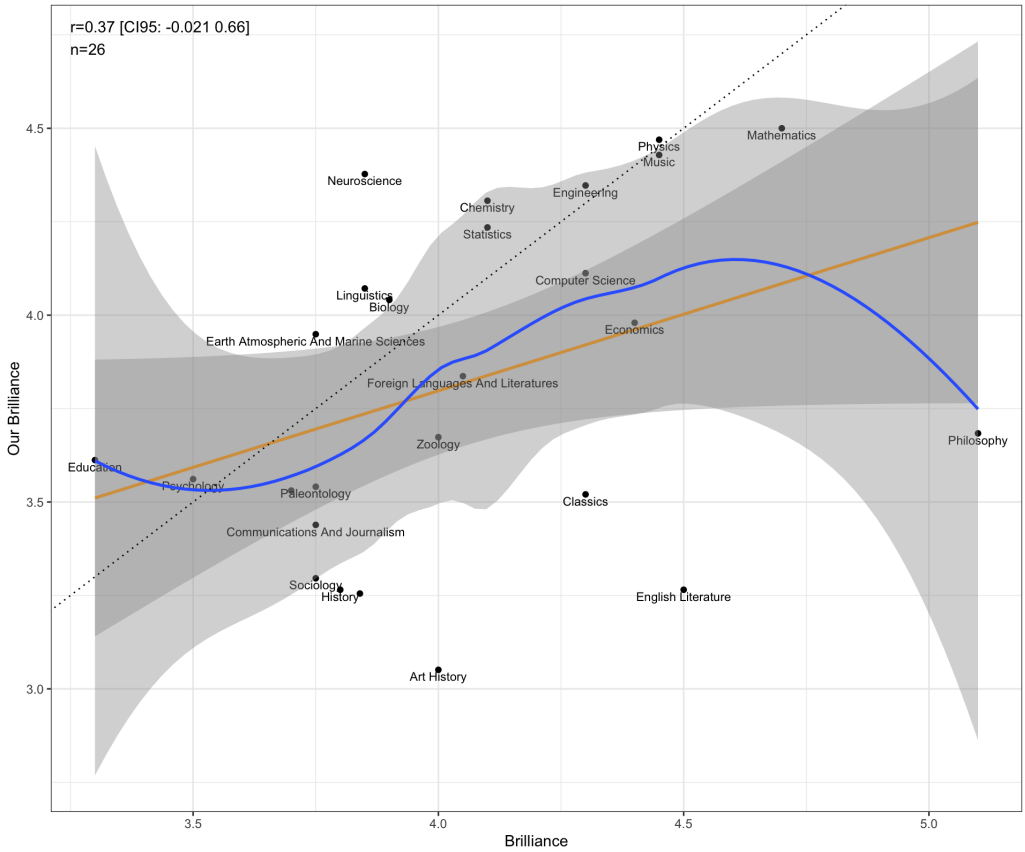

Figure 5: Brilliance ratings of own field vs. brilliance rating among doctorate holders.

Table 4: Regression Model on Percent Black Predicting with Brilliance and GRE.

	Est.	S.E.	p
(Intercept)	-0.232	0.136	0.106
GRE	-0.210	0.181	0.262
Brilliance	-0.162	0.137	0.254
Model Fit	$R^2 = 0.33$	F(2,18) = 4.42	p = 0.03

Table 5: Sequential ANOVA of GRE and Brilliance for Percent Black.

	p
GRE	0.014*
Brilliance	0.254
Brilliance	0.013*
GRE	0.262

Table 6: Type II ANOVA Predicting Percent Black with Brilliance, GRE-M, and GRE-V.

	ϵ^2	β	p
GRE Math	0.006	0.077	0.305
GRE Verbal	-0.018	0.000	0.434
Brilliance	0.015	0.122	0.268

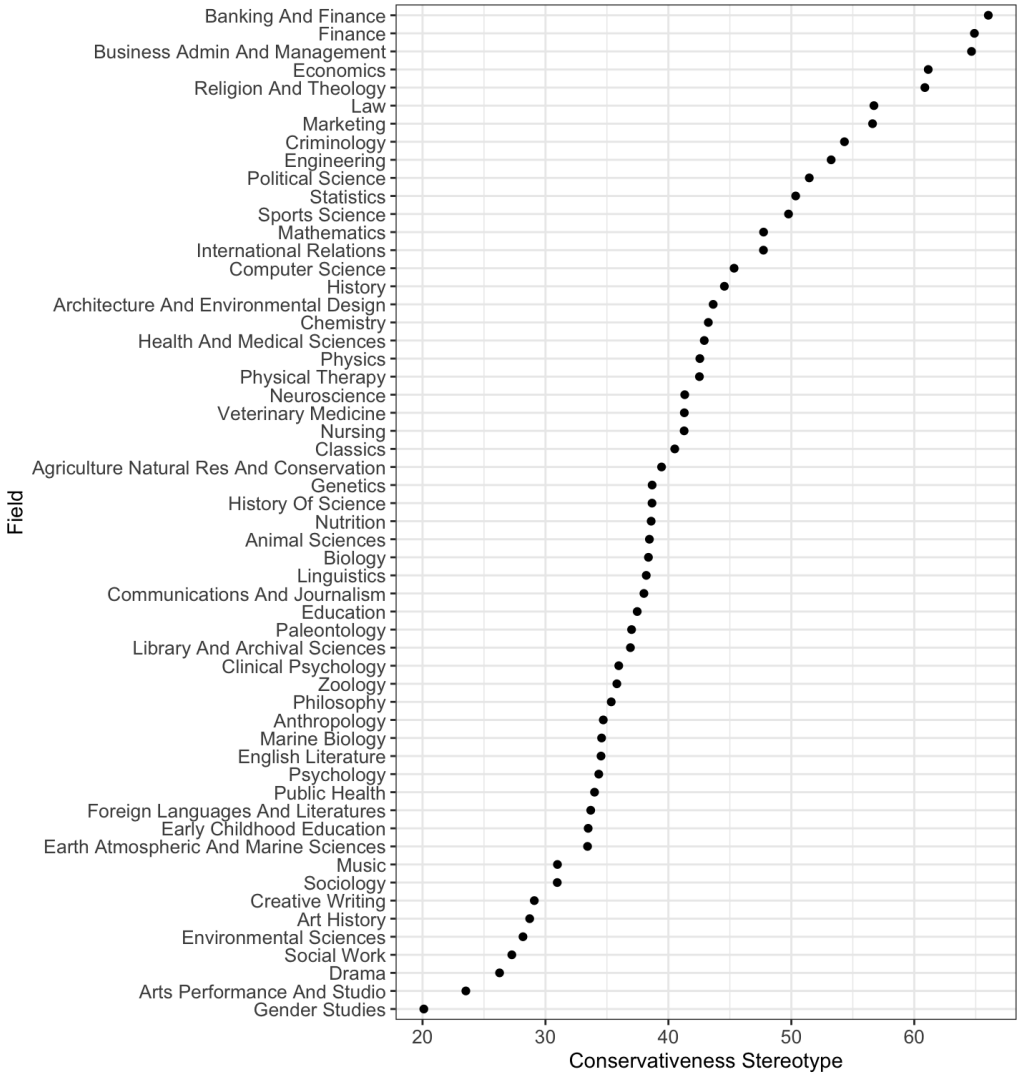

Figure 6: Conservativeness stereotype by field.

performance between US Blacks and Whites, favoring Whites. The GRE is a g-loaded test (Templer & Tomeo, 2002), and the fact that it is essentially replaceable with the brilliance metric, in the context of the Black-White IQ gap, is indicative that the source of differences in percent Black by field probably has to do with average Black performance on cognitive tasks, including intelligence tests like the GRE. More than this, generally speaking, degrees lead to jobs, and there is data about race gaps in various fields of occupation. These tend to follow the expected pattern, with higher IQ jobs have fewer Black people (Pesta & Poznanski, 2016).

4.3 Stereotype Accuracy

Our data is consistent with an accurate-stereotypes model. GRE was highly predictive of "brilliance", GRE math tilt was moderately predictive of scientificness ($r = 0.36$) while GRE verbal was moderately, negatively ($r = -0.28$) predictive of scientificness, suggesting that survey respondents are thinking quantitative subjects wherein rhetorical skills are of low importances, and percent Republican was highly predictive ($r = 0.55$) of conservativeness stereotype.

This matches previous findings regarding the general accuracy of stereotypes, and casts doubt on any model that suggests that the root cause of disparities are inaccurate stereotypes. Stereotypes may be associated with disparities, but this is, prior to any new observations given the current state of evidence, probably because the stereotypes reflect underlying discrepancies in population attributes. Brilliance reflects underlying cognitive test performance, scientificness reflects quantitative tilt, and conservativeness reflects the actual voting patterns of field practitioners.

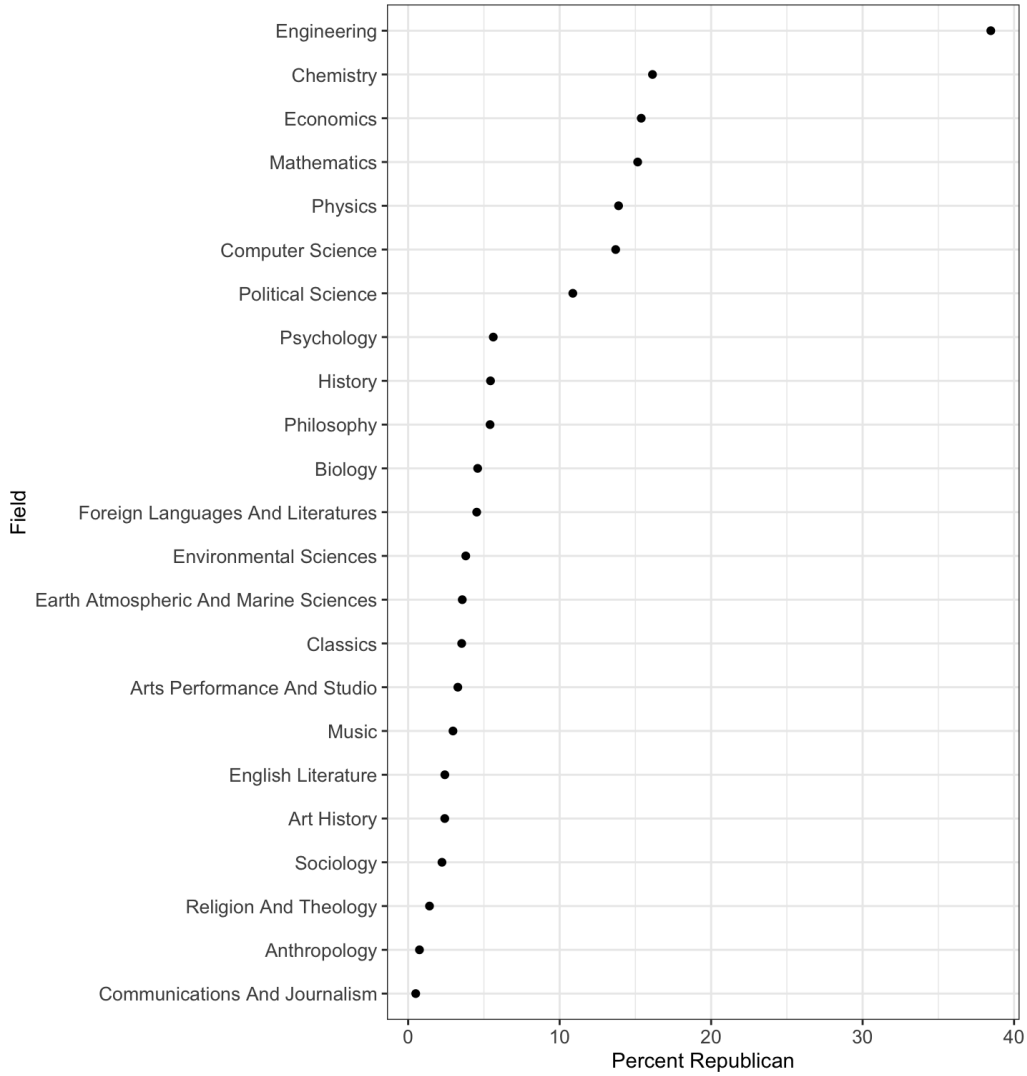

Figure 7: Percent republican by field.

The stereotype is systematically biased towards higher estimates than are accurate, towards believing that academic fields have more Republican professors than they actually have. In other words, there is a severe underestimation of how left-wing academia is. However, it is better at rank-ordering. Our data show that people have some decent idea who is who more Republican, but their estimates are off the mark. The public seems to not fully understand how left-wing academia is.

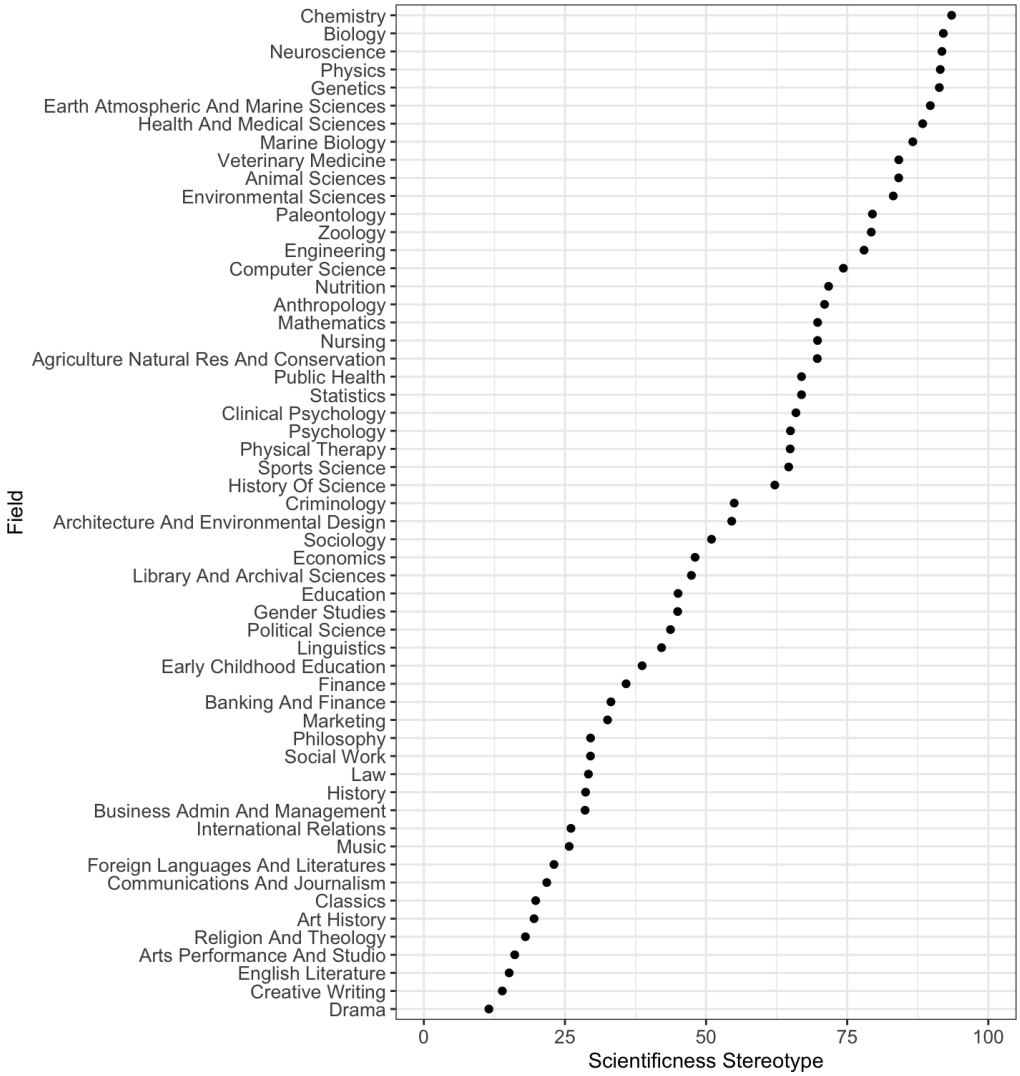

Figure 8: Scientificness stereotype by field.

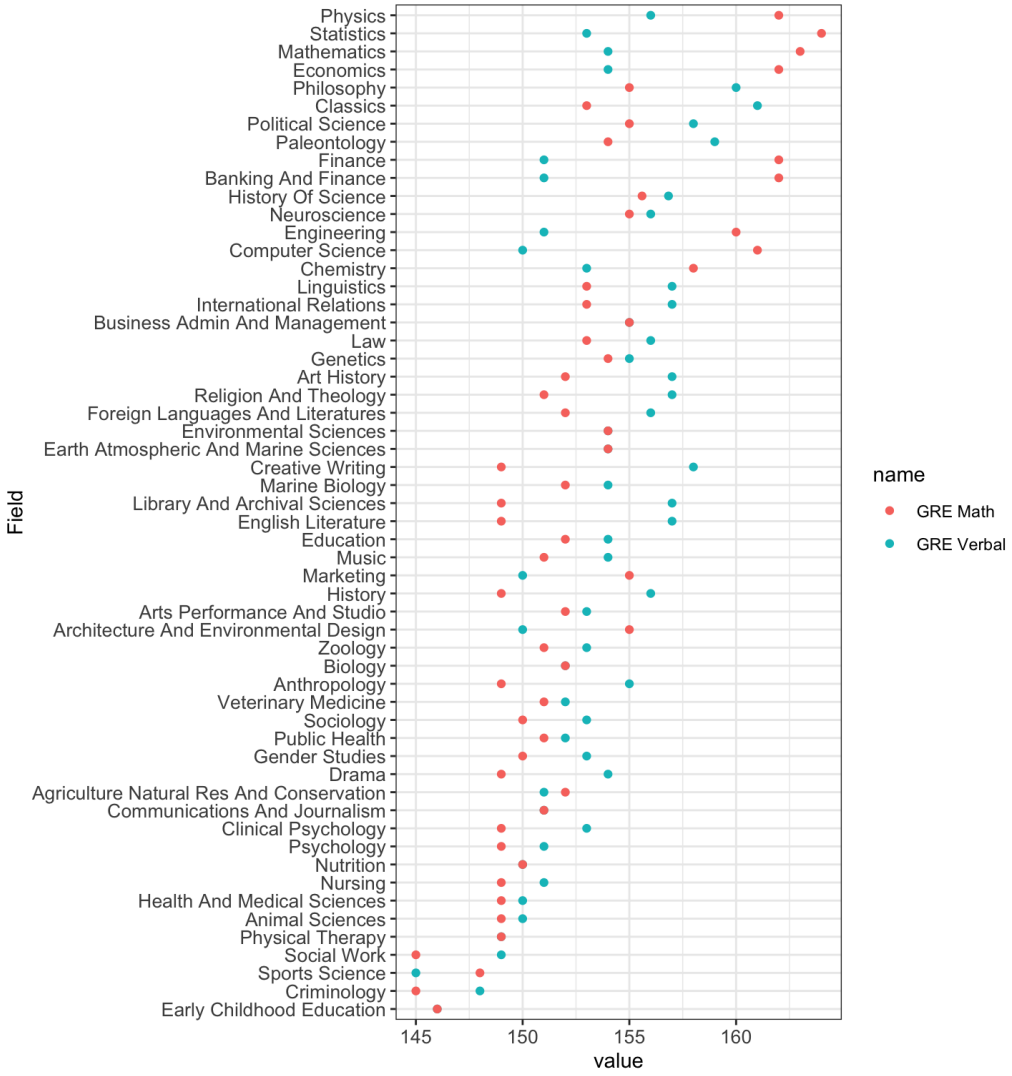

Figure 9: GRE by field.

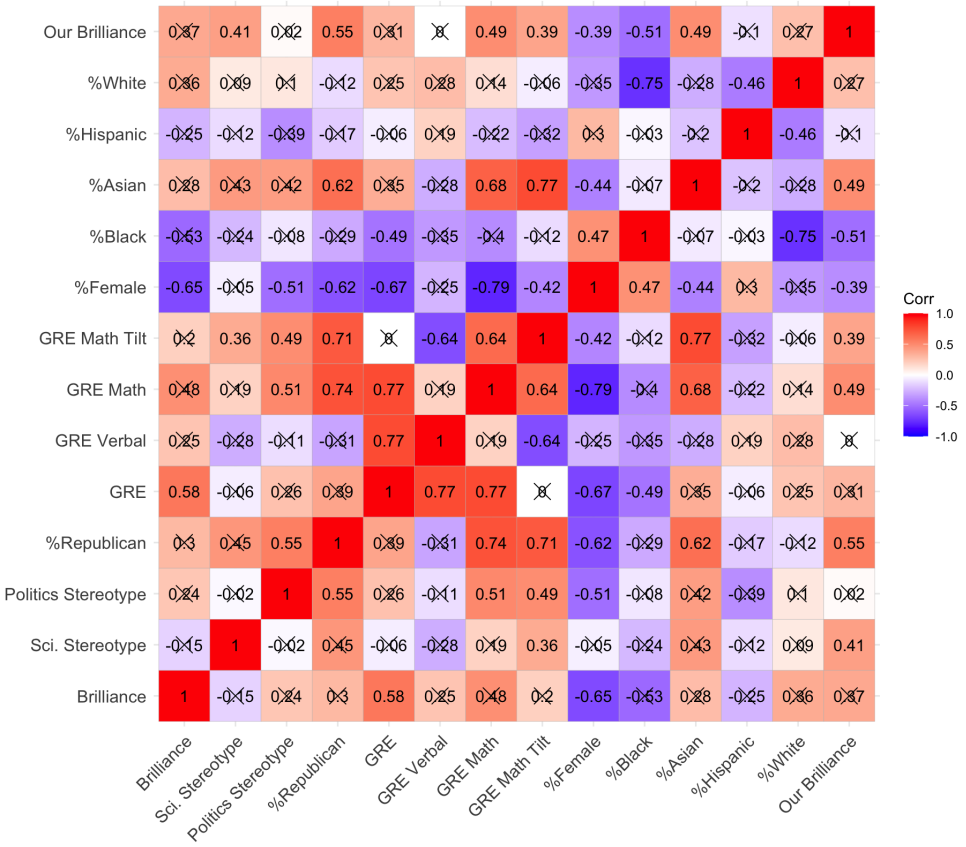

Figure 10: Correlation Matrix of Metrics. Crossed-out coefficients are statistically insignificant ($p > 0.01$).

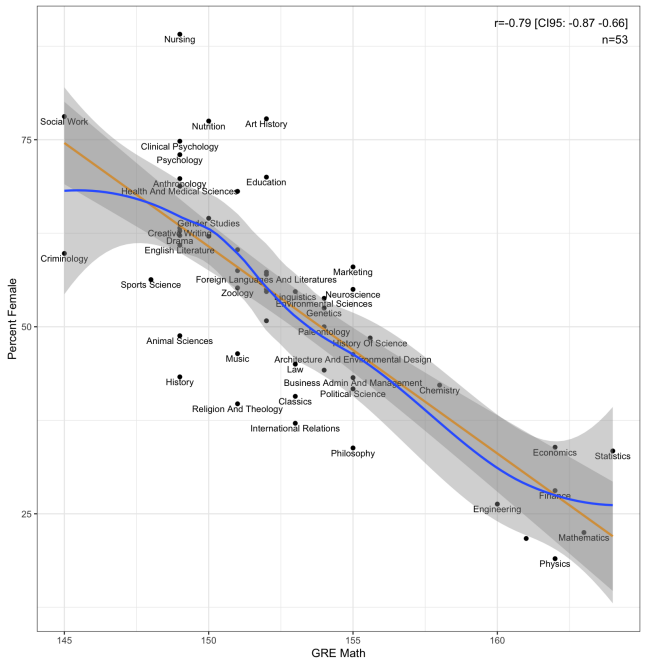

Figure 11: Percent female by field and average GRE Math score. Blue line is LOESS regression.

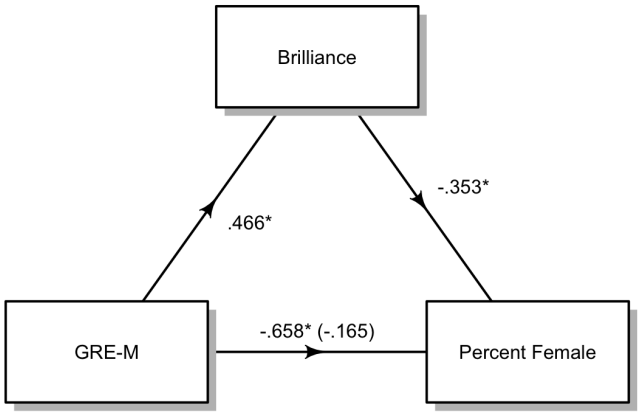

Figure 12: Path model with GRE-M, Brilliance, and Percent Female. -0.165 is. the indirect effect of GRE-M on Percent Female through Brilliance.

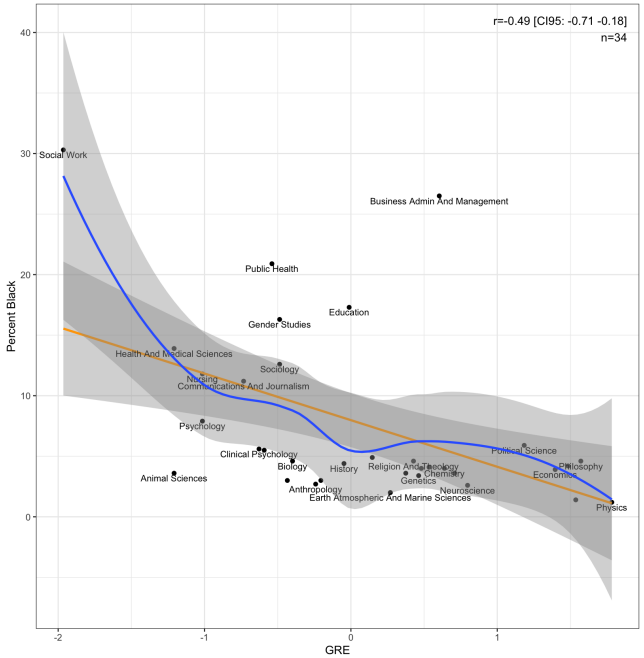

Figure 13: Percent Black by field and average GRE score.

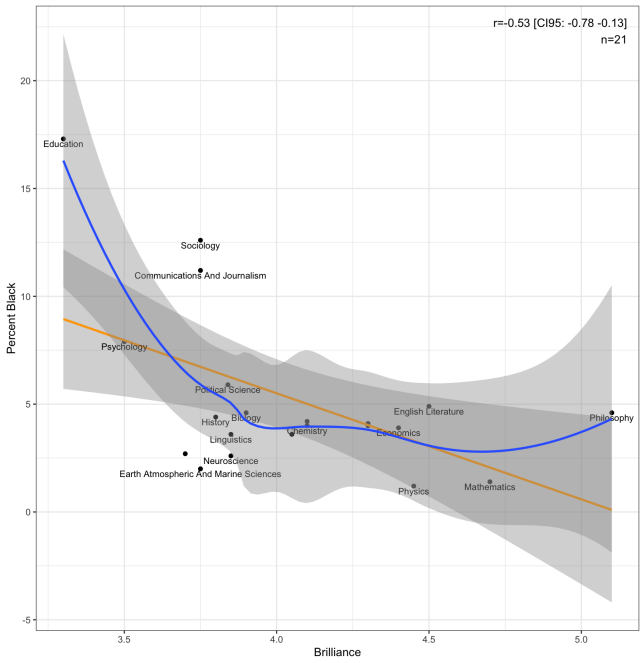

Figure 14: Percent Black by field and Brilliance stereotype.

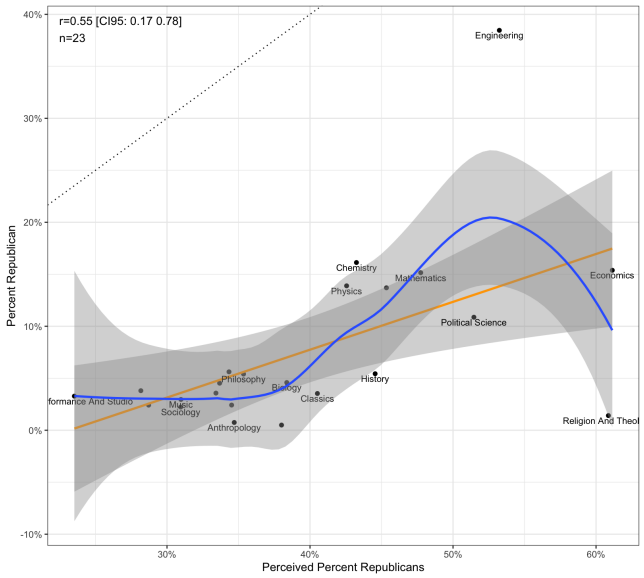

Figure 15: Scatterplot of Percent Republican and Conservativeness Stereotype. Dotted line shows where a perfect stereotype would be.

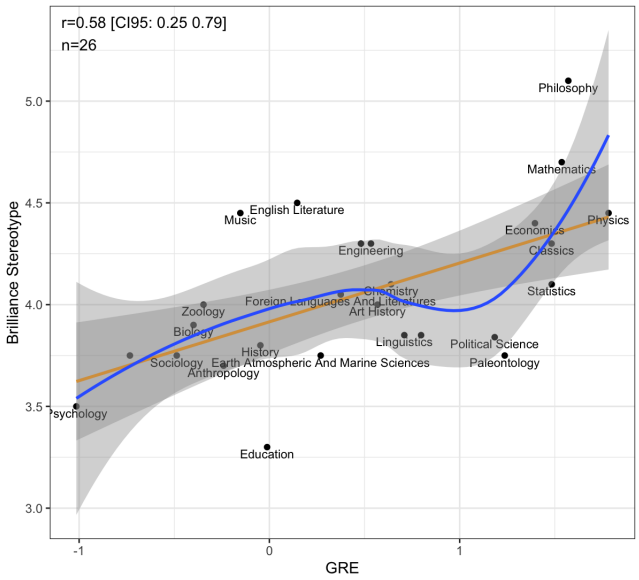

Figure 16: Scatterplot of GRE and Brilliance Stereotype.

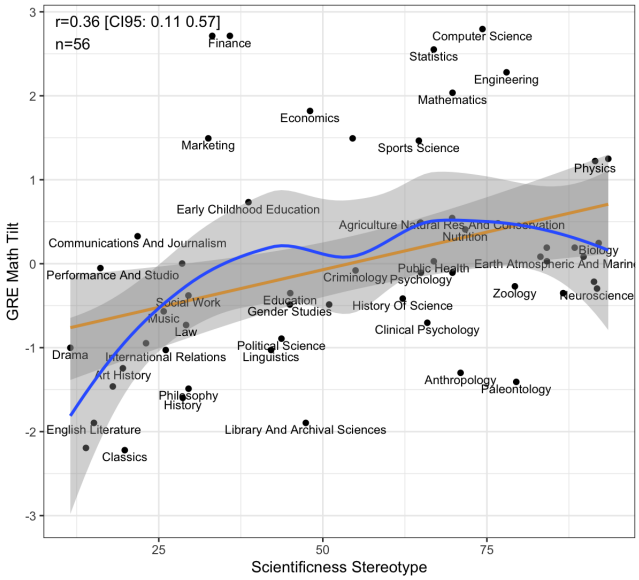

Figure 17: Scatterplot of GRE math tilt and Scientificness Stereotype.

5 Conclusion

Leslie et al. (2015)'s hypothesis that discrepancies in racial and gender diversity among various academic fields is due solely to an inaccurate, irrationally bigoted "brilliance" stereotype has been shown to have some serious issues. When it comes to predicting the proportion of women in a field, average math GRE is a superior predictor compared to the brilliance stereotype. When these metrics are combined in multi-factor models, math GRE remains the superior factor. This throws serious doubt on the idea that gender diversity discrepancies result from baseless stereotypes.

On top of this, the brilliance stereotype is able to be predicted by a field's average GRE score at $r = 0.56$. This, along with our data on the scientificness and conservativeness stereotypes, adds to a growing literature which shows that stereotypes tend to based in reality, not wild fantasy. Brilliance stereotype predicts the proportion of women in a field because stereotypers are noticing real trends about the cognitive requirements of a field, and women score lower, on average, on cognitive tests, including the GRE, which is the main test used by academia to select PhD candidates with the right level of cognitive skill.

The story is similar when it comes to predicting the proportion of Black people in a field. GRE has been shown to be an equivalent predictor to brilliance stereotype, and ANOVA has revealed that these predictors cannot be said to capture non-overlapping variance. The American Black population is known to underperform, on average, on cognitive tests relative to the American White population; the source of racial diversity discrepancies, then, in fields requiring more "brilliance", is not a baseless stereotype but rather is going to ultimately be the same as whatever causes the Black-White IQ gap.

Our two hypotheses have been supported by this study. We have indeed found that 1) GRE scores are superior predictors of the proportion of women in a field relative to stereotype, and 2) stereotypes of fields are generally accurate, being significantly associated with metrics in the expected directions.

5.1 Limitations

A major limitation of this paper is its inability to definitively establish the direction of causality. While a model where GRE-M is mediated by Brilliance is consistent with the data (Figure 12), the data presented here cannot, on its own, falsify a model where *some* unfair stereotyping takes place. Such a model, however, is not supported by the Brilliance data (Table 3), and the literature discussed in the introduction of this paper is unsupported of such a model, instead supporting a model where female talent causes GRE-M discrepancies, which cause discrepancies in the proportion of a field that is female (Figure 12).

References

Baron-Cohen, S. (2005). The essential difference: The male and female brain. *Phi Kappa Phi Forum*, *85*(1), 23–26.

ETS. (2014). *Gre scores.* Retrieved from http://www.ets.org/s/gre/pdf/gre_guide.pdf (Accessed: Date)

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? a meta-analysis. *Journal of School Psychology*, *53*(1), 25–44.

Glick, P., & Fiske, S. T. (2018). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social cognition* (pp. 116–160). Routledge.

Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology*, *77*(2), 350.

Honeycutt, N., & Jussim, L. (2022). *Political bias in the social sciences: A critical, theoretical, and empirical review.* (Unpublished manuscript)

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). *The unbearable accuracy of stereotypes.* (Book chapter)

Kirk, H. R., et al. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in neural information processing systems* (Vol. 34, pp. 2611–2624).

Kirkegaard, E. O., & Gerritsen, A. (2021). *A study of stereotype accuracy in the netherlands: immigrant crime, occupational sex distribution, and provincial income inequality.* (Unpublished manuscript)

Langbert, M. (2018a). Author correction: Homogenous: The political affiliations of elite liberal arts college faculty. *Academic Questions*, *31*(3), 386–387.

Langbert, M. (2018b). Homogenous: The political affiliations of elite liberal arts college faculty. *Academic Questions*, *31*(2), 186–197.

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Murray, C. (2003). *Human accomplishment: The pursuit of excellence in the arts and sciences, 800 bc to 1950*. Publisher Name. (The publisher name needs to be filled in)

Nyborg, H. (2015). Sex differences across different racial ability levels: Theories of origin and societal consequences. *Intelligence*, *52*, 44–62.

Pesta, B. J., & Poznanski, P. (2016). *Putting spearman's hypothesis to work: Job iq as a predictor of employee racial composition.* (Online publication)

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302.

Rushton, J. P. (2012). *No narrowing in mean black–white iq differences—predicted by heritable g.* (Unpublished manuscript)

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for african american-white differences on cognitive tests. *American Psychologist*, *59*(1), 7.

Schmitt, N., Golubovich, J., & Leong, F. T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using big five and riasec measures. *Assessment*, *18*(4), 412–427.

Shields, S. A. (1982). The variability hypothesis: The history of a biological model of sex differences in intelligence. *Signs: Journal of Women in Culture and Society*, *7*(4), 769–797.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593.

Survey of earned doctorates. (2024). Retrieved from https://ncses.nsf.gov/pubs/nsf23300/data-tables (Retrieved from National Center for Science and Engineering Statistics, National Science Foundation)

Templer, D. I., & Tomeo, M. E. (2002). Mean graduate record examination (gre) score and gender distribution as function of academic discipline. *Personality and Individual Differences*, *32*(1), 175–179.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences*, *112*(17), 5360–5365.