THE FREQUENCY OF "BRILLIANT" AND "GENIUS" IN TEACHING EVALUATIONS PREDICTS THE REPRESENTATION OF WOMEN AND AFRICAN AMERICANS ACROSS ACADEMIA

BY

DANIEL SHAFIK STORAGE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Master's Committee:

Associate Professor Andrei Cimpian
Professor Eva Pomerantz

ABSTRACT

Because of the negative stereotypes against women's and African Americans' intellectual abilities, academic fields that prize brilliance and genius might be unwelcoming to members of these stigmatized groups. A recent nationwide survey of academics provided initial support for this possibility, insofar as the fields whose practitioners believed that natural talent is crucial for success in their field were also the fields where women and African Americans were least likely to obtain Ph.D.'s. The present study seeks to replicate this initial finding with a different, and arguably more naturalistic, measure of the extent to which brilliance and genius are prized within a field. Specifically, we measured field-by-field variability in the emphasis on these intellectual qualities by tallying college students' use of the words "brilliant" and "genius" in over 14 million reviews on RateMyProfessors.com. Consistent with prior work, this simple word count predicted both women's and African Americans' representation at the Ph.D. level across the academic spectrum: Fields where the words "brilliant" and "genius" were frequent in undergraduates' evaluations also had fewer female and African American Ph.D.'s. This relationship held even when accounting for a field's intellectual rigor (as indexed by students' average scores on the Quantitative Graduate Record Examination [GRE]), as well as several other explanations concerning group differences in representation. The fact that such a simple, naturalistic measure of a field's focus on brilliance predicted the magnitude of its gender and race gaps speaks to the tight link between ability beliefs and diversity.

Keywords: gender gaps; race gaps; stereotypes; ability beliefs

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The Frequency of "Brilliant" and "Genius" in Teaching Evaluations Predicts the Representation of Women and African Americans across Academia

Why are some academic fields more diverse than others? One possible factor may be that the disciplines in which women and African Americans are underrepresented (e.g., physics, philosophy) are those disciplines whose practitioners believe that a spark of brilliance is required for success. The belief in the importance of untutored genius may make these fields unwelcoming to women and African Americans because current cultural stereotypes portray members of these groups as relatively unlikely to possess genius [1, 2]. Consistent with this hypothesis, a recent survey of academics across 30 disciplines found an inverse relationship between a field's focus on brilliance and its diversity [3, 4; see also 5]. In the present research, we tested this predicted relationship using a different, and arguably more naturalistic, measure of the extent to which a field values brilliance and genius.

Rather than relying on survey methodologies, as in prior work [3, 5], here we measured a field's emphasis on brilliance by analyzing the language used in course reviews on the popular website RateMyProfessors.com. In particular, we tallied the frequency with which college students taking courses in a particular field spontaneously commented on whether their professors were "brilliant" and a "genius." Common use of these terms within a field signals that students taking courses in that field routinely evaluate its members on their intellectual prowess, which might in turn suggest that the field as a whole values this trait. Thus, this simple word count derived from students' anonymous online evaluations can serve as a naturalistic proxy for

a field's emphasis on raw intellectual talent, which in prior work was assessed with survey questions about what is required for success [3, 5]. If this word count is indeed reflective of a field's ability beliefs, it should also (inversely) predict whether women and African Americans pursue advanced degrees in that field. Below, we detail this and other specific predictions that follow from the hypothesized link between a field's brilliance focus and its (lack of) diversity.

Predictions. First, we predict that college students' online reviews will mention the terms "brilliant" and "genius" more often in fields whose members also value these intellectual traits, as determined by Leslie, Cimpian, et al.'s survey of academics [3]. That is, we expect that our naturalistic, language-based measure of college students' beliefs will align with the explicit, survey-based measure administered to academics. Such a result would validate our assumption that the language used by college students in their online reviews captures to a significant extent a field's focus on brilliance and genius.

Second, the frequency of "brilliant" and "genius" in students' evaluations should predict women's and African Americans' likelihood of pursuing a field, operationalized here as Ph.D. attainment: the more brilliance-oriented language, the less diverse the Ph.D.'s. At the same time, however, the career aspirations of groups who are not stereotyped as lacking brilliance, such as Asian Americans, should be unrelated to a field's emphasis on brilliance. Thus, we also predict no significant relationship between the amount of brilliance language within a field and Asian Americans' Ph.D. attainment.

Rather than simply looking at the raw relationship between these variables, we will also compare the predictive power of our linguistic measure of a field's brilliance focus against the available data on several alternative hypotheses concerning diversity in science and beyond. (Because these data are drawn primarily from Leslie, Cimpian, et al. [3], whose main focus was

gender diversity, many of the alternatives concern women's representation specifically.) One such hypothesis suggests that women are underrepresented in fields that require long hours [6]. Another hypothesis suggests that women are underrepresented in fields that privilege thinking systematically and abstractly (termed *systemizing*) over reasoning intuitively about thoughts and emotions (termed *empathizing*; e.g., [7]). A third competing hypothesis suggests that women are less likely than men to possess extreme intellectual ability and are thus underrepresented in fields that are extremely selective [8, but see 9, 10]. Notably, a similar hypothesis has been put forward to explain African Americans' underrepresentation as well [11, but see 12]. The final alternative is that women and African Americans are underrepresented in fields that are math-intensive, as measured by their applicants' Quantitative Graduate Record Exam (GRE) scores (for evidence of gender and race gaps in mathematics, see [13, 14]). Contrary to these alternatives, we expect that use of "brilliant" and "genius" in online evaluations will predict the field-by-field variability in Ph.D. diversity above and beyond these other measures.

Finally, we expect that the superlative language used in online evaluations will be particularly predictive of gender and race gaps *when it pertains to intellectual ability*. Other superlatives should not have the same predictive relationship with diversity. To test this idea, we will compare the terms "brilliant" and "genius" with the similarly positive terms "excellent" and "amazing." Finding that use of these other superlatives does not predict the underrepresentation of stigmatized groups would highlight a field's tendency to idolize brilliance as a potential influence on its diversity.

CHAPTER 2

METHODS

Data on Ph.D. Representation. The dependent variables in this study were the proportions of female, African American, and Asian American Ph.D.'s in a field, as determined by the National Science Foundation [15].[1]

Brilliance Language Measure. The main independent variable—our new language-based measure of a field's emphasis on raw intellectual talent—was calculated using the online Gendered Language Tool [16], which reports the number of uses of any given word per million words in RateMyProfessors.com reviews. More precisely, the tool reports a word's frequency in each of 25 fields, separately for reviews of male and female instructors (see Figure 1). The tool searches over 14 million reviews from hundreds of different colleges and universities. The top three contributors to RateMyProfessors.com (and thus to the frequencies reported by the Gendered Language Tool) are the University of Central Florida, Miami Dade College, and San Diego State University. The data collected specifically for this study (namely, the word counts from the Gendered Language Tool) are completely anonymous and publicly available. Thus, the process of collecting them was exempt from review by an ethics committee.

We computed a *brilliance language score* for each discipline by (1) standardizing the frequencies of the words "brilliant" and (separately) "genius" for male and female instructors across the fields (which resulted in two *z*-scored variables, one for "brilliant" and one for "genius"), and then (2) averaging male and female instructors' standardized scores for "brilliant" and "genius" within each field (4 scores) to derive a single number—the field's brilliance language score.

[1] These data are not broken down by gender × race and thus cannot be used to investigate the intersection of these dimensions.

The words “brilliant” and “genius” were chosen because they map most directly onto the intellectual traits that are prized in fields such as mathematics, physics, philosophy, etc. [3]. We found the same results, however, when we included the weaker term “smart” in the set of words denoting a brilliance focus. Thus, our results do not hinge on a particular configuration of search terms. Finally, it is worth noting that other terms were considered but could not ultimately be used because they appeared very infrequently in the reviews. For example, “gifted” was only used an average of 5.81 times per million words, as compared with 75.10 for “brilliant” and 27.27 for “genius.”

We should point out that, because the brilliance language score is an average of male and female instructors’ separate averages, it weights the two gender-specific scores equally, and it is thus not influenced by whether there are more male or female instructors in a field. As a result, any relationships we identify between this score and women’s representation at the Ph.D. level are not trivial—they are not simply the artifacts of correlating two different measures of gender diversity.

The same algorithm was used to construct the composite usage score for the control superlatives “excellent” and “amazing,” which were selected because they were roughly matched in intensity with the focal terms “brilliant” and “genius” (all being very positive) and were also used relatively frequently by students. However, similar results were found for analogous, but less frequent, control superlatives such as “fantastic” and “wonderful.” Thus, the results reported below are not specific to a particular set of control terms.

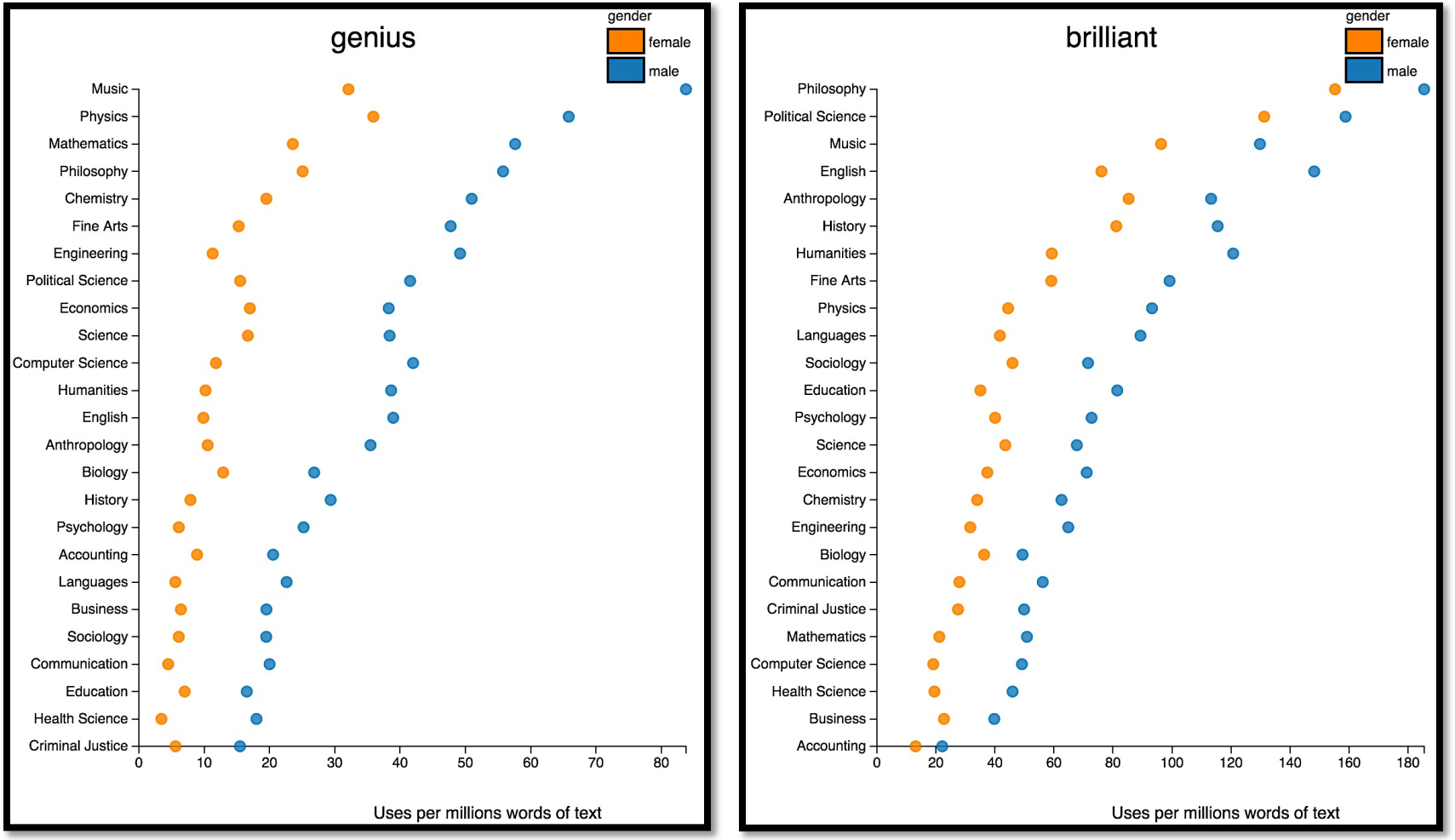

Figure 1. Frequency of "genius" and "brilliant" per millions of words of text on RateMyProfessors.com, split by gender & discipline.

Academics' Ability Beliefs; Competing Hypotheses. The data on academics' ability beliefs, as well as three of the four competing hypotheses (concerning a field's workload, relative emphasis on systemizing vs. empathizing, and selectivity) were taken from Leslie, Cimpian, et al.'s study of academics [3]. In this study, 1820 academics from 30 disciplines were asked a battery of questions designed to assess various characteristics of these disciplines (see Table A1 in Appendix A). To assess the final alternative hypothesis (concerning the math-intensive content of a field), we obtained field-level Quantitative GRE averages from the Educational Testing Service [17].

It is a conservative feature of this study that the data for three of the four competing hypotheses (workload, systemizing vs. empathizing, and selectivity) were obtained from graduate students and faculty—rather than undergraduates, like our brilliance language measure. Since the dependent variable of this study is Ph.D. attainment, variables that are measured on participants who are actually affiliated with Ph.D. programs (such as graduate students and faculty) should in principle be more predictive than variables that are measured on undergraduates. Thus, it would be particularly striking if the word count derived from undergraduates' online reviews predicted Ph.D. attainment above and beyond these other control variables.

CHAPTER 3

RESULTS

Our analyses used the 18 fields from the Gendered Language Tool that could be matched with the fields in Leslie, Cimpian, and colleagues' dataset ([3]; see Table A2 in Appendix A). The data used for the main analyses in this research are provided in Appendix A (see Table A3).

Do Undergraduates Use "Brilliant" and "Genius" more for Male than for Female Instructors? We first highlight the bias in college students' use of "brilliant" and "genius" for their male vs. female instructors (see also Figure 1). The average usage ratios across fields were 1.81:1 for "brilliant" and 3.10:1 for "genius" (male:female instructors), both of which were significantly different from a 1:1 ratio, one-sample *t*s(17) > 7.99, *p*s < .001.[2] However, this bias did not extend to all dimensions of competence evaluation. We found little evidence of gender bias in college students' use of "excellent" and "amazing" in their online evaluations, with male:female ratios of 1.08:1 and 0.91:1, respectively. Both of these ratios were significantly less male-skewed than the ratios for "brilliant" and "genius," paired-sample *t*s(17) > 8.03, *p*s < .001. Thus, it is not the case that female instructors are viewed in an overall negative light. The female disadvantage seems specific to superlatives about intellectual ability, consistent with the existence of pervasive stereotypes against women on this dimension [2].

Does Undergraduates' Use of "Brilliant" and "Genius" Track Academics' Ability Beliefs? Turning to our main predictions, we first tested whether the explicit ability beliefs of a field's practitioners (from Leslie, Cimpian, et al. [3]) agree with the naturalistic, language-based measure derived from college students' use of the words "genius" and "brilliant" in their RateMyProfessors.com reviews (averaged across male and female instructors' evaluations). We

[2] These differences cannot explain the predicted relationship between the brilliance language score and the gender diversity of Ph.D.'s. They could only do so if male instructors' evaluations, which contain more brilliance language, were weighted more heavily in fields where there are more men. This was not the case.

indeed found a tight link between the practitioners' explicit emphasis on raw intellectual aptitude in their survey answers and the frequency of college students' comments about their professors' brilliance and genius in their online reviews, $r(16) = .62$ [.22, .85], $p = .006$. (Throughout, we present 95% confidence intervals in square brackets.) The more strongly academics endorsed the importance of intellectual talent for success in their field, the more frequently undergraduate students used the terms "brilliant" and "genius" to evaluate members of that field.

Does Undergraduates' Use of "Brilliant" and "Genius" Predict Ph.D. Diversity?

Second, we examined whether the amount of brilliance language used in course evaluations for a field (which is a measure of that field's focus on brilliance) predicts the likelihood that women pursue Ph.D.'s in that field. Indeed, the fields with more brilliance language in college students' evaluations were fields where women were less likely to pursue Ph.D.'s, $r(16) = -.49$ [−.78, −.02], $p = .041$ (see Figure 2). This relationship was significant, $\beta = -.48$ [−.88, −.07], $p = .025$, even after adjusting for the four aforementioned competing hypotheses (namely, a field's workload, relative emphasis on systematizing vs. empathizing, selectivity, and average Quantitative GRE score; see Table 1). Although most of these controls are individually predictive of female representation [3], they nonetheless failed to predict significant additional variance beyond our naturalistic measure of a field's focus on brilliance (see Table 1). Finally, note that brilliance language scores computed separately from male and female instructors' evaluations were also predictive of gender gaps in Ph.D. conferral above and beyond these four alternatives (see Table A4 in Appendix A).

Next, we tested whether the representation of African Americans at the Ph.D. level [15] might be similarly explained by the field-level variability in college students' brilliance language. Consistent with our prediction, fields in which college students mentioned "brilliant"

and "genius" more often in their online evaluations were also less likely to have African American Ph.D.'s, $r(16) = -.53$ [−.80, −.09], $p = .023$ (see Figure 3). Moreover, this simple word count remained a significant predictor of race gaps in representation, $\beta = -.65$ [−1.15, −0.14], $p = .016$, even when adjusting for a field's work demands, selectivity, and average Quantitative GRE scores, none of which were themselves significant in the model (see Table 2).[3] Regression models using the separate brilliance language scores computed from male and female instructors' evaluations found these scores to also explain unique variance in African Americans' representation (see Table A5 in Appendix A).

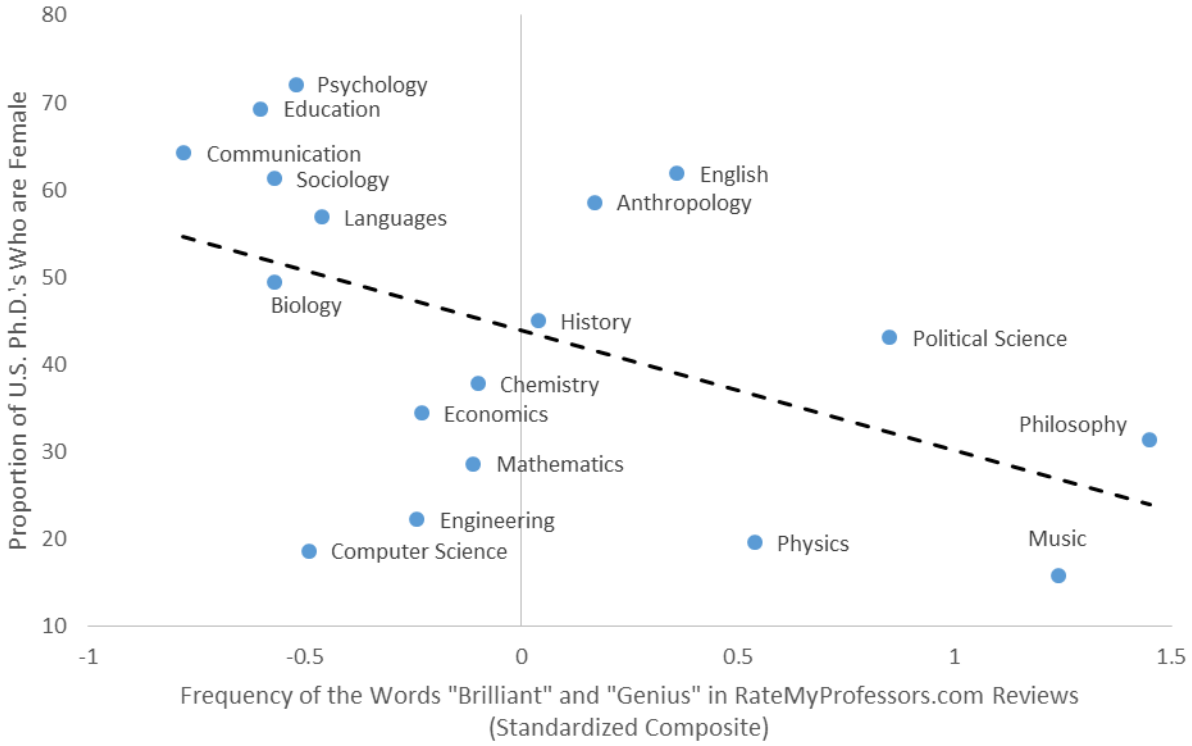

Figure 2. Use of the words "brilliant" and "genius" in course evaluations on RateMyProfessors.com predicts the proportion of 2011 U.S. Ph.D.'s who are female.

[3] Brilliance language was a significant predictor even in a model that included systemizing vs. empathizing (which was omitted from the main analysis because it seemed uniquely relevant to the male vs. female contrast).

Table 1

Multiple regression analysis predicting female representation at the Ph.D. level

Predictor	β	t	p
STEM indicator variable	−.39	−1.27	.230
Brilliance language score	−.48*	−2.60	.025
Hours worked (on-campus)[a]	.26	0.98	.348
Systematizing vs. empathizing	.01	0.04	.971
Selectivity	.10	0.54	.597
Quantitative GRE	−.53	−1.62	.134
R^2		77.9%	

* $p < .05$.

Note. $N = 18$ disciplines. "STEM" stands for "(Natural) Science, Technology, Engineering, and Mathematics."

[a] Although Leslie, Cimpian, et al. [3] collected data on the number of hours worked off campus as well, they found that the number of hours worked on campus was a better predictor of female representation than the total number of hours worked. Thus, to be conservative, we included this stronger competitor in our regression analyses. However, the brilliance language score remains a significant predictor even when the total number of hours worked (on- plus off-campus) is used in the regression.

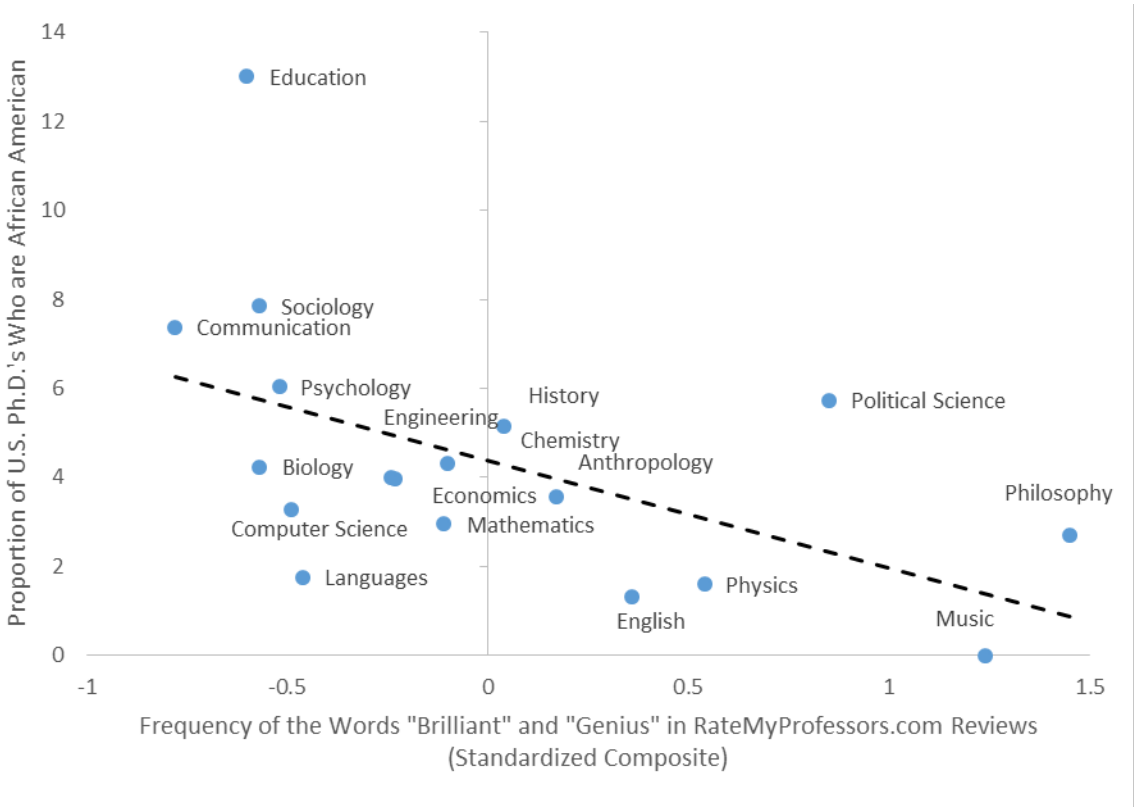

Figure 3. Use of the words “brilliant” and “genius” in course evaluations on RateMyProfessors.com predicts the proportion of 2011 U.S. Ph.D.’s who are African American.

Table 2

Multiple regression analysis predicting African American representation at the Ph.D. level

Predictor	*β*	*t*	*p*
STEM indicator variable	−.32	–0.79	.447
Brilliance language score	−.65*	–2.80	.016
Hours worked (on-campus)	−.20	–0.53	.607
Selectivity	−.37	−1.40	.186
Quantitative GRE	−.09	–0.25	.806
R^2		49.0%	

* $p < .05$.

Note. $N = 18$ disciplines.

It is worth noting that the relationship between brilliance-oriented language in course evaluations and African Americans' representation speaks against a possible alternative interpretation of the results concerning women's representation: Perhaps fields that have more mentions of "brilliant" and "genius" in their online evaluations do so just because more undergraduate men take courses in them, and men may be more likely than women to value and comment on these traits (whereas women may be correspondingly more focused on the level of effort put in by their instructors; e.g., [18]). If so, the relationship between this language-based measure and women's Ph.D. attainment would simply amount to predicting fewer women at the Ph.D. level based on observing fewer women in college. However, this alternative cannot explain why the frequency of "brilliant" and "genius" also predicts the representation of African Americans at the Ph.D. level; no empirically documented differences in valuing brilliance vs.

effort distinguish African Americans from other groups. Thus, the most parsimonious explanation for this set of findings is that our word-count measure indeed taps into the beliefs shared by potential members of a field concerning the qualities that ensure success. When these beliefs emphasize the need for brilliance, members of groups stereotypically portrayed as lacking such a quality might be discouraged from persisting. Consistent with this interpretation, prior studies found that adjusting for the gender composition of the respondents from each discipline did not affect the predictive relationship between disciplines' ability beliefs and their diversity [3, 5]. Although such an adjustment is not possible here (since the gender of the students filling out evaluations on RateMyProfessors.com is not recorded), there is no reason to suppose that it would have any more of an effect on these data.

Next, we tested whether the brilliance language measure is a significant predictor of Asian Americans' Ph.D. attainment. We expected it would not be: The career aspirations of groups who are not targeted by negative stereotypes about intelligence shouldn't be strongly affected by a field's emphasis on brilliance. Indeed, the relationship between the brilliance language score and the representation of Asian Americans at the Ph.D. level was not significant, $r(16) = –.25$ [–.64, .24], $p = .315$. Brilliance language did not significantly predict Asian Americans' representation beyond our controls either, $\beta = −.22$ [−.64, .20], $p = .275$ (see Table 3). This null result, combined with the positive results for women's and African Americans' representation, supports the claim that groups who are the targets of negative stereotypes about their intelligence are particularly likely to be underrepresented in fields that cherish brilliance and genius.

Table 3

Multiple regression analysis predicting Asian American representation at the Ph.D. level

Predictor	*β*	*t*	*p*
STEM indicator variable	.31	0.91	.379
Brilliance language score	−.22	−1.14	.275
Hours worked (on-campus)	−.06	−0.20	.844
Selectivity	.15	0.66	.521
Quantitative GRE	.60~	2.06	.062
R^2		65.1%	

~ $p < .10$.

Note. $N = 18$ disciplines.

The Gendered Language Tool allows word searches to be performed separately for positive vs. negative reviews (i.e., reviews that scored higher vs. lower than the midpoint of the "overall quality" rating on RateMyProfessors.com, respectively). In a final set of analyses, we explored whether brilliance language scores computed separately over the positive and negative reviews predicted women's and African Americans' representation. A priori, there is little reason to expect an asymmetry between these two language scores, since frequent use of "brilliant" and "genius" in reviews indicates a focus on intellectual ability regardless of whether these words are used to say something positive or negative about the instructor.[4] Indeed, the brilliance language scores derived from positive and negative reviews were significantly correlated with each other, $r(16) = .51$ [.06, .79], $p = .029$, and both were also correlated with women's Ph.D. representation

[4] It is worth noting, however, that the most common reasons for negative reviews are probably unrelated to the instructor's intelligence (e.g., "he's a genius, but he can't teach").

(positive reviews: $r(16) = -.45$ [−.76, .02], $p = .061$; negative reviews: $r(16) = -.65$ [−.86, −.27], $p = .003$) and African Americans' Ph.D. representation (positive reviews: $r(16) = -.49$ [−.78, −.03], $p = .039$; negative reviews: $r(16) = -.56$ [−.81, −.12], $p = .016$). The separate brilliance language scores obtained from positive and negative reviews also predicted unique variance in Ph.D. diversity above and beyond the relevant competing hypotheses (βs < −.50, ps < .024; see Tables A6 and A7 in Appendix A). The only exception here was the regression predicting women's representation based on the brilliance language from negative reviews, in which the coefficient for the brilliance language score was not significant, $\beta = -.28$ [−.89, .32], $p = .322$ (see Table S6). One possible reason for this result is that "brilliant" and "genius" were about three times less frequent in negative than in positive reviews; thus, the word tally based on the negative reviews was likely noisier.

Does Undergraduates' Use of "Excellent" and "Amazing" Predict Ph.D. Diversity?

Finally, we investigated the specificity of the link between the language used in teaching evaluations and the underrepresentation of stigmatized groups: Does the frequency of other superlatives (beyond "brilliant" and "genius") also predict gaps in representation, or is this link specific to brilliance-related evaluative terms? Consistent with our hypothesis, the frequency of the adjectives "excellent" and "amazing" was not significantly correlated with either women's Ph.D. representation, $r(16) = .22$ [−.27, .62], $p = .378$, or African Americans' Ph.D. representation, $r(16) = .21$ [−.29, .61], $p = .413$. This pattern of results suggests that it is the fields where people are judged on their *brilliance*—not just their competence—that have a problem attracting members of stigmatized groups.

CHAPTER 4

DISCUSSION

A focus on brilliance in college students' course evaluations within a field consistently predicted lower involvement of women and African Americans—but not Asian Americans—in that field, even when taking into account other possible explanations for race and gender gaps in Ph.D. attainment. These results provide a compelling conceptual replication of the earlier work that used academics' explicit beliefs as a measure of their field's brilliance focus [3, 4], as well as the beliefs of non-academics who had at some point taken college courses in the field [5]. Aside from providing a replication of these prior results, which would be a worthwhile goal in and of itself [19], the present study is valuable because it relies on a wholly naturalistic measure of a field's emphasis on brilliance. The college students whose reviews we used here were not filling out a questionnaire as part of a research study; rather, they were simply expressing their opinions about their instructors in an anonymous online forum. Yet, the frequency with which these students spontaneously commented on whether their instructors were "brilliant" and "geniuses" tracked not only academics' own beliefs about the importance of these traits but also the magnitude of gender and race gaps across much of academia.

Although this research does not speak to the causal mechanisms by which a focus on raw intellectual ability might discourage the involvement of stigmatized groups, several such mechanisms are possible. For instance, members of fields that cherish brilliance might be more likely to discriminate against students and colleagues from groups that are stereotypically seen as lacking such ability, offering them less support [20, 21] and fewer opportunities [22]. At the same time, the evaluative atmosphere in these fields might cause women and stigmatized minorities to worry that they will be judged on the basis of the stereotypes against their

intelligence. This state of "stereotype threat" lowers the motivation and performance of those it affects [23, 24] and could thus lead women and African Americans to look for careers elsewhere. Exploring some of these causal pathways is a crucial next step in this program of research.

This work raises a number of other questions that could be pursued in future work. First, it would be worthwhile to explore how a field's brilliance focus relates to its diversity at various career stages. The present study focused on the diversity of Ph.D. recipients, but would we see similar relationships with diversity of assistant professors, tenured professors, endowed chairs, etc.? To speculate, given that women are likely to encounter additional, non-discipline-specific obstacles as their careers progress (e.g., inadequate childcare support; [25]), it is possible that the relationship between a field's focus on brilliance and its gender diversity might attenuate with time. Second, it is important to examine the developmental origins of the beliefs relevant to this phenomenon. When do children, for example, start believing that women's intellectual abilities are inferior to men's? What are the sources of this belief? Answers to these questions would be useful in part because they could inform interventions to encourage girls' pursuit of "brilliance required" fields. Another interesting, though perhaps less tractable, question concerns the reasons for the variability among fields in their beliefs about brilliance and genius. Why is it that some fields view these traits as essential for success and others do not? To what extent are these beliefs rooted in reality,[5] and to what extent are they merely byproducts of a field's history?

Conclusion. The present study suggests that a focus on inherent intellectual abilities may discourage participation by groups who are stereotypically portrayed as lacking these abilities. In light of these data, it seems likely that turning the spotlight away from sheer brilliance—and toward the importance of sustained effort in achieving professional success [26, 27]—may bring

[5] Importantly, even if these beliefs do track reality, they may nevertheless be discouraging for members of groups that are the targets of negative stereotypes about their intelligence.

about improvements in the diversity of many fields.

REFERENCES

[1] Steele CM, Aronson J (1995) Stereotype threat and the intellectual test performance of African Americans. J Pers Soc Psychol 69(5): 797-811. doi: 10.1037/0022-3514.69.5.797

[2] Stephens-Davidowitz, S (2014, January 18) Google, Tell Me. Is My Son a Genius? The New York Times. Available from: http://tinyurl.com/p3k2998

[3] Leslie SJ, Cimpian A, Meyer M, Freeland E (2015) Expectations of brilliance underlie gender distributions across academic disciplines. Science 347(6219): 262-265. doi: 10.1126/science.1261375

[4] Cimpian A, Leslie SJ (2015) Response to comment on "Expectations of brilliance underlie gender distributions across academic disciplines." Science 349(6246): 391. doi: 10.1126/science.aaa9892

[5] Meyer M, Cimpian A, Leslie SJ (2015) Women are underrepresented in fields where success is believed to require brilliance. Front Psychol 6: 235. doi: 10.3389/fpsyg.2015.00235

[6] Ferriman K, Lubinski D, Benbow CP (2009) Work preferences, life values, and personal views of top math/science graduate students and the profoundly gifted: Developmental changes and gender differences during emerging adulthood and parenthood. J Pers Soc Psychol 97(3): 517-532. doi: 10.1037/a0016030

[7] Baron-Cohen S (2002) The extreme male brain theory of autism. Trends Cogn Sci 6(6): 248-254. doi: 10.1016/S1364-6613(02)01904-6

[8] Hedges LV, Nowell A (1995) Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science 269(5220): 41-45. doi: 10.1126/science.7604277

[9] Hyde JS (2005) The gender similarities hypothesis. Am Psychol 60(6): 581-592. doi: 10.1037/0003-066X.60.6.581

[10] Penner AM (2008) Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. Am J Sociol 114: S138-S170.

[11] Herrnstein RJ, Murray C (1994) The bell curve: Intelligence and class structure in American life. New York: The Free Press.

[12] Nisbett, RE (2009) Intelligence and how to get it: Why schools and culture count. New York: W. W. Norton.

[13] Reardon SF, Robinson-Cimpian JP, Weathers ES (2015) Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In: Ladd HA, Goertz ME, editors. Handbook of Research in Education Finance and Policy. New York: Routledge; p. 491-509.

[14] Robinson JP, Lubienski SL (2011) The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. Am Educ Res J 48(2): 268-302. doi: 10.3102/0002831210372249

[15] National Science Foundation (2011) Survey of Earned Doctorates. Available from: http://www.nsf.gov/statistics/srvydoctorates/

[16] Schmidt B (2015) Gendered Language in Teaching Reviews. Available from: http://benschmidt.org/profGender/

[17] Educational Testing Service (2014) GRE: Guide to the use of scores: Available from: http://www.ets.org/s/gre/pdf/gre_guide.pdf

[18] Goodwin RD, Gotlib IH (2004) Gender differences in depression: The role of personality factors. Psychiatry Res 126(2), 135-142. doi: 10.1016/j.psychres.2003.12.024

[19] Open Science Collaboration (2015) Estimating the reproducibility of psychological science.

Science 349(6251): aac4716. doi: 10.1126/science.aac4716

[20] Milkman KL, Akinola M, Chugh D (2012) Temporal distance and discrimination: An audit study in academia. Psychol Sci 23(7): 710-717. doi: 10.1177/0956797611434539

[21] Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012) Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci 109(41): 16474-16479. doi: http://10.1073/pnas.1211286109

[22] Wennerås C, Wold A (1997) Nepotism and sexism in peer review. Nature 387(6631): 341-343.

[23] Emerson K, Murphy M (2015) A Company I Can Trust? Organizational Lay Theories Moderate Stereotype Threat for Women. Pers Soc Psychol Bull 41(2): 295-307. doi: 10.1177/0146167214564969

[24] Steele CM (2013) Whistling Vivaldi: How stereotypes affect us and what we can do. New York: W. W. Norton.

[25] Mason MA (2012, November 29) Title IX and babies: The new frontier? The Chronicle of Higher Education. Available from: http://chronicle.com/article/Title-IXBabies-The-New/135936/

[26] Dweck CS (1999) Self-theories: Their role in motivation, personality and development. Philadelphia, PA: Psychology Press.

[27] Dweck CS (2006) Mindset: The new psychology of success. New York: Random House.

APPENDIX A

Table A1
The measures from Leslie, Cimpian, et al.'s (2015) study that were used in the present research

Field-specific Ability Beliefs[a]
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]. (R)
When it comes to [discipline], the most important factors for success are motivation and sustained effort; raw ability is secondary. (R)
Hours Worked[b]
Approximately how many hours a week do you spend working:
In your office, lab, classroom, or otherwise on campus?
Off campus (e.g., home, coffee shop, other remote site)?
Systemizing vs. Empathizing[c]
Please rate the extent to which the following processes are involved in doing scholarly work in [discipline]:
Identifying the abstract principles, structures, or rules that underlie the relevant subject matter (Systemizing)
Analyzing the relevant subject matter and constructing a systematic understanding of it (Systemizing)
Having a refined understanding of human thoughts and feelings (Empathizing)
Recognizing and responding appropriately to people's mental states (Empathizing)
Selectivity[d]
Roughly what percentage of applicants are accepted into your department's Ph.D. program in a typical year? (R)

Note. (R) indicates items that were reverse scored.
[a] Responses to these items were given on a 7-point scale (1 = strongly disagree to 7 = strongly agree).
[b] Responses to these items were given on an 8-point scale (1 to 8, 1-7 corresponding to 10-hour increments, and 8 corresponding to >70 hours).
[c] Response to these items were given on a 7-point scale (1 = never involved to 7 = highly involved).
[d] Responses to these items were given on a 10-point scale (1 to 10, each number corresponding to a 10% increment). There were two additional options for "don't know" and "no Ph.D. program." This variable was reversed for analysis so that higher values indicate greater selectivity.

Table A2
The fields matched between the Gendered Language Tool and Leslie, Cimpian, et al.'s (2015) dataset

Gendered Language Tool Fields	Leslie, Cimpian, et al. (2015) Fields
Accounting	N/A[a]
Anthropology	Archaeology[b], Anthropology
Biology	Biochemistry, Evolutionary Biology, Molecular Biology, Neuroscience
Business	N/A[a]
Chemistry	Chemistry
Communication	Communication
Computer Science	Computer Science
Criminal Justice	N/A[a]
Economics	Economics
Education	Education
Engineering	Engineering
English	Comparative Literature[b], English Literature
Fine Arts	N/A[a]
Health Science	N/A[a]
History	History
Humanities	N/A[a]
Languages	Classics[b], Linguistics, Spanish
Mathematics	Mathematics, Statistics
Music	Music Theory & Composition
Philosophy	Philosophy
Physics	Astronomy, Physics
Political Science	Political Science
Psychology	Psychology
Science	N/A[a]
Sociology	Sociology

Note. The matching was performed using the categories provided by the Educational Testing Service (2014) as a guide. Weighted averages of different fields' values were computed where appropriate.

[a] "N/A" denotes that a field from the Gendered Language Tool was not matched with any of the fields from Leslie, Cimpian, et al.'s (2015) dataset (n = 7).

[b] We performed a second set of analyses in which these fields were excluded, for a tighter match between the two datasets (e.g., some readers may disagree about whether Comparative Literature belongs under English). All significant results remain as reported in the main text.

Table A3
The raw data used in the present research

Field	STEM	% Female PhDs	% Afr. Am. PhDs	Quant GRE	"Brilliant" and "genius" composite	"Excellent" and "amazing" composite	FAB	Hours Worked	S vs. E	Selectivity
Anthropology	0	58.60	3.57	149	0.17	-0.37	3.73	3.35	1.33	1.73
Biology	1	49.48	4.22	154	-0.57	-0.23	3.96	5.13	3.30	2.68
Chemistry	1	37.80	4.32	158	-0.10	-0.59	4.11	5.73	3.82	4.00
Comm.	0	64.20	7.38	149	-0.78	-0.34	3.79	3.38	1.26	1.84
Comp. Sci.	1	18.60	3.27	157	-0.49	-0.12	4.29	3.84	3.15	1.64
Economics	0	34.40	3.96	160	-0.23	-1.12	4.37	4.09	2.83	2.18
Education	0	69.30	13.02	149	-0.60	1.49	3.32	3.12	1.01	3.20
Engineering	1	22.20	4.00	159	-0.24	0.19	4.29	4.55	3.38	3.38
English	0	61.87	1.32	149	0.36	-0.02	4.36	2.79	1.27	2.01
History	0	45.00	5.15	148	0.04	-0.02	3.90	2.87	1.16	2.24
Languages	0	56.89	1.76	150	-0.46	1.08	4.11	3.45	2.26	1.77
Mathematics	1	28.60	2.95	162	-0.11	-0.15	4.57	3.72	4.53	2.59
Music	0	15.80	0.00	150	1.24	1.16	4.45	3.22	2.18	3.40
Philosophy	0	31.40	2.70	153	1.45	0.01	5.11	2.71	3.01	1.29
Physics	1	19.56	1.59	161	0.54	-0.88	4.33	4.68	3.98	3.27
Political Sci.	0	43.10	5.73	151	0.85	-0.14	3.94	3.60	2.56	2.18
Psychology	0	72.10	6.04	149	-0.52	0.35	3.55	3.79	1.43	1.59
Sociology	0	61.30	7.86	149	-0.57	-0.28	3.78	3.33	2.37	2.38

Field	Brilliant M	Brilliant F	Genius M	Genius F	Excellent M	Excellent F	Amazing M	Amazing F
Anthropology	113.26	85.33	35.47	10.53	243.49	258.75	421.11	439.11
Biology	49.36	36.33	26.84	12.91	312.48	301.86	319.58	364.07
Chemistry	62.57	33.99	50.96	19.51	283.42	277.61	296.18	329.55
Comm.	56.20	27.95	20.01	4.48	268.22	257.70	375.04	451.24
Comp. Sci.	49.15	19.11	41.99	11.80	434.86	353.21	172.25	186.86
Economics	71.10	37.41	38.25	17.00	276.35	249.94	201.39	231.24
Education	81.44	35.08	16.52	7.02	465.77	355.85	525.19	564.52
Engineering	64.85	31.64	49.18	11.30	462.24	375.14	202.07	207.91
English	148.20	76.10	38.95	9.88	301.47	277.90	421.85	443.75
History	115.49	81.12	29.36	7.90	328.99	282.76	404.68	392.70
Languages	89.32	41.67	22.62	5.59	395.22	367.01	479.82	534.12
Mathematics	50.84	21.16	57.62	23.57	338.03	322.51	267.81	355.52
Music	129.80	96.29	83.77	32.10	313.75	338.23	589.58	704.44
Philosophy	185.45	155.28	55.76	25.08	293.31	298.84	407.59	444.14
Physics	93.23	44.45	65.82	35.90	284.05	256.46	265.89	259.88
Political Sci.	158.82	131.26	41.53	15.51	315.26	271.31	382.86	406.73
Psychology	72.75	40.06	25.22	6.11	312.10	303.19	459.25	513.04
Sociology	71.54	45.95	19.50	6.11	260.36	261.71	383.27	480.68

Note. FAB = academics' field-specific ability beliefs. Hours Worked = hours worked on campus. S vs. E = systematizing vs. empathizing score. The values for FAB, Hours Worked, S vs. E, and Selectivity were all taken from Leslie, Cimpian, et al.'s (2015) dataset. The composite scores were calculated by (1) standardizing the frequencies of the two relevant terms (separately) across all fields, and then (2) averaging male and female instructors' standardized scores for the two relevant terms within each field.

Table A4

*Multiple regression analysis predicting **female representation** at the Ph.D. level based on separate word counts for the male and the female instructors*

	Male instructors' evaluations			**Female** instructors' evaluations		
Predictor	*β*	*t*	*p*	*β*	*t*	*p*
STEM indicator variable	−.34	−1.15	0.276	−.43	−1.31	0.217
Brilliance language score	**−.48***	**−2.69**	**0.021**	**−.45***	**−2.32**	**0.040**
Hours worked (on-campus)	.21	0.80	0.441	.33	1.27	0.229
Systematizing vs. empathizing	−.05	−0.14	0.894	.05	0.13	0.900
Selectivity	.07	0.38	0.712	.15	0.80	0.438
Quantitative GRE	−.47	−1.46	0.172	−.60	−1.72	0.114
R^2		78.5%			76.1%	

* $p < .05$.

Table A5

*Multiple regression analysis predicting **African American representation** at the Ph.D. level based on separate word counts for the male and the female instructors*

	Male instructors' evaluations			**Female** instructors' evaluations		
Predictor	*β*	*t*	*p*	*β*	*t*	*p*
STEM indicator variable	−.29	−0.80	0.440	−.33	−0.72	0.487
Brilliance language score	**−.75****	**−3.46**	**0.005**	**−.51~**	**−2.05**	**0.063**
Hours worked (on-campus)	−.32	−0.91	0.378	−.05	−0.12	0.906
Selectivity	−.45~	−1.82	0.094	−.28	−0.98	0.347
Quantitative GRE	−.02	−0.07	0.949	−.18	−0.47	0.645
R^2		57.7%			37.4%	

~ $p < .10$. * $p < .05$. ** $p < .01$.

Table A6

*Multiple regression analysis predicting **female representation** at the Ph.D. level based on separate word counts for positive and negative reviews*

	Positive evaluations			**Negative** evaluations		
Predictor	*β*	*t*	*p*	*β*	*t*	*p*
STEM indicator variable	−.38	−1.32	0.213	−.29	−0.75	0.468
Brilliance language score	**−.50***	**−2.95**	**0.013**	**−.28**	**−1.04**	**0.322**
Hours worked (on-campus)	.23	0.94	0.366	.51	1.76	0.106
Systematizing vs. empathizing	<.01	0.01	0.990	−.26	−0.55	0.593
Selectivity	.11	0.62	0.546	.16	0.69	0.504
Quantitative GRE	−.55	−1.77	0.104	−.34	−0.84	0.419
R^2		80.1%			67.5%	

* $p < .05$.

Table A7

*Multiple regression analysis predicting **African American representation** at the Ph.D. level based on separate word counts for positive and negative reviews*

	Positive evaluations			**Negative** evaluations		
Predictor	*β*	*t*	*p*	*β*	*t*	*p*
STEM indicator variable	−.30	−0.71	0.490	−.42	−0.99	0.343
Brilliance language score	**−.62***	**−2.62**	**0.022**	**−.77***	**−2.61**	**0.023**
Hours worked (on-campus)	−.19	−0.51	0.622	.14	0.40	0.699
Selectivity	−.35	−1.28	0.225	−.42	−1.50	0.158
Quantitative GRE	−.15	−0.42	0.679	.24	0.56	0.584
R^2		46.4%			46.1%	

* $p < .05$.