Gender Inequalities and Peer Review Disparities in the Academic Workforce

by

Nicholas LaBerge

B.A., Claremont McKenna College, 2019

M.S., University of Colorado Boulder, 2022

A dissertation submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2024

Committee Members:

Aaron Clauset, Co-Chair

Daniel Larremore, Co-Chair

Cassidey Sugimoto

Bailey Fosdick

Daniel Acuna

LaBerge, Nicholas (Ph.D., Computer Science)

Gender Inequalities and Peer Review Disparities in the Academic Workforce

Thesis directed by Prof. Aaron Clauset and Prof. Daniel Larremore

Academic faculty make important contributions to the scientific ecosystem by educating future researchers, advancing knowledge, and shaping the direction of their fields. Despite a continued emphasis on broadening participation, women faculty in the U.S. remain underrepresented in many scientific disciplines. This thesis presents three distinct research projects which each focus on women's representation in science from different perspectives. First, I quantify variations in women's representation across computing subfields, producing a detailed picture of past and likely future trends and inequalities in computing. Next, I investigate the relative impacts of faculty hiring versus faculty attrition on women's representation across academia more broadly, quantifying how these two processes shaped gender representation across the academy over the decade 2011–2020 and projecting how specific interventions to hiring and attrition patterns could impact the future representation of women in academia. Finally, I develop and investigate a comprehensive dataset on peer review outcomes at two elite multidisciplinary journals, *Science* and *Science Advances*, finding that gender disparities exist at nearly every stage of the peer review process. I conclude by discussing the limitations of my findings and by suggesting avenues for future research.

Dedication

To a more diverse and innovative scientific ecosystem.

Acknowledgements

First, I would like to thank my advisors, Aaron Clauset and Dan Larremore. You have both contributed profoundly to my growth as a scientist, and this thesis would have been impossible without your guidance and support. Aaron, your mentorship has been invaluable to me. I am particularly appreciative of all the detailed feedback that you provided on paper drafts over the years, which accelerated my development as a scientific writer. Dan, your enthusiasm for science and your dedication to teaching inspire me. I am especially grateful for the hours that we spent whiteboarding together–yeehaw!

I would also like to thank Cassidey Sugimoto, Bailey Fosdick, and Daniel Acuna for their feedback and support. I cannot believe how lucky I am to have such a talented and supportive dissertation committee.

Thank you Adam Landsburg, for introducing me to research and for encouraging me to pursue graduate school. It has been one of the best decisions that I have ever made.

To my friends and collaborators in the Larremore and Clauset labs, thank you for making my time in graduate school so enjoyable. Allie, Hunter, Sam, Tzu-Chi, Erik, Ian, Mike, Katie, Casey, Kate, Kate, Lucy, Eun, Upasana, Alex, Ellen, Kath, Carolina, Chethan, Marilena, and Bisman, I am deeply grateful for all the conversations, the shenanigans, and the laughs that we shared. I hope we share many more.

Thank you to my family for your unwavering support.

Finally, to Jake, thank you for your love and support. Working towards this thesis defense over the past five years would have been less meaningful and less joyful without you by my side.

Contents

Chapter 1

Background & Introduction

Academic faculty play an important role in society by training future generations of scientists and by contributing to discovery and innovation. Demographic diversity in science accelerates innovation and has been shown to improve group problem-solving [1, 2]. Additionally, people of different racial and gender identities can be drawn to differing distributions of topics, such that more diversity could distribute the rewards of scientific innovation more broadly (e.g., reducing health disparities that are a product of historically understudied topics) [3, 4, 5, 6]. Despite general support for a more diverse, equitable, and inclusive academic ecosystem, progress remains slow with respect to equitable gender representation.

Past work

This thesis proposal builds on past work that has identified existing barriers and inequities that act to constrain both the number of women and the success of women's careers in science. By studying the mechanisms driving underrepresentation and inequity, my hope is that these inequalities can be reduced, and that science will become more innovative and more broadly beneficial to society as a result.

Underrepresentation is not ubiquitous or even across fields. For example, women make up roughly 37% of faculty in the biological, agricultural, and environmental life sciences, but only 23% in physical sciences, geosciences, atmospheric sciences, and ocean sciences [7]. A number of explanations provided in the literature, outlined in this introduction, aim to partially explain the variation in women's representation across fields.

1.0.1 Achievement and Confidence

Women's lower self-confidence in their math and science abilities is a common explanation for their lower representation in STEM [8, 9, 10, 11]. Those who persist in spite of low confidence are often negatively impacted by imposter syndrome, which can lead students and faculty to attribute their successes to chance, unnecessarily compare themselves against their high-achieving peers, and over-prepare for simple tasks [12, 13].

Math heavy classes and careers are often considered to be particularly challenging, which can limit the participation of students who receive lower grades or do not have confidence that they will succeed in future course, especially in math and science courses. However, past achievement does not inform students' confidence in a vacuum; cultural stereotypes, including the notion that boys are more likely to possess raw, innate talent than girls, are suspected to influence students' beliefs in their abilities. As early as 6 years old, girls are less likely than boys to indicate that people of their gender are "really, really smart," and they begin to avoid activities that are said to be for "really, really smart" children [8]. The immediate effect of these stereotypes on girls' self esteem is believed to play an early role in shaping girls' interests away from scientific topics.

Fields vary in terms of what is believed to be required to succeed. In some fields, it is commonly held that hard work and dedication are sufficient for success, while in others, it is believed that raw, innate talent is required (which women can be stereotypically considered less likely to possess). It is hypothesized that field specific ability beliefs, paired with these negative gendered stereotypes, influence women's career choices and can partially explain gender differences between fields. Leslie, Cimpian, Meyer, and Freeland identify a striking correlation between such beliefs and the gender compositions of academic fields, both within STEM and in the Humanities [9].

Women's lower self confidence in math and science has been attributed to low grades in past courses [12, 14]. Recent work corroborates these findings, but also identifies differences in how men and women are impacted by low scores [15, 16]. Using a nationally representative, longitudinal dataset that tracked students though 7 years of high school and college, Cimpian, Kim, and McDermott find that high school boys and girls who are high achieving in STEM intend to pursue a college major in physics, engineering, and computer science (PEC) at similar rates. On the other hand, lower achieving girls are far less likely to intend to major in a PEC major compared to boys of equal achievement [16]. While scores indeed appear to correlate with future STEM participation for girls, the relative lack of impact that low scores have on boys' intentions to pursue STEM exists as a substantial source of gender disparity at the STEM pipeline junction between high school and college.

1.0.2 Inhospitable Climate

Inhospitable working and learning conditions for women, sometimes referred to as a "chilly climate," are another common explanation for women's underrepresentation in science. For example, in grade school, boys have been recorded to call out more frequently in class than girls, and when girls do call out answers, teachers are more likely to reprimand girls for not waiting their turns to speak [17], which may disrupt academic self-confidence. Due to the low representation of women among both students and tenure track faculty, women STEM students are also less likely to have same-gender peers and role models [18, 19].

Less favorable assessments of women as scientists, likely driven by cultural stereotypes, may not only impact womens' self perceptions but can also contribute to an external climate that is not hospitable to women's career success. A Canadian funding agency piloted two parallel grant programs: one that explicitly included the caliber of the principal investigator as a part of the evaluation criteria, and one that solely focused on the proposed research material. Analyzing application success rates between these programs, Witteman, Hendricks, Straus, and Tannenbaum find that women were significantly less likely, relative to men, to be awarded a grant when they themselves were explicitly being evaluated. These findings suggest observed gender gaps in funding may be attributable to less favorable views of women as researchers and not to the lower quality of their proposed research [20].

Women are also more likely than men to report being sexually harassed both as students and as faculty [21, 22]. A survey conducted by the University of Texas System in 2018 revealed that roughly a quarter of female STEM students experienced sexual harassment from faculty or staff [21]. These experiences create a hostile environment that makes it difficult to learn and feel comfortable.

1.0.3 Parenthood

The role that some women play in parenthood is another commonly raised explanation for their underrepresentation in science. As early as the first year of college, women students are more likely than their male peers to anticipate that conflicts between work and parenting will be a barrier to a career in engineering [23], and indeed, women researchers are more likely to report being the primary caregiver to their children than male researchers, which is correlated with lower publication productivity and article placement in top journals for both women and men primary caregivers [24, 25]. Cech and Blair-Loy find that 43% of women leave full-time STEM employment after their first child [26]. Considering faculty more specifically, Goulden, Frasch, and Mason conduct a survival analysis using the NSF's survey of doctoral recipients, and find that "women in the sciences who are married with children are 35 percent less likely to enter a tenure track position after receiving a PhD than married men with children. And they are 27 percent less likely than their male counterparts to achieve tenure upon entering a tenure-track job." In contrast, they find that "single women without young children are roughly as successful as married men with children in attaining a tenure-track job, and a little more successful than married women with children in achieving tenure [27]." These findings are consistent with additional studies on parenthood, gender, and faculty positions [28, 29]. One factor believed to contribute to this disparity is that the availability and duration of paid maternity leave varies significantly by institution [30] and is especially limited for mothers who are graduate students and postdoctoral fellows [31].

1.0.4 Promotion and Tenure

The tenure track is an up-or-out system in which assistant professors must impress their departments and universities by meeting the often subjective criteria for tenure within a probationary period or face termination. Several empirical studies have examined the tenure process as a potential explanation for women's lower representation, especially among associate and full processors. Findings report that women are granted tenure at lower rates than men, even after controlling for relevant covariates like publication productivity [32, 29, 33, 34, 35], which is lower on average for women [36]. On average, women who are eventually granted tenure tend to receive tenure later than men who are granted tenure [37, 38, 39], which is in part due to women's higher likelihood to stop the tenure clock than their male peers [39]. The literature on the effects of gender on promotion to tenured faculty has yielded mixed results, as a similarly designed study identified no significant gender differences in promotion rates [40].

1.0.5 The Matilda Effect and Epistemic Injustice

Contributing explanations for women's underrepresentation in some fields are instances of the "Matilda Effect," a broad term used to categorize the "many cases, historical and contemporary, of women scientists who have been ignored, denied credit or otherwise dropped from sight [41]." In particular, there is evidence that women's research contributions are underrecognized by the scientific community, reflected by lower citation counts, fewer prizes, unfair attribution of women's contributions, and a broad devaluation of topics to which women are more likely to contribute knowledge.

Publications and citations are valuable in academia for career progression and recognition, yet recent work has observed gendered disparities in peer review outcomes [42, 43, 44, 45, 46] and in citations [47, 48, 49, 50, 51, 52]. The career impacts of publications are especially prominent for publications in the most elite scientific journals (which tend to receive the most citations [53, 54]), such that the presence of gender disparities at these venues in particular have the greatest potential to make or break academic careers (see Chapter 4). Articles by women first and last authors are under-cited with respect to their prevalence in the field, especially in the reference lists of papers written by men [47, 52]. While there is presently no work testing causal or correlative links between women's under-citation and heightened attrition rates, I believe this will be a worthwhile avenue for future work.

Women are also underrepresented among prize recipients [55], especially among recipients of the most prestigious and high-paying prizes. Ma, Oliveira, Wodruff, and Uzzi identified gender disparities in prize prestige and payout by gender that resulted in women prize winners receiving only 64.4 cents for every dollar that a male prizewinner received [56]. There is also a relationship between prizes being awarded to researchers in a given area, and that area's growth with respect to future publication productivity and citation impact [55]. This relationship, paired with findings that men and women have different distributions over topics of study within fields, suggests that women's underrepresentation among prize winners may have broader negative impacts on women in science, beyond the individual careers of the would-be women prize winners.

In collaborative projects, team members can take on different roles, which are all important, but the relative importance between roles can be subjective. There is evidence of gendered differences in researchers' roles in different stages of their careers, where women tend to be more likely to play supporting and specialized roles, while men are more likely to play the role of leader, especially in early and mid-career stages [57]. If the Matilda Effect is at play in the collaborative process, one would hypothesize that the roles that women are more likely to play are undervalued. This hypothesis is consistent with survey findings, which reveal that women are more likely than men to have encountered disagreements with co-authors on naming, ordering, and acknowledgment in papers and that women are more likely to report being under-credited for their contributions [58]. In many fields, women are underrepresented among first and last authors in academic papers, which are typically considered the highest credited contributors [59]. Notably, papers with women last authors are substantially more likely to have women first authors (46%), relative to papers with men last authors (32%) [60]. This gendered pattern in collaboration suggests that the underrepresentation of women in more senior positions may contribute to lower counts of women first authored publications (see also [61, 62]).

1.1 Presented work

Building on this extensive body of literature, I have identified and explored several directions for research, which I present in this dissertation.

First, in chapter 2, I consider how the subfield structure of a particularly homogeneous field—computing—may limit women's participation. One key motivation for this study is that, because faculty searches are typically at the subfield level, field-level demographic changes are in fact driven by subfield hiring dynamics. In this work, I quantify and forecast variations in the demographic composition of the subfields of computing using a comprehensive database of training and employment records for 6882 tenure-track faculty from 269 PhD-granting computing departments in the United States, linked with 327,969 publications. I find that subfield prestige correlates with gender inequality, such that faculty working in computing subfields with more women tend to hold positions at less prestigious institutions. In contrast, I find no significant evidence of racial or socioeconomic differences by subfield. Tracking representation over time, I find steady progress toward gender equality in all subfields, but more prestigious subfields tend to be roughly 25 years behind the less prestigious subfields in gender representation. These results illustrate how the choice of subfield in a faculty search can shape a department's gender diversity.

In chapter 3, I frame my analysis around the reality that there are two main ways in which representation of academic faculty can change: through hiring and through attrition. I quantify the relative impact and importance of hiring and attrition for the past and future gender diversity of faculty using a comprehensive database of training and employment records for 268,769 tenure-track faculty from 12,112 U.S. PhD-granting departments, spanning 111 academic fields. Over this time, I find that hiring had a far greater impact on women's representation among faculty than attrition in the majority (90.1%) of academic fields, even as academia loses a higher share of women faculty relative to men at every career stage due to gendered differences in attrition rates.

Finally, I model the impact of five specific policy interventions on women's representation, and project that eliminating attrition differences between women and men only leads to a marginal increase in women's overall representation—in most fields, successful interventions will need to make substantial and sustained changes to hiring in order to reach gender parity.

In chapter 4, I quantify gender and prestige disparities in peer review outcomes at two elite journals, *Science* and *Science Advances.* Elite scientific journals have broad influence, authoritative reputations, wide multidisciplinary readerships, and a highly selective peer review process, yet little is understood about the extent and mechanisms through which nonmeritocratic social factors drive peer review outcomes at these journals. This lack of integrated understanding of the elite peer review process has been largely driven by a lack of data, since these journals maintain strict standards of confidentiality with authors and reviewers. Alongside my collaborators, I have prepared six years of anonymized data from the entire publishing process at *Science* and *Science Advances*, spanning 110,303 submissions to both journals, and including information about the manuscript, authors, reviewers, editors, and advisors at multiple stages, including the initial editor screen, peer review, and the final decision. I quantify disparities in peer review across several stages of the editorial process, including desk rejection decisions, reviewer recommendations, and final acceptance decisions. For both journals, we observe lower success rates at most editorial stages for smaller teams, certain topics, lower prestige authors, Chinese authors, and women corresponding authors.

Finally, in Chapter 5, I discuss conclusions and directions for future research.

Chapter 2

Subfield Prestige and Gender Inequality Among U.S. Computing Faculty

This chapter is adapted from:

N. LaBerge, K. H. Wapman, A. C. Morgan, S. Zhang, D. B. Larremore, A. Clauset, "Subfield Prestige and Gender Inequality among U.S. Computing Faculty." *Communications of the ACM* 65(12), 46–55 (2022).

Introduction

In computing, faculty play many critical roles, including training the next generation of researchers, advancing scientific research across a diverse array of computing topics, and translating that research into practice. The composition of the academic workforce thus shapes what advances are made and who benefits from them [63, 64, 65, 66, 67], in part because demographic diversity in science is known to accelerate innovation and improve problem solving [1, 2].

Despite a continued emphasis on broadening participation, women faculty in the U.S. remain underrepresented relative to women's share of the U.S. population by over a factor of two, and Black, Hispanic, and Native faculty by over a factor of five [68, 69]. Women's underrepresentation among computing researchers also persists internationally. For example, women are estimated to comprise less than 10% of contributors to international computer science journals [70].

Explanations for this persistent pattern generally fall into two categories. On the one hand, there are generational problems, in which faculty diversity changes slowly because it takes many years for increases in diversity at the earliest stages of training to propagate up to more senior levels [71, 72]. On the other hand, there are structural and social climate problems in the U.S. [73], in which members of underrepresented groups who aspire to or have a faculty career are pushed or pulled out of the community, which may counteract efforts to address generational problems. Thus, in concert, these two effects may lead to a persistent overrepresentation of majority groups [12, 74, 22, 75] in spite of efforts to the contrary.

We consider a third class of problem, which exists because most faculty are hired via searches that focus on a particular subfield of computing, e.g., a search in the area of artificial intelligence. As a result, field-level demographic dynamics like gender, racial, and socioeconomic representation are in fact driven by diversity differences across computing's subfields [76], and the representation of those subfields among the suppliers of future faculty [77]. For example, faculty searches in subfields with fewer women than other subfields are less likely to increase a department's gender diversity. Similarly, if more gender or racially diverse subfields are underrepresented at elite departments—the ones that produce the majority of future faculty [77]—then the diversity of those subfields is unlikely to be reflected in new faculty hires. While there is some evidence that job searches that do not focus on a particular subfield can attract more diverse candidates [78, 79], most searches in computing remain subfield specific.

In practice, faculty hiring closely follows a prestige hierarchy, in which more prestigious departments produce a disproportionate share of all computing faculty [77], and a department's position within this hierarchy can be inferred directly from where its graduates were hired as faculty [77, 80]. In this way, high prestige departments exert a correspondingly large influence over the field's demographics [81], and efforts to understand patterns, trends, and causes of demographic diversity in computing must account for effects of prestige.

What are the implications of subfield structure and prestige in faculty hiring for diversity and demographic trends in computing? Here, we address this question by studying the intersections of gender, race, socioeconomic status, prestige, and subfield structure in computing. Our analysis uses a comprehensive database of training and employment records for 6882 tenure-track faculty from 269 PhD-granting computing departments in the United States, linked with 327,969 publications.

We first quantify variation in gender, racial, socioeconomic, and prestige across the subfields of computing. We then develop simple forecasts of future gender diversity for the field as a whole, which account for diversification trends over time at the subfield level. We close with a discussion of the particular patterns and trends in faculty diversity we observe, how they relate to more general patterns in academia, and we highlight a few specific implications of our findings for long and short term efforts to increase demographic diversity among computing faculty.

Data

Our analysis spans 6882 tenured or tenure track faculty at U.S. PhD-granting computing departments between 2010 and 2018, and includes faculty names, academic rank, institution, and the year and institution from which they received their PhD training. The underlying data are derived from a larger census-style dataset obtained under a Data Use Agreement with the Academic Analytics Research Center (AARC). For this study, we define the field of computing to include computer science departments and joint departments between computer science and information sciences, computer engineering, and other closely related departments.

To these basic education and employment variables, we add information on gender, race, childhood socioeconomic status, faculty subfield, and institutional prestige using a combination of institutional covariates, automated tools, detailed publication information, and a large survey of faculty, which we describe below.

2.0.1 Gender

We use a set of name-based tools to match faculty with the genders that are culturally associated with their names. This methodology assigns only binary (woman/man) labels to faculty, even as we recognize that gender is nonbinary. This approach is a compromise due to the technical limitations of name-based gender methodologies and is not intended to reinforce the gender binary. We assess the reliability of our gender labeling methodology using self-labeled genders from a representative survey of computing faculty that we conducted in 2017. Comparing these gender labels, our name-based methodology agrees with self-identified genders 97% of the time ($N = 985$).

2.0.2 Race and Childhood Socioeconomic Status

Faculty race is known only for the 608 faculty (8.8%) who self-reported their race in our survey. Our survey question's design followed the U.S. Office of Management and Budget's standards for collecting race data [82], which facilitates comparisons between the computing professoriate and aggregated U.S. census data. We recognize that these categories are imperfect socially constructed representations of racial, ethnic, and place-of-origin identities. For example, the census category "Asian" is broad, and includes South Asians, Southeast Asians, and East Asians, among others, each of which themselves contain diverse groups.

In our survey, 633 faculty (9.2%) report the highest level of education achieved by their parents or legal guardians, which we use as a simple indicator of faculty's childhood socioeconomic status, following Ref. [83].

2.0.3 Subfields

We assign each professor to a distribution over computing research subfields, based on their publications in the DBLP computer science bibliography. Using unique matches to faculty names, we algorithmically linked 5472 faculty (80%) to their listed publications, leading to a set of 327,969 author-linked publications.

Publications were then assigned to computing research areas using a topic model of paper titles. We first manually identified 35 computing research areas grouped into 8 computing subfields using domain knowledge and advice from subfield specialists. For each research area, we algorithmically extracted a set of "anchor words" that are highly informative of publication topic as measured by mutual information [84]. These anchor words were then used to parameterize a topic model, guiding the clustering of publication titles to be aligned with our intended delineation of research areas [85]. We then checked the topic assignments by manually verifying that the final, larger set of words the model learned to associate with each topic aligned with commonly agreed upon

	N	% women
Theory of Computer Science	573.2	13.1
Programming Languages	181.1	14.2
Numerical & Scientific Computing	478.8	14.5
Systems	1486.1	14.6
Computational Learning	950.8	17.9
Software Engineering	317.6	18.9
Interdisciplinary Computing	904.1	19.7
Human-Computer Interaction	580.4	20.0
All of Computing (total)	5472.0	16.7
Computer Science PhDs (effective)	–	19.0
U.S Population (effective)	–	51.1

Table 2.1: Number of tenured or tenure-track faculty and the corresponding gender compositions for 8 computing subfields, along with the gender compositions of two reference populations, the population of computing science PhDs [7] and the United States population [86, 87, 69], each adjusted for changes over time over the years that faculty were trained.

computing research areas, and that the assigned research areas for a set of well-known computing scientists agree with their known expertise.

While computing research can be divided into a multiplicity of fine-grained topics, faculty hiring typically takes place at a higher level. For example, departments aiming to hire in the subfield of human-computer interaction may consider applicants who specialize in any of a variety of its nested research areas. Under our taxonomy for computing research, each of the 35 identified research areas belong to exactly one of the 8 subfields: computational learning, systems, theory of computer science, numerical & scientific computing, human-computer interaction, interdisciplinary computing, programming languages, and software engineering.

Because faculty often publish in a wide variety of areas, we assign a distribution over subfields to each professor, in proportion to the share of their publications classified into each subfield. Under this assignment, faculty belong to multiple subfields, meaning that our subsequent estimates of subfield sizes can take on non-integer values. This “soft” assignment scheme better captures the range of research topics that faculty work on across the boundaries of multiple subfields, compared to a “hard” assignment into a single subfield which we explore in the SI Appendix of the published manuscript.

2.0.4 Institutional Prestige

There are many ways to quantify institutional prestige in computing, including authoritative rankings like the U.S. News & World Report rankings of computer science graduate departments, or the older National Research Council rankings. Such rankings have been widely criticized for their subjective selection of institutional characteristics, and for largely measuring only the inputs to the educational and research process [88, 89]. In contrast, publication-based approaches like that of CSRankings.org at least measure outputs of the education and research process [90], but nevertheless depend on subjective choices and values, and is sensitive to pathologies in the academic publishing system [91, 92, 93]. We use an alternative output-based ranking, based on institutional placement power, which quantifies prestige according to how well an institution is able to place its graduates as faculty at other institutions [80]. This approach avoids many of the weaknesses of other measures of institutional prestige. Notably, the prestige rankings produced by this approach strongly correlate with other computer science rankings including the U.S. News & World Report rankings, the National Research Council (NRC) rankings, and related methods based on faculty hiring [80, 77], and are representative of hiring patterns across all 8 computing subfields, indicating that all of these measures are capturing aspects of the underlying social processes that drive measures of prestige.

2.0.5 Demographic Reference Data

Finally, we compare the demographic composition of current computing faculty to two reference populations: the U.S. population and the population of U.S. computer science PhD recipients. We reconstruct the demographics of these reference populations using the U.S. Census [94, 95, 86, 87, 69] and the National Science Foundation's Survey of Earned Doctorates (SED) [96].

Most current computing faculty received their PhD within the past 40 years, but over that time period, the demographics of these two reference populations have changed substantially. A simple comparison of the diversity of current faculty to the diversity observed in a reference population at some particular point in time can be misleading. Instead, we construct a time-adjusted reference population, based on the demographics of the year each professor received their degree.

For the U.S. population, we match each professor to the U.S. census year nearest to the year of their PhD and construct from the set of such years a weighted-average demographic distribution of the U.S. Similarly, we calculate a weighted average demographic distribution of U.S. computing PhD recipients by matching faculty to the closest year recorded by the SED's records of computer and information sciences doctoral recipients, which date back to 1980. While most faculty match to the survey for their exact PhD year, 11% match to 1980, the earliest SED year, meaning they received their PhD in or prior to 1980. This procedure will tend to slightly overestimate the true diversity in the reference population. Using this methodology, we also construct reference populations for each computing subfield, which account for different age demographics across subfields.

Results

Using these augmented data, we first quantify the gender, racial, and socioeconomic representation of faculty across computing subfields and provide a quantitative view of the demographic composition at stages prior to becoming faculty. We then ask if computing departments' choices of which subfields to hire in is predictive of overall departmental gender diversity. Then, we measure differences in subfield representation across the hierarchy of institutional prestige, and quantify how subfield prestige covaries with subfield gender diversity. Finally, we use trends in subfield diversification and growth over time to forecast the future gender diversity of the field as a whole.

2.0.6 Gender, Race, and Socioeconomic Status

We find wide differences in gender composition across the 8 computing subfields (Fig. 2.1A, Table 2.1; $\chi^2 = 20.65$, $N = 4421$, $p < 0.01$), ranging from theory of computer science (13.1% women) and programming languages (14.2%), to interdisciplinary computing (19.7%) and human-computer interaction (20.0%). No computing subfield is close to being representative of gender in

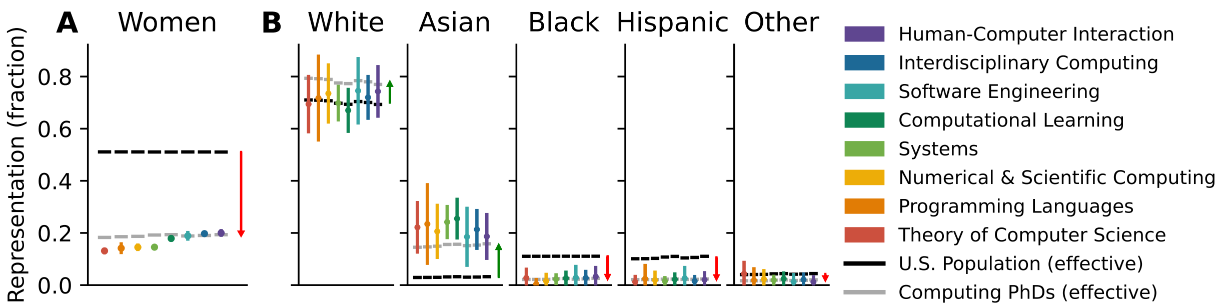

Figure 2.1: (A) Gender and (B) racial representation among faculty by computing subfield, with 95% confidence intervals, along with expected levels of representation according to two time-adjusted reference populations (see text), the U.S. population (black lines) and computing PhDs (grey lines). Gender proportions differ significantly across subfields ($\chi^2 = 20.65$, $N = 4421$, $p < 0.01$), but no computing subfield is representative of women in the U.S. reference population and all but two subfields have fewer women than expected based on the reference population of computing PhDs. Racial proportions do not differ significantly across subfields ($\chi^2 = 30.88$, $N = 547$, $p = 0.67$), but do differ significantly from the U.S. reference population. Across racial groups, representation among PhDs varies from the U.S. population (with overrepresentation denoted by upward arrows), however, representation levels among Black, Hispanic and Other faculty are close to those expected based on their representation among recipients of computing PhDs, indicating that the largest systemic source of underrepresentation occurs prior to the transition from PhD to faculty.

the U.S. reference population (51.1% women). However, the proportions of women faculty in both interdisciplinary computing and human-computer interaction modestly exceed the proportion we would expect based on the time adjusted share of women PhD recipients (19.0%). This subfield-level heterogeneity suggests that problems for gender diversity are not monolithic, and some subfields may address them more successfully than others.

In contrast, we do not find significant differences in racial composition across subfields (Fig. 2.1B; $\chi^2 = 30.88$, $N = 547$, $p = 0.67$). Rather, across all subfields, we find that some racial groups are systematically underrepresented among faculty, while others are overrepresented. To better elucidate these differences across groups, we decompose the professional pathway to becoming faculty into two stages.

The first stage spans all steps up to and including obtaining a PhD. Hence, by comparing the proportions of different racial groups in the reference U.S. population to those in the reference population of recipients of computing PhDs, we may quantify the relative rates of racial enrichment or depletion over this stage. Over this first stage, we find that White and Asian representation is enriched by factors of 1.1 and 5.0, respectively, while Black, Hispanic, and Native representation is depleted by factors of 4.5, 4.8, and 4.0 (Fig. 2.1B). For comparison, women's representation at this stage is depleted by a factor of 2.7.

The second stage spans all steps between obtaining the PhD and becoming faculty in a computing department. By comparing the racial proportions of the PhD recipient reference population with those of our faculty population, we can quantify the racialized rates of progression into the faculty workforce. Over this second stage, we find that White representation is depleted relative to the PhD recipients, perhaps because White PhDs are less likely to remain in academia (e.g., choosing positions in industry) or because they are less likely to receive and accept a faculty position. The enrichment of White representation in the first stage of the pathway to becoming faculty is largely compensated by their depletion in the second stage, so that White representation among computing faculty is very close to the expected levels, given the U.S. population overall. Conversely, Asian representation is enriched in both the first and second stages, leading to a substantial overrepresentation of Asian faculty in computing relative to the U.S. reference population. Black, Hispanic, and Native representation sees no significant enrichment or depletion in the second stage.

This evidence suggests that the largest systematic source of racial underrepresentation occurs in the first stage of the pathway to becoming faculty, prior to the transition from PhD to faculty. This first stage includes graduate admissions and retention, which are stages known to magnify racial disparities [68, 97, 98]. We note that the data we use in this study are not equipped to determine the causes of the observed population level patterns, but observing these patterns nevertheless helps to quantify how demographics change along the professional pathway.

Past analysis found that computer science faculty tend to come from highly educated families and are between 14.5 and 28.8 times more likely to have at least one parent with a PhD than the general U.S. population. Faculty in that category of high socioeconomic status are also more likely to hold a position at a prestigious institution: faculty at institutions ranked in the top 20% by U.S. News & World Report are 57.4% more likely to have a parent who holds a PhD than faculty at the least prestigious institutions [83]. Examining childhood socioeconomic status, as measured by parental educational attainment, we find no significant differences across subfields (see online supplement; $\chi^2 = 9.91$, $N = 570$, $p = 0.99$).

Faculty at the intersection of underrepresented identities are noticeably absent within our faculty sample. Black, Hispanic, and Native men comprise 3.3% of faculty who are men, while Black, Hispanic, and Native women comprise only 0.2% of women faculty. These small proportions preclude a detailed intersectional analysis. We return to this point in the discussion.

2.0.7 Departments

Most faculty are hired via searches that focus on a particular subfield of computing, e.g., a search in the area of artificial intelligence. The choice of subfield for a search may be driven by various factors, including practical needs related to the department's curriculum or strategic goals related to its research ambitions, e.g., to build on existing areas of strength or to build up a less well-established area.

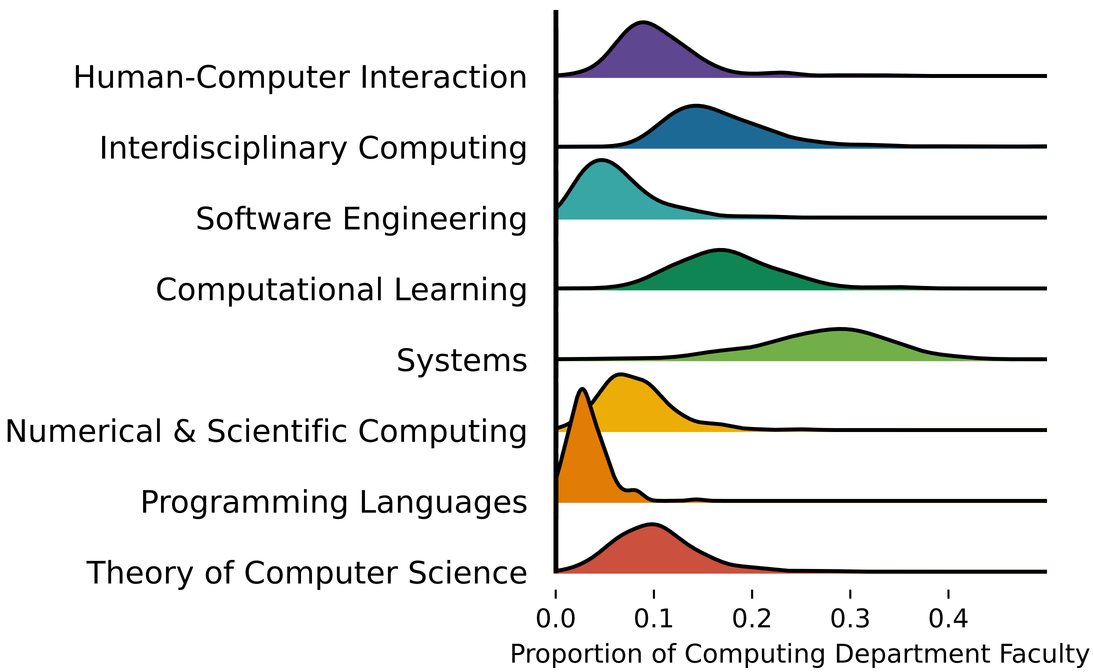

Figure 2.2: Distributions of subfield representation across computing departments showing that systems is the largest subfield (mean of departments 27%), followed by computational learning (mean 17%) while the smallest subfields are programming languages (mean 3%) and software engineering (mean 6%). Right skew in representation reflects a small number of departments that tend to specialize in that subfield.

Grouping the faculty in our data by department, we find that subfields have varying representation within computing departments. The largest subfield overall focuses on systems ($N = 1486$), and also tends to include the largest share of faculty within a typical department (mean 27%). In contrast, the smallest subfield focuses on programming languages ($N = 181$; Table 2.1) and it, likewise, tends to include the smallest share of faculty within a typical department (mean 3%). While most departments have some representation in each of the eight subfields, there are nevertheless departments that exhibit an unusually high degree of subfield specialization (Fig. 2.2), particularly when universities form departments dedicated to specific subfields. For instance, Carnegie Mellon University's Machine Learning department has the highest concentration of faculty in computational learning among all departments with 10 or more faculty in our dataset. Similarly, the University of Washington's Human Centered Design and Engineering department has the highest concentration faculty studying human-computer interaction.

Some of the most gender diverse departments heavily specialize in subfields that have more women researchers. For example, the University of Washington's Human Centered Design and Engineering department and Rochester Institute of Technology's Interactive Games and Media School are among those with the highest representation of women in our dataset, and are also the most specialized in human-computer interaction and interdisciplinary computing, the two subfields with the highest proportion of women faculty (Table 2.1). These examples highlight a connection between a department's particular subfield hiring strategy and the observed gender compositions of their faculty. In the online supplement, we show that departments' subfield compositions can meaningfully improve predictions of their gender compositions.

2.0.8 Prestige

The subfields of computing are correlated with prestige. On the high end, faculty in programming languages are 5.9 times more likely to be found in the most prestigious departments than in the least prestigious departments, while on the low end, software engineering faculty are only 2.3 times more concentrated at high prestige departments (Fig. 2.3A).

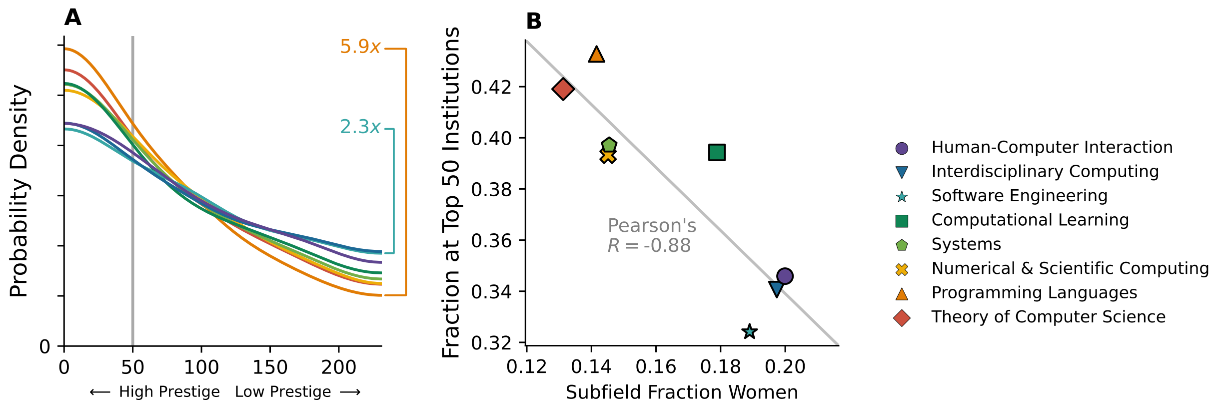

Figure 2.3: (A) Concentration of faculty across the computing department prestige hierarchy, for each computing subfield. Some subfields, such as programming languages and theory of computer science, have relatively high concentrations of faculty at highly prestigious departments, while others, including interdisciplinary computing, have higher faculty concentrations among less prestigious departments. Because more prestigious departments also tend to be larger, the highest concentrations occur among the high-prestige departments for all 8 subfields. To mitigate boundary effects in the kernel density estimation, the data is reflected across the minimum and maximum prestige points. (B) Fraction of faculty in each subfield working at the 50 most prestigious institutions, as a function of women's representation in that subfield. The strong negative correlation (Pearson's $R = -0.88$, $p = 0.004$) indicates that subfields with more women tend to have a smaller share of their faculty at these elite departments.

In fact, subfield prestige correlates with subfield gender representation (Fig. 2.3B), such that more male-dominated subfields tend to have a greater share of their researchers located at higher prestige departments. In other words, men tend to be overrepresented within the more prestigious subfields, and women are more likely to be working in the less prestigious subfields. Reflecting this pattern, we find strong correlations between a field's fraction of women faculty with both the average departmental prestige for a faculty working in a given subfield (Pearson's $R = -0.95$, $p = 0.0003$) and the fraction of faculty in the top 50 ranked institutions (Pearson's $R = -0.88$, $p = 0.004$, Fig. 2.3B). Even after adjusting for a professor's PhD-granting institution's prestige, their publication productivity, and their gender, a multiple linear regression shows that faculty who study more male-heavy topics are still more likely to hold positions at higher prestige departments (Pearson's $R = -0.82$, $p = 0.01$), such that faculty fully specialized in the most prestigious subfield (programming languages) are expected to be located 12 ranks higher than faculty fully specialized in the least prestigious subfield (human-computer interaction). Because the most prestigious institutions train the majority of future faculty [77, 81], the underrepresentation of gender diverse subfields among these institutions may act as a structural barrier to the gender diversity of computing as a whole.

We note that in this model, the coefficient associated with faculty gender is not statistically distinguishable from 0 ($p = 0.13$). This fact suggests that both women **and men** in subfields with more women are expected to hold faculty positions lower in the prestige hierarchy. For more details on the regression findings, including regression tables, see the online supplement.

2.0.9 Trends

Over the past 40 years, both the sizes and demographics of subfields have changed substantially. We can estimate the temporal dynamics of these variables by assigning current computing faculty to cohorts, according to the year they received their PhD, and then track how demographic and subfield distributions change over cohorts. Many faculty do not start their first faculty position until several years after completing their PhD, a pattern which would induce systematic undersam-

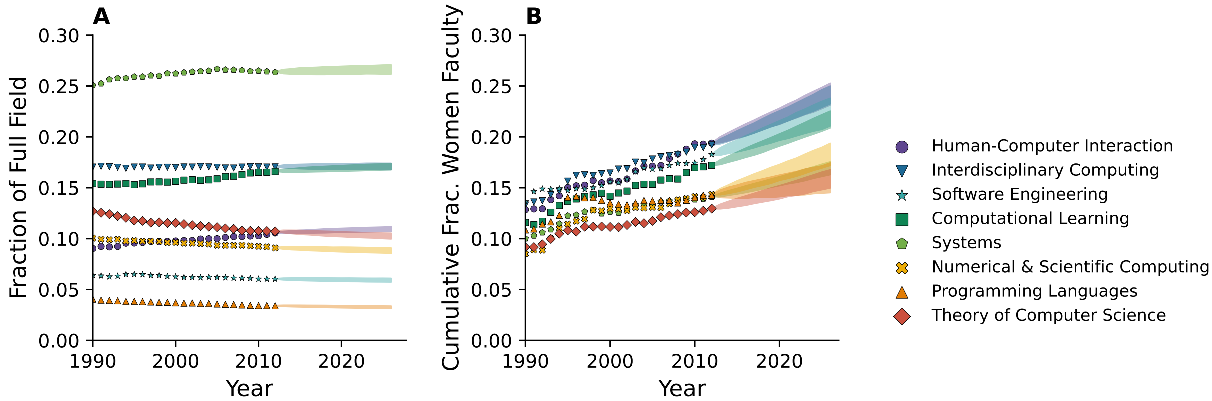

Figure 2.4: (A) Yearly subfield size relative to computing as a whole for 1990 to 2012 and (B) cumulative fraction of faculty that are women from 1990 to 2012, for each of the eight subfields of computing, along with 95% confidence interval forecasts, projected out to 2027, showing that we may expect the bimodal distribution of gender across subfields to continue into the foreseeable future.

pling of the most recent cohorts. To control for this effect, we report size and demographic estimates only up to the 2012 cohort. We then forecast subfield sizes and demographics 15 years into the future by extrapolating the historic trends in subfield faculty hiring over time and the yearly gender compositions of new hires.

Analyzing these data, we find that subfields' relative sizes have remained relatively stable over time (Fig. 2.4A), even as the field as a whole has grown substantially in absolute terms. In the 22 years between 1990 and 2012, the largest increase in relative size is in human-computer interaction (+1.5%), and the largest decline is in theory of computer science (-2.0%). Despite enormous parallel changes in the field of computing itself since the year 2000, trends in relative subfield size appear largely stable over the past 20 years (Fig. 2.4A).

All computing subfields have increased their representation of women faculty over time, though at varying rates, and some subfields are substantially closer to parity than others (Fig. 2.4B). Women's representation in each annual faculty cohort has increased an average of 0.43% per year. However, because of the generational nature of the field's composition, these annual increases in each additional cohort's gender diversity accumulate to a more modest field-level average increase of 0.2% per year [71, 72]. These rates of change are in close agreement with past estimates for computing [81, 99]. In the 22 years between 1990 and 2012, programming languages and theory of computer science have increased women's representation by the lowest relative amounts—3.4% and 3.7%, respectively—while interdisciplinary computing and human-computer interaction have increased by 5.9% and 6.5%, respectively. Although subfields' increases in women's representation have been relatively steady over time, their slow pace produces forecasts that predict only two fields—interdisciplinary computing and human-computer interaction—are likely to reach 25% women faculty by 2027, assuming historical trends continue.

Our data indicate that women's representation in the four most diverse subfields (human-computer interaction, interdisciplinary computing, software engineering, and computational learning) is roughly 25 years ahead of women's representation in the four least diverse subfields, and this gap is projected to persist over the next 15 years.

Discussion

Using comprehensive data on the education, employment, research subfield, and demographic variables of tenure-track faculty at U.S.-based PhD-granting computing departments, we quantified the intersection of multiple forms of demographic diversity by computing subfield, producing a detailed picture of past and likely future trends and inequalities.

Although we find little variation in racial representation across subfields, our analysis reveals several interesting patterns about the pathway to becoming faculty, by comparing racial representation among current faculty to that of computing PhDs and the U.S. population as a whole (Fig. 2.1B). These comparisons divide the faculty pathway into two stages. The first spans all steps prior to obtaining a PhD in computing, including doctoral training, and the second spans all steps between obtaining a PhD and becoming faculty in a computing department. Over the first stage, White and Asian representation is enriched, while Black, Hispanic, and Native representation is depleted. Over the second stage, White representation is depleted, while Asian representation is further enriched, and Black, Hispanic, and Native representation remains low without substantial change.

These patterns are consistent with racialized factors influencing retention at multiple points in the pathway to becoming faculty in computing, and the direction and magnitude of that influence is not necessarily uniform across stages. For instance, the up and then down pattern of White representation indicates a substantial decrease in White retention after the PhD. The availability of attractive non-academic careers for computing PhDs, e.g., in the computing industry, is one plausible explanation for this decrease. White PhDs may also be more likely to pursue tenure-track jobs at non-PhD granting institutions, which are not included in our data. The down and then stable pattern of Black, Hispanic, and Native representation indicates that a large portion of the systemic effects occur prior to receiving a PhD in computing, e.g., in graduate school and college—a pattern that is well-documented by studies of race in academia [100, 101]. In contrast, Asian representation in the second stage (post-PhD) increases by about the same amount that White representation decreases, suggesting additional racialized differences in achieving a faculty career after a PhD.

Patterns underlying the underrepresentation of women computing faculty are similar to patterns of Black, Hispanic and Native underrepresentation, where the largest share of depletion occurs prior to receiving a PhD in computing (Fig. 2.1A,B). In contrast to racial diversity, we find that gender diversity varies substantially across subfields, even as overall gender diversity also remains low (16.7%, Table 2.1), and is increasing at only about 0.2% per year. The subfields of human-computer interaction, interdisciplinary computing, computational learning (which includes artificial intelligence), and software engineering exhibit substantially greater gender diversity among faculty (19.0%). At the current rate of gender diversification, this level of gender diversity places them roughly 25 years ahead of the remaining four subfields (14.2%).

Past work has developed a number of interacting explanations for gender, racial, and intersectional underrepresentation among U.S. computing faculty, including culturally pervasive gendered and racialized stereotypes, which may shape career decisions [73, 102, 9], inhospitable educational and professional climates [103, 104, 105, 18, 11, 20, 22, 21], structural disparities in education and socioeconomic status [106, 107, 83, 108, 109] and the unequal impact of parenthood [25, 23, 26, 27].

Our results do not identify any specific underlying mechanisms for differential representation, and instead quantify patterns in ways that support further research in this direction. Our results suggest that more work is needed to understand how interactions between industry and academia shape the demographic diversity of computing faculty. These interactions are likely important early in the faculty pathway, and later, e.g., where gendered or racialized hiring rates of senior faculty into industry can effectively increase the demographic diversity in academia [110].

The four most gender diverse subfields represent fully half (50.4%, Table 2.1) of all computing faculty. They are also substantially underrepresented among high prestige departments (Fig. 2.3A,B), which exert substantial influence over field-level norms, culture, and research agendas due to their status and their role in training the majority of computing faculty [77, 111]. This difference holds even after controlling for factors like doctoral institution prestige, productivity, and gender itself, such that faculty working in more gender diverse subfields work at institutions, on average, 12 ranks lower than faculty working in less gender-diverse subfields. This gender-prestige pattern illustrates a kind of systemic devaluing of women's contributions to computing overall, and the substantial size of the more diverse but less prestigious group of subfields raises the question of whether they are adequately represented among departmental curricula and degree requirements. Realigning institutional practices to reflect the true diversity of computing's subfields may help institutionalize efforts to broaden participation.

Our retrospective analysis of subfield growth and gender shows that gender diversity is increasing at similar rates across all eight computing subfields. However, current gender diversity is essentially bimodal, with four of the eight subfields (human-computer interaction, interdisciplinary computing, software engineering, and computational learning) being substantially more gender diverse than the other four (systems, numerical & scientific computing, programming languages, and theory of computer science). Our forecasting exercise indicates that these differences are likely to continue into the foreseeable future (Fig. 2.4A,B), even as some of the less gender-diverse subfields appear to be shrinking (theory of computer science) while some of the more gender-diverse subfields are growing (computational learning, and human-computer interaction). As a result, the overall trend of slow gender diversification is highly robust to minor changes in hiring patterns among subfields.

Our methodology for analyzing demographic patterns and trends among subfields of computing is general, and could be applied to any other academic field, given an appropriate subfield taxonomy. Applied to many fields, this approach could elucidate the systemic role that subfields play in driving field-level demographic patterns, and help identify new insights into field-specific systematic barriers to broadening participation.

There are a number of limitations to our methodology. Although DBLP provides good general coverage of computing publications, our analysis inherits DBLP's publication inclusion bias over areas of computing, which is largest in older and in more interdisciplinary areas of research [112]. We also use the year in which a faculty received their PhD to estimate the relative sizes and gender compositions of subfields over time. This assignment assumes that faculty in our sample started their faculty jobs immediately after their PhDs and that they are representative of faculty who left jobs prior to 2010, the first year observed in our data. As a result, we are likely underestimating the historic participation of women in computing (Fig. 2.4B), because women faculty have historically left their positions at higher rates than men [12, 22, 26]. This historical underestimate would imply that our estimate of gender diversification rates are likely upper bounds. Our data are limited to tenure-track faculty employed by PhD granting institutions, and do not support an analysis of contingent faculty, who make up a growing share of faculty [113, 114] , or faculty at non-PhD granting institutions, who may exhibit different demographic compositions. We do not separately analyze faculty who hold multiple minority identities. Past research shows that people at the intersection of multiple identities often experience discrimination and exclusion beyond what would be expected from simply adding the individual elements of their identities [115]. The small sample of faculty for whom we have race data limits our ability to conduct a detailed quantitative analysis of the least represented groups, and in particular, Black, Hispanic, and Native faculty, or to conduct intersectional analyses.

We now return to the idea that explanations for slow rates of diversification in computing can be divided into categories. On the one hand, generational problems introduce a lag in faculty diversity, where, if the pathway to faculty positions were to suddenly become equitable, it would still take many years for this change to manifest as equitable representation among faculty [71, 72]. On the other hand, there are structural and social climate problems that tend to push or pull members of underrepresented groups away from faculty positions, sometimes in different magnitudes and directions depending on the career stage [73].

Our findings identify and quantify a third type of explanation, where the diversity of computing is driven by diversity differences across its subfields. The computing community must explore several questions before these findings can be translated into concrete policy recommendations. For example, the differences in diversity and prestige that we find across the subfield structure of computing suggest a simple departmental strategy for enhancing the probability of hiring women faculty: increase hiring in the subfields with greater gender diversity, such as human-computer interaction and interdisciplinary computing (20% women). While this strategy may be an effective way to increase women's representation for computing as a whole, it is unlikely to reduce the heterogeneity in gender diversity across subfields.

Future research could help shape how we design policy to increase diversity in computing, by identifying the causal mechanisms driving gender differences across subfields. On one hand, some subfields may be particularly inhospitable to women, effectively pushing women away. In this case, policy should aim to make these subfields more accessible and inclusive. On the other hand, women may, on average, be more interested in topics belonging to some subfields over others, i.e., some subfields exert stronger pulls [66]. In this case, policy should respect the validity of women's interests by expanding the subfields that have greater pulls, instead of pushing to increase representation where there is less interest.

An additional causal understanding of the relationship between subfield gender diversity and subfield prestige would provide further context for policy recommendations. The tendency for male-dominated areas of work to be assigned greater prestige, and hence for areas of work with greater gender diversity to be less valued, is not a phenomenon special to computing. Gendered patterns are also observed in medical subspecialties, in different areas of law [116, 117], and even in less specialized positions [118, 119]. One explanation of this pattern posits a direct causal relationship between an occupation's diversity and its prestige [120]. If this explanation applies to computing, then it may not be feasible to simultaneously increase both a subfield's prestige and its gender diversity without first making more foundational changes to collective values and beliefs. This relationship remains untested in computing, but is an important question for diversity because the departments at the top of the prestige hierarchy tend to train the majority of future computing faculty [77].

A subfield-focused hiring strategy alone is unlikely to increase racial or socioeconomic diversity, as we find that these faculty characteristics do not appear to correlate with subfield in our sample. Different approaches will be needed to improve representation along these dimensions, and our findings suggest these should include interventions that increase representation among PhD recipients. Some programs are available as models for future work in this direction, including the Distributed Research Experiences for Undergraduates (DREU) and the Collaborative Research Experiences for Undergraduates (CREU), two funded research programs intended to broaden participation in computing, with participants twice as likely to attend graduate school than standard REU participants [121]. Academic institutions are also turning to the University of Maryland, Baltimore County's Meyerhoff Scholars Program as a model for their own scholarship programs, which have been shown to markedly improve undergraduate retention and STEM graduate school matriculation for underrepresented minorities [122]. Doctoral programs can additionally establish partnerships with minority serving institutions (MSIs), as modeled by the highly successful Fisk–Vanderbilt Masters-to-PhD Bridge Program, which substantially contributes to the number PhDs earned by underrepresented minorities in a number of STEM fields, but has yet to expand to computing [123, 98]. These are a few examples of programs that can be implemented or expanded to additonal academic institutions to increase accessibility for underrepresented groups, in conjunction with other efforts to mitigate the social climate problems in computing [75, 97, 124, 125].

For computing departments to benefit from the innovative scientific research that diverse scientists produce [1], diversity and inclusion efforts must contend with generational, social climate, and subfield problems. For example, structural improvements to recruitment, like those suggested here, are by themselves no guarantee that diverse faculty will be adequately included and supported once they begin their faculty jobs [126, 127, 104]. Cultural change can also be slow, and also does not guarantee diverse representation among faculty. The empirical patterns and trends shown here provide new insights that can inform and support multifaceted efforts to make computing more diverse, equitable, and inclusive.

The work presented in this chapter was co-authored by K. Hunter Wapman, Allison C. Morgan, Sam Zhang, Daniel B. Larremore, and Aaron Clauset. I would additionally like to thank Bor-Yuh Evan Chang, Leysia Palen, Ben Shapiro, Huck Bennett, Joshua Grochow, and Jed Brown for helpful comments, and all survey participants for providing their valuable time.

Chapter 3

Gendered hiring and attrition on the path to parity for academic faculty

This chapter is adapted from:

N. LaBerge, K. H. Wapman, A. Clauset, D. B. Larremore, "Gendered hiring and attrition on the path to parity for academic faculty." *eLife* (2024) 13:RP93755.

Introduction

Faculty play a crucial role in educating future researchers, advancing knowledge, and shaping the direction of their fields. Diverse representation of social identities and backgrounds within the professoriate improves educational experiences for students [19, 128], accelerates innovation and problem-solving [1, 2, 129], and expands the benefits of scientific advances to a broader range of society [4, 66]. Gender equality in particular is a foundational principle for a fair and just society in which individuals of any gender identity are free to achieve their full potential. However, the composition of the professoriate has never been representative of the broader population, in part because higher education has remained unattractive or inaccessible to large segments of society [130, 131, 83].

Over the past 50 years, U.S. higher education has made substantial gains in women's representation at the undergraduate and PhD levels, but progress towards greater representation among tenure-track faculty has been much slower. Women have earned more than 50% of bachelor's degrees since 1981 [132], and now receive almost half of doctorates in the U.S. (46% in 2021) [133]. However, women comprise only 36% of all U.S. tenured and tenure-track faculty [130], and there are significant differences in women's representation across disciplines. For example, many fewer women earn PhDs in Science, Technology, Engineering, and Mathematics (STEM) fields (38%), compared to PhD recipients in non-STEM fields (59%) [133].

There are two primary ways by which women faculty representation changes: through hiring and through attrition. In our analysis of faculty demographics, faculty attrition refers to "all-cause attrition," which encompasses all the reasons that may lead someone to leave the professoriate, including retirement or being drawn to non-academic activities in the commercial sector. If the proportion of women among incoming hires is greater than the proportion of women among current faculty, then new hires will slightly increase the field's representation of women. On the other hand, if the proportion of women among faculty who leave their field is greater than the proportion of women among current faculty, then attrition will slightly decrease the field's representation of women. Because faculty often have very long careers, a trend in a field's overall gender representation is a cumulative integration, over many years, of the net differences in representation caused by a mixture of hiring and attrition.

Policies targeting gender parity can focus on changes to hiring, attrition, or both, and many have been tried. Hiring-focused policies include grants for diverse faculty recruitment [134], efforts to reduce bias in the hiring processes [78], and a range of other measures intended to increase women's representation at earlier educational and career stages. Attrition-focused policies include initiatives to reduce gender bias in the promotion and evaluation of faculty [135], efforts to diminish the gender wage gap among faculty [136], and diversity-focused grants for early and mid-career faculty research [137]. Some policies simultaneously impact hiring and attrition. For example, improvements to parental leave and childcare policies may lessen the attrition of women who become parents as faculty and also encourage more prospective women faculty to consider faculty careers.

Although both types of strategy can be important, their impact alone or together on historical trends in gender diversity remain unclear. Also, we lack a clear prediction of how gender diversity may change in the future and whether current trends, and the policies that support them, may ultimately achieve gender parity. Some studies have used empirically informed models of faculty hiring, attrition, and promotion to estimate the effectiveness of certain specific policy interventions [72, 138, 139, 140, 141, 142]. However, most focus on a single institution, which tends to limit their generalizability to whole fields or to other institutions. Field-wide assessments and cross-field comparisons are necessary to provide a clear understanding of the overall patterns and their variations. Such broad comparative analyses would support evidence-based approaches to policy work and would shed new light on the causes and consequences of persistent gender inequalities among faculty.

In this study, we aim to quantify the individual and relative impacts of faculty hiring and attrition on the historical, counterfactual, and future representation of women faculty across fields and institutions. Our models and analyses are guided by a census-level dataset of faculty employment records spanning nearly all U.S.-based PhD-granting institutions, including in 111 academic fields across the Humanities, Social Sciences, Natural Sciences, Engineering, Mathematics & Computing, Education, Medicine, Health, and Business. This wide coverage allows us to quantify broad patterns and trends in both hiring and attrition, across institutions and within fields and develop model-based extrapolations under a variety of possible policy interventions.

Results

We take three distinct approaches in our analysis of the relative importance of faculty hiring and faculty attrition for women's representation among tenure track faculty. First, we characterize the relative contributions of hiring and attrition to changes in women's representation across a range of academic fields over 2011–2020. Second, we model a hypothetical historical scenario over this same period in which we preserve demographic trends in hiring, but we eliminate "gendered attrition" by assigning equal attrition rates to men and women at each career stage. Here, gendered attrition refers only to the differences in the rates in which men and women at the same career stage leave academia. It does not refer to the absolute magnitude of the rates, which increases for both men and women in the late career as faculty approach an age where retirement is common. This counterfactual model provides data driven estimates of what different fields' gender diversity could have been, and hence provides a general estimate of the loss of diversity due to gendered attrition over this time. Finally, we use our hiring and attrition model to make quantitative projections of the potential impact of specific changes to faculty hiring and faculty attrition patterns on the future representation of women in academia, allowing us to assess the relative impact of practical or ambitious policy changes for achieving gender parity among faculty by field and in academia overall.

For these analyses, we use a census-level dataset of employment and education records for tenured and tenure-track faculty in 12,112 PhD-granting departments across 392 PhD granting institutions in the United States from 2011-2020 [143]. We organize these data into annual department-level faculty rosters. In turn, each department belongs to at least one of 111 academic fields (e.g., Chemistry and Sociology) and one of 11 high-level groupings of related fields that we call domains (e.g., Natural Sciences and Social Sciences), enabling multiple levels of analysis. This dataset was obtained under a data use agreement with the Academic Analytics Research Center (AARC), and was extensively cleaned and preprocessed to support longitudinal analyses of faculty hiring and attrition [130] (see Methods for data cleaning details).

We added gender annotations to faculty using nomquamgender, an open source name-based gender classification package that is comparable in performance to the most reliable paid name-based gender classification services [144]. Gender annotations were applied to faculty names with high cultural name-gender associations (88%) [144], resulting in a dataset of $n = 268,769$ unique faculty, making up 1,768,118 person-years. The methodology we use assigns only binary (woman/-man) labels to faculty, even as we recognize that gender is nonbinary. This approach is a compromise due to the technical limitations of name-based gender methodologies and is not intended to reinforce a gender binary.

We define faculty hiring and faculty attrition to include all cases in which faculty join or leave a field or domain within our dataset. For example, hires include first-time tenure track faculty, and mid-career faculty who transition from an out-of-sample institution (e.g., from a non-U.S. or non PhD granting institution, or from industry). Examples of faculty attritions include faculty who leave for another job in academia to an institution outside the scope of our dataset (e.g., non-U.S. or non PhD granting), faculty who leave the tenure-track, faculty who move to another sector, and faculty who retire. Faculty who transition from one field to another are counted as an attrition from the first field and a hire into the new field. Finally, faculty who switch institutions but remain in-sample and in the same field are not counted as hires or attritions.

3.0.1 Historical impacts of hiring and attrition

Our data show a clear increase in women's representation between 2011 and 2020, increasing by an average of 4.8 percentage points (*pp*) across fields. Trends among new hires can drive increases in women faculty's representation, and in many fields women's representation among PhD graduates has been growing for many years [145]. At the same time, attrition can also drive increases in women's faculty representation if the trends run in the opposite direction, e.g., in many fields retiring faculty are more likely to be men than women (Fig. A.8) [130]. However, attrition in the early- or mid-career stages may have the opposite effect if it is gendered, e.g., when women comprise a greater proportion of those leaving academia at these career stages [12, 146]. The balance of these inflows and outflows, relative to a field's current composition, determines whether women's overall representation will increase, decrease, or hold steady over time.

To investigate the effects of hiring and attrition on changes in women's representation between 2011 and 2020, we decomposed the total change in representation into separate hiring effects and attrition effects for each of our studied fields (Fig. 3.1, Table A.2; see Methods). A total of 106 (95%) of 111 fields saw an increase in women's representation overall, with hiring contributing to increases in 106 of 111 fields (95%), and attrition contributing to increases in 82 of 111 fields (74%). In general, hiring was the larger cause of increases to women's representation, with greater effects in the majority (87.4%) of academic fields.

The effects of hiring do not always increase women's representation, nor do they always dominate the effects of attrition. For instance, hiring has contributed to negative changes in women's representation in 5 fields, including the majority-women fields of Nursing and Gender

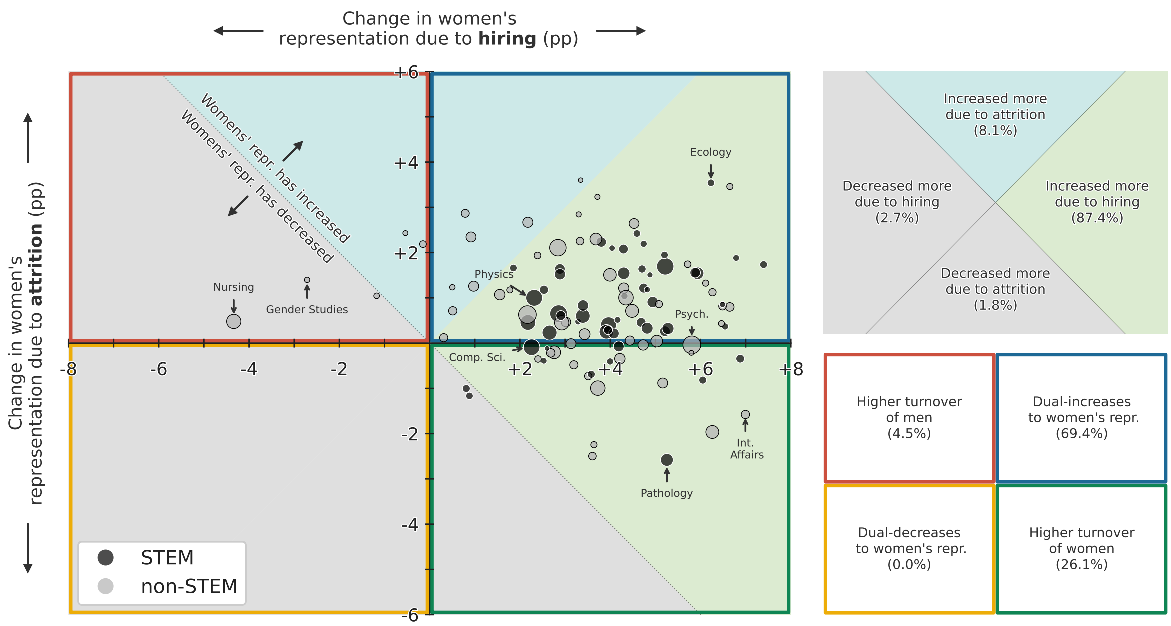

Figure 3.1: Change in women's overall faculty representation for 111 academic fields between 2011–2020, decomposed into change due to hiring (horizontal axis) and change due to attrition (vertical axis, see Supplementary Sec. A.1), showing that hiring increased women's representation for a large majority (87.4%) of fields, while it decreased women's representation for five fields. Point size represents the relative size of each field by number of faculty in 2020, and points are colored by STEM (black) or non-STEM (gray).

Studies. Our analysis also finds that 9 fields (8.1%) saw increases driven more by attrition than hiring, of which all are non-STEM fields.

The decomposition into hiring and attrition effects also shows that attrition can decrease women's representation, even if women's representation increases overall. In fact, faculty attrition decreased women's representation in 29 fields (26%) between 2011 and 2020 (Fig. 3.1, Table A.2). These attrition-driven changes likely reflect a failure to retain early and mid-career women, as they are unlikely to be driven by retirements: among faculty, men are more likely to be at or near retirement age than women faculty due to historical demographic trends (Fig. A.8). Indeed, of these fields, 25 (83%) of 29 are majority men, such that differential losses of women due to attrition move such fields away from parity. Nevertheless, despite net losses of women faculty to attrition, the effects of hiring were large enough to see overall increases in women's representation in 27 (93%) of 29 of these fields.

3.0.2 Quantifying the impact of gendered attrition

The fact that men are systematically more likely to be at or near retirement age than women across fields (Fig. A.8; see also [130]) means that it is possible for all-cause attrition to cause a net-increase in women's representation (Fig. 3.1), even if women leave the academy at higher rates at every career stage. Indeed, recent work has shown that this is the case [146], a phenomenon termed *gendered attrition*. These findings together suggest that in some fields, women's representation might be higher had attrition been gender neutral over the past decade.

To estimate the potential impacts of gender-neutral attrition, we created a counterfactual model in which we fixed men's and women's attrition risks to be equal at every career age (see Methods). By initializing the model at our 2011 faculty census data, and preserving demographic trends in hiring, we simulated $n = 500$ counterfactual demographic trajectories for 2011-2020 under gender-neutral attrition for each field. Here, our model bears an important resemblance to the seminal Leslie Matrix model used and adapted by demographers and ecologists (see, e.g. [147]), with a few notable differences to ensure the total faculty population size and the distribution of

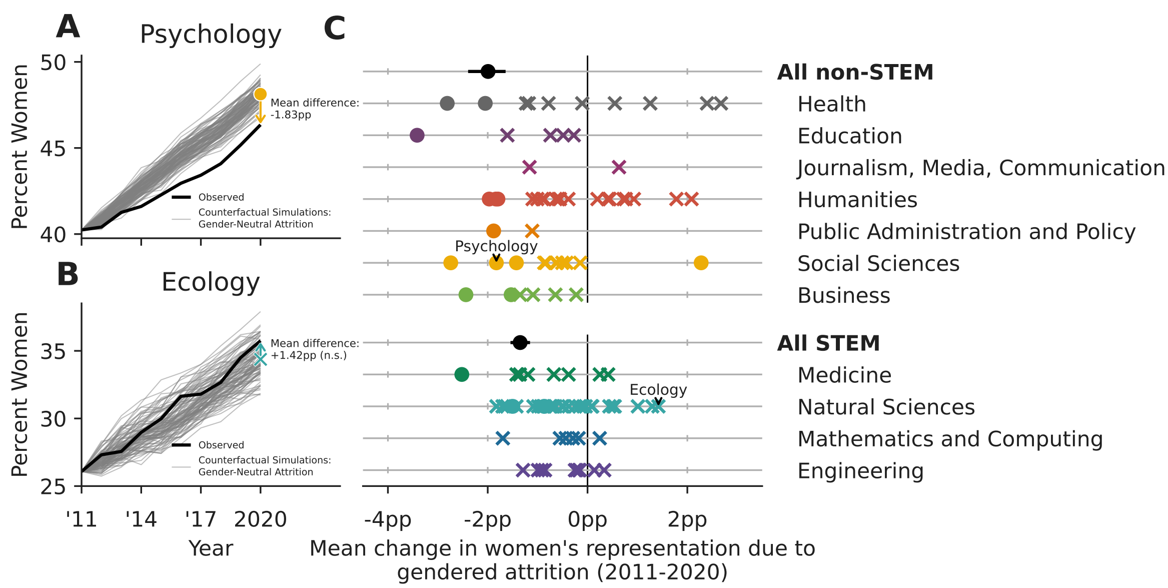

Figure 3.2: Gendered faculty attrition has caused a differential loss of women faculty in both STEM and non-STEM fields. (A) Gendered attrition in psychology has caused a loss of -1.83 *pp* ($p < 0.01$) of women's representation between 2011-2020, relative to a counterfactual model with gender-neutral attrition (see Methods Sec. 3.0.6). In contrast, (B) gendered attrition in Ecology has not caused a statistically significant loss ($+1.42$ *pp*, $p = 0.24$). Relative to their field-specific counterfactual simulations, 15 academic fields and the STEM and non-STEM aggregations exhibit significant losses of women faculty due to gendered attrition (circles on figure; two-sided test for significance relative to the gender-neutral null distribution derived from simulation, $\alpha = 0.1$). The differences in the remaining 95 fields were not statistically significant (crosses on figure), but we note that their lack of significance is likely partly attributable to their smaller sample sizes at the field-level compared to the all STEM and all non-STEM aggregations, which exhibited large and significant differences. Error bars for the non-STEM and STEM aggregations contain 95% of $n = 500$ stochastic simulations. No bars are included for field-level points to preserve readability.

career ages at hiring match historical data (see Methods). From these trajectories, we quantified the effect of gendered attrition as the difference in women's representation between the real 2020 census and gender-neutral simulations, and labeled effects as significant only if 95% of simulations ended with either higher or lower representation of women. For instance, there were 1.83 *pp* fewer women in 2020 in Psychology due to a decade of significant gendered attrition (Fig. 3.2A; $p < 0.01$), whereas we find no significant gendered attrition in Ecology (Fig. 3.2B; $p = 0.24$).

A total of 16 fields exhibited significantly gendered attrition between 2011 and 2020 (circles, Fig. 3.2C). Of these, 15 fields, including Psychology, Philosophy, Chemistry, and Sociology, ended 2020 with fewer women than our gender-neutral attrition model predicted, while just one ended 2020 with fewer men (Gender Studies). Counterfactual simulations for the remaining 95 fields provided inconsistent outcomes, either towards greater or lesser representation of women faculty, for at least 5% of simulations (crosses in Fig. 3.2C, Table A.2). In general, simulations for smaller fields tended to exhibit more variable outcomes which were consequently less often statistically significant.

To evaluate the effects of gendered attrition at a higher level of aggregation, we also estimated the impacts of gendered attrition for all STEM and all non-STEM fields, respectively. Gendered attrition was significant in both cases, decreasing women's representation by -1.35 *pp* (STEM, $p < 0.01$) and -1.99 *pp* (non-STEM, $p < 0.01$) relative to counterfactual simulations (Fig. 3.2C). Aggregating all fields, our counterfactual model estimates that gendered attrition has caused a net loss of 1378 women faculty from the PhD-granting sector of the U.S. tenure track between 2011 and 2020. Assuming 19.2 faculty per department (the mean department size in our dataset), this is an asymmetric loss of approximately 72 entire departments.

3.0.3 Projecting future gender representation

Proposed strategies to increase faculty gender diversity often emphasize changes to hiring or retention. However, administrators and policymakers typically lack any ability to quantitatively evaluate a policy's long-term impact or to compare its outcomes against alternatives. In our third

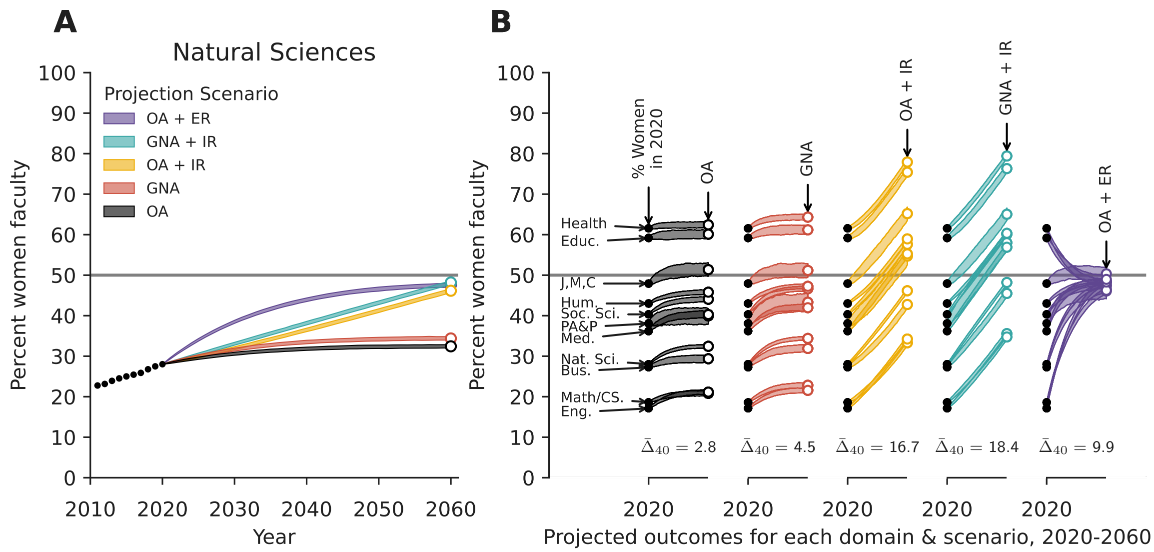

Figure 3.3: (A) Observed (dotted line, 2011-2020) and projected (solid lines, 2021-2060) faculty gender diversity for Natural Sciences over time and (B) projections for 11 academic domains over 40 years under five policy scenarios. Line widths span the middle 95% of $N = 500$ simulations and $\bar{\Delta}_{40}$ gives the mean change in women's representation across domains over the 40-year period. Educ. = Education, J,M,C = Journalism, Media & Communications, Hum. = Humanities, Soc. Sci. = Social Sciences, and PA&P = Public Administration & Policy., Med. = Medicine, Nat. Sci. = Natural Sciences, Bus. = Business, Eng. = Engineering. See text for scenario explanations. OA = observed attrition, GNA = gender-neutral attrition, IR = increasing representation of women among hires (+0.5 *pp* each year), ER = equal representation of women and men among hires.

analysis, we therefore turn our counterfactual model, previously used to investigate the past, to investigate five projections capturing future hiring and attrition scenarios. These scenarios are not intended to predict or forecast precisely what will occur in the future, but instead serve to illustrate what could happen if a set of specific assumptions were held fixed. We operationalize a particular policy intervention by altering two parameters: faculty attrition risks, which can be gendered or gender-neutral, and the fraction of women among new hires, which can be maintained at 2011–2020 levels or increased over time. These scenarios are described in detail in Methods Sec. 3.0.8. Across all scenarios, we hold the sizes of academic domains fixed at their 2020 values, initialize projections at empirical 2020 values, and project women's representation through 2060.

Even in the absence of interventions, our baseline projection using current hiring patterns and observed attrition (OA) shows a mean increase in women's representation of 2.8 *pp* (Fig. 3.3). Although the fraction of women among new hires has been increasing over time in some academic domains (Supplementary Sec. A.2), here we do not extrapolate that trend into the future. Thus, this projection reflects demographic inertia, where it takes roughly a full career-length of time for the most recent, more gender diverse cohorts of newly hired faculty to fully replace all the older and less gender-diverse faculty cohorts [71, 72].

The gender-neutral attrition (GNA) scenario maintains the same assumption about hiring as the OA scenario, but alters the attrition risks to be equal for men and women at each career stage. This scenario therefore represents a set of policy interventions that entirely close the retention gap between women and men faculty. The resulting projected fraction of women faculty in the Natural Sciences domain in 2060 is 34.4%, representing a 6.4 *pp* increase from 2020 (Fig. 3.3A). Across all domains, the mean increase in women's representation relative to 2020 is 4.5 *pp* (Fig. 3.3B), exceeding the mean increase in the OA scenario by 1.7 *pp*. Nevertheless, in this scenario women are projected to remain underrepresented in most academic domains in 2060.

The increased representation (IR) scenarios alter new faculty hiring such that women's representation among new hires grows by +0.5 *pp* each year. This rate of growth ultimately increases women's representation among new hires by +20 *pp* by 2060, and is close to the median observed change 2011-2020 (+0.58 *pp*; Supplementary Table A.1). Thus, for some domains, this scenario may not represent new action, but rather continued effects of current policies.

When increased representation among new hires is combined with observed attrition (OA+IR), the projected mean increase in women's overall representation is 16.7 *pp* (Fig. 3.3). The magnitude of this increase is the largest among these first three scenarios, exceeding the OA scenario's mean increase by 12.2 *pp*. We note, however, that these increases would grow women's representation beyond gender parity in Medicine, Social Sciences, Humanities, and Public Administration & Policy, and would maintain representation above parity in Health and Education (Fig. 3.3).

When increased representation among new hires is instead combined with gender-neutral attrition (GNA + IR), the projected mean increase in women's overall representation is 18.4 *pp* (Fig. 3.3), exceeding the OA + IR scenario by only 1.7 *pp*. These additional modest increases do not push any additional academic domains beyond gender parity (Fig. 3.3).

Finally, we consider a more radical intervention in which universities immediately and henceforth hire men and women faculty at equal rates (ER) but do not change attrition patterns (OA + ER). Overall, the projected mean increase in women's representation between 2020 and 2060 is 9.9 *pp* (Fig. 3.3), a result of all domains moving markedly toward parity. However, none of the academic domains is projected to achieve stable gender parity under this scenario because attrition remains gendered. In domains that are particularly male-dominated, such as Natural Sciences, this gender-parity hiring scenario causes the most rapid progress toward greater representation of women faculty of any of the scenarios (Fig. 3.3A).

Intentionally omitted is the scenario of equal-rate hiring and gender-neutral attrition (ER + GNA), for which conclusions can be drawn without simulation. Any academic domain with hiring at parity and equal retention rates between women and men is guaranteed to achieve stable parity, modulo stochastic effects, after one complete academic generation.

The five projection scenarios show that changes to hiring drive larger increases in the long-term representation of women faculty of a field. Most fields, and particularly those with the least gender diversity, cannot achieve equal representation by changing attrition patterns alone. In fact, our results suggest that even relatively modest annual changes in hiring tend to accumulate to substantial field-level changes over time. In contrast, eliminating gendered attrition leads to only modest changes in women's projected representation (Fig. 3.3).

Discussion

In this study, we used a decade of census-level employment data on U.S. tenured and tenure track faculty at PhD-granting institutions to investigate the relative impacts of faculty hiring versus all-cause faculty attrition on women's representation across academia. Toward this end, we answer three broad questions: (i) How have these two processes shaped gender representation across the academy over the decade 2011–2020? (ii) How might women's representation today have been different if gendered attrition among faculty were eliminated in 2011? (iii) And, how might we expect gender diversity among faculty to change over time under different future hiring and attrition scenarios?

The effects of hiring were stronger than the effects of attrition in changing women's representation among faculty over the decade 2011–2020. By separating hiring and attrition effects, we found that hiring served to increase women's representation in 96% of fields, while attrition (including retirement) served to increase women's representation in 74% of fields. However, these effects were not always synergistic, such that 26% of fields saw increases due to hiring amidst decreases due to attrition, while 5% of fields saw decreases due to hiring and increases due to attrition. In total, hiring effects dominated attrition effects in 90% of fields, 97% of which saw net increases in women's representation (Fig. 3.1). In contrast, attrition effects were stronger than hiring effects in just 10% of fields, yet 9 of these 11 nevertheless saw net increases in women's representation, including Linguistics, Theological Studies, and Art History. Only 5 (4.5%) fields saw decreased women's representation over same period, including Nursing and Gender Studies, both fields where women are over-represented (Fig. 3.1).

Our counterfactual analyses of the past decade's gender diversity trends, in which we preserved historical hiring trends but eliminated the effects of gendered attrition, indicate that U.S.

academia as a whole has lost approximately 1378 women faculty because of gendered attrition. This number is both a small portion of academia (0.67% of the professoriate), and a staggering number of individual careers, enough to fully staff 72 nineteen-person departments. Because women are more likely to say they felt "pushed" out of academic jobs when they leave [146], each of these careers likely represents an unnecessary loss, both to the individual and to society, in the form of lost discoveries [1, 4, 66, 129], missing mentorship [19, 128], and many other contributions that scholars make. Moreover, although our study focuses narrowly on gender, past studies of faculty attrition [103, 148, 75] lead us to expect that a disproportionate share of these lost careers would be women of color.

The two findings above—that all-cause attrition served to increase women's representation in 74% of fields, while gendered attrition led to the net loss of an estimated 1378 women from the academy—seem at first glance to be at odds. However, the former reflects primarily the turnover and retirement of a previous generation of faculty [130], while the latter stems from observations that in most areas of the academy, women are at higher risk of leaving their faculty jobs than men of the same career age, i.e., gendered attrition [146]. Reconciling and quantifying these separate effects through models, parameterized by gender and career-age stratified empirical data, is therefore a key contribution of this work.

We find large and statistically significant effect sizes for gendered attrition at high levels of aggregation in our data, (e.g., all of academia, among all STEM or all non-STEM fields, and within domains), yet we often find smaller effects that are not statistically significant in constituent individual fields (Fig. 3.2; see also Ref. [146]). This pattern implies that there must be gendered effects within at least some of the constituent fields which may statistically indistinguishable from noise due to smaller population sizes and greater relative fluctuations in hiring. This perspective may therefore explain why field-specific studies of gendered faculty attrition sometimes reach conflicting conclusions [141, 40, 38, 149, 146], and suggests that future studies should seek larger samples sizes whenever possible.

While this study cannot identify specific causal mechanisms that drive gendered attrition, the literature points to a number of possibilities, including disparities in the levels of support and value attributed to women and the scholarly work that women produce [150, 58], sexual harassment [21], workplace culture [146], work-life balance [151, 152], and the unequal impacts of parenthood [26, 25]. Even in fields where we found no significant evidence that women and men leave academia at different rates, the reasons women and men leave may nevertheless be strongly gendered. For instance, past work has shown that men are more likely to leave faculty jobs due to attractive alternate opportunities ("pulls"), while women are more likely to leave due to negative workplace culture or work-life balance factors ("pushes") [146].

The relative importance of hiring vs attrition is also borne out in our projection scenarios. Indeed, we find that eliminating the gendered attrition gap, in isolation, would not substantially increase representation of women faculty in academia. Rather, progress toward gender parity depends far more heavily on increasing women's representation among new faculty hires, with the greatest change occurring if hiring is close to gender parity (Fig. 3.3).

A limitation to this study is that it only considers tenured and tenure-track faculty at PhD granting institutions in the U.S. Non-tenure track faculty, including instructors, adjuncts, and research faculty, are an increasing portion of the professoriate [114, 113], and they may experience different trends in hiring and attrition, as may faculty outside the U.S. and faculty at institutions that do not grant PhDs. The results presented in this work will only generalize to these other populations to the extent that they share similar attrition rates, similar hiring rates, and similar current demographics. Notably, understanding how faculty hiring, faculty attrition, and faculty promotion [153] are shaping the gender composition of these populations is an important direction for future work.

Another important limitation of this work is that we focused our analysis at the level of entire fields and academic domains. However, hiring and attrition may play different roles in specific departments or in specific types of departments. These differences may be elucidated by future analyses of mid-career moves, in which faculty change institutions but stay in the same fields.

Faculty who belong to multiple marginalized groups (e.g. women of color) are particularly underrepresented and face unique challenges in academia [?]. The majority of women faculty in the U.S. are white [154], meaning the patterns in women's hiring and retention observed in this study are predominantly driven by this demographic. While attrition may not be the primary challenge for women's representation overall, it could still be a significant barrier for women of color and those from less privileged socioeconomic backgrounds. Additional data are needed to study trends in faculty hiring and faculty attrition across racial and socioeconomic groups.

More broadly, while this study has focused on a quantitative view of men's and women's relative representations, we note that equal representation is not equivalent to equal or fair treatment [126, 127]. Pursuing more diverse faculty hiring without also mitigating the causes that sustain existing inequities can act like a kind of "bait and switch," where new faculty are hired into environments that do not support their success, a dynamic that is believed to contribute to higher turnover rates for women faculty [104].

Our study's detailed and cross-disciplinary view of hiring and attrition, and their relative impacts on faculty gender diversity, highlights the importance of sustained and multifaceted efforts to increase diversity in academia. Achieving these goals will require a deeper understanding of factors that shape the demographic landscape of academia.

Methods

3.0.4 Data cleaning and preparation

Our analysis is based on a comprehensive dataset of U.S. faculty at PhD granting institutions, obtained through a data use agreement with the Academic Analytics Research Center (AARC). This dataset includes the employment records for all tenured and tenure-track faculty at all 392 U.S. doctoral-granting universities from 2011 to 2020, along with the year of each professor's terminal degree. To ensure the consistency and robustness of our measurements, we cleaned and preprocessed this dataset according to the following steps. For additional technical details relating to these steps, see Ref. [130]. For a manual audit to assess potential attrition errors in this dataset, see Ref. [146].

We first de-duplicated departments. This involved combining records due to: (i) variations in department names (e.g., "Computer Science Department" vs. "Department of Computer Science") and (ii) departmental renaming events (e.g., "USC School of Engineering" vs. "USC Viterbi School of Engineering").

Next, we annotated departments according to a two-level department taxonomy with lower level fields and higher level domains assignments to departments. While most departments were assigned a single annotation for each level of the taxonomy, some interdisciplinary departments received multiple annotations. This deliberate choice can reflect how a "Department of Physics and Astronomy" is relevant to both "Physics" within the "Natural Sciences" domain and "Astronomy" within the same domain. Therefore, we included all applicable annotations for such departments to capture their full scope, but note that domain-level analyses included such departments only once.

Then, we addressed certain interdisciplinary fields which could conceptually reside in multiple domains, e.g., Computer Engineering (potentially belonging to domains of either Mathematics & Computing or Engineering), and Educational Psychology (potentially belonging to domains of either Education or Social Sciences). To address this ambiguity, we employed a heuristic approach. Fields were assigned to the domain containing the largest proportion of faculty members whose doctoral universities housed a department within that domain. Thus, we grouped fields based on the domain where their faculty were most likely to have been trained. Supplementary Data Table A.2 contains a complete list of fields and domains.

In rare instances, faculty members temporarily disappeared from the dataset before reappearing in their original departments. We treated these as likely data errors and imputed the missing employment records. Missing records were filled in if the faculty member's department had data for the missing years. Employment records were not imputed if they were associated with a department that did not have any employment records in the given year. Imputations affected 1.3% of employment records and 4.7% of faculty.

Next, we limited our analyses to departments consistently represented in the AARC data across the study period (2011-2020). This exclusion was necessary because not all departments were consistently recorded by AARC. Departments appearing in the majority of years within the study period were retained, resulting in the removal of 1.8% of employment records, 3.4% of faculty, and 9.1% of departments. This exclusion also resulted in the removal of 24 institutions (6.1%), primarily seminaries.

We next filtered the data to include only tenure-track faculty. This involved removing temporary faculty, including non-tenure-track instructors holding titles such as “lecturer,” “instructor,” or “teaching professor” at any rank, individuals with missing rank information, and faculty classified as “research,” “clinical,” or “visiting.” This filtering process resulted in a dataset solely comprised of tenured and tenure-track faculty holding the titles of “assistant professor,” “associate professor,” and “full professor.”

Finally, we defined career age for each person-year record in our dataset as the difference between the given year for the record and the year that the faculty member received their doctoral degree. However, doctoral year was missing for 29,872 faculty members (9.8% of faculty), necessitating their exclusion from the counterfactual analysis (results Sec. 3.0.2) and the forecasting analysis (results Sec. 3.0.3).

3.0.5 Decomposition of changes in representation into hiring and attrition

We decomposed the annual changes in women’s representation (in units of proportion per year) within each field into hiring (δ_{hiring}) and attrition ($\delta_{attrition}$) components as follows. First, we define n_w and n_m as the counts of women and men faculty in the field in a given index year. Next, we let h_w and h_m be the counts of women and men that were hired between the index year and the following year, and similarly let x_w and x_m be the counts of women and men among faculty “all-cause” attritions between the index year and the following year. We then approximate δ_{hiring} and $\delta_{attrition}$ as

$$\delta_{hiring} = \frac{n_w + h_w}{n_w + h_w + n_m + h_m} - \frac{n_w}{n_w + n_m}$$
$$\delta_{attrition} = \frac{n_w - x_w}{n_w - x_w + n_m - x_m} - \frac{n_w}{n_w + n_m} \ .$$

These equations are developed in Supplementary Text A.1. The annual changes δ_{hiring} and $\delta_{attrition}$ were summed, respectively, over 2011-2020 for each field, to construct Fig. 3.1.

3.0.6 Model of faculty hiring and attrition

We developed a model of annual faculty hiring and attrition structured by faculty career age (years since PhD) and gender, allowing model parameters to vary as a function of both. This model was used for both the counterfactual historical analysis, which investigated gendered attrition, and the projection of future scenarios, which investigated five stylized futures. Details of the particular parameterizations for each investigation follow this structural model description.

In this model, we track the number of people with a given career age a and gender g, with annual updates given by

$$n(a, g) \leftarrow n(a-1, g) + h(a, g) - x(a-1, g) \ , \tag{3.1}$$

where h and x are hires and attritions, respectively. In each stochastic model realization, both h and x are drawn according to distributions that allow control over the extent to which hiring and attrition are (or are not) gendered processes.

We stochastically draw $h(a, g)$ and $h(a, \tilde{g})$ by first specifying the total number of hires that year $H(a)$. For each of the $H(a)$ hires, we draw gender annotations from independent Bernoulli trials with parameter $\psi(a, g)$. In this way, the ψ parameters control the extent to which hiring is gendered across career ages.

To calculate attritions $x(a-1, g)$, we subject each of the $n(a-1, g)$ sitting faculty to a career age and gender-stratified annual attrition risk $\phi(a-1, g)$, realizing the actual number of attritions from a binomial draw with n trials and parameter ϕ. The relative values of $\phi(a, g)$ and $\phi(a, \tilde{g})$ therefore control the extent to which attrition is gendered for a particular career age.

Thus, this model is stochastic, and after specifying initial conditions for the values of $\mathbf{n}$, one needs only values for the total hires $\mathbf{H}$, gendered hiring parameters ψ, and gendered attrition risk parameters ϕ to simulate stochastically, by iterating Eq. (3.1).

3.0.7 Parameters for counterfactual model of 2011–2020

The goal of this model was to quantify the impact of gendered attrition over the period 2011–2020. As such, this model manipulated the parameters ϕ which control the extent to which attrition is gendered. To model a counterfactual scenario in which attrition was not gendered, we set women's and men's attrition risks to be identical $\phi(a, w) = \phi(a, m)$, taking on values estimated from empirical data in which gender was ignored, for each field (see below).

For each field, the model's faculty counts $\mathbf{n}$ were initialized using our 2011 faculty roster data, thereby matching all empirical 2011 values for both gender and career age. To ensure that a field's total faculty size grew or shrunk each year in a manner that exactly matched empirical changes, we first drew all attrition values x and then set the total number of new hires accordingly. Those new hires were assigned initial career ages drawn from the empirical age distribution of new hires, averaged over 2012-2020. Gendered hiring parameters ψ were set to values estimated from empirical data for each field (see below).

To parameterize $\psi(a, w, t)$, the probability that a new hire of career age a in year t is a woman, we used a logistic regression model fit to empirical hiring data for each field. Because the probability that a new hire is a woman varies non-linearly with career age (see Supplementary Fig. A.4), the dependent variables in this model include new hires' career ages up to their fifth exponents and a linear term for the calendar year, which ranges from 2012 to 2020.

To parameterize $\phi(a, g, t)$, the probability that faculty of career age a and gender g experience attrition in year t, we used a logistic regression model trained on empirical attrition and retention data for each field. Because faculty attrition rates vary non-linearly with career age (see Supplementary Fig. A.3), we include new hires' career ages up to their fifth exponents as dependent variables in the regression model, in addition to a linear term for the year, which ranges from 2012 to 2020. This regression model is fit to all observed cases of attrition and retention for both men and women together, for each field. Accordingly, men and women in our gender-neutral counterfactual model were subjected to the same age-varying attrition risks, eliminating the gendered aspect of these patterns, while preserving the rises and declines in attrition rates across faculty career ages.

To capture the distribution of counterfactual historical outcomes under gender neutral attrition, we drew 500 simulations of the years 2012–2020 for each field, recording women's representation at the end of each. See Supplementary Sec. A.2 for details of model validation. This model was run for each field, independently. It was also run again for all STEM fields, and all non-STEM fields, with respective parameters estimated at those respective higher levels of data aggregation (Fig. 3.2).

3.0.8 Parameters for projection model of 2020–2060

The goal of this model was to quantify the effects of a set of five highly stylized scenarios for how hiring and attrition might evolve over the 40 years spanning 2020-2060. For scenarios with gender-neutral attrition (GNA), we set women's and men's attrition risks to be identical $\phi(a, w) = \phi(a, m)$, taking on values estimated from empirical data in which gender was ignored, for each academic domain. For scenarios with observed attrition (OA), we set women's and men's attrition risks to values estimated from empirical data in which gender was included, for each domain. For scenarios with equal representation of women and men among hires (ER), we set all ψ parameters to 0.5. And, for scenarios with increasing representation of women among hires (IR), we let the hiring parameters vary over time, such that the expected proportion of women among new hires increases by $0.5pp$ per year for each career age starting in 2020, i.e.,

$$\psi(a, w, t) = \psi(a, w, 2020) + 0.005(t - 2020) \ .$$

In the absence of one of the above manipulations, ϕ and ψ parameters were set to their empirical values, estimated from aggregated 2011-2020 data for each domain.

For each academic domain, the model's faculty counts $\mathbf{n}$ were initialized using our 2020 faculty roster data, thereby matching all empirical 2020 values for both gender and career age. Each domain's total faculty size was held fixed at 2020 values by setting the total number of new hires to be equal to the number of stochastically drawn attritions. This model was run for each domain, independently, using parameters estimated at the domain level accordingly.

To capture the distribution of outcomes in each projection scenario, we drew 500 simulations of the years 2020–2060 for each academic domain, recording women's representation at the end of each (Fig. 3.3). See Supplementary Sec. A.2 for details of model validation.

The work presented in the chapter was developed in collaboration with my advisors Aaron Clauset and Daniel Larremore. The dataset used was cleaned by Hunter Wapman. I additionally thank Katie Spoon, Ian Van Buskirk, and Bailey Fosdick for helpful comments.

Chapter 4

Quantifying disparities in peer review outcomes at elite journals

Introduction

Peer review is the process by which scientific experts critically evaluate a completed study to assess whether it surpasses certain standards for becoming part of the scientific record. As a social process, peer review is susceptible to non-meritocratic idiosyncrasies and biases in these decisions, e.g., based on author attributes like gender or nationality, which can distort the scientific record and limit scientific progress. Scientists are invested in the legitimacy of peer review, which depends on reliability [155, 156], correctness [157, 158], proper allocation of attention [159], and the minimization of bias [160]. Despite its importance, we lack an integrated understanding of the influence of these non-meritocratic factors across fields and elite venues, largely due to an unavailability of peer review data, especially at elite venues. An integrated understanding of biases in elite peer review would inform efforts to improve the reliability, accuracy, and utility of the scientific record.

Elite journals disproportionately influence the scientific record [157]. These journals have broad influence, authoritative reputations, wide multidisciplinary readerships, and a highly selective peer review process (for example, *Science* accepts only 6.1% of submissions). Publications at these journals often garner the most citations, count the most toward tenure and promotion, grant funding decisions, and the allocation of prestigious awards [161, 162]. As a result, even small biases in the decisions made at elite journals can shape individual scientific careers, the composition of the scientific community [163, 164, 162], and the direction of scientific discovery [161, 165].

We develop and investigate a comprehensive dataset on peer review outcomes at two elite multidisciplinary journals, *Science* and *Science Advances*, for which we have permission from the American Association for the Advancement of Science (AAAS) to anonymize and publicly release. This dataset includes all manuscripts submitted to either journal in the period 2015–2020. This data enables researchers to study elite peer review while protecting reviewer and author confidentiality. We analyze this anonymized data and determine the extent to which team size, author prestige, author region, and author gender predict success at every step of the peer review process, and the extent to which reviewer gender interacts with author gender during peer review.

Data

The anonymized data (Table 4.1) contains manuscript metadata; editorial decisions; demographics for editors, authors, and reviewers; where possible, associated institutional prestige and nationality information; and statistical summaries of the peer reviews for 110,303 manuscripts submitted to *Science* and *Science Advances* between January 1st 2015 and January 1st 2021. The only measure of the multiple dimensions of manuscript quality is the peer reviews, which we summarize using statistical measures of their sentiment. We additionally use modern natural language processing techniques to add high-level topic variables to articles based on title and abstract text, and to measure the narrative structures of review text.

To make the data anonymous, we drop all identifying and confidential information, including article titles and abstracts, the names and other identifiers of all authors, reviewers, advisors, and editors, and the peer review text itself. Records with multiple rounds of peer review or with many reviewers stand out in the data as particularly re-identifiable. To mitigate the associated risk of re-identification, we preserved only the first round of reviews, and randomly selected two first-round reviewers. The anonymized data does not link individuals across submissions, further minimizing potential harm from re-identification. The most sensitive data that could be revealed for a given manuscript is a binned prestige score and gender for the two randomly selected first-round reviewers, the gender and numerical rating of the Board of Reviewing Editors (BoRE) member who evaluated the submission, and the gender of up to two editors.

Results

On this anonymized data, we analyze the effects of team size, topic choice, the corresponding author's nationality, gender, and prestige at three stages of the editorial process: (1) whether a submission is desk rejected or sent to review, (2) the sentiment of the external reviews, and (3) the final editorial decision to accept or reject the manuscript (Fig. 4.1). Team size is statis-

Variable	What it applies to	Possible values
Publication	All submissions	*Science* (61.7%) / *Science Advances* (38.3%)
Is accepted	All submissions	TRUE (7.8%) / FALSE (92.2%)
Number of authors	All submissions	[1–5] (39.2%), [6–9] (29.6%), [10+] (31.2%)
Topic	All submissions	0–N for $N = 10, 30, 100$
Gender	BoRE advisors	Man (76.6%) / Woman (23.4%)
	Editors	Man (65.8%) / Woman (34.2%)
	First authors	Man (66.6%) / Woman (33.4%)
	Corresponding authors	Man (78.5%) / Woman (21.5%)
	Reviewers	Man (77.2%) / Woman (22.8%)
Prestige	Corresponding authors	[1–2] (17.0%); [3–4] (17.5%); [5–6] (24.2%); [7–10] (28.2%); Null (13.0%)
	Reviewers	[1–2] (28.0%); [3–4] (23.1%); [5–6] (21.5%); [7–10] (18.7%); Null (8.7%)
Region	Corresponding authors	U.S. + Canada (31.3%), China (19.7%), Europe (24.5%), Other (17.2%), Null (7.4%)
BoRE rating	Most *Science* submissions	(lowest) 1–10 (highest)
BoRE confidence	Most *Science* submissions	(lowest) 1–5 (highest)
Review sentiment	Reviewed manuscripts	Z-scores $[-3.5, 3.5+]$, coarsened and capped
Review length	Reviewed manuscripts	Z-scores $[-3.5, 3.5+]$, coarsened and capped
Review trajectory	Reviewed manuscripts	DF, DFU, DU, DUD, DUDF, F, FD, U, UD, UDF, UF, Null

Table 4.1: **Overview of data.** Rows show variables included in the anonymized dataset. BoRE is the Board of Reviewing Editors. Topics are created through topic models of 10, 30, or 100 topics. Review lengths are the z-score of the numbers of words. Reviewer sentiment trajectory values are concatenations of (D)own, (F)lat, and (U)p (see SI).

tically significant at all stages at both *Science* and *Science Advances*, with teams of six or more authors consistently achieving higher success rates and sentiment scores than teams of fewer than six authors (Fig. 4.1A). Moreover, the topic of a submission correlates with outcomes, with editors rejecting manuscripts of different topics at statistically significantly different rates (e.g., Topic 0 vs Topic 1, see SI Tables B.1 to B.4), illustrating the distinct mixture of editorial preferences across the two journals. Corresponding authors with Chinese or other non-European affiliations experience less success at every stage of the editorial process compared to corresponding authors with a North American affiliation (Fig. 4.1B). Thus manuscript-level covariates and author nationality correlate strongly with submission success at every stage of the editorial process, even after controlling for individual demographics, such as gender.

A central theme in the literature on peer review is the question of outcome disparity based on author identity, with gender being the most commonly studied of these variables [166, 167]. At both journals, manuscripts with women corresponding authors are less likely to be sent to review compared with men, and receive lower sentiment reviews on average (Fig. 4.1C). However, at the last stage of review, when reviews have been received, gender does not exhibit a consistent effect on the final editorial decision. At *Science*, the gender of the first author does not have a consistent effect in any stage of the review process. Using our model to remove gender disparity from desk rejection, review sentiment, and acceptance stages of the peer review process (see Materials and Methods Sec. 4.0.2), we estimate that there would have been 92 more papers over the 5 year period published in *Science* by women corresponding authors (Table 4.5). This change would represent a 12% increase in publications with a woman corresponding author for *Science*, or roughly one more paper by a woman corresponding author every two and half weeks.

Several distinct mechanisms may lead to a disadvantage for women corresponding authors during the peer-review process: (1) evaluators may demonstrate homophily, favoring authors of their own gender, e.g., men reviewers favor men authors and women reviewers favor women authors, (2) men evaluators may favor men authors while women evaluators remain unbiased with respect to gender, (3) women evaluators may favor men authors while men evaluators remain unbiased, or (4)

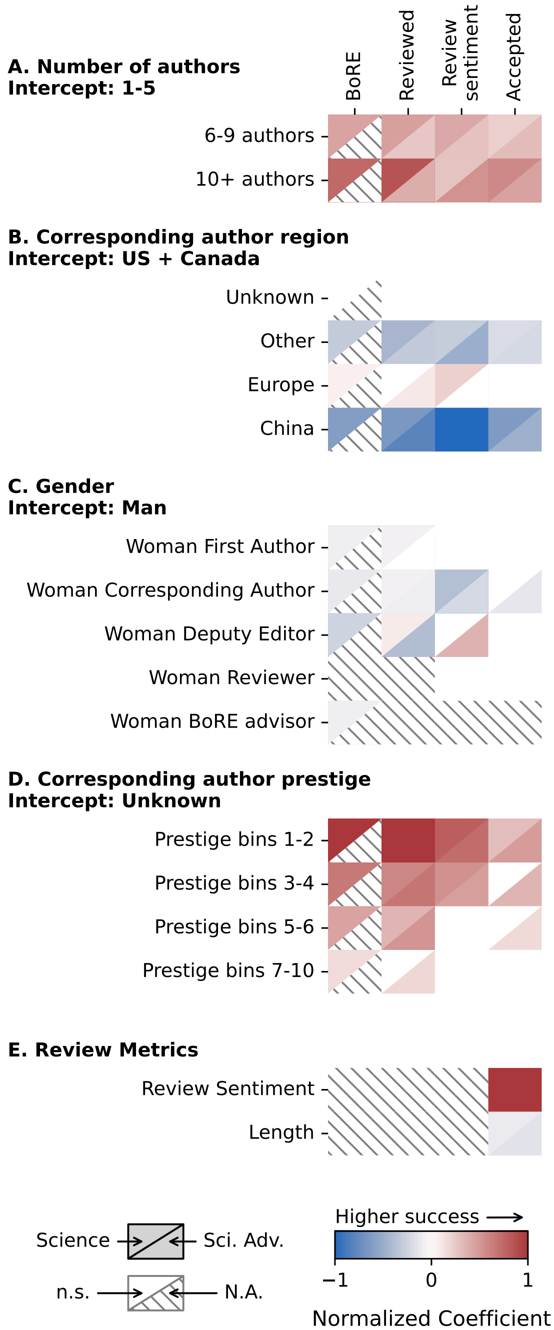

Figure 4.1: Overview of the effects of (A) number of authors, (B) region, (C) gender, (D) prestige, (E) review sentiment, and review length on each applicable stage of peer review for *Science* (upper triangles) and *Science Advances* (lower triangles). The colors display linear regression coefficients for the Board of Reviewing Editors (*BoRE*) and *review sentiment* stages, and logistic regression coefficients for the *reviewed* and *accepted* stages (controlling for topic; see SI Tables B.1 to B.4 for full regression tables). Positive coefficients (red hues) are associated with evaluations and success rates that are higher than the corresponding intercept (indicated in bold text), while negative coefficients (blue hues) are associated with lower evaluations and success rates. A dashed cell indicates an excluded coefficient-stage pair, and non-significant effects were colored white. Coefficients are normalized by dividing by the absolute value of the highest magnitude coefficient within each publication and stage.

Comparisons		Theorized outcomes				*Science*			*Science Advances*		
Pair 1	Pair 2	M1	M2	M3	M4	Obs.	Coef.	p	Obs.	Coef.	p
MM	MW	>	n.s.	n.s.	n.s.	n.s.	−0.012	0.483	n.s.	−0.006	0.774
MM	WM	>	>	n.s.	>	>	0.075	0.000	n.s.	0.048	0.026
MM	WW	n.s.	n.s.	>	>	>	0.120	0.002	n.s.	0.062	0.107
MW	WM	n.s	>	n.s.	>	>	0.099	0.000	n.s.	0.045	0.076
MW	WW	<	n.s.	>	>	>	0.092	0.007	n.s.	0.064	0.086
WM	WW	<	<	>	n.s.	n.s.	−0.053	0.106	n.s.	0.000	0.995

Table 4.2: **Testing drivers of women corresponding author disadvantage.** Each mechanism for women's disadvantage in review sentiment predicts a set of specific outcomes for each comparison between corresponding author gender and reviewer gender pairings. (M1) represents gender homophily between reviewers and corresponding authors. In (M2), men reviewers favor men authors while women evaluators remain unbiased with respect to gender, while in (M3), women evaluators favor men authors while men evaluators remain unbiased. (M4) predicts that both women and men reviewers preference teams with men authors. In this table, the first letter for a pair represents the corresponding author gender, and the second letter represents the reviewer gender (M=Man, W=Woman). Pairwise comparisons for *Science* based on propensity score weighting analyses align with outcomes predicted by M4, and comparisons for *Science Advances* were all insignificant, based on a significance threshold of 0.008, by the Bonferroni correction for multiple hypothesis testing within each journal (see Materials and Methods Sec. 4.0.1). Theorized outcomes that align with observed outcomes in *Science* are highlighted.

evaluators may tend to preference teams with men authors regardless of evaluator gender [167, 166, 168]. By analyzing pairwise differences in sentiment scores across combinations of reviewer and author genders, we test these theories individually (Table 4.2). At *Science*, reviewers preference teams with men corresponding authors regardless of reviewer gender (Mechanism 4), but results are less conclusive for *Science Advances* (Table 4.2, Materials and Methods Sec. 4.0.1).

In past work, prestige has been shown to have a large effect on peer review, whereby authors from elite institutions are more likely to have their submissions accepted [169]. Indeed, a consistent and large effect for corresponding author prestige appears at both journals and at each stage of the review process. The effect of prestige is stronger at the initial editor's desk than during or after peer review (Fig. 4.1D). For instance, at *Science*, only corresponding authors from the highest prestige universities are statistically significantly associated with higher acceptance rates at the final acceptance stage after controlling for the sentiment of reviews. The fact that prestige has a stronger effect at the initial editor's desk than the post-review decision is consistent with the idea that editors rely on prestige signals to guide their investment of expert time, and that they rely less on those signals once the expert opinion is in hand. Indeed we find that expert opinion, as measured by the sentiment of the review text, is one of the most predictive signals of whether a reviewed manuscript is accepted (Fig. 4.1E).

Discussion

Elite journals shape scientific discourse and make scientific careers, yet the dearth peer review data has made it difficult for the broader scientific public to answer important questions such as whether and where disparities appear during their editorial processes. At both *Science* and *Science Advances*, the following characteristics of manuscripts correlate most strongly with editorial outcomes: prestige, nationality, team size, topic, and corresponding author gender (Fig. 4.1). These correlations typically compound across the entire editorial process, but there are exceptions. At *Science*, the corresponding author's gender is significantly correlated with the outcomes of each stage of peer review, except for the stage of the editor's final decision. By charting exactly where each disparity occurs, we open up the questions of how much each disparity is the result of effective peer review versus undesirable biases, and how to intervene to address biases where they do occur.

Our publicly released data will allow the scientific community to perform novel analyses of the elite peer review process, creating a foundation of knowledge that can guide future experiments and journal policies. To protect author and reviewer confidentiality, we dropped data that could have been useful for certain analyses. For example, the exclusion of submission and publication dates eliminates the possibility of performing temporal analyses such as event studies or differences-in-differences. Nevertheless, the current data can shed light on many additional topics, such as evaluation idiosyncrasies and disagreements among the BoRE and reviewers, peer-review differences across fields and subfields, intersectional effects across available author covariates (e.g., region, gender, and prestige), and the processes through which editors assign reviewers to manuscripts.

The current observational data is not fully sufficient to isolate the causal mechanisms driving disparities in peer review. For example, at *Science*, teams with men corresponding authors receive more positive reviews than teams with woman corresponding authors, regardless of the gender of the reviewer, but our data cannot untangle whether this is the result of a boost for men corresponding authors or a penalty for women corresponding authors (Table 4.2). Moreover, the degree to which corresponding author gender is a causal factor in peer review outcomes remains inconclusive due to possible confounding between gender and seniority: our data does not measure corresponding author seniority, and the career age of researchers correlates with gender due to historic exclusion of women from faculty positions [170]. Another set of confounding variables relate to the suitability of each submission for publication, such as quality and relevance to a broad audience. When certain fields value publication in elite interdisciplinary journals more than others, and nations or institutions implement financial reward programs for publishing in elite journals, scientists subject to those incentives experience a higher expected reward for publication, driving them to submit works of lower publishability and tolerate higher risks of rejection [171, 172].

Three actions from elite journals could advance the community's understanding of elite peer review. First, AAAS has set an example that other publishers and journals could follow, of collaborating with academic partners to anonymize and publicly release data for their flagship journals. Most effects were consistent across *Science* and *Science Advances*, but there were some differences, and extending analyses to a wider set of elite journals would allow the scientific community to adjudicate which finding is more generalizable, while identifying biases that may be idiosyncratic to specific journals. Second, journals can gather, anonymize, and release more data related to authors so that future analyses can control for a larger set of potential confounds. For example, more complete seniority information from corresponding authors would have allowed stronger causal identification of the gendered effects in peer review. Lastly, journals can run experiments. A randomized controlled experiment assigning papers to single blind vs. double blind review could determine the extent to which differences in peer review outcomes are the result of preferences for (or bias against) authors with particular sets of characteristics vs. preferences for (or bias against) the research submitted by these authors.

Preferences for certain topics and methods are more defensible than biases against authors with specific characteristics, but the complete elimination of such biases still would not imply an effective or equitable peer review process. Important contributions may face desk rejections, while less impactful publications are accepted. Moreover, author identity can correlate with research topics, and topics that are likely to be studied by groups of scientists that have been historically excluded from science often receive fewer citations, regardless of the authors' identity [63, 66]. The determination of which research is considered significant, relevant, and novel enough to be published in elite scientific journals is a high stakes social process carried out by editors and reviewers that deserves continuous attention. Transparency at elite journals is necessary for the scientific community at large to participate in the allocation of collective scientific attention, and the release of open data at elite journals is an important first step toward that transparency.

Materials and Methods

We augment the raw author and reviewer data with institutional prestige labels based on self-reported affiliation or affiliation derived from an email addresses (Sec. 4.0.3) and with name-based gender annotations when self-reported gender is missing [144]. The methodology we use assigns only binary (woman/man) labels to authors and reviewers, even as we recognize that gender is nonbinary. This approach is a compromise due to the technical limitations of name-based gender methodologies and is not intended to reinforce a gender binary. We additionally use modern natural language processing techniques to add topic variables to articles based on title and abstract text (Sec. 4.0.4) and to measure the sentiment and narrative structures of review text (Sec 4.0.5).

4.0.1 Testing drivers of women author disadvantage

To measure potential interactions between reviewer gender and author gender at *Science* and *Science Advances*, we estimate differences in review sentiment assigned to articles, grouping by corresponding author gender and reviewer gender pairings (Table 4.2). For example, we compare the sentiment scores for articles authored by a man and reviewed by a man ("MM") to articles authored by a man but reviewed by a women ("MW").

We use propensity score weighting to control for covariate differences between comparison groups, and we check the robustness of our results by separately running the analysis using both regular logistic regression to estimate propensity scores [173] (as presented in Table 4.2 and Table 4.3) and using gradient boosted logistic regression (see Table 4.4) [174]. To estimate propensity scores, we consider first author gender, deputy editor gender, corresponding author prestige, corresponding author region, number of authors, review length, topic (one-hot encoded into 10 categorical topic variables), reviewer gender, and reviewer prestige.

We then use average treatment effect (ATE) weighting in a weighted linear regression, with review sentiment scores as the outcome variable. For this regression, we once again control for the same set of covaraites that were used to estimate propensity scores, to be doubly-robust [175]. This regression estimates the ATE, which captures the direction and magnitude of the average difference in sentiment score between groups (net of the controls).

We use this approach to estimate the average difference in sentiment score and the corresponding statistical significance for each unique author-reviewer gender pairing, MM vs. MW, MM vs. WM, MM vs. WW, MW vs. WM, MW vs. WW, and WM vs. WW. This results in six estimates for each journal, as presented in Table 4.2. We present results that correct for multiple hypothesis testing using the Bonferroni correction by dividing the significance threshold (0.05) by the number of statistical hypotheses conducted for each publication–6 in this case–yielding a corrected threshold of 0.0083. Under this correction, the results for *Science* were unchanged, and the uncorrected results for *Science Advances* changed to become entirely non-significant. In tables 4.3 and 4.4, we additionally present results under the the Benjamini-Hochberg correction [176], which is more permissive of significant results.

While the uncorrected and Benjamini-Hochberg corrected results based on propensity scores estimated using gradient boosted logistic regression (Table 4.4) match the results based on regular logistic regression (Table 4.3, the two of the estimated pair-wise differences based on gradient boosted scores for *Science* become insignificant under the Bonferroni correction: MM vs. WW, and MW vs. WW (Table 4.4).

Pair 1	Pair 2	Pub.	Not Corrected	Bonferroni	Benjamini-Hochberg	Coef.	p	Pub.	Not Corrected	Bonferroni	Benjamini-Hochberg	Coef.	p
MM	MW	Science	n.s.	n.s.	n.s.	-0.012	0.483	Science Advances	n.s.	n.s.	n.s.	-0.006	0.774
MM	WM	Science	>	>	>	0.075	0.000	Science Advances	>	n.s.	>	0.048	0.026
MM	WW	Science	>	>	>	0.120	0.002	Science Advances	n.s.	n.s.	n.s.	0.062	0.107
MW	WM	Science	>	>	>	0.099	0.000	Science Advances	n.s.	n.s.	n.s.	0.045	0.076
MW	WW	Science	>	>	>	0.092	0.007	Science Advances	n.s.	n.s.	n.s.	0.064	0.086
WM	WW	Science	n.s.	n.s.	n.s.	-0.053	0.106	Science Advances	n.s.	n.s.	n.s.	0.000	0.995

Table 4.3: **Testing differences in sentiment using regular logistic regression to estimate propensity scores.** We use propensity score weighting to control for covariate differences between comparison groups [173]. The first letter for a pair represents the corresponding author gender, and the second letter represents the reviewer gender (M=Man, W=Woman). Coefficients represent differences in sentiment score across between groups, while the significance and direction for each pairwise comparison is reported under three conditions: Not corrected (significance threshold = 0.05), the Bonferroni correction (significance threshold = 0.0083), and the Benjamini-Hochberg correction.

4.0.2 Counterfactually removing gender disparity from peer review

To estimate the effect sizes of disparities by corresponding author gender at *Science* and *Science Advances*, we conducted several analyses that modeled counterfactually gender neutral outcomes at select stages of peer review (Table 4.5).

Pair 1	Pair 2	Pub.	Not Corrected	Bonferroni	Benjamini-Hochberg	Coef.	p	Pub.	Not Corrected	Bonferroni	Benjamini-Hochberg	Coef.	p
MM	MW	Science	n.s.	n.s.	n.s.	-0.011	0.539	Science Advances	n.s.	n.s.	n.s.	-0.007	0.715
MM	WM	Science	>	>	>	0.100	0.000	Science Advances	>	n.s.	>	0.056	0.025
MM	WW	Science	>	n.s.	>	0.104	0.021	Science Advances	n.s.	n.s.	n.s.	0.050	0.240
MW	WM	Science	>	>	>	0.116	0.000	Science Advances	n.s.	n.s.	n.s.	0.055	0.051
MW	WW	Science	>	n.s.	>	0.097	0.031	Science Advances	n.s.	n.s.	n.s.	0.071	0.104
WM	WW	Science	n.s.	n.s.	n.s.	-0.050	0.143	Science Advances	n.s.	n.s.	n.s.	-0.001	0.988

Table 4.4: **Testing differences in sentiment using gradient boosted logistic regression to estimate propensity scores.** We use propensity score weighting to control for covariate differences between comparison groups [174]. The first letter for a pair represents the corresponding author gender, and the second letter represents the reviewer gender (M=Man, W=Woman). Coefficients represent differences in sentiment score across between groups, while the significance and direction for each pairwise comparison is reported under three conditions: Not corrected (significance threshold = 0.05), the Bonferroni correction (significance threshold = 0.0083), and the Benjamini-Hochberg correction.

First, we ran regression models at the desk rejection, review sentiment, and manuscript acceptance stages of the peer review process separately for *Science* and *Science Advances* to estimate how article attributes correlate with peer review outcomes. The decision to send an article to review and the decision to ultimately accept an article each have binary outcomes, so we ran logistic regression at these stages. Peer review sentiment scores are continuous, so we ran linear regression at the peer review stage. For each regression, we included first author gender, corresponding author gender, deputy editor gender, corresponding author prestige, corresponding author region, number of authors, and topic as categorical predictors. For the regression analyses that correspond with the editor's final decision to accept or reject each paper, we additionally included the review sentiment from one of the reviews as a continuous predictor.

Then, we used the parameters inferred by these models to compute the estimated outcomes for each submitted article at each stage of the peer-review process, regardless of whether the articles actually made it beyond the editor's desk. To estimate the probability of desk rejection, we applied the corresponding regression equation, which was empirically fit based on observed desk rejection outcomes, to all submissions. To estimate review sentiment scores for each article, we probabilistically draw from the prediction interval for each record based on the OLS regression fit to peer review sentiment outcomes. Here, we discard the actual peer review sentiment scores, for the cases in which a given article actually did go to peer-review. Finally, we estimate the probability of acceptance (conditioned on being reviewed), based on observed covairates and the model-based review sentiment scores drawn in the prior step.

To estimate gender-neutral outcomes at a given stage of peer review, we modify records with women corresponding authors to counterfactually have man corresponding authors. This counterfactual switch has the effect of increasing the estimated probability of success (or increasing review sentiment scores) for women authored publications because the coefficient associated with women corresponding authors was negative across stages for both *Science* and *Science Advances*.

We use this process of inferring success probabilities and of drawing review sentiment scores to run 1000 Monte Carlo simulations for each scenario at each journal, as outlined in Table 4.5. We first compute the estimated probabilities that each article goes to review, considering whether women corresponding author's records should be counterfactually altered at this stage, based on the given scenario. Then, to prevent the journals from counterfactually publishing more papers, we scale the estimated probabilities, such that the sum of probabilities is equal to the observed number of reviewed articles. Then, desk rejections are simulated based on Bernoulli distributed weighted coin-flips. Review sentiments are then drawn for articles that were not desk rejected, which are used as one of the covariates for the model fitted to final acceptance decisions. Before being used to weight another set of coin-flips to determine simulated acceptances, the final acceptance decision probabilities are scaled to sum to the total number of observed acceptances. Scaling the probabilities causes the simulated number of acceptances to match the observed number of acceptances in expectation, across stochastic trials.

At the end of each trial, we record the simulated number of articles with woman corresponding authors, which we use to calculate the mean counterfactual number of articles with women corresponding authors, along with confidence intervals spanning the middle 95% of Monte Carlo outcomes (Table 4.5).

4.0.3 Adding prestige variables to authors

Our approach to add prestige variables to authors is motivated by the following two requirements. First, we require an international definition of prestige, since only 31.3% of authors are from the US and Canada Table 4.1. This prevents us from using rankings derived from the US faculty hiring network, since it excludes non-US faculty and we do not have the international faculty hiring data necessary to construct the full global network [177]. Second, we seek to extend the definition of prestige beyond academic universities to encompass other institutions that frequently submit papers to *Science* and *Science Advances*, such as hospitals, national research institutes, private research centers, and certain companies. This excludes the direct use of global university rankings, such as those provided by the US News and World Report. To meet both of these requirements, we perform a hybrid approach that first employs algorithmic linking to prior university rankings, and then annotates non-universities by soliciting a diverse range of expert feedback on the rankings.

We first algorithmically link author institutions to the rankings of the US News and World (USNW) Top 2000 Global Universities [1] . This approach requires cleaning self-reported institutions using common natural language processing techniques, such as removing punctuation, parentheticals, and certain stopwords. Helpfully, AAAS provided us with a dictionary mapping 16,715 out of the 48,436 unique self-reported affiliations to a canonical institution name, which does not necessarily match the name provided by the USNW rankings. To assist the linkage, we create a further set of manual synonyms for each university. Viewed as a network where nodes are potential institution names, and edges are known connections between names (such as a canonical linkage provided by AAAS, a name that is a preprocessed version of another name, or a synonym that we generated ourselves), we deduplicate the names by approximately inferring connected components of names on this network. In particular, we first check for direct matches between self-reported institutions and the names provided by the USNW rankings. Then, we expand the set of canonical names for any given institution by associating with a given USNW institution the set of aliases provided by AAAS and ourselves. Lastly, we match any remaining self-reported institutions with aliases that

[1] https://www.usnews.com/education/best-global-universities/search

are now linked to USNW institutions. This approach matches 74.3% of author affiliations to the USNW rankings. We sample 54 unmatched institutions, and find that 15 (26%) should have been matched to the USNW rankings, suggesting that our algorithmic approach has covered 92% of possible university-based matches.

To expand our rankings to include affiliations of submitting authors that were omitted from our algorithmic linkage, we can focus on the most common institutions that remained unmatched. By binning submissions by institution, we find that the top 1% of the unranked institutions in terms of manuscripts submitted (292 institutions), accounted for 30.6% of the total unranked author-institution pairs. The top 5% of unranked institutions (1463 institutions) would only expand the total coverage to 50.9% of author-institution pairs. Thus labeling the top 292 institutions would increase our coverage of prestige from 74.3% to 82% of all author-institution pairs. These 292 institutions included a mix of internationally renowned research institutes and institutions with virtually no international reputation, such as universities without a ranking in the USNW Top 2000 Global Universities.

We label these remaining institutions, and check the validity of the USNW rankings, by sending a preliminary set of rankings to an international group of experts. First, we binned the USNW rankings into deciles, such that there were roughly an equal number of submissions from each decile. Thus the top of the hierarchy had fewer institutions than the bottom of the hierarchy, since the population of scholars who submit to *Science* and *Science Advances* are disproportionately based in the most prestigious institutions. Second, we assigned preliminary rankings to the missing institutions through the following set of rules. For foreign national labs, we assigned them the highest value of prestige from universities from that country in USNW. When an institution had a clear university partner, we assigned that university's prestige to the institution. Well-known US federal agencies were generally placed into the top two bins. Some private institutions that we recognized as having exceptional reputations were placed into the top bin. The remaining unrecognized institutions, such as universities that were omitted from the USNW Top 2000 rankings, were placed in bottom bins.

These preliminary binned rankings were sent to 22 international experts, spanning a range of disciplines including but not limited to earth sciences, microbiology, cognitive science, applied mathematics, public health, data science, and physiology. Eleven of the experts were based in the US or Canada, six were based in Europe, three were based in Asia, and the remaining two were based in Australia and New Zealand, respectively. From this panel of experts, we received responses from 12, for a total of 195 proposed re-binning of institutions. We generally incorporated their suggestions, and when there were conflicts, we took the suggestion of the scholar who was geographically or topically closer to the disputed institution, using seniority as a tie-breaker.

4.0.4 Adding topic variables to articles

We use submission titles and abstracts to derive article topic labels at three levels of granularity. First, we applied the publicly available SPECTER2 embedding model to submission titles and abstracts, yielding a vector of length 768 for each paper [178]. The SPECTER2 model was designed specifically to generate effective embeddings for downstream applications on scientific articles, e.g., two similar articles will produce vectors that are separated by a relatively small euclidean distance, while two dissimilar papers will produce vectors that are farther apart.

We apply K-means clustering to each article's embedding vector in order to group nearby papers into the same topics. We run this clustering for k=10, k=30, and k=100, resulting in three sets of topic labels. The k=10 clustering assigns each article to one of 10 coarse grained topics, while the k=100 clustering assigns each article to one of 100 more fine-grained topics. The resulting topic labels do not necessarily adhere to a strict tree structure, meaning that two articles from a fine-grained topic can belong to different coarse grained topics. By providing multiple levels of granularity, we allow data users to make analysis appropriate choices. For example, in our broad analysis of peer review presented in Fig. 4.1, we used the k=10 topic clustering as fixed effects in our regression models. To preserve the anonymity of these data, we do not disclose how the reported topics (e.g., topic 21) corresponds to which field (e.g., cryptography vs. particle accelerator physics vs. epidemiology).

4.0.5 Adding sentiment annotations to reviews

We derive two measures of sentiment from the text provided by reviewers: overall sentiment and sentiment trajectories. The overall sentiment aims to reflect the reviewer's evaluation of a manuscript, while the sentiment trajectory reflects the review's narrative structure. To motivate the creation of a sentiment trajectory variable with an example, consider two hypothetical reviews that are both generally negative, and both ultimately recommend rejection. One review is negative throughout its entire length, and accordingly has a flat sentiment trajectory. The other review begins on a positive note before enumerating the papers many flaws in a more negative tone, resulting in a downward sentiment trajectory. A natural question that follows is whether these two review narrative structures have differing impacts on editor decisions. By estimating review sentiment trajectories, future work may be able to address these types of questions.

To extract overall review sentiments, we applied a Bidirectional Encoder Representations from Transformers (BERT) model that was fine-tuned for sentiment analysis on product reviews [179]. This model takes review text as input and returns the estimated probabilities that the review text is associated with a 1-star, 2-star, 3-star, 4-star, or 5-star review. We multiply these five estimated probabilities by their corresponding star-values, and sum the resulting products to construct a sentiment score that can range continuously from 1 (lowest possible sentiment) to 5 (highest possible sentiment). This model only considers the first 512 words of each review for evaluation. Even though this model was fine-tuned on product reviews, it surprisingly performs well in the context of peer-review sentiment. We hand-annotated a random sample of 50 reviews for general sentiment on a scale from 1-10, considering the entire length of each review, and found that the sentiment analysis model's annotations strongly correlated ($r = 0.80$, $p < 0.001$).

To extract review sentiment trajectories, we used the same BERT sentiment analysis model as before on multiple 3-sentence segments of each review [179]. These segments were created by moving a 3-sentence window across the entire review, with each segment overlapping the next by 2 sentences. Thus, a 30-sentence review would yield 28 sequential sentiment measurements which form its sentiment trajectory.

To assign each review's sentiment trajectory to one of 11 narrative structures, we first normalize the trajectories by subtracting the sentiment associated with the first three sentences of the review from each point in the sentiment trajectory, so each trajectory starts with a sentiment score of 0. This step aims to isolate the review's trajectory from its overall sentiment, which is measured separately. Next, we normalized the lengths of the reviews by linearly interpolating 1000 evenly spaced points over the span of the trajectory. Accordingly, even if a trajectory is only 10 sentences long yielding only 8 sentiment estimates, the path of these 8 points are linearly interpolated to construct a 1000-point trajectory of the same shape. Finally, we performed K-means clustering with $k = 30$ on the normalized trajectories, yielding 30 narrative structure clusters. We visualized the mean trajectory for each of the clusters by plotting their centroids. We then further coarsened each cluster by assigning each of the 30 to one of 11 coarse-grained trajectories, defined by up (U), down (D) and flat (F) segments that we observed (Fig. 4.2). We note that this clustering was performed on all first round reviews, even if some of these reviews are later dropped from the dataset to limit each paper to two reviews each.

To allow data users to make more informed decisions about whether and how to include these sentiment variables in their analyses, we provide code that can be used to apply the sentiment trajectory classification on user-defined input text.

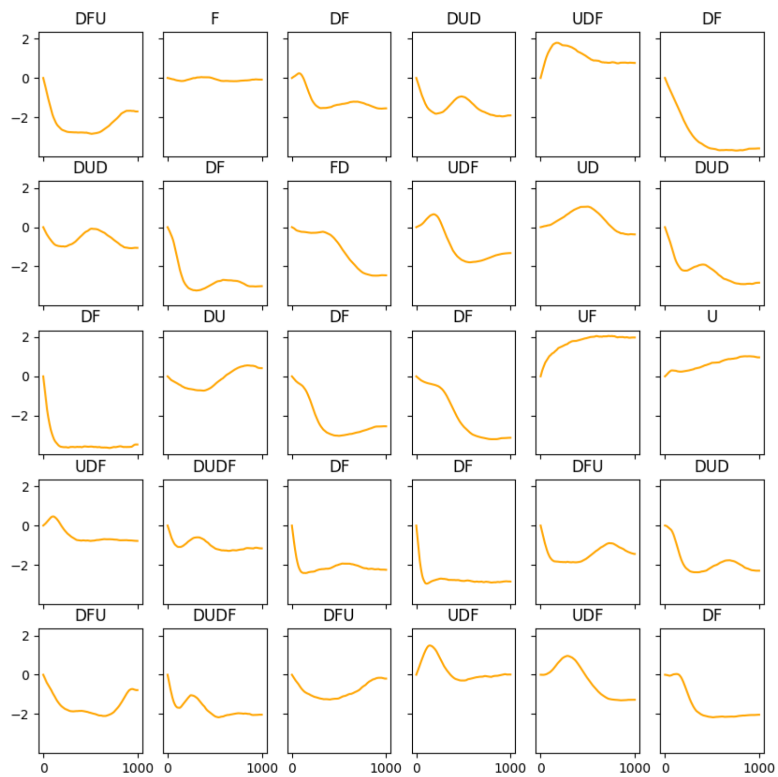

Figure 4.2: K-means centroids for the 30 narrative structure clusters to which reviews are assigned. We coarsen these narrative structures by hand-labeling their various shapes into 11 groups, based on upward trends in sentiment (U), downward trends in sentiment (D) and flat spans of sentiment (F).

Journal	Stages of Peer Review with Gender Disparities Counterfactually Removed	Observed Num. Articles with Women CAs	Counterfactual Num. Articles with Women CAs (95% C.I.)	Change Factor Under Intervention (95% C.I.)	
Science	None	750	752 (702, 800)	1.00 (0.94, 1.07)	
Science	Reviewed	750	788 (739, 839)	1.05 (0.99, 1.12)	
Science	Review Sentiment	750	765 (716, 816)	1.02 (0.95, 1.09)	
Science	Accepted	750	789 (740, 838)	1.05 (0.99, 1.12)	
Science	Review Sentiment, Accepted	750	803 (755, 856)	1.07 (1.01, 1.14)	*
Science	Reviewed, Accepted	750	827 (775, 877)	1.10 (1.03, 1.17)	*
Science	Reviewed, Review Sentiment, Accepted	750	842 (791, 889)	1.12 (1.05, 1.19)	*
Science Advances	None	898	902 (848, 953)	1.00 (0.94, 1.06)	
Science Advances	Reviewed	898	940 (889, 993)	1.05 (0.99, 1.11)	
Science Advances	Review Sentiment	898	912 (857, 965)	1.02 (0.95, 1.07)	
Science Advances	Accepted	898	952 (898, 1002)	1.06 (1.00, 1.12)	
Science Advances	Review Sentiment, Accepted	898	965 (914, 1017)	1.07 (1.02, 1.13)	*
Science Advances	Reviewed, Accepted	898	995 (942, 1045)	1.11 (1.05, 1.16)	*
Science Advances	Reviewed, Review Sentiment, Accepted	898	1005 (946, 1064)	1.12 (1.05, 1.18)	*

Table 4.5: **Counterfactual analysis of gender.** CA = Corresponding Author. For each journal and subset of stages, we simulate predicted numbers of accepted papers after counterfactually removing gender disparities in that subset of stages. Shown are the observed number of women CAs, the counterfactual number of women CAs (with 95% CIs from Monte Carlo simulation), and the change factor under intervention with 95% Monte Carlo CIs. Change factor is the ratio of counterfactual women CAs to observed women CAs. Asterisks in the right-hand column indicate statistical significance, meaning that the 95% confidence interval of the change factor excludes 1.0.

Chapter 5

Conclusions and Future Directions

This thesis presented three distinct research projects, which each focused on diversity among scientists and their research outputs, but each from different perspectives and each yielding unique insights. In chapter 2, I used data on the education, employment, research subfield, and demographics of tenure-track faculty at U.S.-based PhD-granting computing departments to quantify the intersection of multiple forms of demographic diversity by computing subfield with particular focus on faculty gender diversity, producing a detailed picture of past and likely future trends and inequalities. In chapter 3, I used a decade of census-level employment data on U.S. tenured and tenure track faculty to investigate the relative impacts of faculty hiring versus faculty attrition on women's representation across academia, answering three broad questions: (i) How have these two processes shaped gender representation across the academy over the decade 2011–2020? (ii) How might women's representation today have been different if gendered attrition among faculty were eliminated in 2011? (iii) And, how might we expect gender diversity among faculty to change over time under different future hiring and attrition scenarios? Finally, in chapter 4, I worked alongside collaborators to develop and investigate a comprehensive dataset on peer review outcomes at two elite multidisciplinary journals, *Science* and *Science Advances*, quantifying the effects of gender and additional relevant covariates at three key stages of the editorial process: (i) whether a submission is desk rejected or sent to review, (ii) the sentiment of the external reviews, and (iii) the final editorial decision to accept or reject the manuscript.

Across the three projects, I used large datasets to quantify the extent to which a variety of processes tended to either enrich women's representation or deplete women's representation. In chapter 2, I showed that the largest share of depletion in women's representation along the path to becoming tenure track faculty occurs prior to receiving a PhD in computing, yet additional depletion occurs between receiving a PhD and starting a faculty position to varying degrees, for all but two subfields (human-computer interaction and interdisciplinary computing, Fig. 2.1A). In chapter 3, I considered the extent to which gendered attrition has acted to enrich or deplete women's representation, finding that gendered faculty attrition caused a loss of women in both STEM and non-STEM fields, even as the direction and statistical significance of this effect varied at the more fine-grained field level (Fig. 3.2). In chapter 4, I estimated the impact of author, editor, and reviewer gender across several stages of peer review at *Science* and *Science Advances*. At *Science*, the corresponding author's gender was significantly correlated with the outcomes of each stage of peer review, except for the stage of the editor's final decision. Across these three separate analyses, I worked towards locating where each disparity occurs. By identifying the stages (e.g., pre-PhD/post-PhD, desk-reject/review-sentiment) and processes (e.g., gendered attrition) that broadly cause a depletion of women, I teed up future research to better understand these particular stages and processes.

In chapters 2 and 3, I made projections for women's future representation, finding that several academic domains, including engineering, math and computer science, and business, are not projected to reach gender parity within the next 40 years under scenario conditions that align with historical trends (Fig. 3.3). Even if every academic domain reached gender parity, our projections for the subfields of computing would suggest that quantifiable disparities would persist, but at the subfield level. Assuming historical trends in computing faculty representation continue, I predicted that only two of computing's subfields are likely to reach 25% women faculty by 2027 (Fig. 2.4).

The nested structures of academic domains, fields, and subfields, and the differences in demographic representation across these structures, can complicate authoritative claims regarding the (in)equitable representation of women. In the hypothetical situation where women faculty are representative of women's share of the greater population–that is, roughly at parity with men–there are still many configurations in which men and women faculty may be distributed at the domain, field, and subfield level. Woman and men may tend concentrate in some subfields and participate less in other subfields, as I observe in computing, or parity could permeate down the nested taxonomy of science. Identifying the causal mechanisms driving gender differences across fields and subfields would help determine the most equitable configuration; on one hand, some fields or subfields may be particularly inhospitable to women, effectively pushing women away. In this case, policy should aim to make these subfields more accessible and inclusive. On the other hand, women may, on average, be more interested in topics belonging to some fields or subfields over others. In this case, women's interests should govern their representational distribution across topics of study.

5.1 Limitations

All three of these projects utilized large observational datasets, and while these data were sometimes supplemented by survey responses (chapter 2), this survey data was primarily used to attain additional faculty demographic data. While these data supported the presented analyses, they were not well suited to dig deeply into the causal mechanisms driving the disparities that the presented analyses identified. For example, in chapter 3, I showed that gendered attrition has led to a differential loss of women in academia in the past decade, however, I could not explain why this was the case without additional data. One way to overcome this limitation is by collecting additional data. I contributed to a project led by Katie Spoon, which deployed a detailed survey of faculty specifically targeted to better understand the causal mechanisms that can lead faculty to leave their jobs. One key finding of this work is that reasons for leaving often differed by faculty gender: men were more likely to leave their jobs due to "pulls" to more desirable (non-faculty) positions, while women were more likely to leave their jobs due to "pushes" out of the tenure track [32].

Other major limitations to this work were that it only considered binary faculty gender identities, and its consideration of intersectional faculty identities were limited. These limitations were driven by a number of factors. First, in all three projects, I use name-based tools to estimate faculty gender, and these tools are not well suited to making nonbinary gender annotations. Second, I relied on survey responses to collect additional demographic information, including faculty race and measures of faculty socioeconomic status. Because a relatively small subset of faculty provided survey responses (less than 20%), and because faculty at the intersections of underrepresented identities are especially underrepresented among faculty, our sample sizes often became prohibitively small for robust intersectional statistical analyses. This point speaks to the importance of ongoing efforts to make science more inclusive and to collect more detailed faculty data to support these efforts.

Another limitation to this work was its relatively narrow scope. Even as the analyses presented in this thesis include some of the largest sample sizes of faculty in the related literature, the populations considered nevertheless represent an incomplete slice of the academic and scientific ecosystems—U.S. based tenure track faculty at PhD granting institutions for chapters 2 and 3, and authors submitting to *Science* and *Science Advances* for chapter 4. In addition to students and non-tenure track faculty who share many overlapping roles and responsibilities with tenure track faculty, there are researchers in industry and in government who also train future researchers and contribute to the scientific record. Moreover, science is an international collaboration, and the researchers and research output from other countries should be considered in a wholistic analysis of the scientific ecosystem. Similarly, not all research is submitted to elite journals like *Science* and *Science Advances*, and a more holistic understanding of peer-review, and the gender disparities therein, should consider the peer review outcomes across a breadth of publication venues.

5.2 Future Research

Researchers can conduct experiments to answer some key questions raised by this work. For example, in chapter 4, I find statistically significant gender disparities at several key stages of the peer review process at elite journals. Are these disparities a result of confounding variables, like career age? Are these disparities a result of editors and retailers knowing the gender of the authors (i.e., explicit or implicit bias)? Or are these disparities a result of gendered differences in the submitted articles? Consider a randomized controlled experiment at one of these journals, in which the journal randomly assigns double-blind versus single-blind review to manuscripts. A researcher could compare gender disparities in peer-review outcomes under the single-blind and double-blind conditions. The extent to which disparities are reduced in the double-blind case represents the degree of explicit and implicit bias against women authors themselves, assuming appropriate precautions are taken to ensure the robustness of the double-blinding.

An additional way to make progress on important questions raised in this thesis is by collecting and combining additional observational data. Data on a broader range of faculty and scientists, international faculty, or more expanded coverage for detailed demographic variables like race, ethnicity, socioeconomic status, self-reported gender identity, and sexuality, would facilitate novel research to address new questions and to evaluate the extent to which prior research on tenure track faculty in the U.S. generalizes to other scientists. Unfortunately, data collection can be expensive and logistically challenging. Future work may be able to address some of these challenges by collaborating with researchers at the federal government who have access to rich federal tax and demographic data (i.e., Decennial Census responses).

Finally, the elite peer review data that I anonymized and presented in chapter 4 is well-suited for several novel analyses. These include an analysis of editors' assignments of reviewers to articles that explores how editor attributes like prestige and gender correlate with the prestige and gender of the assigned BoRE evaluators and reviewers. It would also be valuable to develop a better understanding of the idiosyncrasies and disagreements among reviewers. When do the sentiment of peer review evaluations tend to align between reviewers, and how do agreements and disagreements relate to author and reviewer demographics, and article attributes like topic and number of authors? Finally, this data could support a more in depth exploration of how peer review outcomes vary by topic at *Science* and *Science Advances*. Do the differential acceptance rates between topics tend to differentially filter for or against women authors, or are the topic acceptance rates gender-neutral? Does prestige matter for some topics, but much for others? These analyses are ready to be explored and do not require any additional data collection.

Bibliography

[1] Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. The diversity–innovation paradox in science. Proceedings of the National Academy of Sciences USA, 117(17):9284–9291, 2020.

[2] Scott E Page. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press, 2008.

[3] Diego Kozlowski, Vincent Larivière, Cassidy R. Sugimoto, and Thema Monroe-White. Intersectional inequalities in science. Proceedings of the National Academy of Sciences, 119(2), 2022.

[4] Rembrand Koning, Sampsa Samila, and John-Paul Ferguson. Who do we invent for? Patents by women focus more on women's health, but few women get to invent. Science, 372(6548):1345–1348, 2021.

[5] Sam S Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E Barcelo, Marquitta J White, Danielle M de Bruin, Ruth M Greenblatt, Kirsten Bibbins-Domingo, Alan HB Wu, et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. PLoS medicine, 12(12):e1001918, 2015.

[6] Nicholas Laberge, K Hunter Wapman, Allison C Morgan, Sam Zhang, Daniel B Larremore, and Aaron Clauset. Subfield prestige and gender inequality among us computing faculty. Communications of the ACM, 65(12):46–55, 2022.

[7] National Science Foundation. Survey of doctorate recipients: 2019. National Center for Science and Engineering Statistics, 2019.

[8] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. Gender stereotypes about intellectual ability emerge early and influence children's interests. Science, 355(6323):389–391, 2017.

[9] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. Science, 347(6219):262–265, 2015.

[10] Jessica Ellis, Bailey K Fosdick, and Chris Rasmussen. Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. PLOS ONE, 11(7):e0157447, 2016.

[11] Casey A Shapiro and Linda J Sax. Major selection and persistence for women in STEM. New Directions for Institutional Research, 2011(152):5–18, 2011.

[12] Alice N Pell. Fixing the leaky pipeline: Women scientists in academia. Journal of Animal Science, 74(11):2843–2848, 1996.

[13] Jane Margolis, Allan Fisher, and Faye Miller. The anatomy of interest: Women in undergraduate computer science. Women's Studies Quarterly, 28(1/2):104–127, 2000.

[14] Jillian Kinzie. Women's paths in science: A critical feminist analysis. New Directions for Institutional Research, 2007(133):81–93, 2007.

[15] Sandra Katz, David Allbritton, John Aronis, Christine Wilson, and Mary Lou Soffa. Gender, achievement, and persistence in an undergraduate computer science program. ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 37(4):42–57, 2006.

[16] Joseph R Cimpian, Taek H Kim, and Zachary T McDermott. Understanding persistent gender gaps in STEM. Science, 368(6497):1317–1319, 2020.

[17] Myra Sadker and David M. Sadker. Failing at fairness: how our schools cheat girls. Touchstone, New York, 1994.

[18] David M Marx and Jasmin S Roman. Female role models: Protecting women's math test performance. Personality and Social Psychology Bulletin, 28(9):1183–1193, 2002.

[19] Penelope Lockwood. "Someone like me can be successful": Do college students need same-gender role models? Psychology of women quarterly, 30(1):36–46, 2006.

[20] Holly O Witteman, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. The Lancet, 393(10171):531–540, 2019.

[21] National Academies of Sciences, Engineering, and Medicine. Sexual harassment of women: climate, culture, and consequences in academic sciences, engineering, and medicine. National Academies Press, 2018.

[22] Jacob Clark Blickenstaff. Women and science careers: leaky pipeline or gender filter? Gender and education, 17(4):369–386, 2005.

[23] Brenda K Hawks and Joan Z Spade. Women and men engineering students: Anticipation of family and work roles. Journal of Engineering Education, 87(3):249–256, 1998.

[24] Gemma E Derrick, Adam Jaeger, Pei-Ying Chen, Cassidy R Sugimoto, Thed N van Leeuwen, and Vincent Larivière. Models of parenting and its effect on academic productivity: preliminary results from an international survey. In ISSI, pages 1670–1676, 2019.

[25] Allison C Morgan, Samuel F Way, Michael JD Hoefer, Daniel B Larremore, Mirta Galesic, and Aaron Clauset. The unequal impact of parenthood in academia. Science Advances, 7(9):eabd1996, 2021.

[26] Erin A Cech and Mary Blair-Loy. The changing career trajectories of new parents in STEM. Proceedings of the National Academy of Sciences, 116(10):4182–4187, 2019.

[27] Marc Goulden, Karie Frasch, Mary Ann Mason, et al. Staying competitive: Patching america's leaky pipeline in the sciences. Berkeley, CA: Center for American Progress, 2009.

[28] Laura L Risler. Female STEM Doctorate Holders in the Academic Workforce: An Event-History Analysis. Ohio University, 2019.

[29] Nicholas H Wolfinger, Mary Ann Mason, and Marc Goulden. Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. The Journal of Higher Education, 79(4):388–405, 2008.

[30] A. C. Morgan, S. F. Way, Daniel Larremore, and Aaron Clauset. Paid parental leave at us and canadian universities. The ANNALS of the American Academy of Political and Social Science, Retreived from https://aaronclauset.github.io/parental-leave.

[31] Marc Goulden, Mary Ann Mason, and Karie Frasch. Keeping women in the science pipeline. The ANNALS of the American Academy of Political and Social Science, 638(1):141–162, 2011.

[32] Katie Spoon, Nicholas LaBerge, K. Hunter Wapman, Sam Zhang, Allison C. Morgan, Mirta Galesic, Daniel B. Larremore, and Aaron Clauset. Gender and retention patterns among U.S. faculty. Science Advances, 9(42), 2023.

[33] J Scott Long, Paul D Allison, and Robert McGinnis. Rank advancement in academic careers: Sex differences and the effects of productivity. American Sociological Review, pages 703–722, 1993.

[34] Sandra Groeneveld, Kea Tijdens, and Daphne van Kleef. Gender differences in academic careers: Evidence for a dutch university from personnel data 1990-2006. Equality, Diversity and Inclusion: An International Journal, 2012.

[35] Janet M Box-Steffensmeier, Raphael C Cunha, Roumen A Varbanov, Yee Shwen Hoh, Margaret L Knisley, and Mary Alice Holmes. Survival analysis of faculty retention and promotion in the social sciences by gender. PLOS ONE, 10(11):e0143093, 2015.

[36] Jonathan R Cole and Harriet Zuckerman. The productivity puzzle. Advances in Motivation and Achievement. Women in Science, 1984.

[37] Helen S Astin and Christine M Cress. A national profile of academic women in research universities. In Equal Rites, Unequal Outcomes, pages 53–88. Springer, 2003.

[38] Marcia Gumpertz, Raifu Durodoye, Emily Griffith, and Alyson Wilson. Retention and promotion of women and underrepresented minority faculty in science and engineering at four large land grant institutions. PLOS ONE, 12(11):e0187285, 2017.

[39] National Research Council et al. Gender differences at critical transitions in the careers of science, engineering, and mathematics faculty. National Academies Press, 2010.

[40] Deborah Kaminski and Cheryl Geisler. Survival analysis of faculty retention in science and engineering by gender. Science, 335(6070):864–866, 2012.

[41] Margaret W Rossiter. The matthew matilda effect in science. Social studies of science, 23(2):325–341, 1993.

[42] Anne E Lincoln, Stephanie Pincus, Janet Bandows Koster, and Phoebe S Leboy. The matilda effect in science: Awards and prizes in the us, 1990s and 2000s. Social Studies of Science, 42(2):307–320, 2012.

[43] Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Gender bias in scholarly peer review. eLife, 6:e21718, 2017.

[44] Dakota Murray, Kyle Siler, Vincent Lariviere, Wei Mun Chan, Andrew M Collings, Jennifer Raymond, and Cassidy R Sugimoto. Author-reviewer homophily in peer review. BioRxiv, page 400515, 2018.

[45] Charles W Fox and CE Timothy Paine. Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. Ecology and Evolution, 9(6):3599–3619, 2019.

[46] Laura Hospido and Carlos Sanz. Gender gaps in the evaluation of research: Evidence from submissions to economics conferences. Oxford Bulletin of Economics and Statistics, 83(3):590–618, 2021.

[47] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. The extent and drivers of gender imbalance in neuroscience reference lists. Nature neuroscience, 23(8):918–926, 2020.

[48] Neven Caplar, Sandro Tacchella, and Simon Birrer. Quantitative evaluation of gender bias in astronomical publications from citation counts. Nature Astronomy, 1(6):0141, 2017.

[49] Paula Chatterjee and Rachel M Werner. Gender disparity in citations in high-impact journal articles. JAMA Netw Open, 4(7):e2114509, 2021.

[50] Maxwell A. Bertolero, Jordan D. Dworkin, Sophia U. David, Claudia López Lloreda, Pragya Srivastava, Jennifer Stiso, Dale Zhou, Kafui Dzirasa, Damien A. Fair, Antonia N. Kaczkurkin, Bianca Jones Marlin, Daphna Shohamy, Lucina Q. Uddin, Perry Zurn, and Danielle S. Bassett. Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. bioRxiv, 2020.

[51] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. Gendered citation patterns across political science and social science methodology fields. Political Analysis, 26(3):312–327, 2018.

[52] Xinyi Wang, Jordan D Dworkin, Dale Zhou, Jennifer Stiso, Erika B Falk, Dani S Bassett, Perry Zurn, and David Martin Lydon-Staley. Gendered citation practices in the field of communication. Annals of the International Communication Association, 45(2):134–153, 2021.

[53] Ludo Waltman. A review of the literature on citation impact indicators. Journal of informetrics, 10(2):365–391, 2016.

[54] Vincent A Traag. Inferring the causal effect of journals on citations. abbrev-type=" full" Quantitative Science Studies, 2(2):496–504, 2021.

[55] Ching Jin, Yifang Ma, and Brian Uzzi. Scientific prizes and the extraordinary growth of scientific topics. Nature Communications, 12(1):1–11, 2021.

[56] Yifang Ma, Diego FM Oliveira, Teresa K Woodruff, and Brian Uzzi. Women who win prizes get less money and prestige, 2019.

[57] Nicolas Robinson-Garcia, Rodrigo Costas, Cassidy R Sugimoto, Vincent Larivière, and Gabriela F Nane. Task specialization across research careers. ELife, 9, 2020.

[58] Chaoqun Ni, Elise Smith, Haimiao Yuan, Vincent Larivière, and Cassidy R Sugimoto. The gendered nature of authorship. Science Advances, 7(36):eabe4639, 2021.

[59] Jevin D West, Jennifer Jacquet, Molly M King, Shelley J Correll, and Carl T Bergstrom. The role of gender in scholarly authorship. PLOS ONE, 8(7):e66212, 2013.

[60] Cassidy R Sugimoto and Vincent Larivière. Equity for women in science: Dismantling systemic barriers to advancement. Harvard University Press, 2023.

[61] Luke Holman and Claire Morandin. Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. PloS one, 14(4):e0216128, 2019.

[62] Marek Kwiek and Wojciech Roszka. Gender-based homophily in research: A large-scale study of man-woman collaboration. Journal of Informetrics, 15(3):101171, 2021.

[63] Mathias Wullum Nielsen, Jens Peter Andersen, Londa Schiebinger, and Jesper W Schneider. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. Nature human behaviour, 1(11):791–796, 2017.

[64] Cassidy R Sugimoto, Yong-Yeol Ahn, Elise Smith, Benoit Macaluso, and Vincent Larivière. Factors affecting sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. The Lancet, 393(10171):550–559, 2019.

[65] Rembrand Koning, Sampsa Samila, and John-Paul Ferguson. Who do we invent for? Patents by women focus more on women's health, but few women get to invent. Science, 372(6548):1345–1348, 2021.

[66] Diego Kozlowski, Vincent Larivière, Cassidy R Sugimoto, and Thema Monroe-White. Intersectional inequalities in science. Proceedings of the National Academy of Sciences, 119(2), 2022.

[67] James A Evans, Jae-Mahn Shim, and John PA Ioannidis. Attention to local health burden and the global disparity of health research. PloS one, 9(4):e90147, 2014.

[68] Stuart Zweben and Betsy Bizot. 2020 Taulbee survey. Computing Research Association, 2020.

[69] U.S. Census Bureau. Quickfacts. U.S. Department of Commerce, 2019.

[70] Sandra Mattauch, Katja Lohmann, Frank Hannig, Daniel Lohmann, and Jürgen Teich. A bibliometric approach for detecting the gender gap in computer science. Communications of the ACM, 63(5):74–80, 2020.

[71] Lowell L Hargens and J Scott Long. Demographic inertia and women's representation among faculty in higher education. The Journal of Higher Education, 73(4):494–517, 2002.

[72] Robyn Marschke, Sandra Laursen, Joyce McCarl Nielsen, and Patricia Rankin. Demographic inertia revisited: An immodest proposal to achieve equitable gender representation among faculty in higher education. The Journal of Higher Education, 78(1):1–26, 2007.

[73] Monica Adya and Kate M Kaiser. Early determinants of women in the IT workforce: a model of girls' career choices. Information Technology & People, 2005.

[74] Tracy Camp. The incredible shrinking pipeline. Communications of the ACM, 40(10):103–110, 1997.

[75] Sin-Ning C Liu, Stephanie EV Brown, and Isaac E Sabat. Patching the "leaky pipeline": Interventions for women of color faculty in stem academia. Archives of Scientific Psychology, 7(1):32, 2019.

[76] Betsy Bizot. Expanding the pipeline: Gender and ethnic differences in PhD specialty areas. Computing Research Association, 2019.

[77] Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. Systematic inequality and hierarchy in faculty hiring networks. Science Advances, 1(1):e1400005, 2015.

[78] Elizabeth Carlson and Christopher Zorn. Toward better hiring practices. PS: Political Science and Politics, pages 1–4, 2021.

[79] Jeffrey Mervis. NIH's new cluster hiring program aims to help schools attract diverse faculty. Science, 2020.

[80] Caterina De Bacco, Daniel B Larremore, and Cristopher Moore. A physical model for efficient ranking in networks. Science Advances, 4(7):eaar8260, 2018.

[81] Samuel F Way, Daniel B Larremore, and Aaron Clauset. Gender, productivity, and prestige in computer science faculty hiring networks. In Proceedings of the 25th International Conference on World Wide Web, pages 1169–1179, 2016.

[82] Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity. Federal Register, 62(210):58782–58790, 1997.

[83] Allison C Morgan, Nicholas LaBerge, Daniel B Larremore, Mirta Galesic, Jennie E Brand, and Aaron Clauset. Socioeconomic roots of academic faculty. Nature Human Behaviour, pages 1–9, 2022.

[84] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 204–213, 2012.

[85] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. Transactions of the Association for Computational Linguistics, 5:529–542, 2017.

[86] Frank Hobbs and Nicole Stoops. Demographic trends in the 20th century. U.S. Department of Commerce, 2002.

[87] Lindsay M Howden and Julie A Meyer. Age and sex composition: 2010. U.S. Department of Commerce, 2011.

[88] Michael N Bastedo. How to make college rankings less self-serving. The Chronicle of Higher Education, 2011.

[89] Alvin P Sanoff. The U.S. news college rankings: a view from the inside. College and University Ranking Systems, 9, 2007.

[90] Emery Berger. CSRankings: Computer science rankings, 2017-2021. https://csrankings.org/.

[91] Behrooz Parhami. Low acceptance rates of conference papers considered harmful. Computer, 49(4):70–73, 2016.

[92] Mario Biagioli. Quality to impact, text to metadata: publication and evaluation in the age of metrics. KNOW: A Journal on the Formation of Knowledge, 2(2):249–275, 2018.

[93] Michael Fire and Carlos Guestrin. Over-optimization of academic publishing metrics: observing Goodhart's law in action. GigaScience, 8(6):giz053, 2019.

[94] Campbell Gibson and Kay Jung. Historical Census Statistics on Population Totals by Race, 1790 to 1990, and by Hispanic origin, 1790 to 1990, for the United States, Regions, Divisions, and States. U.S. Census Bureau, 2002.

[95] Karen R Humes, Nicholas A Jones, and Roberto R Ramirez. Overview of Race and Hispanic Origin: 2010. U.S. Census Bureau, 2011.

[96] National Science Foundation. Survey of earned doctorates: 1980-2019. National Center for Science and Engineering Statistics, 2020.

[97] Ebony Omotola McGee. Interrogating structural racism in stem higher education. Educational Researcher, 49(9):633–644, 2020.

[98] Julie R Posselt. Equity in science. Stanford University Press, 2020.

[99] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. The gender gap in science: How long until women are equally represented? PLoS Biology, 16(4):e2004956, 2018.

[100] Yingyi Ma and Yan Liu. Race and STEM degree attainment. Sociology Compass, 9(7):609–618, 2015.

[101] Catherine Riegle-Crumb, Barbara King, and Yasmiyn Irizarry. Does stem stand out? examining racial/ethnic gaps in persistence across postsecondary fields. Educational Researcher, 48(3):133–144, 2019.

[102] Eileen M Trauth, Curtis C Cain, Kshiti D Joshi, Lynette Kvasny, and Kayla M Booth. The influence of gender-ethnic intersectionality on gender stereotypes about IT skills and knowledge. ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 47(3):9–39, 2016.

[103] Kimberly A Griffin, Meghan J Pifer, Jordan R Humphrey, and Ashley M Hazelwood. (re)defining departure: Exploring black professors experiences with and responses to racism and racial climate. American Journal of Education, 117(4):495–526, 2011.

[104] Kelly E Slay, Kimberly A Reyes, and Julie R Posselt. Bait and switch: Representation, climate, and tensions of diversity work in graduate education. The Review of Higher Education, 42(5):255–286, 2019.

[105] M Kevin Eagan Jr and Jason C Garvey. Stressing out: Connecting race, gender, and stress with faculty productivity. The Journal of Higher Education, 86(6):923–954, 2015.

[106] Dalton Conley. Being black, living in the red: Race, wealth, and social policy in America. Univ of California Press, 2010.

[107] EdBuild. $23 billion, 2019. https://edbuild.org/content/23-billion/full-report.pdf.

[108] Jason A Grissom and Christopher Redding. Discretion and disproportionality: Explaining the underrepresentation of high-achieving students of color in gifted programs. Aera Open, 2(1):2332858415622175, 2015.

[109] Samuel R Lucas and Mark Berends. Sociodemographic diversity, correlated achievement, and de facto tracking. Sociology of Education, pages 328–348, 2002.

[110] Roman Jurowetzki, Daniel Hain, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. The privatization of ai research(-ers): Causes and potential consequences – from university-industry interaction to public research brain-drain?, 2021 Preprint available at https://arxiv.org/abs/2102.01648, version 2.0.

[111] Allison C Morgan, Dimitrios J Economou, Samuel F Way, and Aaron Clauset. Prestige drives epistemic inequality in the diffusion of scientific ideas. EPJ Data Science, 7(1):40, 2018.

[112] Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. The misleading narrative of the canonical faculty productivity trajectory. Proceedings of the National Academy of Sciences USA, 114(44):E9216–E9223, 2017.

[113] Jon McNaughtan, Hugo A García, and Kim Nehls. Understanding the growth of contingent faculty. New Directions for Institutional Research, 2017(176):9–26, 2017.

[114] Ashley Finley. Women as contingent faculty: The glass wall. On Campus with Women, 37(3), 2009.

[115] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. The University of Chicago Legal Forum, page 139, 1989.

[116] Rafael A Buerba, Armin Arshi, Danielle C Greenberg, and Nelson F SooHoo. The role of gender, academic affiliation, and subspecialty in relation to industry payments to orthopaedic surgeons. Journal of the National Medical Association, 112(1):82–90, 2020.

[117] Jo Dixon and Carroll Seron. Stratification in the legal profession: Sex, sector, and salary. Law and Society Review, pages 381–412, 1995.

[118] Linda M Blum. Between Feminism and Labor: The Significance of the Comparable Worth Movement. University of California Press, 1991.

[119] Tony Tam. Sex segregation and cccupational gender inequality in the united states: Devaluation or specialized training? American Journal of Sociology, 102(6):1652–1692, 1997.

[120] Paula England. Comparable worth: Theories and evidence. Routledge, 2017.

[121] Burçin Tamer. Twice as many CREU/DREU students attend graduate school, compared to other reu students. Computing Research News, 28, 2016.

[122] Mariano R. Sto. Domingo, Starlette Sharp, Amy Freeman, Thomas Freeman, Keith Harmon, Mitsue Wiggs, Viji Sathy, Abigail T. Panter, Leticia Oseguera, Shuyan Sun, Mary Elizabeth Williams, Joseph Templeton, Carol L. Folt, Eric J. Barron, Freeman A. Hrabowski, Kenneth I. Maton, Michael Crimmins, Charles R. Fisher, and Michael F. Summers. Replicating meyerhoff for inclusive excellence in STEM. Science, 364(6438):335–337, 2019.

[123] Kendall Powell. Higher education: On the lookout for true grit. Nature, 504(7480):471–473, 2013.

[124] Bettina J Casad, Jillian E Franks, Christina E Garasky, Melinda M Kittleman, Alanna C Roesler, Deidre Y Hall, and Zachary W Petzel. Gender inequality in academia: Problems and solutions for women faculty in STEM. Journal of Neuroscience Research, 99(1):13–23, 2021.

[125] Jane Stout and Tracy Camp. Now what? action items from social science research to bridge the gender gap in computing research. ACM SIGCAS Computers and Society, 44(4):5–8, 2014.

[126] Marta Tienda. Diversity≠inclusion: Promoting integration in higher education. Educational Researcher, 42(9):467–475, 2013.

[127] Laurel Smith-Doerr, Sharla N Alegria, and Timothy Sacco. How diversity matters in the U.S. science and engineering workforce: a critical review considering integration in teams, fields, and organizational contexts. Engaging Science, Technology, and Society, 3:139–153, 2017.

[128] Jane G Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A McManus. STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (stem). Journal of Personality and Social Psychology, 100(2):255, 2011.

[129] Yang Yang, Tanya Y Tian, Teresa K Woodruff, Benjamin F Jones, and Brian Uzzi. Gender-diverse teams produce more novel and higher-impact scientific ideas. Proceedings of the National Academy of Sciences, 119(36):e2200841119, 2022.

[130] K. Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B. Larremore. Quantifying hierarchy and dynamics in us faculty hiring and retention. Nature, 610(7930):120–127, 2022.

[131] Leslie Davis and Richard Fry. College faculty have become more racially and ethnically diverse, but remain far less so than students. Pew Research Center, 2019.

[132] C De Brey, TD Snyder, A Zhang, and SA Dillow. Digest of education statistics. National Center for Education Statistics, Institute of Education Sciences, US Department of Education, 2021.

[133] National Center for Science and National Science Foundation Engineering Statistics. Doctorate recipients from U.S. universities: 2021, 2022.

[134] Arrman Kyaw. NIH awards $16 million grants for diverse faculty recruitment to Northwestern and UC San Diego. Diverse Issues in Higher Education, 2022.

[135] Bailey Margaret and Sean Rommel. Recommendations for reducing unconscious bias in promotion and tenure review processes. Rochester Institute of Technology Office of the Provost, 2018.

[136] Charissa Samaniego, Peggy Lindner, Maryam A Kazmi, Bobbie A Dirr, Dejun Tony Kong, Evonzia Jeff-Eke, and Christiane Spitzmueller. Higher research productivity = more pay? gender pay-for-productivity inequity across disciplines. Scientometrics, pages 1–13, 2023.

[137] National Science Foundation. Launching early-career academic pathways in the mathematical and physical sciences, 2023.

[138] Nicole R Thomas, Daniel J Poole, and Joan M Herbers. Gender in science and engineering faculties: Demographic inertia revisited. PLOS One, 10(10):e0139767, 2015.

[139] Alex James and Ann Brower. Levers of change: using mathematical models to compare gender equity interventions in universities. Royal Society Open Science, 9(9):220785, 2022.

[140] Ann Brower and Alex James. Research performance and age explain less than half of the gender pay gap in new zealand universities. PLOS ONE, 15(1):e0226392, 2020.

[141] Barbara Lawrence and Xiao Chen. The effect of demographic inertia on increasing the percent of women on university faculty. UCLA CCPR Population Working Papers, 2015.

[142] Allison K Shaw and Daniel E Stanton. Leaks in the pipeline: separating demographic inertia from ongoing gender differences in academia. Proceedings of the Royal Society B: Biological Sciences, 279(1743):3736–3741, 2012.

[143] Academic Analytics Research Center. Academic analytics research center list of available data elements. https://aarcresearch.com/available-data-elements.pdf, 2021.

[144] Ian Van Buskirk, Aaron Clauset, and Daniel B Larremore. An open-source cultural consensus approach to name-based gender classification. In Proceedings of the International AAAI Conference on Web and Social Media, volume 17, pages 866–877, 2023.

[145] National Center for Science and Engineering Statistics. Women, minorities, and persons with disabilities in science and engineering: 2021. special report NSF 21–321, 2021.

[146] Katie Spoon, Nicholas LaBerge, K. Hunter Wapman, Sam Zhang, Allison C. Morgan, Mirta Galesic, Daniel B. Larremore, and Aaron Clauset. Gender and retention patterns among U.S. faculty. Science Advances, 9(42), 2023.

[147] Hal Caswell. Matrix population models: construction, analysis, and interpretation. Sinauer Associates, Inc., 2nd edition, 2001.

[148] Uma M Jayakumar, Tyrone C Howard, Walter R Allen, and June C Han. Racial privilege in the professoriate: An exploration of campus climate, retention, and satisfaction. The Journal of Higher Education, 80(5):538–563, 2009.

[149] Phyllis L Carr, Anita Raj, Samantha E Kaplan, Norma Terrin, Janis L Breeze, and Karen M Freund. Gender differences in academic medicine: retention, rank, and leadership comparisons from the National Faculty Survey. Academic Medicine: Journal of the Association of American Medical Colleges, 93(11):1694, 2018.

[150] Benoit Macaluso, Vincent Larivière, Thomas Sugimoto, and Cassidy R Sugimoto. Is science built on the shoulders of women? a study of gender differences in contributorship. Academic Medicine, 91(8):1136–1142, 2016.

[151] Francine M Deutsch and Beier Yao. Gender differences in faculty attrition in the usa. Community, Work & Family, 17(4):392–408, 2014.

[152] Larry R Martinez, Katharine R O'Brien, and Michelle R Hebl. Fleeing the ivory tower: Gender differences in the turnover experiences of women faculty. Journal of Women's Health, 26(5):580–586, 2017.

[153] Rickard Danell and Mikael Hjerm. Career prospects for female university researchers have not improved. Scientometrics, 94:999–1006, 2013.

[154] National Center for Education Statistics. Fast facts: Race/ethnicity of college faculty, 2022.

[155] L. Bornmann, R. Mutz, and H.-D. Daniel. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. PLOS ONE, 5(12):e14331, 2010.

[156] D. Cicchetti. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14(1):119–135, 1991.

[157] Kyle Siler, Kirby Lee, and Lisa Bero. Measuring the effectiveness of scientific gatekeeping. Proc. Natl. Acad. Sci. USA, 112(2):360–365, 2015.

[158] Christian Greiffenhagen. Checking correctness in mathematical peer review. Social Studies of Science, 54(2):184–209, 2024.

[159] Travis A Hoppe, Aviva Litovitz, Kristine A Willis, Rebecca A Meseroll, Matthew J Perkins, B Ian Hutchins, Alison F Davis, Michael S Lauer, Hannah A Valantine, James M Anderson, and George M Santangelo. Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. Science Advances, 5(10), 2019.

[160] Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1):2–17, 2013.

[161] Ludo Waltman. A review of the literature on citation impact indicators. Journal of Informetrics, 10(2):365–391, 2016.

[162] Erin C McKiernan, Lesley A Schimanski, Carol Munoz Nieves, Lisa Matthias, Meredith T Niles, and Juan P Alperin. Use of the journal impact factor in academic review, promotion, and tenure evaluations. eLife, 8:e47338, 2019.

[163] E. Samuel Reich. Science publishing: The golden club. Nature, 502:291–293, 2013.

[164] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. Science Advances, 1:e1400005, 2015.

[165] V. A. Traag. Inferring the causal effect of journals on citations. Quantitative Science Studies, 2(2):496–504, 2021.

[166] Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Research: Gender bias in scholarly peer review. Elife, 6:e21718, 2017.

[167] Dakota Murray, Kyle Siler, Vincent Larivière, Wei Mun Chan, Andrew M Collings, Jennifer Raymond, and Cassidy R Sugimoto. Gender and international diversity improves equity in peer review. BioRxiv, page 400515, 2019.

[168] David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Are referees and editors in economics gender neutral? The Quarterly Journal of Economics, 135(1):269–327, 2020.

[169] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. Proc. Natl. Acad. Sci. USA, 114(48):12708–12713, 2017.

[170] Cassidy R Sugimoto and Vincent Larivière. Equity for Women in Science: Dismantling Systemic Barriers to Advancement. Harvard University Press, 2023.

[171] Alison Abritis and Alison McCook. Cash bonuses for peer-reviewed papers go global. Science, 357(6351):541, 2017.

[172] Wei Quan, Bikun Chen, and Fei Shu. Publish or impoverish: An investigation of the monetary reward system of science in china (1999-2016). Aslib Journal of Information Management, 69(5):486–502, 2017.

[173] Noah Greifer. Weightit: weighting for covariate balance in observational studies. R package version 1.1.0, 2020.

[174] Greg Ridgeway, Daniel F McCaffrey, Andrew R Morral, Matthew Cefalu, Lane F Burgette, Joseph D Pane, and Beth Ann Griffin. Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the R TWANG package. Rand Santa Monica, Calif, 2022.

[175] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973, 2005.

[176] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.

[177] K Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B Larremore. Quantifying hierarchy and dynamics in us faculty hiring and retention. Nature, 610(7930):120–127, 2022.

[178] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In Conference on Empirical Methods in Natural Language Processing, 2022.

[179] NLP Town. bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment, 2020. Accessed: 2024-05-24.

Appendix A

Supplementary Information for Chapter 3

A.1 Decomposition of Change in Gender Diversity

We develop a method to decompose the change in each field's gender diversity into its two main components: change due to hiring, and change due to attrition. First, the fraction of faculty in a field that are women in any given year can be written as

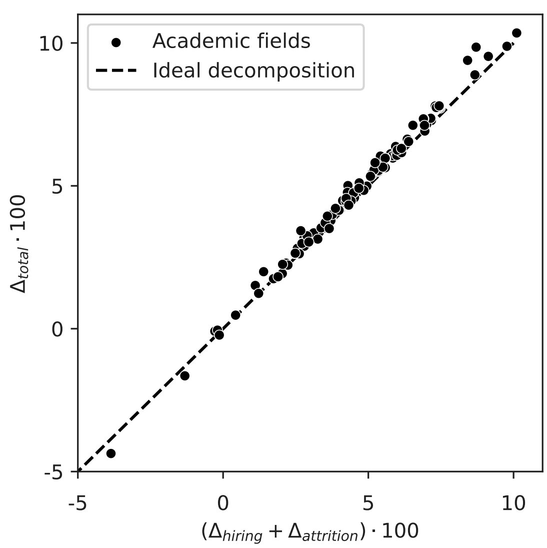

Figure A.1: The change in gender diversity between 2011 and 2020 can be approximately decomposed into parts due to hiring and attrition for each academic field, but there is a leftover residual term. In practice, we find that the residual term tends to be very small, such that the decomposition is nearly ideal. The dotted line represents an ideal decomposition, where the change in women's representation among faculty due to hiring and attrition perfectly matches the total observed change.

$$\frac{n_w}{n_w + n_m}$$

where n_w and n_m are the number of women and men faculty in the field, respectively. An additional term could be added to account for nonbinary faculty representation if our data included nonbinary gender annotations, however the methodology we use [144] assigns only binary (woman / man) labels to faculty.

Then, in the following year, the new fraction of women faculty can be written as

$$\frac{n_w + h_w - x_w}{n_w + h_w - x_w + n_m + h_m - x_m}$$

where h_w and h_m are the numbers of women and men that were hired in the following year, respectively, and x_w and x_m are the number of women and men among faculty "all-cause" attritions (whether for retirement, or otherwise). The total change in women's representation between two academic years δ_{total}, can thus be written as

$$\delta_{total} = \frac{n_w + h_w - x_w}{n_w + h_w - x_w + n_m + h_m - x_m} - \frac{n_w}{n_w + n_m}$$

We decompose this total change into change due to hiring, δ_{hiring}, and change due to attrition, $\delta_{attrition}$, as follows:

$$\delta_{hiring} = \frac{n_w + h_w}{n_w + h_w + n_m + h_m} - \frac{n_w}{n_w + n_m}$$
$$\delta_{attrition} = \frac{n_w - x_w}{n_w - x_w + n_m - x_m} - \frac{n_w}{n_w + n_m}$$

This decomposition behaves intuitively. For example, if the share of women that are hired exceeds the fraction of women in the field prior to hiring, then δ_{hiring} will be positive. On the other hand, if the share of women that are hired is lower than the fraction of women in the field prior to hiring, then δ_{hiring} will be negative. If there are no hires, $\delta_{hiring} = 0$. Similar intuition can be applied for $\delta_{attrition}$.

We sum the change in representation due to hiring and attrition over each year between 2011 to 2020 to get the overall change in representation due to hiring, Δ_{hiring}, and attrition, $\Delta_{attrition}$.

One potential limitation of this decomposition is that Δ_{hiring} and $\Delta_{attrition}$ do not perfectly sum to the exact observed change in women's representation over a given range of years, Δ_{total}. Instead, there is a residual term

$$\Delta_{residual} = \Delta_{total} - \Delta_{hiring} - \Delta_{attrition}$$

Intuitively, we know that there should be a residual term, because the change in representation that results from a given cohort of new hires can depend upon the number of attritions observed in that year, and vice versa. If the residual terms are large, this decomposition would not be a good approximation of the total change in representation, and Fig. 3.1 could be misleading. In Fig. A.1 we show that the residual terms are small (ranging from -0.51 *pp* to 1.14 *pp*, median = 0.2 *pp*), and thus the decomposition is a good approximation of the total change in women's representation.

A.1.1 Model Validation & Sensitivity Analysis

One way that we validate this model of faculty hiring and attrition is by starting the model in 2011, and comparing the resulting gender composition of faculty with the observed gender composition of faculty in 2020. In this validation, we use the same set of model parameters as in the gender-neutral counterfactual analysis (Sec. 3.0.2), except attrition patterns are inferred separately for women faculty and men faculty. In other words, the attrition probabilities in this validation are not gender-neutral. We find that the observed outcomes are statistically indistinguishable from the model-based outcomes for all 111 fields, and for STEM and non-STEM aggregations (Fig. A.2). This finding is not surprising, because the model is fit to the observed data, but it serves to validate the methods that we used to set the model's parameters (e.g., fitting logistic regression models to infer attrition risks and to infer the fraction of women faculty among new hires, as described in Methods Sec. 3.0.7).

We additionally validate the model by comparing the projected 2060 faculty career age distributions for Natural Sciences from Fig. 3.3 with the observed career age distribution for Natural Sciences in 2020 (Fig. A.3). We find that the projected 2060 career age distributions are similar to the observed 2020 career age distribution for Natural Sciences (shown in Fig. A.3) and for the additional academic domains.

We perform a sensitivity analysis to test the robustness of our results to changes in the model's parameters. In particular, we test the robustness of our counterfactual results to different models of attrition risks and to different models of the fraction of women among new hires. First, we fit several alternative models to the empirical attrition risks (Fig. A.4) and to the fractions of women among new hires (Fig. A.5) to validate our choice of including career age up to its fifth power as a predictor in these logistic regression models. Then, we test the robustness of our counterfactual results to changes in the model's parameters by running the counterfactual analysis with the alternative model in which we only include career age up to its third power. This alternate model is less likely to overfit the data, and even tends to underfit the observed correlation structures between attrition risk and career age (Fig. A.4, Fig. A.5). Nevertheless, we find that the results are robust to these changes (Fig. A.6).

A.2 Gender diversity of hires over time

In academia overall, the fraction of women faculty among hires has been increasing on average over the past decade, at a rate of around 0.91 $pp/year$ (Fig. A.7), however, these rates of change are not uniform across academic domains. Table A.1 shows regression results for trends in women's representation among hires for 11 academic domains. While women's representation has been increasing in 6 of the 11 domains over time at rates up to 1.30 $pp/year$, the remaining 5 domains have not exhibited significant trends (Table A.1).

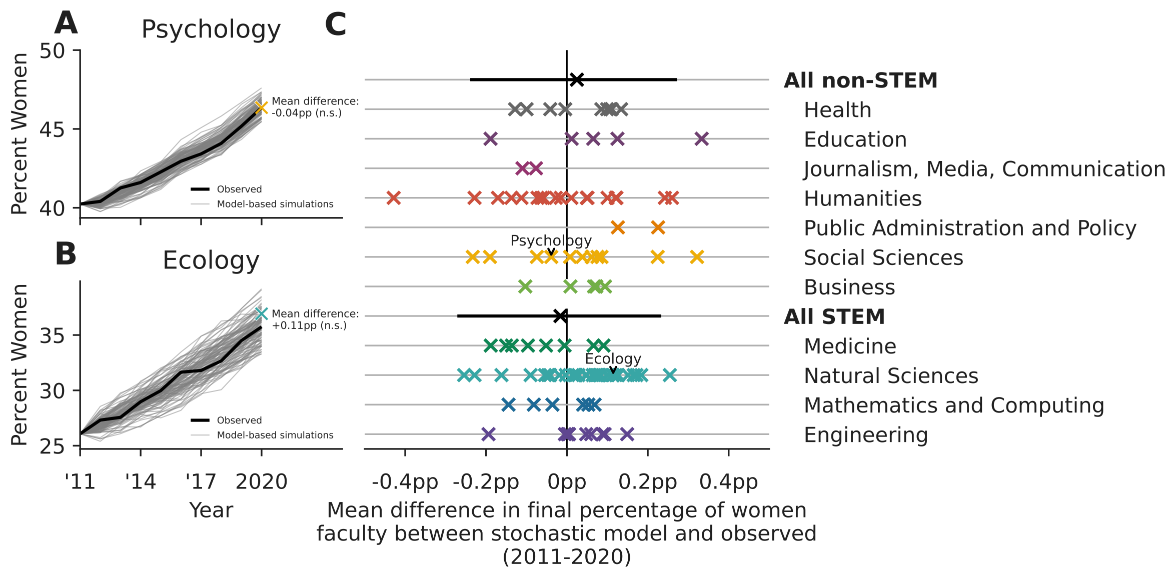

Figure A.2: Model validation: Differences between observed gender diversity outcomes and model-based outcomes. (A) The mean outcomes of model-based simulations in psychology differ from the observed outcomes by $-0.04pp$, and (B) in Ecology by $+0.11pp$, but these differences are not statistically significant. (C) Gender diversity outcomes from model-based simulations of hiring and attrition are statistically indistinguishable from observed gender diversity outcomes for all 111 fields, and for STEM and non-STEM aggregations, based on a two-sided test for significance relative to the model-based null distribution derived from simulation, $\alpha = 0.1$). Error bars for the non-STEM and STEM aggregations contain 95% of stochastic simulations. No bars are included for field-level points to preserve readability.

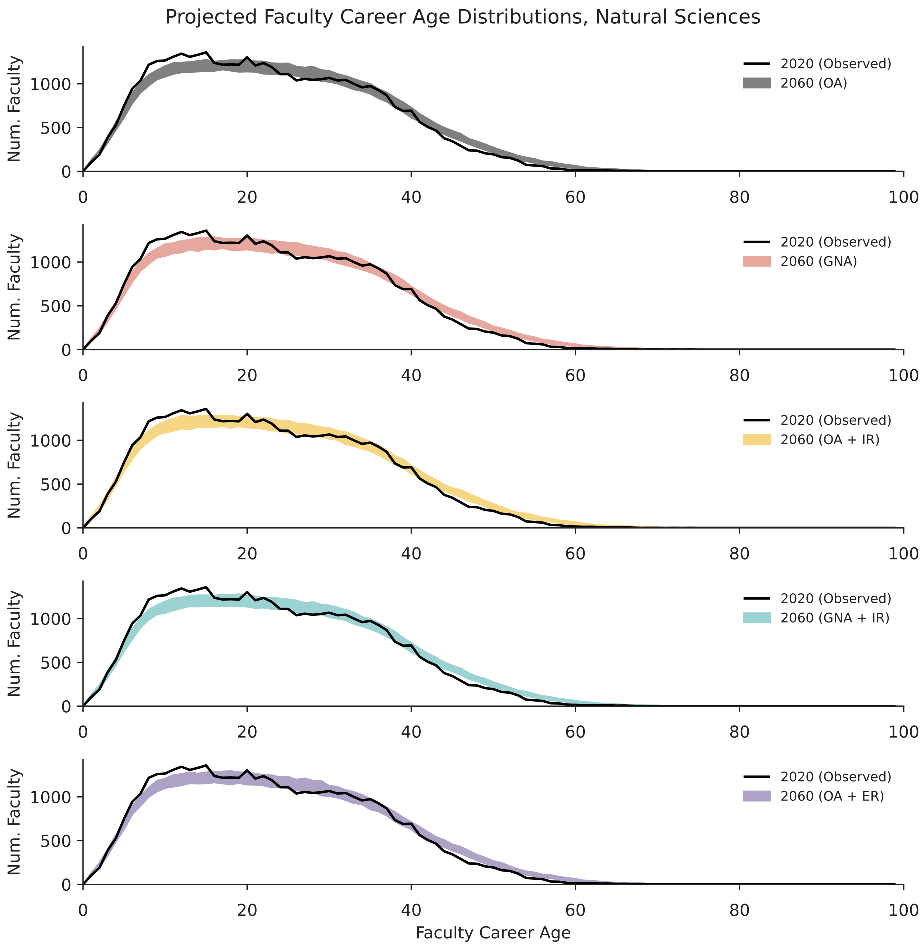

Figure A.3: Model validation: Projected 2060 faculty career age distributions for Natural Sciences from Fig. 3.3 are similar to the observed career age distribution for Natural Sciences in 2020, for each projection scenario. Line widths for the simulated scenarios span the middle 95% of simulations. OA = observed attrition, GNA = gender-neutral attrition, IR = increasing representation of women among hires (+0.5 *pp* each year), ER = equal representation of women and men among hires.

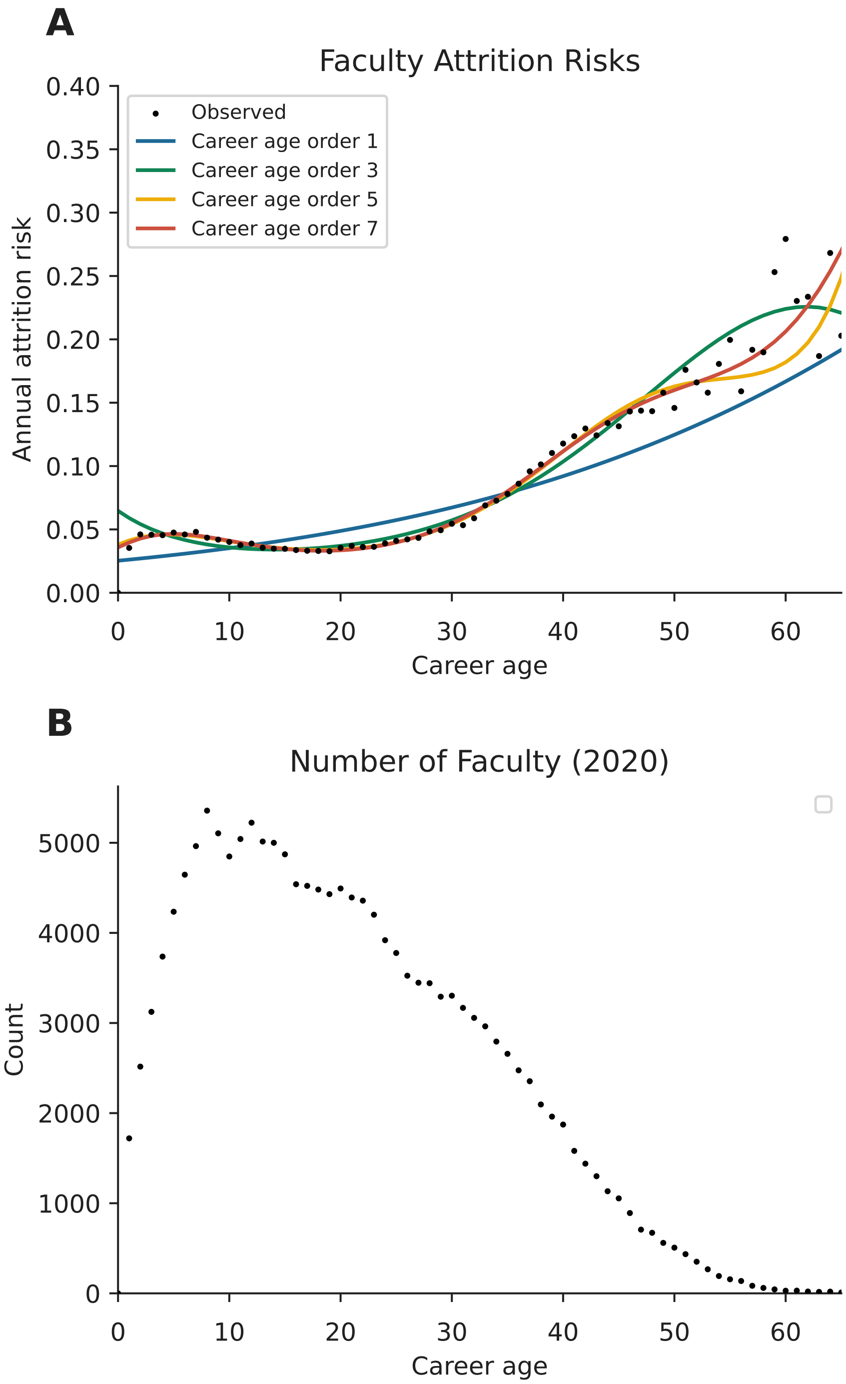

Figure A.4: Model selection. (A) Four logistic regression models fit to observed faculty attrition data. Each model includes career age up to a different power, e.g., the model labeled "Career age order 3" includes career age up to its third power: $logit(p) = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 a^3 + \beta_4 t$ where a represents career age and t represents year (see Methods Sec. 3.0.7 for details). The pattern in observed attrition risk becomes more noisy at higher career ages, because (B) there are relatively low numbers of faculty at the highest observed career ages.

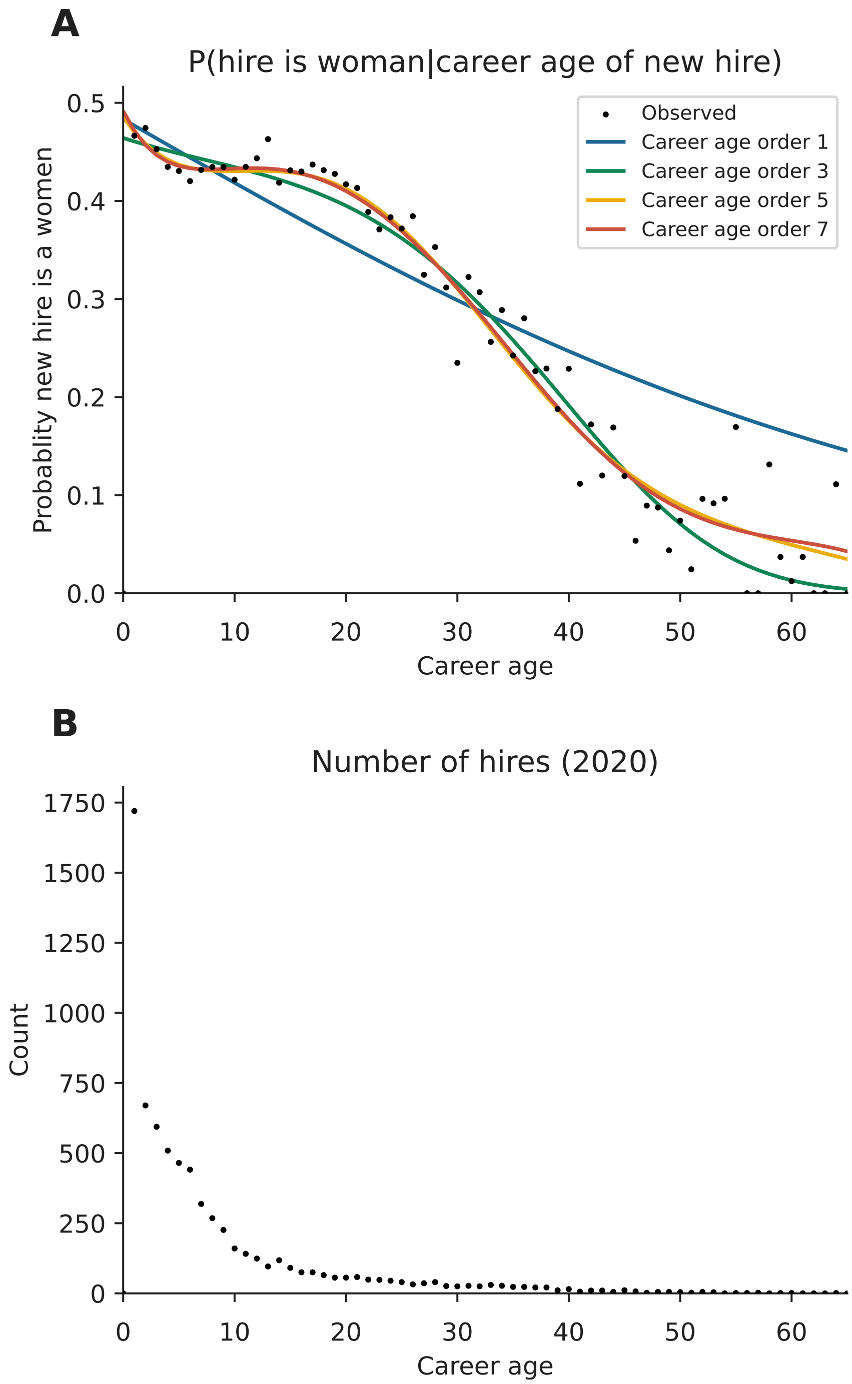

Figure A.5: Model selection. (A) Four logistic regression models fit to observed faculty hiring data, where the outcome variable is the gender of the faculty hire (1 = woman, 0 = man). Each model includes career age up to a different power, e.g., the model labeled "Career age order 3" includes career age up to its third power: $logit(p) = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 a^3 + \beta_6 t$ where a represents career age and t represents year (see Methods Sec. 3.0.7 for details). The pattern in the gender representation among new faculty hires becomes more noisy at higher career ages, because (B) there are relatively low numbers of faculty hired at higher career ages.

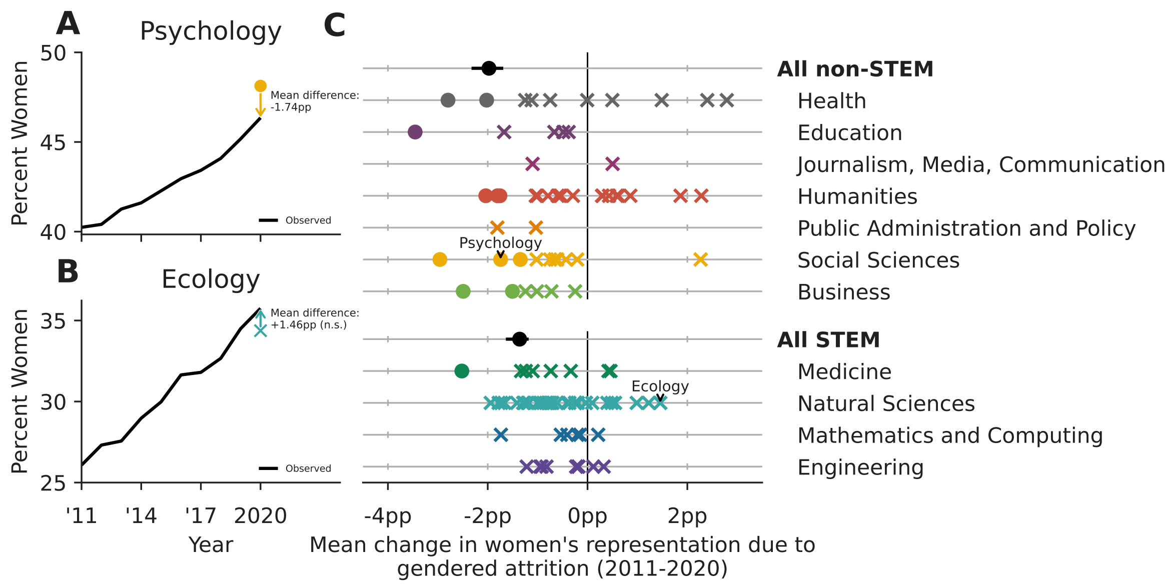

Figure A.6: Sensitivity analysis: Replicating the counterfactual analysis from results Sec. 3.0.2 using career age up to its third power in the associated logistic regressions model, instead of the fifth power (see Supplementary Sec. A.1.1 for details). Findings under this parameterization are qualitatively very similar to those presented in Fig. 3.2, indicating that the results are robust to modest changes to model parameterization.

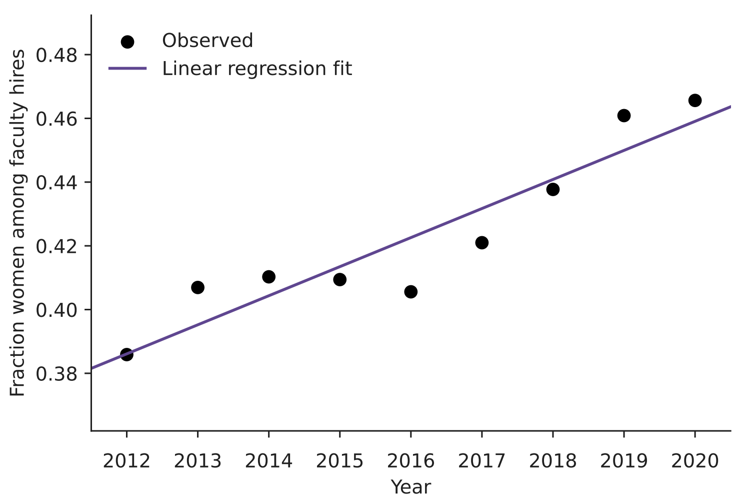

Figure A.7: Fraction of women among tenure-track faculty hires over time at U.S. PhD granting institutions. Women's share of new hires is observed to increase at around 0.91 *pp* annually (t-test, $p < 0.001$), measured by an ordinary least squares regression fit (shown in purple).

Academic Domain	Estimated Annual Change (pp)	p
Mathematics and Computing	0.25	0.095
Social Sciences	1.18**	0.006
Natural Sciences	1.30***	0.000
Engineering	0.95***	0.000
Health	0.58*	0.020
Humanities	1.18**	0.002
Public Administration and Policy	0.22	0.717
Business	0.38	0.089
Medicine	0.98**	0.009
Journalism, Media, Communication	0.49	0.173
Education	0.34	0.057
Academia Overall	0.91***	0.000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.1: Trends in women's representation among new hires from 2012 to 2020 for 11 academic domains, along with academia overall. We use linear regression to measure the expected change in women's concentration among new hires each year, and find that women's representation has been increasing in 6 of the 11 domains over time, at rates ranging from 0.58 *pp* to 1.30 *pp* per year. The remaining 5 domains have not exhibited significant linear trends. Overall, the fraction of women among hires has been increasing in academia over time (Fig. A.7). These findings are qualitatively replicated using logistic regression, so we present the linear regression results here for enhanced interpretability.

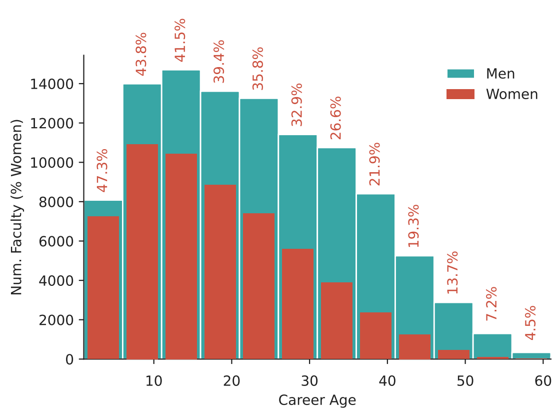

Figure A.8: Career age distribution of women (red) and men (blue) tenured and tenure-track faculty across all academic fields. Career age is measured as the number of years since earning a PhD. There are substantially more men faculty with high career ages than women faculty.

Table A.2: **Changes in Women's Representation through Hiring, Attrition, and Gendered Attrition in Academic Fields (2011-2020).** Observed changes in women's representation resulting from hiring and attrition, expressed in percentage points (*pp*), based on data from Fig. 3.1, and the estimated average change in women's representation due to gendered attrition as depicted in Figure 3.2, accompanied by the 2.5 percentile and 97.5 percentiles of simulations in parentheses. The analysis covers 111 academic fields.

Field	Observed Δ_{hiring}	Observed $\Delta_{\text{attrition}}$	Est. $\Delta_{\text{gendered attrition}}$
Health			
Environmental Health Sciences	+6.65	+3.46	+2.40 (−1.03, +5.60)
Nursing	−4.34	+0.48	+0.55 (−0.59, +1.65)
Public Health	+4.48	+0.71	−0.11 (−1.89, +1.60)
Human Development and Family Sciences	+3.19	−0.48	−2.81 (−5.44, −0.25)
Speech and Hearing Sciences	+3.30	+2.84	+2.67 (−0.81, +6.31)
Exercise Science, Kinesiology, Rehab, Health	+6.26	−1.96	−2.05 (−3.63, −0.27)
Nutrition Sciences	+0.31	+0.12	−1.18 (−3.62, +1.37)
Communication Disorders and Sciences	+6.11	+1.32	+1.26 (−1.77, +4.99)
Health, Physical Education, Recreation	+2.40	−0.35	−0.78 (−4.34, +2.52)
Social Work	+4.73	−0.04	−1.23 (−3.21, +0.60)
Education			
Education	+1.55	+1.07	−0.28 (−2.21, +1.62)
Special Education	+3.64	−2.24	−3.41 (−6.84, −0.28)
Education Administration	+0.91	+2.34	−0.74 (−2.84, +1.51)
Counselor Education	+5.08	+0.86	−0.49 (−3.36, +2.49)
Curriculum and Instruction	+0.97	+1.26	−1.61 (−3.70, +0.22)
Journalism, Media, Communication			

Continued on next page

Table A.2 – Continued from previous page

Field	Observed Δ_{hiring}	Observed $\Delta_{\text{attrition}}$	Est. $\Delta_{\text{gendered attrition}}$
Communication	+3.43	+0.20	−1.16 (−3.01, +0.80)
Mass Communications and Media Studies	+4.30	+1.22	+0.63 (−1.61, +2.73)
Humanities			
Theological Studies	+0.51	+0.71	−0.91 (−3.07, +1.02)
Asian Languages	−0.54	+2.43	+1.78 (−1.91, +5.32)
Slavic Languages and Literatures	+3.34	+3.60	+0.93 (−3.00, +5.18)
Classics and Classical Languages	+3.33	+2.25	+0.20 (−2.14, +2.69)
French Language and Literature	+0.50	+1.23	−0.38 (−3.69, +3.30)
Germanic Languages and Literatures	+3.72	+3.23	+2.08 (−1.44, +5.76)
Theatre Literature, History and Criticism	+0.79	+2.87	+0.40 (−4.04, +4.58)
Art History and Criticism	+2.18	+2.67	+0.44 (−1.85, +2.86)
Asian Studies	−1.17	+1.04	+0.72 (−2.84, +3.90)
History	+2.84	+2.11	−0.64 (−1.69, +0.44)
Urban and Regional Planning	+5.71	+1.74	−1.01 (−4.53, +2.34)
Linguistics	−0.15	+2.19	+0.44 (−2.12, +3.29)
English Language and Literature	+2.17	+0.63	−1.80 (−2.80, −0.80)
Near and Middle Eastern Languages and Cultures	+5.80	−0.22	−0.87 (−4.87, +3.41)
Music	+3.72	−0.99	−1.10 (−2.54, +0.47)
Philosophy	+5.03	+0.04	−1.84 (−3.37, −0.24)
Religious Studies	+2.68	−0.22	−1.97 (−4.28, +0.08)
Comparative Literature	+1.78	+1.17	−0.58 (−4.04, +2.40)
Spanish Language and Literature	+2.39	+1.94	−0.55 (−3.09, +1.98)
Architecture	+4.53	+2.64	+0.77 (−2.31, +3.72)
Public Administration and Policy			
Public Policy	+5.16	−0.88	−1.88 (−3.85, +0.23)
Public Administration	+6.49	+0.82	−1.11 (−3.71, +1.62)
Social Sciences			

Continued on next page

Table A.2 – Continued from previous page

Field	Observed Δ_{hiring}	Observed $\Delta_{\text{attrition}}$	Est. $\Delta_{\text{gendered attrition}}$
Sociology	+3.99	+1.51	−1.43 (−2.92, +0.03)
Gender Studies	−2.71	+1.40	+2.28 (−0.56, +5.25)
Anthropology	+3.68	+2.30	−0.42 (−1.94, +1.31)
Political Science	+4.35	+1.00	−0.15 (−1.41, +0.98)
International Affairs	+6.99	−1.58	−2.74 (−5.31, −0.33)
Geography	+6.65	+0.80	−0.51 (−2.26, +1.82)
Psychology	+5.81	−0.04	−1.83 (−2.82, −0.74)
Agricultural Economics	+6.46	+0.43	−0.87 (−4.10, +1.76)
Educational Psychology	+3.51	−0.73	−0.84 (−3.75, +2.13)
Economics	+2.93	+0.45	−0.61 (−1.66, +0.50)
Criminal Justice and Criminology	+6.26	+1.12	−0.63 (−3.47, +2.28)
Business			
Accounting	+3.02	+0.47	−1.35 (−3.37, +0.61)
Marketing	+4.43	+0.06	−1.09 (−3.11, +0.99)
Management Information Systems	+3.61	−2.50	−2.44 (−4.61, −0.10)
Finance	+3.13	−0.01	−0.64 (−2.26, +0.88)
Business Administration	+4.21	−0.34	−0.23 (−1.98, +1.67)
Management	+2.76	−0.20	−1.53 (−2.94, +0.17)
Medicine			
Genetics	+4.31	+1.04	+0.25 (−2.77, +3.00)
Pharmaceutical Sciences	+6.88	−0.34	−1.42 (−3.78, +0.96)
Epidemiology	+3.95	+0.29	−0.67 (−3.04, +1.59)
Pharmacology	+2.91	+0.61	−0.38 (−2.19, +1.36)
Pharmacy	+10.25	−1.54	−1.36 (−4.17, +1.52)
Physiology	+5.21	+0.26	−1.19 (−3.05, +0.72)
Veterinary Medical Sciences	+10.34	−1.92	−2.52 (−4.31, −0.74)
Immunology	+3.80	+2.23	+0.41 (−1.55, +2.35)

Continued on next page

Table A.2 – Continued from previous page

Field	Observed Δ_{hiring}	Observed $\Delta_{\text{attrition}}$	Est. $\Delta_{\text{gendered attrition}}$
Natural Sciences			
Entomology	+6.49	+0.85	−1.42 (−4.55, +1.21)
Soil Science	+4.04	+2.10	+0.52 (−2.46, +3.38)
Anatomy	+6.05	−0.82	−1.61 (−4.11, +0.84)
Natural Resources	+4.70	+1.64	+0.09 (−2.23, +2.49)
Plant Sciences	+4.74	+2.19	+1.01 (−1.61, +3.68)
Plant Pathology	+4.88	+1.51	−1.70 (−4.92, +1.69)
Biophysics	+4.00	−0.41	−0.75 (−3.31, +1.71)
Food Science	+4.16	+0.51	−1.08 (−4.07, +2.03)
Pathology	+5.25	−2.58	−1.50 (−3.14, +0.04)
Horticulture	+2.60	−0.12	−1.66 (−5.07, +1.69)
Biostatistics	+3.58	−0.69	−0.68 (−3.49, +2.02)
Agronomy	+4.82	+1.18	−0.93 (−3.96, +1.68)
Animal Sciences	+7.40	+1.74	−0.51 (−2.68, +1.63)
Forestry and Forest Resources	+6.54	+0.37	−1.82 (−4.83, +0.71)
Geology	+5.93	+1.54	−0.38 (−1.85, +1.03)
Biological Sciences	+5.22	+1.69	−0.04 (−1.07, +0.92)
Physics	+2.32	+1.00	−0.25 (−1.00, +0.50)
Chemistry	+3.96	+0.40	−0.87 (−1.77, +0.02)
Biochemistry	+3.91	+0.25	−0.89 (−2.22, +0.40)
Chemical Engineering	+4.08	+0.21	−0.87 (−2.38, +0.63)
Environmental Sciences	+5.88	+1.56	−0.64 (−2.62, +1.16)
Atmospheric Sciences and Meteorology	+4.59	+2.42	+1.30 (−1.00, +3.30)
Biomedical Engineering	+4.29	+2.08	+0.44 (−1.38, +2.40)
Microbiology	+5.28	+0.32	−0.99 (−2.51, +0.59)
Cell Biology	+4.82	+0.33	−0.48 (−2.06, +1.11)
Marine Sciences	+5.20	+1.95	−0.08 (−2.85, +2.44)

Continued on next page

Table A.2 – Continued from previous page

Field	Observed Δ_{hiring}	Observed $\Delta_{\text{attrition}}$	Est. $\Delta_{\text{gendered attrition}}$
Astronomy	+2.89	+1.52	+0.08 (−1.22, +1.35)
Evolutionary Biology	+6.79	+1.88	+0.54 (−2.10, +3.21)
Ecology	+6.23	+3.54	+1.42 (−0.88, +3.92)
Neuroscience	+4.19	−0.07	−0.75 (−2.63, +1.02)
Molecular Biology	+3.40	+0.83	−0.16 (−1.83, +1.53)
Mathematics and Computing			
Statistics	+2.89	+1.64	+0.24 (−1.43, +1.72)
Mathematics	+2.86	+0.65	−0.44 (−1.22, +0.41)
Computer Engineering	+2.66	+0.23	−0.18 (−1.08, +0.79)
Computer Science	+2.26	−0.09	−0.55 (−1.51, +0.32)
Information Technology	+0.88	−1.17	−0.29 (−2.68, +2.01)
Information Science	+0.81	−1.00	−1.70 (−4.20, +1.05)
Engineering			
Mechanical Engineering	+3.39	+0.60	−0.18 (−1.09, +0.75)
Systems Engineering	+3.29	+0.48	−1.29 (−3.85, +1.15)
Aerospace Engineering	+2.53	+1.18	+0.33 (−1.27, +1.93)
Electrical Engineering	+2.18	+0.46	−0.22 (−1.11, +0.66)
Agricultural Engineering	+4.74	+1.22	−0.16 (−1.96, +1.44)
Operations Research	+2.53	−0.39	−0.99 (−3.44, +1.32)
Environmental Engineering	+4.94	+0.91	−0.91 (−2.30, +0.52)
Civil Engineering	+4.30	+1.54	−0.25 (−1.51, +1.07)
Materials Engineering	+4.68	+0.46	−0.85 (−2.76, +0.80)
Industrial Engineering	+1.86	+1.66	+0.14 (−2.25, +2.54)

Table A.3: **Number of faculty by field and gender, 2020.** Estimated counts of women and men faculty based on 2020 faculty rosters and name-based gender inference [144].

Field	Women	Men	Pct. Women
Health			
Environmental Health Sciences	285	430	39.9
Nursing	3531	515	87.3
Public Health	1813	1555	53.8
Human Development and Family Sciences	765	486	61.2
Speech and Hearing Sciences	352	184	65.7
Exercise Science, Kinesiology, Rehab, Health	1612	1555	50.9
Nutrition Sciences	722	604	54.4
Communication Disorders and Sciences	450	215	67.7
Health, Physical Education, Recreation	356	427	45.5
Social Work	1308	696	65.3
Education			
Education	1301	857	60.3
Special Education	452	277	62.0
Education Administration	986	840	54.0
Counselor Education	566	405	58.3
Curriculum and Instruction	1204	654	64.8
Journalism, Media, Communication			
Communication	1054	1208	46.6
Mass Communications and Media Studies	947	1093	46.4
Humanities			

Table A.3 – Continued from previous page

Field	Women	Men	Pct. Women
Theological Studies	324	958	25.3
Asian Languages	189	257	42.4
Slavic Languages and Literatures	198	186	51.6
Classics and Classical Languages	513	596	46.3
French Language and Literature	291	253	53.5
Germanic Languages and Literatures	244	264	48.0
Theatre Literature, History and Criticism	574	615	48.3
Art History and Criticism	1006	912	52.5
Asian Studies	237	344	40.8
History	2071	3008	40.8
Urban and Regional Planning	335	506	39.8
Linguistics	404	467	46.4
English Language and Literature	2968	2918	50.4
Near & Middle Eastern Langs. & Cultures	160	268	37.4
Music	1239	2747	31.1
Philosophy	788	1773	30.8
Religious Studies	446	900	33.1
Comparative Literature	354	364	49.3
Spanish Language and Literature	438	409	51.7
Architecture	606	1205	33.5
Public Administration and Policy			
Public Policy	687	1146	37.5
Public Administration	446	645	40.9
Social Sciences			

Table A.3 – Continued from previous page

Field	Women	Men	Pct. Women
Sociology	1501	1483	50.3
Gender Studies	474	82	85.3
Anthropology	1305	1291	50.3
Political Science	1345	2702	33.2
International Affairs	426	851	33.4
Geography	482	933	34.1
Psychology	2826	3215	46.8
Agricultural Economics	171	532	24.3
Educational Psychology	555	463	54.5
Economics	804	3039	20.9
Criminal Justice and Criminology	466	588	44.2
Business			
Accounting	536	1186	31.1
Marketing	516	1069	32.6
Management Information Systems	231	867	21.0
Finance	377	1503	20.1
Business Administration	546	1473	27.0
Management	880	2136	29.2
Medicine			
Genetics	324	612	34.6
Pharmaceutical Sciences	444	912	32.7
Epidemiology	778	747	51.0
Pharmacology	512	1268	28.8
Pharmacy	567	639	47.0

Table A.3 – Continued from previous page

Field	Women	Men	Pct. Women
Physiology	620	1446	30.0
Veterinary Medical Sciences	956	1281	42.7
Immunology	677	1303	34.2
Natural Sciences			
Entomology	191	494	27.9
Soil Science	163	506	24.4
Anatomy	387	763	33.7
Natural Resources	340	877	27.9
Plant Sciences	250	656	27.6
Plant Pathology	166	446	27.1
Biophysics	223	689	24.5
Food Science	353	474	42.7
Pathology	1199	1868	39.1
Horticulture	117	394	22.9
Biostatistics	457	676	40.3
Agronomy	149	526	22.1
Animal Sciences	337	749	31.0
Forestry and Forest Resources	229	655	25.9
Geology	770	2098	26.8
Biological Sciences	1895	3656	34.1
Physics	800	4364	15.5
Chemistry	1024	3671	21.8
Biochemistry	1001	2884	25.8
Chemical Engineering	405	1692	19.3

Table A.3 – Continued from previous page

Field	Women	Men	Pct. Women
Environmental Sciences	624	1416	30.6
Atmospheric Sciences and Meteorology	267	757	26.1
Biomedical Engineering	426	1284	24.9
Microbiology	884	1815	32.8
Cell Biology	844	1699	33.2
Marine Sciences	279	722	27.9
Astronomy	416	1937	17.7
Evolutionary Biology	293	521	36.0
Ecology	370	662	35.9
Neuroscience	721	1443	33.3
Molecular Biology	732	1669	30.5
Mathematics and Computing			
Statistics	474	1598	22.9
Mathematics	1072	4688	18.6
Computer Engineering	584	3581	14.0
Computer Science	885	4291	17.1
Information Technology	211	759	21.8
Information Science	404	723	35.8
Engineering			
Mechanical Engineering	562	3428	14.1
Systems Engineering	152	654	18.9
Aerospace Engineering	209	1364	13.3
Electrical Engineering	613	3914	13.5
Agricultural Engineering	378	1386	21.4

Table A.3 – Continued from previous page

Field	Women	Men	Pct. Women
Operations Research	149	632	19.1
Environmental Engineering	517	1854	21.8
Civil Engineering	585	2217	20.9
Materials Engineering	340	1439	19.1
Industrial Engineering	212	838	20.2

Appendix B

Supplementary Figures and Tables for Chapter 4

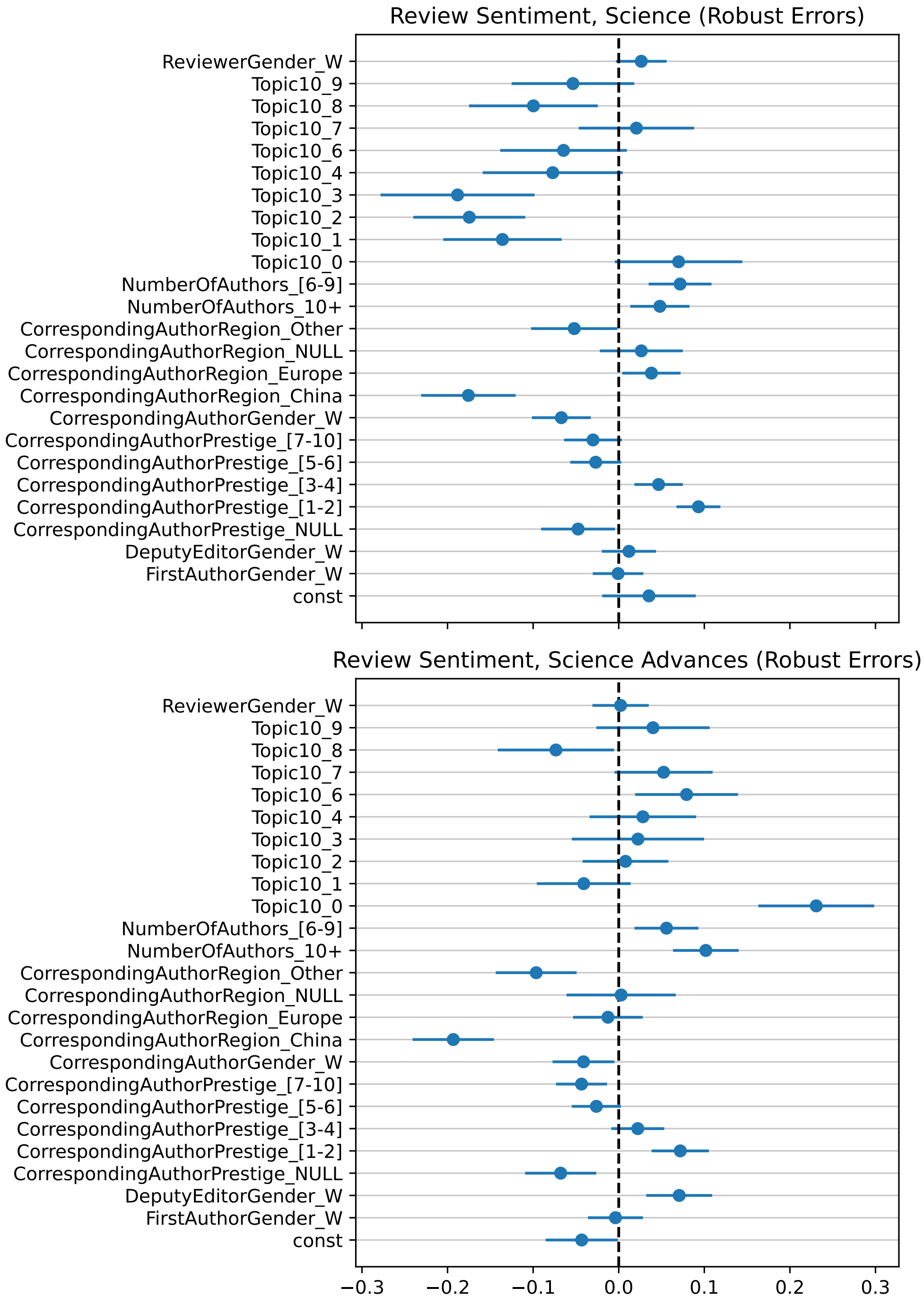

Figure B.1: **Review sentiment regression with cluster-robust standard errors.** OLS regression on review sentiment at *Science* and *Science Advances* where each record is a single review-manuscript pair, and standard errors are cluster-robust for each manuscript.

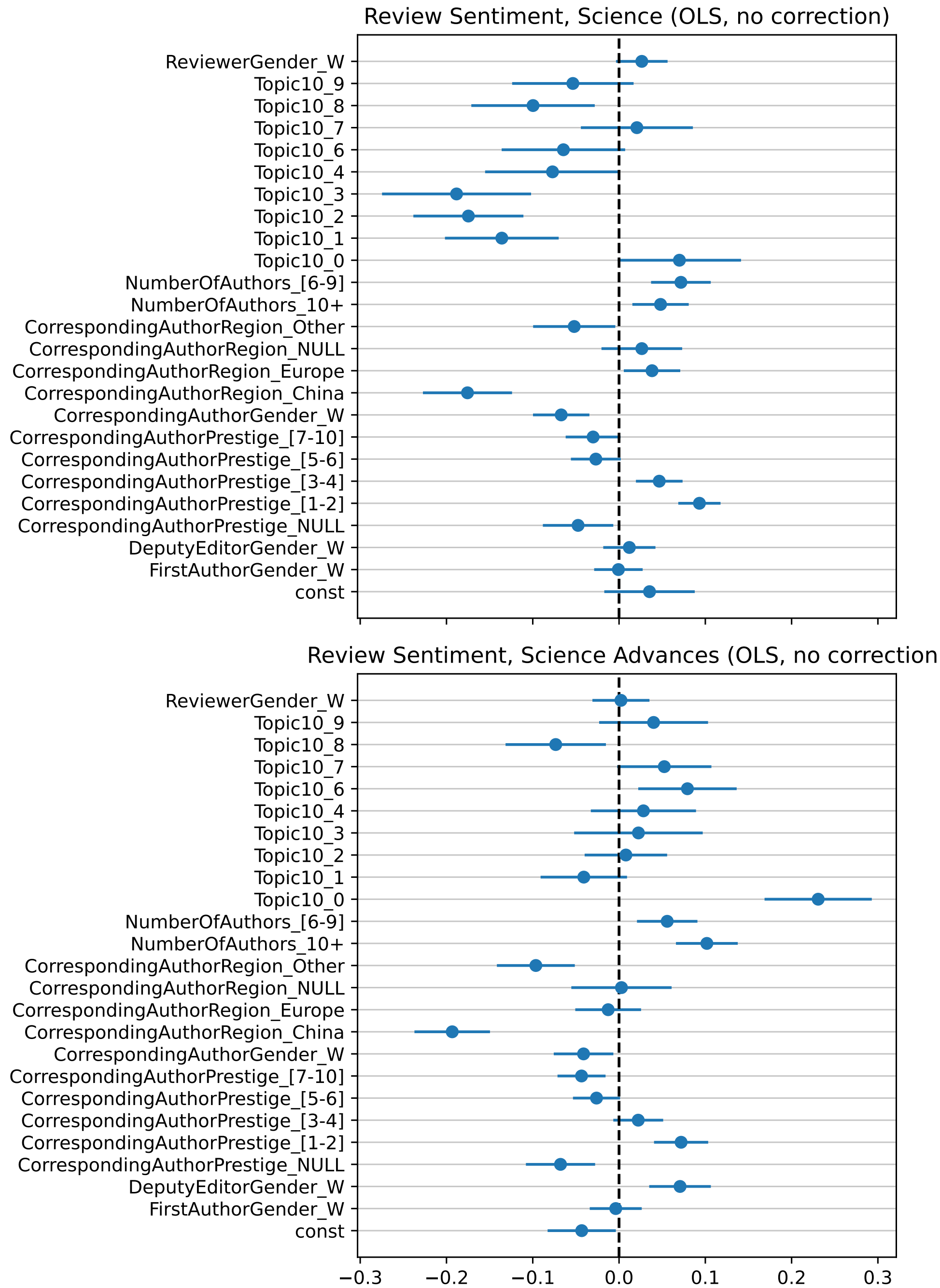

Figure B.2: **Review sentiment regression without cluster-robust standard errors.** OLS regression on review sentiment at *Science* and *Science Advances* where each record is a single review-manuscript pair. Standard errors are not adjusted.

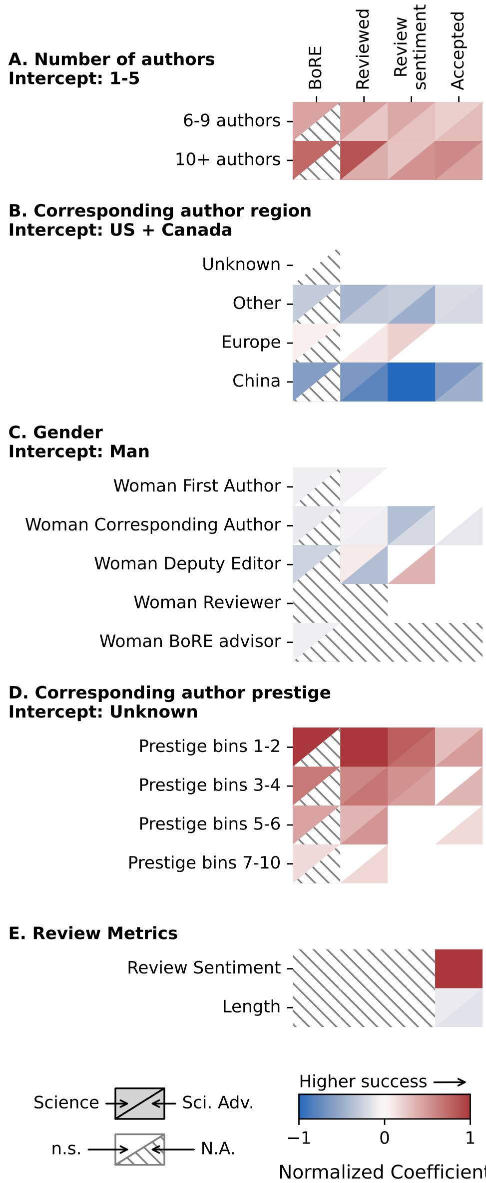

Figure B.3: Alternative regression analysis omitting manuscript topic. Overview of the effects of (A) number of authors, (B) region, (C) gender, (D) prestige, (E) review sentiment, and review length on each applicable stage of peer review for *Science* (upper triangles) and *Science Advances* (lower triangles). The colors display linear regression coefficients for the Board of Reviewing Editors (*BoRE*) and *review sentiment* stages, and logistic regression coefficients for the *reviewed* and *accepted* stages (see SI for complete description of controls and results). Positive coefficients (red hues) are associated with evaluations and success rates that are higher than the corresponding intercept (indicated in bold text), while negative coefficients (blue hues) are associated with lower evaluations and success rates. A dashed cell indicates an excluded coefficient-stage pair, and non-significant effects were colored white. Coefficients are normalized by dividing by the absolute value of the highest magnitude coefficient within each publication and stage.

Outcome: BoRE advisor rating	*Science*	
Variable	Coef (95% CI)	p
[Intercept]	5.12 (5.03, 5.20)	**0.000
Topic 1 (vs. Topic 0)	−0.14 (−0.21, −0.06)	**0.000
Topic 2	0.10 (0.04, 0.17)	**0.001
Topic 3	−0.36 (−0.45, −0.27)	**0.000
Topic 4	0.14 (0.05, 0.24)	**0.003
Topic 5	−0.10 (−0.18, −0.02)	*0.013
Topic 6	0.51 (0.43, 0.59)	**0.000
Topic 7	0.39 (0.33, 0.46)	**0.000
Topic 8	−0.71 (−0.80, −0.62)	**0.000
Topic 9	−0.02 (−0.10, 0.05)	0.546
Woman First Author (vs. man)	−0.07 (−0.10, −0.03)	**0.000
Woman Deputy Editor (vs. man)	−0.18 (−0.22, −0.15)	**0.000
6-9 authors (vs. 1-5)	0.32 (0.28, 0.36)	**0.000
10+ authors	0.54 (0.50, 0.58)	**0.000
Prestige bins 1-2 (vs. unranked)	0.73 (0.66, 0.79)	**0.000
Prestige bins 3-4	0.48 (0.42, 0.55)	**0.000
Prestige bins 5-6	0.32 (0.26, 0.39)	**0.000
Prestige bins 7-10	0.11 (0.05, 0.18)	**0.000
CA Nationality China (vs. US + Canada)	−0.45 (−0.51, −0.39)	**0.000
CA Nationality Europe	0.04 (0.00, 0.09)	*0.041
CA Nationality Unknown	0.02 (−0.05, 0.08)	0.582
CA Nationality Other	−0.22 (−0.27, −0.16)	**0.000
Woman Corresponding Author (vs. man)	−0.09 (−0.13, −0.05)	**0.000
Woman BoRE advisor (vs. man)	−0.07 (−0.10, −0.03)	**0.001
$N = 65048$; $R^2 = 0.06$; * $p < 0.05$, ** $p < 0.005$		

Table B.1: Linear regression coefficients predicting BoRE advisor rating. Only *Science* coefficients are shown because *Science Advances* does not use the BoRE system. Intercepts for categorical covariates are shown in parentheses. CA = Corresponding Author.

Outcome: Manuscript sent to review	*Science*		*Science Advances*	
Variable	Coef (95% CI)	p	Coef (95% CI)	p
[Intercept]	−2.29 (−2.40, −2.18)	**0.000	−1.16 (−1.28, −1.04)	**0.000
Topic 1 (vs. Topic 0)	0.21 (0.11, 0.30)	**0.000	−0.12 (−0.23, −0.01)	*0.028
Topic 2	−0.18 (−0.26, −0.10)	**0.000	−0.36 (−0.46, −0.26)	**0.000
Topic 3	−0.75 (−0.87, −0.63)	**0.000	−0.66 (−0.80, −0.53)	**0.000
Topic 4	−0.10 (−0.21, 0.02)	0.094	−0.06 (−0.18, 0.06)	0.322
Topic 5	−0.48 (−0.59, −0.37)	**0.000	0.00 (−0.10, 0.11)	0.935
Topic 6	−0.01 (−0.11, 0.10)	0.879	0.41 (0.29, 0.53)	**0.000
Topic 7	−0.02 (−0.11, 0.06)	0.599	−0.19 (−0.30, −0.07)	**0.001
Topic 8	0.10 (−0.01, 0.20)	0.079	−0.15 (−0.27, −0.04)	*0.010
Topic 9	−0.18 (−0.28, −0.08)	**0.000	−0.28 (−0.41, −0.16)	**0.000
Woman First Author (vs. man)	−0.08 (−0.13, −0.03)	**0.001	−0.02 (−0.07, 0.03)	0.398
Woman Deputy Editor (vs. man)	0.09 (0.04, 0.14)	**0.000	−0.34 (−0.40, −0.28)	**0.000
6-9 authors (vs. 1-5)	0.50 (0.45, 0.56)	**0.000	0.22 (0.16, 0.28)	**0.000
10+ authors	0.94 (0.88, 0.99)	**0.000	0.33 (0.27, 0.39)	**0.000
Prestige bins 1-2 (vs. unranked)	1.10 (1.02, 1.18)	**0.000	0.85 (0.76, 0.95)	**0.000
Prestige bins 3-4	0.65 (0.57, 0.73)	**0.000	0.58 (0.49, 0.66)	**0.000
Prestige bins 5-6	0.39 (0.30, 0.47)	**0.000	0.43 (0.35, 0.52)	**0.000
Prestige bins 7-10	0.06 (−0.02, 0.15)	0.155	0.16 (0.07, 0.24)	**0.000
CA Nationality China (vs. US + Canada)	−0.72 (−0.80, −0.64)	**0.000	−0.68 (−0.75, −0.61)	**0.000
CA Nationality Europe	0.01 (−0.04, 0.07)	0.604	0.08 (0.02, 0.15)	*0.013
CA Nationality Unknown	0.07 (−0.01, 0.15)	0.076	−0.09 (−0.19, 0.01)	0.069
CA Nationality Other	−0.49 (−0.56, −0.41)	**0.000	−0.26 (−0.33, −0.19)	**0.000
Woman Corresponding Author (vs. man)	−0.08 (−0.13, −0.02)	**0.005	−0.08 (−0.13, −0.02)	*0.008
	$N = 68047$		$N = 42256$	
	pseudo $R^2 = 0.08$		pseudo $R^2 = 0.04$	
			* $p < 0.05$, ** $p < 0.005$	

Table B.2: Logistic regression coefficients predicting whether a manuscript is sent to review. Intercepts for categorical covariates are shown in parentheses. CA = Corresponding Author.

Outcome: Reviewer sentiment	*Science*		*Science Advances*	
Variable	Coef (95% CI)	p	Coef (95% CI)	p
[Intercept]	0.06 (−0.02, 0.13)	0.123	0.12 (0.04, 0.20)	**0.004
Topic 1 (vs. Topic 0)	−0.21 (−0.27, −0.15)	**0.000	−0.27 (−0.34, −0.20)	**0.000
Topic 2	−0.24 (−0.30, −0.19)	**0.000	−0.22 (−0.29, −0.16)	**0.000
Topic 3	−0.26 (−0.34, −0.18)	**0.000	−0.21 (−0.30, −0.12)	**0.000
Topic 4	−0.15 (−0.22, −0.07)	**0.000	−0.20 (−0.28, −0.13)	**0.000
Topic 5	−0.07 (−0.14, 0.00)	0.066	−0.23 (−0.30, −0.16)	**0.000
Topic 6	−0.13 (−0.20, −0.07)	**0.000	−0.15 (−0.23, −0.08)	**0.000
Topic 7	−0.05 (−0.10, 0.01)	0.082	−0.18 (−0.25, −0.11)	**0.000
Topic 8	−0.17 (−0.24, −0.10)	**0.000	−0.30 (−0.39, −0.22)	**0.000
Topic 9	−0.12 (−0.19, −0.06)	**0.000	−0.19 (−0.27, −0.11)	**0.000
Woman First Author (vs. man)	−0.00 (−0.03, 0.03)	0.961	−0.00 (−0.04, 0.03)	0.815
Woman Deputy Editor (vs. man)	0.01 (−0.02, 0.04)	0.460	0.07 (0.03, 0.11)	**0.000
Prestige bins 1-2 (vs. unranked)	0.14 (0.09, 0.20)	**0.000	0.14 (0.08, 0.20)	**0.000
Prestige bins 3-4	0.09 (0.04, 0.15)	**0.001	0.09 (0.03, 0.15)	**0.002
Prestige bins 5-6	0.02 (−0.04, 0.08)	0.475	0.04 (−0.01, 0.10)	0.131
Prestige bins 7-10	0.02 (−0.04, 0.08)	0.564	0.02 (−0.03, 0.08)	0.381
Woman Corresponding Author (vs. man)	−0.07 (−0.10, −0.03)	**0.000	−0.04 (−0.08, −0.00)	*0.026
CA Nationality China (vs. US + Canada)	−0.18 (−0.23, −0.12)	**0.000	−0.19 (−0.24, −0.15)	**0.000
CA Nationality Europe	0.04 (0.00, 0.07)	*0.028	−0.01 (−0.05, 0.03)	0.543
CA Nationality Unknown	0.03 (−0.02, 0.07)	0.286	0.00 (−0.06, 0.07)	0.931
CA Nationality Other	−0.05 (−0.10, −0.00)	*0.043	−0.10 (−0.14, −0.05)	**0.000
10+ authors	0.05 (0.01, 0.08)	*0.006	0.10 (0.06, 0.14)	**0.000
6-9 authors (vs. 1-5)	0.07 (0.03, 0.11)	**0.000	0.06 (0.02, 0.09)	**0.003
Woman Reviewer (vs. man)	0.03 (−0.00, 0.06)	0.080	0.00 (−0.03, 0.04)	0.896
	$N = 23532$		$N = 20965$	
	$R^2 = 0.02$		$R^2 = 0.02$	
* $p < 0.05$, ** $p < 0.005$				

Table B.3: Linear regression coefficients predicting reviewer sentiment. Intercepts for categorical covariates are shown in parentheses. CA = Corresponding Author.

Outcome: Manuscript accepted	*Science*		*Science Advances*	
Variable	Coef (95% CI)	p	Coef (95% CI)	p
[Intercept]	−0.90 (−1.13, −0.67)	**0.000	−0.58 (−0.82, −0.34)	**0.000
Review Sentiment	0.82 (0.77, 0.86)	**0.000	0.95 (0.90, 1.00)	**0.000
Length	−0.09 (−0.13, −0.05)	**0.000	−0.14 (−0.19, −0.10)	**0.000
Topic 1 (vs. Topic 0)	0.43 (0.26, 0.61)	**0.000	0.22 (0.03, 0.42)	*0.022
Topic 2	−0.31 (−0.47, −0.15)	**0.000	0.25 (0.06, 0.43)	*0.009
Topic 3	−0.39 (−0.66, −0.13)	**0.004	−0.43 (−0.69, −0.17)	**0.001
Topic 4	−0.10 (−0.33, 0.13)	0.375	−0.14 (−0.36, 0.07)	0.189
Topic 5	0.31 (0.09, 0.53)	*0.005	0.14 (−0.04, 0.32)	0.137
Topic 6	0.61 (0.41, 0.80)	**0.000	−0.29 (−0.50, −0.08)	*0.006
Topic 7	0.52 (0.35, 0.68)	**0.000	0.14 (−0.06, 0.34)	0.177
Topic 8	0.36 (0.15, 0.56)	**0.001	−0.16 (−0.37, 0.06)	0.151
Topic 9	−0.08 (−0.28, 0.11)	0.416	−0.11 (−0.33, 0.12)	0.344
Woman First Author (vs. man)	−0.08 (−0.17, 0.01)	0.075	0.00 (−0.09, 0.09)	0.999
Woman Deputy Editor (vs. man)	0.01 (−0.09, 0.10)	0.902	−0.04 (−0.15, 0.06)	0.428
10+ authors	0.47 (0.37, 0.57)	**0.000	0.42 (0.32, 0.53)	**0.000
6-9 authors (vs. 1-5)	0.18 (0.07, 0.29)	**0.001	0.30 (0.20, 0.41)	**0.000
Prestige bins 1-2 (vs. unranked)	0.25 (0.09, 0.41)	**0.002	0.46 (0.29, 0.63)	**0.000
Prestige bins 3-4	0.16 (−0.01, 0.33)	0.062	0.33 (0.17, 0.50)	**0.000
Prestige bins 5-6	−0.03 (−0.20, 0.14)	0.762	0.16 (0.00, 0.31)	*0.048
Prestige bins 7-10	−0.10 (−0.28, 0.07)	0.253	0.00 (−0.16, 0.16)	0.993
CA Nationality China (vs. US + Canada)	−0.52 (−0.69, −0.35)	**0.000	−0.46 (−0.59, −0.33)	**0.000
CA Nationality Europe	−0.01 (−0.11, 0.10)	0.913	−0.03 (−0.14, 0.08)	0.574
CA Nationality Unknown	0.02 (−0.12, 0.16)	0.778	−0.13 (−0.31, 0.04)	0.123
CA Nationality Other	−0.16 (−0.31, −0.01)	*0.035	−0.21 (−0.34, −0.07)	**0.002
Woman Corresponding Author (vs. man)	−0.10 (−0.20, 0.01)	0.066	−0.13 (−0.23, −0.03)	*0.013
Woman Reviewer (vs. man)	−0.06 (−0.13, 0.01)	0.092	−0.07 (−0.14, 0.01)	0.075
Reviewer prestige 1-2 (vs. unranked)	−0.19 (−0.30, −0.08)	**0.001	−0.01 (−0.13, 0.11)	0.879
Reviewer prestige 3-4	−0.20 (−0.31, −0.08)	**0.001	−0.08 (−0.20, 0.04)	0.203
Reviewer prestige 5-6	−0.16 (−0.27, −0.04)	*0.007	−0.03 (−0.16, 0.09)	0.572
Reviewer prestige 7-10	−0.08 (−0.20, 0.04)	0.168	0.00 (−0.12, 0.12)	0.990
	$N = 23532$		$N = 20965$	
	pseudo $R^2 = 0.09$		pseudo $R^2 = 0.10$	
				* $p < 0.05$, ** $p < 0.005$

Table B.4: Logistic regression coefficients predicting whether a manuscript is accepted after being sent to review. Intercepts for categorical covariates are shown in parentheses. CA = Corresponding Author.