Self-image Bias and Lost Talent

Marciano Siniscalchi
Northwestern University

Pietro Veronesi
University of Chicago, NBER, and CEPR

June 28, 2022

Abstract

We propose an overlapping-generation model wherein researchers belong to two groups, M or F, and established researchers evaluate new researchers. Group imbalance obtains even with group-neutral evaluations and identical productivity distributions. Evaluators' self-image bias and mild between-group heterogeneity in equally productive research characteristics lead the initially dominant group, say M, to promote scholars similar to them. Promoted F-researchers are few and similar to M-researchers, perpetuating imbalance. Consistently with the data, our mechanism also predicts stronger and widening group imbalance in top institutions; higher quality of accepted F-researchers; clustering of M- and F-researchers across different fields; greater imbalance for seniors than juniors; less credit for F-researchers in co-authored work; and established researchers' false perception that increasing F-representation reduces quality. Policy-wise, mentorship reduces group imbalance, but increases F-group talent loss. Affirmative action reduces both.

Keywords: gender discrimination, self-image bias, affirmative action

JEL codes: A11, J16, J7

[*] Veronesi acknowledges financial support from the Fama-Miller Center for Research in Finance and by the Center for Research in Security Prices at the University of Chicago Booth School of Business. For useful comments, we thank Gadi Barlevi, Marco Bassetto, Alberto Bisin, Shereen Chaudhry, Hui Chen, Alexia Delfino, James Dow, Nicola Gennaioli, Lars P. Hansen, Christopher Hennessy, Alex Imas, Jessica Jaffers, Seema Jayachandran, Emir Kamenica, Elisabeth Kempf, Nadia Malenko, Lubos Pastor, Jane Risen, Antoinette Schoar, Oleg Urminsky, Laura Veldkamp, Adrien Verdhelan, participants at Behavioral Science workshop, the Finance workshop, and the Economics workshop at the University of Chicago, and seminar participants at the Bocconi University, Chicago Fed, Hitotsubashi University, Imperial College, LBS, MIT, Northwestern University, the 2021 NBER Economics of Culture meetings, the 2021 Adam Smith conference, the 2021 WFA meetings, and the 2022 AEA meetings. Errors are our own. The views expressed in this paper are our own and do not necessarily reflect the views of our respective employers. We declare that we have no relevant or material financial interests that relate to the research described in this paper. Contact information: Marciano Siniscalchi: marciano@northwestern.edu; Pietro Veronesi: pietro.veronesi@chicagobooth.edu.

1. Introduction

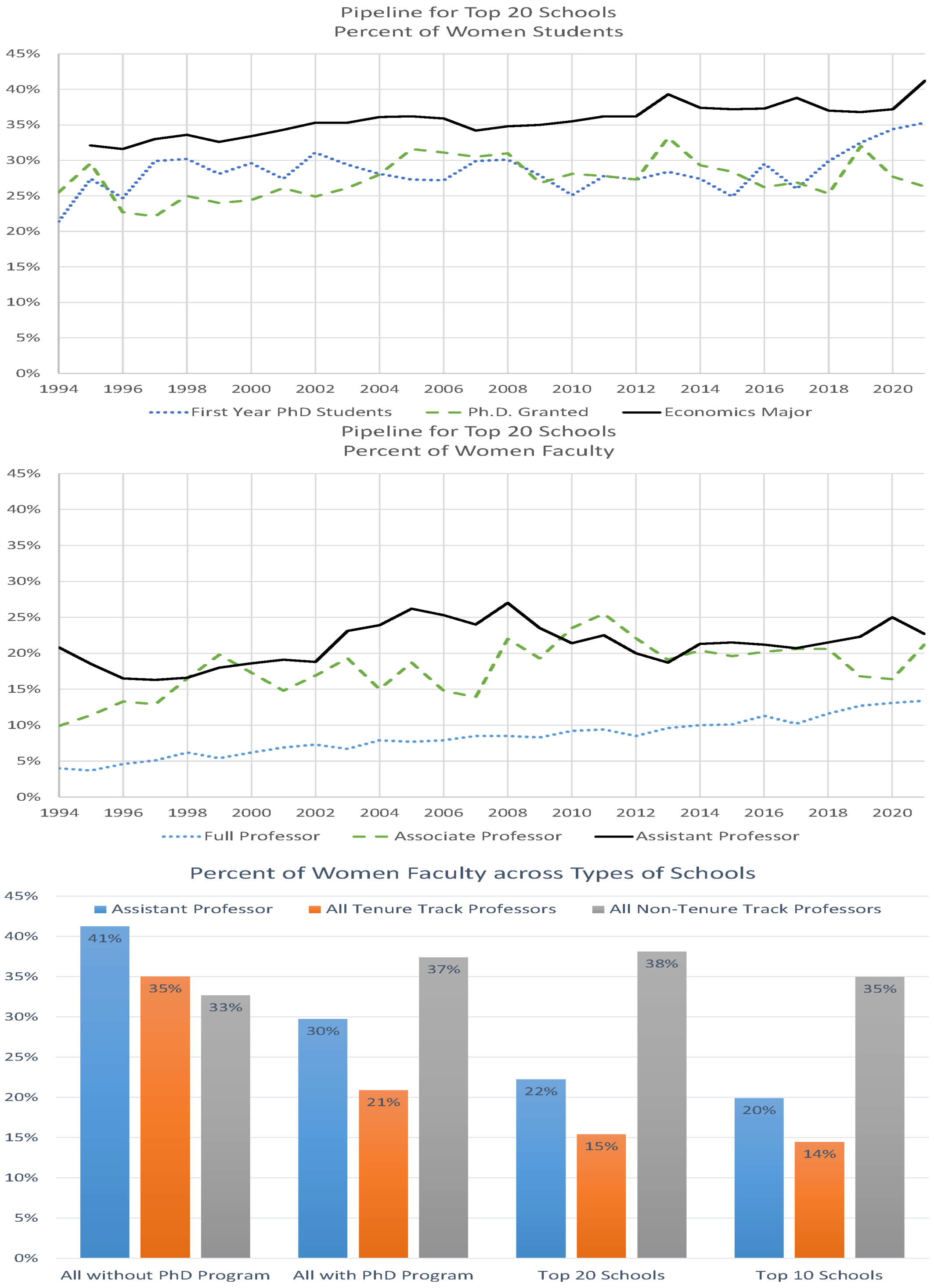
The economics profession has long been male-dominated. The Committee on the Status of Women in the Economics Profession (CSWEP), a standing committee of the AEA since 1971,[1] has been regularly documenting the progress of female economists (or lack thereof): see Chevalier (2020). This phenomenon has recently received renewed attention, possibly due to the very slow progress attained in the last 25 years. The top panel of Figure 1 shows that, while in this time span the fraction of women in undergraduate majors increased to over 40% in the top-20 schools, the fraction of women PhD students has been flat at around 30%. More troubling, perhaps, is the middle panel, which shows that, among assistant professors—i.e. the intake for the academic career—the fraction of women has been flat at around 22% since 1994. The bottom panel shows a striking difference between schools with and without a PhD program, with the latter hiring over 40% of female tenure-track faculty while the former below 30%, and with the top-10 schools only 20%. In sharp contrast, the share of women among teaching faculty is quite uniform across schools at around 37%, a difference that indicates that the bias is specific to research.

The lack of progress is puzzling given the initiatives aimed at increasing female representation in the economics profession over the past several decades. Many of these interventions are however informed by existing theories of discrimination, such as taste-based and statistical discrimination, implicit bias, and stereotyping, which we review in Section 6. From this perspective, recent empirical evidence may suggest that efforts to remove such sources of discrimination or bias have only partially succeeded. For instance, Card, DellaVigna, Funk, and Iriberri (2020) documents that acceptance rates for women-authored papers is lower conditional on quality (proxied by future citations); Sarsons (2017) and Sarsons, Gërxhani, Reuben, and Schram (2021) show that female coauthors tend to receive less credit for published papers that are joint with male coauthors; Dupas, Modestino, Niederle, and Wolfers (2021) document a bias against female presenters in economics seminars. Large differences in women representation exist across fields, however (e.g. Chari and Goldsmith-Pinkham, 2018 and Lundberg and Stearns, 2019), which would then suggest that gender-bias is more prominent in some economic fields than others.

We propose a novel theory that is consistent with the empirical evidence above but that does not depend on stereotypes or gender discrimination, whether taste-based or statistical. In our model, gender imbalance is due to the combination of self-image bias, i.e. the tendency of individuals to place more weight on their own positive attributes when judging others, and

[1] See `https://www.aeaweb.org/about-aea/committees/cswep/about`.

Figure 1: Percentage of Women in Academia



Source: CSWEP Report, 2021. The data in the bottom panel are averages over the 2017-2021 sample.

mild population heterogeneity in equally-valuable research characteristics. Both assumptions have strong empirical and experimental support, as we discuss below. Our model, which we calibrate to the data, yields several additional predictions, that are also verified in the data.

Specifically, our model features overlapping generations of agents that belong to one of two groups, labelled M and F. A new cohort of young M- and F-researchers appears in every period, in equal proportions. Each researcher is endowed with a set of characteristics. Examples of such characteristics include research approach (e.g. empirical or theoretical), methodology (e.g. structural versus reduced form), field, topic, type of questions asked, depth vs. breadth, writing style, ties to reality, policy relevance, and so on. Research characteristics are randomly distributed in the population of young researchers, with some of them slightly more common in the F-group and others symmetrically slightly more common in the M-group. As in the data, we let between-group heterogeneity be far smaller than within-group heterogeneity. Moreover, all research characteristics are equally valuable: each has the same positive effect on the likelihood of quality research (i.e., that which achieves its objectives). This implies the distribution of the likelihood of quality research in the M and F populations is the same. We emphasize that we do not make any assumptions about the *origins* of these distributional differences, which can very well be socially determined, but only that some mild differences exist, as documented in the empirical evidence discussed below.

We assume that the quality of a young researcher's output is objective and observable. However, each young researcher who has produced quality work must also be evaluated by a randomly matched member of the established population. This evaluator (hereafter, referee) decides whether or not to accept the young researcher as a member of the established population—and thus as a referee of future cohorts. Each referee's perceptions of young researchers' output reflect self-image bias (Lewicki, 1983): evaluators use their own characteristics as yardstick to assess others' research. Importantly, the referees' evaluation is group-neutral: each given referee uses the same set of research characteristics to assess young M and F researchers. If the referee's evaluation is positive, the latter becomes a recognized, permanent member of the population; otherwise, he or she leaves the model.

Our key finding is that, when research is evaluated on a large number of characteristics, the combination of self-image bias and even mild between-group heterogeneity generates a persistent bias that favors young researchers who belong to the group that is initially larger, say the M-group. Moreover, there is no convergence. While researchers from the F-group are also successful, not only are they a minority: they are endogenously selected to be the ones whose research characteristics are closer to the ones that are more prevalent among M-researchers; this perpetuates the bias forward. Intuitively, it is as if the initially larger

M-group decided for society which characteristics are important and worthy of reward, and which are not, despite the fact that all research characteristics are equally conducive to quality research. Thus, valuable characteristics that are (mildly) more common among the F-group, but also very common in the M-group, are vastly underrepresented in the steady state. This implies a persistent loss of talent and knowledge, and a sub-optimal steady state.

The same basic logic delivers a number of additional predictions that we did not originally set out to obtain, but that are verified in the data:

A higher bar for F-researchers. Our model features gender-blind evaluations, and yet M-researchers are more likely to meet with the approval of the profession than F-researchers who are their equal in terms of objective quality. In this sense, the "bar" for F-researchers is higher, consistently with the evidence in Card et al. (2020) that women-authored papers are accepted less frequently conditional on quality (proxied by future citations).[2]

Lower F-representation at more research-intensive institutions. Our model also predicts that institutions with higher research intensity, measured by their ex-post publication record, correlate with lower percentage of F-researchers. That is, top institutions have lower F-representation; this is consistent with the bottom panel of Figure 1.

Widening gap in F vs. M representation. In addition, our model predicts that the gap in F vs. M representation between top institutions and all institution should widen over time. The data confirms this prediction. For instance, in the 1970s top institution employed about the same fraction of women as assistant professors as all other institutions (9% in 1975). However, the difference has widened substantially in recent years, to 24% vs. 19% in 2001, 30% vs. 23% in 2011, and 32.6% vs. 22.7% in 2021.

Clustering into different fields. Our model predicts that M- and F- researchers tend to cluster around types whose characteristics are (mildly) more common in their own groups. It is plausible that different sets of characteristics may be more valuable in different fields. In this case, our model predicts that M- and F-researchers will be differently represented across fields, as documented e.g. by Chari and Goldsmith-Pinkham (2018) and Lundberg and Stearns (2019). The latter paper in particular also shows no variation over time in the relative percentage of women across fields, which is also a prediction of our model (convergence).

Similarity across countries. Our mechanism does not rely on specific cultural norms. Thus, it also explains why, for instance, the share of women faculty in the U.S. is roughly

[2] The evidence in Card et al. (2020) is more nuanced and we discuss it in Section A1.1. This "higher bar" for F-researcher is also evident in the empirical finding that female presenters are subject to more frequent and more hostile questioning than "equivalent" male presenters in economics seminars (Dupas et al., 2021).

similar to the one in e.g. Northern European countries (see e.g. Auriol, Friebel, Weinberger, and Wilhem, 2022 or Carlsson, Finseraas, Midtbøen, and Rafnsdóttir, 2021), despite the fact that the latter score far higher than the U.S. on other measures of gender equality.[3]

Loss in research innovation due to F under-representation. Borrowing from the literature on the "science of science" (see e.g. Carnehl and Schneider, 2021), we interpret researchers' characteristics in our model as those that enable them to solve real-world problems. F-talent loss then has a negative impact on aggregate welfare-improving research. This is reminiscent of similar conclusions of recent empirical research by e.g. Bell, Chetty, Jaravel, Petkova, and Van Reenen (2019), who suggest that "increasing exposure to innovation among women, minorities, and children from low-income families may have greater potential to spark innovation and growth than traditional approaches" (p. 647).

Mistakenly perceived trade-off between "quality and diversity." In our model, F- and M-researchers have ex-ante identical objective quality; moreover, accepted F- researchers are of higher objective quality, on average, than accepted M-researchers. Yet our model also predicts that, on average, established researchers will counterfactually believe that M-applicants are of higher quality than F-applicants. Hence, they will mistakenly perceive a trade-off between diversity and quality (First Round Review, 2022), or "merit" (Crosby, Iyer, Clayton, and Downing, 2003).

We assess different policy interventions through the lens of our model. We first investigate the impact of mentorship, and highlight an unintended consequence. We assume that young researchers are matched with random advisors from the set of established researchers. Given self-image bias, advisors advise young researchers to "become like them"—that is, acquire their advisor's type. Young researchers can do so by paying a cost that increases in the distance between their advisor's type their own. We show that, while mentorship may help reduce (but not necessarily eliminate) gender imbalance, it also accelerates the loss of F-group characteristics. Intuitively, this is because mentors are drawn from the dominant population, which over-represents M-group characteristics.

[3] For instance, *the Economist* places the Northern European countries at the top of their glass-ceiling index `https://www.economist.com/graphic-detail/glass-ceiling-index`, and so does the World Economic Forum on their Gender Gap Index `https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality`. Auriol et al. (2022) also document that European countries display a negative relation between research intensity and the share of women among faculty, and a leaky pipeline, as in U.S. While the authors show that countries with a lower gender gap index correlate with a higher fraction of women in economics, the percentages (in levels) are actually quite similar: For instance, at all faculty levels (resp. senior level), they report 27% (resp. 22%) for US and 31% (resp. 26%) for Northern European countries. In Carlsson et al. (2021), gender imbalance is contrasted with results from a survey-based study suggesting that, on average, female applicants are viewed *more* positively than male applicants with the same qualifications.

Second, we analyze the impact of affirmative-action policies. Specifically, we consider a mandate to accept the same number of F researchers as M researchers each period. Clearly, such policy mechanically leads to gender balance. However, we also find that such a policy additionally ensures that all characteristics are represented in the limit: thus, qualitatively, there is no loss of talent. Intuitively, increasing the F-group representation by mandate also increases heterogeneity in the future pool of referees, which in turn makes it more likely that research characteristics (mildly) more prevalent across F researchers will be accepted.

The Online Appendix analyzes extensions and implications. First, gender imbalance and loss of talent are exacerbated by candidates' career concerns. We endogenize the choice of young researchers to pursue an academic career, or enjoy an outside option. With costly entry, anticipating a bias against their research characteristics, the mass of F-agents who choose academia shrinks over time, and eventually converges to a smaller fraction of "applicants" than their M counterparts. If costs are sufficiently high, characteristics (mildly) more common in the F-group disappear altogether. This intuitive result can help explain why the applications of women to PhD programs in Economics are low to start with. Similar results obtain if hiring institutions bear a cost to hire a young researcher, and receive a payoff from hiring those who later become recognized members of the profession.

Second, we allow for different levels of seniority for established researchers. Senior researchers evaluate junior researchers, and both senior and junior researchers evaluate new entrants. This mimics the career dynamics in academia. Our results about the persistent bias in hiring carry through. Moreover, under suitable parameter configurations, there is a "leaky" pipeline (cf. Chevalier, 2020): senior researchers are even more biased towards characteristics prevalent in the M-group than junior researchers.

Third, while our model does not explicitly allow for co-authorships, its basic force helps explain why female coauthors tend to receive less credit for published papers that are joint with male coauthors (Sarsons, 2017; Sarsons et al., 2021). Intuitively, the referees' population mostly reflects the characteristics of the M-group and thus the positive characteristics of joint research are mostly ascribed to those of the M coauthor.

Our results depend on two main assumptions: mild heterogeneity in research characteristics between M researchers and F researcher, and self-image bias, i.e. the tendency of reviewers to use their own research style to judge the importance and worth of others' research output. Both assumptions are grounded in the empirical and experimental literature.

First, there is a considerable body of research studying gender differences in personality traits, preferences, and attitudes. Regarding personality traits, Hyde and Linn (2006) reviews the literature and concludes that medium-sized effects are found for aggression (Cohen's d between 0.40 and 0.60) and activity level in the classroom ($d = 0.49$)[4]. Similarly, Hyde (2014) reports the following d statistics of gender differences in the "big-5 personality traits," earlier studied by Costa, Terracciano, and McCrae (2001): among US subjects, there are small-to-moderate differences in neuroticism ($d = -0.40$), extraversion ($d = -0.21$), openness ($d = 0.30$) and agreeableness (-0.31), but a trivial difference in conscientiousness ($d = -0.05$). Within economics, Croson and Gneezy (2009) provide a review of the experimental literature and find "robust differences in risk preferences, social (other-regarding) preferences, and competitive preferences." Borghans, Golsteyn, Heckman, and Meijers (2009) also find differences in risk aversion, but less so on ambiguity aversion. Dittrich and Leipold (2014) find that women tend to be more patient than men, and Dreber and Johannesson (2008) that males are more likely to lie in order to secure a monetary gain; see also Betz, O'Connell, and Shepard (1989). Goldin (2014) discusses the higher gender pay gap in professions where "working long hours" is rewarded, and suggests a (possibly socially determined) preference for flexible work hours on the part of women. Finally, Andre and Falk (2021) survey nearly 10,000 economists' opinion about the current state and preferred direction of economic research. They find that female scholars are significantly more likely to emphasize multidisciplinarity, disruptive research, and policy relevance (cf. Table 3.)

As mentioned, we do not need to take a stand on the *origins* of these (small) distributional differences. Indeed, the evidence suggests that many of the traits for which a gender difference exists may be socially determined—they are the result of cultural attitudes and gender stereotyping. Guiso, Monte, Sapienza, and Zingales (2008) argue that gender differences in math scores across countries, as measured by the PISA assessment, are largely explained by broad measures of gender equality in those countries. Falk, Becker, Dohmen, Enke, Huffman, and Sunde (2018) document variation in preference traits across 76 countries and find that women are more risk-averse than men in most countries; however, for trust and patience, the correlation with gender is only significant for a subset of countries. This suggests that cultural factors may partly account for gender differences in preference traits. Andersen, Ertac, Gneezy, List, and Maximiano (2013) provide experimental evidence indicating that the gender gap in competitiveness does not arise in a matriarchal society.

The second important assumption of our model is researchers' self-image bias. The psychological literature on self-image bias (Lewicki, 1983) suggests that, when evaluating others, individuals tend to place more weight on positive attributes that they themselves possess (or believe they possess). Hill, Smith, and Hoffman (1988) show that this is true in

[4] Cohen (2013)'s d measures the standardized mean difference between two populations. $d \approx 0.2$ is considered "small" and $d \approx 0.5$ is considered "medium."

particular when subjects are asked to select a partner in a competitive game. Dunning, Perie, and Story (1991) argue that a similar principle is at work when judging social categories by means of prototypes (e.g., what makes a good economist?): "people may expect the 'ideal instantiation' of a desirable social category to resemble the self in its strengths and idiosyncracies" (p. 958). Story and Dunning (1998) document a "rational" source for self-image bias and self-serving prototypes: in their experiment, "those who received success feedback came to perceive a stronger relationship between 'what they had' and 'what it takes to succeed' than did those who received failure feedback" (p. 513). Translated to our environment, established researchers view their personal success in research as evidence that their own research characteristics are the right ones to produce quality research that, in addition, is valuable to society. Hence, they use the same characteristics to evaluate the research of others.

Our assumption that referees accepts young researchers who are similar to them can also be due to referees' preferences (e.g. theorists like theorists, and empiricists like empiricists). However, this interpretation must be subject to two caveats. First, referees' preferences do *not* take group membership into account; thus, even this "homophily" interpretation of our model differs from Becker's taste-based theory of discrimination. Moreover, in this interpretation, referees do not value heterogeneity (e.g., theorists derive no benefit from interacting with empiricists, and conversely), nor the candidate's objective productivity. That is, they completely disregard the benefits that would accrue to a department—or, in fact, from the profession as a whole—from advancing a productive young researcher who however does not share their own characteristics. This strikes us as extreme.

2. The Basic Model

We consider an overlapping-generations model in which unit masses of two groups of young researchers, the M-group and F-group, appear at discrete times $t = 1, 2, \ldots$. Each researcher $i \in M \cup F$ is endowed with a *type* drawn from a set Θ, and distributed heterogeneously across M and F researchers. While systematic, these distributional differences may well be small. Research output fully reflects the researcher's type; in fact, we assume that the characteristics of a paper written by a researcher of type θ *are* θ itself.

We adopt a simple symmetric environment in which each type corresponds to a vector of N characteristics which can only take two values, 0 and 1: that is, $\Theta \equiv \{0, 1\}^N$. (See the Online Appendix for a more general case.) For each agent i of type $\theta^i \in \Theta$, θ^i_n denotes the value of the n-th characteristic. The number N of characteristics is even, characteristics are

Figure 2: Symmetric Distribution of Research Characteristics

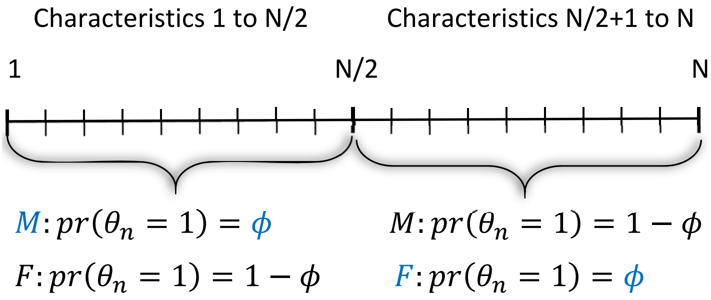

Characteristics 1 to N/2
Characteristics N/2+1 to N
1
N/2
N
$M: pr(\theta_n = 1) = \phi$
$F: pr(\theta_n = 1) = 1 - \phi$
$M: pr(\theta_n = 1) = 1 - \phi$
$F: pr(\theta_n = 1) = \phi$

Table 1: Type Frequencies in a Simple Example

	p_m^θ	p_f^θ
$(0,0)$	$0.2 \times 0.8 = 0.16$	$0.8 \times 0.2 = 0.16$
$\theta^m = (1,0)$	$0.8 \times 0.8 = 0.64$	$0.2 \times 0.2 = 0.04$
$\theta^f = (0,1)$	$0.2 \times 0.2 = 0.04$	$0.8 \times 0.8 = 0.64$
$\theta^* = (1,1)$	$0.8 \times 0.2 = 0.16$	$0.2 \times 0.8 = 0.16$

mutually independently distributed, and their distributions depend on a single parameter $\phi > 0.5$. Our main assumption, illustrated in Figure 2, is that characteristics are distributed symmetrically in the M and F population, in the sense that for $n = 1, \ldots, \frac{N}{2}$, $Pr(\theta_n^i = 1) = \phi$ for M-researchers and $Pr(\theta_n^i = 1) = (1 - \phi)$ for F-researchers, and the opposite for $n = \frac{N}{2} + 1, \ldots, N$. For every $\theta \in \Theta$, let $p^{\theta,f}$ (resp. $p^{\theta,m}$) denote the fraction of types in the F (resp. M) population of young researchers. Also let $p^g = (p^{\theta,g})_{\theta \in \Theta}$ for $g = f, m$. To sum up,

$$p^{\theta,m} = \prod_{n=1}^{N/2} \phi^{\theta_n}(1-\phi)^{1-\theta_n} \cdot \prod_{n=N/2+1}^{N} (1-\phi)^{\theta_n}\phi^{1-\theta_n}, \quad p^{\theta,f} = \prod_{n=1}^{N/2}(1-\phi)^{\theta_n}\phi^{1-\theta_n} \cdot \prod_{n=N/2+1}^{N} \phi^{\theta_n}(1-\phi)^{1-\theta_n}. \tag{1}$$

To illustrate this setting with a simple numerical example, let $N = 2$. In this case, we only have four types of researchers, namely $\Theta = (0,0), (1,0), (0,1), (1,1)$. Let's consider an unrealistically large $\phi = 0.8$ to highlight the impact of this parameter (see Section 3. for a calibration). In this case, 80% of M-researchers have characteristic 1, but only 20% have characteristic 2; conversely, 80% of F-researchers have characteristic 2, but only 20% have characteristic 1. The probability distributions p_m^θ and p_f^θ of types θ are in Table 1. Notice that, ex-ante, the distribution of characteristics among M and F researchers is symmetric.

We model each characteristic as a desirable research attribute, which makes it more likely for the researcher to produce quality research. "Quality" research is one that achieves its stated goals—estimating a parameter of interest, establishing a causal effect, documenting a phenomenon experimentally, or proving a theorem. We assume that whether a research paper achieves its goals is observable and can be objectively determined; this may involve, for instance, checking a formal argument regarding a theoretical claim or the application of a statistical procedure, evaluating an experimental procedure for possible biases or ambiguities, or ensuring that the formal results are clearly explained and interpreted, and that the contribution is correctly placed within its literature.

Again, we adopt a simple symmetric specification: we fix $\gamma_0 \in (0,1)$, $\rho \in [1, \frac{1}{\gamma_0}]$, and assume that type $\theta = (\theta_n)_{n=1}^N$ writes a quality paper with probability

$$\gamma^\theta \equiv \gamma_0 \, \rho^{\frac{1}{N}\sum_n \theta_n}. \tag{2}$$

Thus, $\gamma^{(0,\dots,0)} = \gamma_0$, and the probability of producing quality research depends solely on the number of 1's in $\sum_n \theta_n$, with the maximum attained for $\gamma^{(1,\dots,1)} = \gamma_0 \, \rho \in [\gamma_0, 1]$. A young scholar with many desirable characteristics is more likely to produce quality research than another scholar with fewer desirable characteristics. Still, even scholar type $(0,\dots,0)$ has probability $\gamma_0 > 0$ to produce quality research, perhaps by sheer luck. The parameter ρ reflects the relative impact of characteristics on the probability of producing "quality" research. If $\rho = 1$, for instance, then all types produce quality research with probability γ_0. If $\rho = 4$, instead, it means that the best researcher $(1,\dots,1)$ is four times more likely to produce quality research than the worst researcher, $(0,\dots,0)$.

To sum up, the free parameters in our model are ϕ, γ_0, ρ, and N.

2.1. Objective Refereeing

This section studies a benchmark system where the evaluation by established scholars is objective and only certifies whether the research is of sufficient quality or not. Since each young scholar of type θ produces quality research with probability γ^θ, given in (2), this is also the probability with which the research is "accepted" by referees.

For every type $\theta \in \Theta$, let $a_t^{\theta,m}$ and $a_t^{\theta,f}$ denote the mass of young researchers of group M and, respectively, group F of type θ that produce quality research and are thus "accepted" at the end of period t:

$$a_t^{\theta,g} = \gamma^\theta \cdot p^{\theta,g}, \quad g \in \{f,m\}. \tag{3}$$

Denote the total mass of accepted young researchers by $a_t = \sum_{\theta\in\Theta} \sum_{g\in\{f,m\}} a_t^{\theta,g}$.

Denote by $\lambda_t^{\theta,g}$ the mass of established researchers of type θ and group g at time t. We normalize the initial mass of all established researchers to one: $\sum_\theta \sum_g \lambda_0^{\theta,g} = 1$.[5] In order to keep the mass of referees constant, we assume that each young agent whose research is accepted replaces a randomly drawn established one. This is not necessary for the results but keeps the analysis balanced. As we discuss in Section 2.2. below, this assumption is also geared towards maximizing the impact of young researchers on the evolution of the system.[6] The resulting dynamic is then described by the following equation:

$$\lambda_t^{\theta,g} = (1 - a_t)\lambda_{t-1}^{\theta,g} + a_t^{\theta,g}, \quad g \in \{f, m\}. \tag{4}$$

We then obtain the following proposition:

Proposition 1 *In the benchmark model with objective refereeing, regardless of the composition* $(\lambda_0^{\theta,m}, \lambda_0^{\theta,f})_{\theta\in\Theta}$ *of the initial population of established researchers, we have*

$$\lambda_t^{\theta,m} \to \frac{\gamma^\theta p^{\theta,m}}{a}, \quad \lambda_t^{\theta,f} \to \frac{\gamma^\theta p^{\theta,f}}{a}, \quad \textit{and} \quad \frac{\sum_\theta \lambda_t^{\theta,m}}{\sum_\theta \lambda_t^{\theta,f}} \to 1.$$

where $a = \sum_\theta \gamma^\theta \left(p^{\theta,f} + p^{\theta,m}\right)$.

Proof: This and all subsequent results are proved in the Online Appendix.

That is, in our benchmark model with objective refereeing, initial conditions have no long-run effects. In addition, the system always converges to equal shares of M and F established researchers, and the limiting type distribution is fully characterized by the probability of producing quality research and the relative frequency of each type in the population of young researchers. Given the symmetry of the model, this is intuitive.

2.2. Refereeing with Self-Image Bias

Our main model differs from the benchmark in Section 2.1. in that established researchers (referees) not only evaluate young researchers on whether their research is of sufficient quality (as in previous section), but they also use their personal research styles to guide their subjective judgement as to the "importance" or "relevance" of the candidate's output. Specifically,

[5] The fact that the total mass of established scholars (a stock) equals the mass of young M and F researchers (flows) is of course not realistic, but immaterial for our analysis. Normalizing the stock of established researchers to any positive number K yields the same predictions.

[6] We also considered a similar model with a fix retirement rate of existing researchers to be replaced by cohorts of hired young researchers. The results are similar. The assumption in the text has one less parameter and it is more favorable to an eventual convergence to group balance.

each young researcher $i \in M \cup F$ of type θ^i is now randomly matched to a referee r, who uses his or her own characteristics θ^r to evaluate agent i's work. Importantly, evaluation is anonymous and group-blind: it depends solely upon referee r's own type θ^r and the characteristics of researcher i's output, which by assumption coincides with his of her type θ^i.

Consistently with self-image bias, referee r rejects applicants whose type is far from his/her own set of characteristics. We make in fact a stark assumption: referee r has a positive view of young agent i's research if and only if $\theta^r = \theta^i$. (We relax this assumption in the on-line appendix.) If agent i's output is positively evaluated, i becomes an established researcher, and will serve as referee for future cohorts of young researchers.

As in previous section, each young researcher who enters the population of established researchers randomly replaces an existing one. This assumption is the most favorable to young researchers; in particular, if the initial referee population is predominantly made of M-researchers, this assumption makes it easier for the dynamics to "push out" old M-researchers and replace them with young F-researchers. In other words, this assumption is most conducive to attaining group balance in the limit.

Let $\lambda_t^\theta = \lambda_t^{\theta,f} + \lambda_t^{\theta,m}$ be the total mass of established researchers of type θ at time t; also let $\lambda_t = (\lambda^\theta)_{\theta\in\Theta}$. Retaining the notation of Section 2.1., the dynamics for the mass of young researchers of type θ and group g that are accepted in round t is

$$a_t^{\theta,g} = \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g}. \tag{5}$$

Importantly, whether a young researcher is accepted or not depends solely on the type θ, and not also on the group g. As in Equation (4), the total mass of established researchers of type θ and group g is given by

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}\,(1 - a_t) + a_t^{\theta,g} \tag{6}$$

where as above $a_t = \sum_\theta \sum_g a_t^{\theta,g}$. Equations (5) and (6) indicate that there are two forces at play. On one hand, the distribution of incumbent types impacts which research characteristics are likely to be positively evaluated by referees. On the other hand, even among incumbents, types that are more likely to produce quality research tend to be more prevalent. As we shall demonstrate, the interplay of these two forces determines whether the system ultimately attains the first-best outcome in Section 2.1., or if instead an inefficient outcome, characterized by group imbalance, is reached.

2.3. Type Dynamics

We begin by studying the evolution of the mass of each type in the population. The following proposition identifies the types that can potentially survive (i.e. have positive mass) in the limit. All other types vanish over time.

Proposition 2 *Only three types can potentially survive in the limit: either*

(i) the types most prevalent across M and, respectively, F researchers,

$$\theta^m = (1, \ldots, 1, 0, \ldots, 0) \quad \textit{and} \quad \theta^f = (0, \ldots, 0, 1, \ldots, 1); \; \textit{or} \tag{7}$$

(ii) the type most likely to produce quality research,

$$\theta^* = (1, \ldots, 1). \tag{8}$$

Types θ^m and θ^f have frequency ϕ^N; type θ^* has frequency $\phi^{N/2}(1-\phi)^{N/2}$, and is thus less prevalent among both M and F researchers.

Not all three types can survive simultaneously. Except for knife-edge parameter choices, either θ^* dominates in the limit and all other types (including θ^m and θ^f) disappear, or θ^m and θ^f dominate (and θ^* disappears). Thus, one of the two forces at play—the initial distribution of types and the likelihood of producing quality research—eventually prevails.

In the next proposition, recall that the parameter ρ measures the impact of research characteristics on the probability of producing quality research (see equation (2)).

Proposition 3 *Let $\bar{\lambda}^\theta = \lim_{t \to \infty} \lambda_t^\theta$ for all $\theta \in \Theta$ and*

$$\bar{\rho}(\phi, N) = \frac{1}{4}\left(\left(\frac{1-\phi}{\phi}\right)^{N/2} + \left(\frac{\phi}{1-\phi}\right)^{N/2}\right)^2. \tag{9}$$

(a) If $\rho < \bar{\rho}(\phi, N)$, then only types θ^m and θ^f survive in the limit. In addition, if at time 0, all referees are in the M-group with $\lambda_0 = p^m$, then

$$\bar{\lambda}^{\theta^m} = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1}{2}; \quad \bar{\lambda}^{\theta^f} = 1 - \bar{\lambda}^{\theta^m}. \tag{10}$$

(b) If $\rho > \bar{\rho}(\phi, N)$ then, regardless of the distribution of time-0 referees, only type θ^* survives in the limit.

In part (a), the impact of research characteristics on the probability of producing a quality paper, which is a function of ρ, is comparatively small. In this case, the dynamics of the system are driven primarily by the initial conditions and the flows of young researchers. In particular, if all referees are initially in the M-group, then in the limit M-researchers will represent the majority—despite the fact that an equal mass of young M and F researchers enters the model in every period, and that the research characteristics of both types are equally conducive to quality research.

Interestingly, even type θ^* disappears in this scenario, despite the fact that such type has *all* desirable research characteristics. For instance, when a young researcher of type θ^* is matched with a referee of type θ^m, the latter "disapproves of" the θ^* traits from $N/2+1$ to N, even if they are objectively desirable. Similarly, a referee of type θ^f "disapproves of" characteristics from 1 to $N/2$. To interpret, recall that research characteristics may also include e.g. research topics or methodologies. More generally, the nature of self-image bias is exactly that each reviewer considers his or her traits as the important ones, and discounts the other ones.

Part (b) characterizes a more "meritocratic" scenario in which research characteristics significantly improve the odds of producing quality research. In this case, regardless of the initial conditions, the system reaches an efficient steady state in which all researchers possess every research characteristics—regardless of their group. Self-image bias is still at work in this scenario, but each characteristic is important enough that, over time, referees themselves will tend to possess more and more of them, and hence select in a "virtuous" way.

Taken together, parts (a) and (b) show that our simple symmetric model is capable of generating both long-run outcomes that are affected by group imbalance, as well as meritocratic and balanced outcomes. The next corollary shows, however, that irrespective of parameter values, if the number N of research characteristics is large enough, the biased outcome in part (a) of Proposition 3 will prevail—even if between-group differences are arbitrarily small (i.e. if ϕ is close to 0.5):

Corollary 1 *For any $\phi \in (\frac{1}{2}, 1)$, $\gamma_0 \in (0,1)$, and $\rho \in (1, \frac{1}{\gamma_0})$, if $\lambda_0 = p^m$, then*

1. *there exists N large enough such that outcome (a) of Proposition 3 realizes;*
2. *as the number of characteristics $N \to \infty$, $\bar{\lambda}^{\theta^m} \to 1$.*

Thus, if the number of research characteristics is large and the M–group dominates the initial population, its most prevalent type θ^m will dominate in the steady state. Informally,

M-researchers effectively determine on behalf of society that the only important research characteristics are their own. It follows that F-researchers have no chance to grow to equality, even without any explicit bias against them.

2.4. Model Predictions

In this section, we discuss the model's predictions, namely, convergence to group imbalance (Section 2.4.1.); higher "bar" for F-researchers (Section 2.4.2.); talent loss and clustering across fields (Section 2.4.3.); negative relation between institutions' publication success and F-representation (Section 2.4.4.); misperceived tradeoff between quality and diversity by established researchers (Section 2.4.5.); and, finally, welfare losses due to F-underrepresentation (Section 2.4.6.)

2.4.1. Group Imbalance in the Limit

Proposition 3 mostly concerns the distribution of researcher types irrespective of their group. We now discuss the model's implications for group imbalance.

Proposition 4 *Assume that all referees are initially from the M-group with $\lambda_0 = p^m$.*

(a) If $\rho < \bar{\rho}(\phi, N)$, then the total limit mass of M and F researchers are

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} > 0.5. \tag{11}$$

(b) If $\rho > \bar{\rho}(\phi, N)$, then $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$.

The result in part (a) intuitively follows from the corresponding result in Proposition 3. Eventually, only θ^m and θ^f survive, but θ^m is more common in the M group than θ^f. Thus, the limiting total mass of M-researchers is larger than 0.5. The next corollary illustrates the limiting case as the number of research characteristics N diverges to infinity:

Corollary 2 *For all $\phi \in (\frac{1}{2}, 1)$, $\gamma_0 \in (0, 1)$, and $\rho \in (1, \frac{1}{\gamma_0})$, if $\lambda_0 = p^m$,*

1. *there exist N large enough such that case (a) in Proposition 4 realizes;*
2. *as $N \to \infty$, $\bar{\Lambda}^m \to 1$ and $\bar{\Lambda}^f \to 0$.*

This reinforces and refines tho message of Corollary 1: in particular, for *all* parameter values, as N increases, the fraction of M-researchers always dominates in the limit, and in the limit converges to one.

To visually illustrate Propositions 2 through 4, return to the numerical example in Table 1, where $N = 2$. In the notation of Proposition 2, $\theta^m = (1, 0)$, $\theta^f = (0, 1)$, and $\theta^* = (1, 1)$. In addition to $\phi = 0.8$, assume now $\gamma_0 = 0.2$, and $\rho = 4$. This implies that type θ^* is twice as likely as types θ^m and θ^f to produce quality research; in turn, these types are twice as likely as the worst type $(0, 0)$ to do so. Thus, research characteristics *do* matter in this scenario; however, by Proposition 3 self-image bias prevails:

$$\rho = 4 < 4.51625 = \bar{\rho}(\phi, N) = \frac{1}{4}\left(\left(\frac{0.2}{0.8}\right)^{2/2} + \left(\frac{0.8}{0.2}\right)^{2/2}\right)^2.$$

Panel (a) of Figure 3 displays the evolution of the fractions Λ_t^m and Λ_t^f of M- and F-researchers over 100 periods, assuming that all established researchers at time $t = 0$ are M-researchers ($\lambda_0 = p^m$) and that p^m and p^f are as in Table 1. Established researchers are predominantly M-type in the limit.[7]

2.4.2. Higher "Bar" for F-researchers

If the initial population of referees is entirely from the M group, a basic force in our model implies that young researchers from the F group are, in a sense, held to a higher standard. Recall that, in our parameterization of objective quality γ^θ, all characteristics are equally important. Now consider the set of all types θ that possess exactly L characteristics. All such types have the same objective productivity, independently of group membership. Furthermore, the *same* mass of young researchers in the M and F groups possesses exactly L characteristics. Yet, if the referees are initially all from the M group, the mass of accepted M-group researchers of such types is *always* at least as large as for the F group. This is true even if parameters are consistent with the "meritocratic" regime.

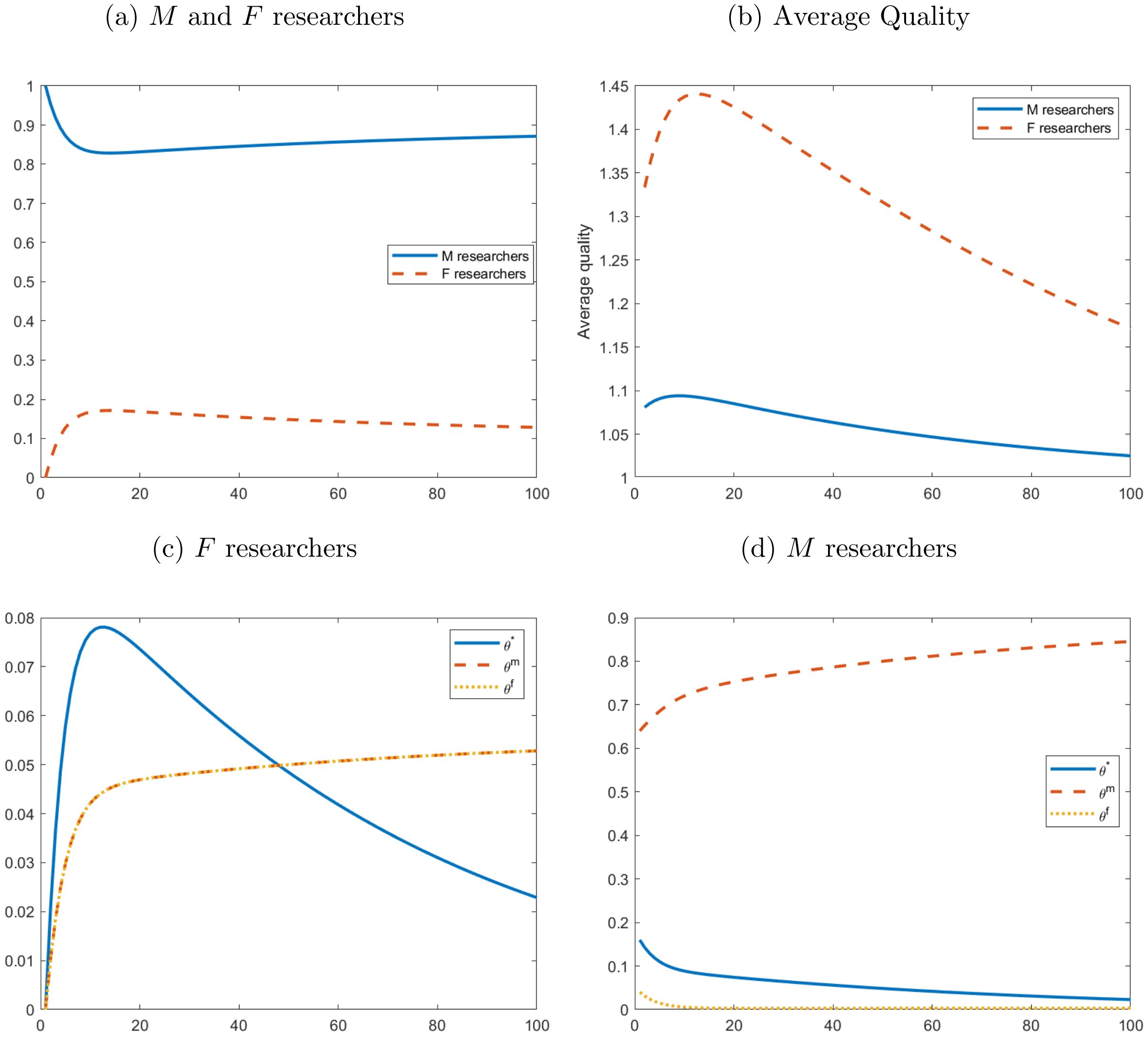
Proposition 5 *Assume that $\lambda_0 = p^m$. For every $L \in \{0, \ldots, N\}$ and $t > 0$, the acceptance rate of M-researchers of quality L is higher than the one of F-researchers of the same quality:*

$$\sum_{\theta:\sum_n \theta_n = L} a_t^{\theta,m} \geq \sum_{\theta:\sum_n \theta_n = L} a_t^{\theta,f}$$

and the inequality is strict if there is $\theta \in \Theta$ with $\sum_n \theta_n = L$ and $\theta_n \neq \theta_{N+1-n}$ for some n.

[7] Indeed, from Eq. (11) with $N = 2$ and $\phi = 0.8$ the fraction of M-researchers in the limit is $\bar{\Lambda}^m \approx 89\%$.

Figure 3: Fraction of M and F Researchers



(a) M and F researchers

(b) Average Quality

(c) F researchers

(d) M researchers

Panel (a) plots the fraction of M and F researchers. Panel (b) plots the average quality of M and F researchers conditional on acceptance, i.e. $\sum_\theta L^\theta w_t^{\theta,g}$ where $L^\theta = \sum_{n=1}^N \theta_n$ and $w_t^{\theta,g} = a_t^{\theta,g}/\sum_{\theta'} a_t^{\theta',g}$, $g = f, m$. Panel (c) plots the types of established F-researchers and panel (d) plots the types of established M- researchers. We show types $\theta^* = (1,1)$, $\theta^m = (1,0)$, and $\theta^f = (0,1)$. Initially $\lambda_0 = p^m$.Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

That is, in aggregate, it is easier for young M-researchers researchers to be accepted than for F-young researchers, controlling for objective quality—the number of desirable characteristics $\sum_n \theta_n = L$. This is in line with the cited evidence in Card et al. (2020) that, conditional on quality (proxied by citations post-publication) women-authored papers tend to be accepted less frequently than men's.[8] Indeed, the following Proposition shows

[8] Card et al. (2020) also show that, unconditionally, men- and women-authored papers are equally likely

that accepted F-researchers are of higher quality than accepted M researchers, on average, for the case $N = 2$. Based on extensive numerical exploration, we conjecture the same conclusions to hold for arbitrary N—but we are unable to prove this at this time.

Proposition 6 *Assume that $\lambda_0 = p^m$.*

(i) Let $N = 2$. The average quality of accepted F-researchers is higher than the one of accepted M-researchers:

$$E[L|f, accepted] = \sum_{\theta} L^{\theta}\ w_t^{\theta,f} > \sum_{\theta} L^{\theta}\ w_t^{\theta,m} = E[L|m, accepted] \tag{12}$$

where $L^{\theta} = \sum_{n=1}^{N} \theta_n$ is the quality type θ (i.e. its number of 1's in θ), and

$$w_t^{\theta,g} = \frac{a_t^{\theta,g}}{\sum_{\theta'} a_t^{\theta',g}}$$

(ii) As $t \to \infty$ the average quality of both F and M converges to either $N/2 = 1$ if only θ^m and θ^f survive in the limit, or $N = 2$ if only θ^* survives in the limit.

Panel (b) of Figure 3 illustrates the result in the numerical example of Table 1. The intuition is as follows. With $N = 2$, referees accept the same mass of M and F researchers of types $(0, 0)$ and $\theta^* = (1, 1)$. However, among types with $L^{\theta} = 1$ (that is, θ^m and θ^f), since established researchers are predominantly from the M group, type θ^m is accepted more frequently than θ^f (see panels (c) and (d) of Figure 3). But this type is more common among young M researchers than among young F researchers. Thus, overall, more M-researchers of quality $L = 1$ are accepted. This implies that the *relative frequency* of type $\theta^* = (1, 1)$ is *higher* among accepted F-researchers than accepted M-researchers. This turns out to imply that the average quality of accepted F-researchers is higher.[9]

2.4.3. Talent Loss and Clustering

One further implication of our model is that, when self-image bias prevails, the characteristics $n = N/2+1, \ldots, N$ that are more common in the F-group are under-represented in the limit.

to be accepted. The model in this section does not generate this finding: summing over $L = 0, \ldots, N$ in the displayed equation of Proposition 5, one readily sees that young M researchers are more likely to be accepted on average. The model with endogenous choice in Section A1.1. yields more uniform unconditional acceptance across genders, and fewer female acceptance overall due to self-selection.

[9] The argument above is incomplete because the fraction of accepted type-$(0, 0)$ researchers is also higher among F rookies than M rookies; the proof of Proposition 6 takes this into account. The same basic forces are at play with $N > 2$. However, in this case there are many different intermediate types, and this makes extending the argument given above non-trivial.

Corollary 3 *In part (a) of Proposition 4, in the limit,*

$$\frac{\bar{\lambda}^{\theta^m,m}}{\bar{\lambda}^{\theta^m,m}+\bar{\lambda}^{\theta^f,m}} > 0.5 = \frac{\bar{\lambda}^{\theta^f,f}}{\bar{\lambda}^{\theta^m,f}+\bar{\lambda}^{\theta^f,f}} = \frac{\bar{\lambda}^{\theta^m,f}}{\bar{\lambda}^{\theta^m,f}+\bar{\lambda}^{\theta^f,f}} > \frac{\bar{\lambda}^{\theta^f,m}}{\bar{\lambda}^{\theta^m,m}+\bar{\lambda}^{\theta^f,m}} \tag{13}$$

In the limiting distribution, the majority of established M researchers are of type θ^m. However, the established $F-$population has the *same* fraction of type θ^f as type θ^m. This result is in stark contrast with θ^f being the prevalent type in each cohort of young F-researchers. The selection mechanism makes the type most prevalent among M-researchers, θ^m, be a frequent type in the established F-researchers (50% of the time), even if such type only has $(1-\phi)^N < 0.5$ frequency in the population of young F-researchers. That is, F-group research characteristics are underrepresented in the limit.

In the numerical example of Table 1, for instance, the fact that successful F-researchers have equal mass of types θ^m and θ^f should be contrasted with the fact that $\frac{\phi^2}{(1-\phi)^2} = \frac{0.64}{0.04} = 16$ times as many θ^f types as θ^m types appear among F-researchers in *every* period. It turns out that this ex-ante difference in the masses of types θ^m and θ^f in the F population is offset by the fact that θ^m types are much more likely to be matched with referees of the same type. In our symmetric model, these two effects exactly offset each other.

Self-image bias also implies clustering of different characteristics within the two groups. In particular, established researchers of type θ^m are more likely to be from the M group; in contrast, type-θ^f researchers are mostly going to be from the F group.

Corollary 4 *In part (a) of Proposition 4, in the limit, M-researchers are relatively more frequent as type θ^m and F-researchers are relatively more frequent as type θ^f:*

$$\frac{\bar{\lambda}^{\theta^m,m}}{\bar{\lambda}^{\theta^m,m}+\bar{\lambda}^{\theta^m,f}} = \frac{\bar{\lambda}^{\theta^f,f}}{\bar{\lambda}^{\theta^f,m}+\bar{\lambda}^{\theta^f,f}} = \frac{1}{1+\left(\frac{1-\phi}{\phi}\right)^N} > 0.5; \tag{14}$$

If, as seems plausible, at least some of the research characteristics are more prevalent, or more valuable, in certain fields than in others, this result implies that the two groups will be differently represented across fields.

This is qualitatively consistent with the evidence documenting large gender differences across economics topics (see e.g. Chari and Goldsmith-Pinkham (2018) and Lundberg and Stearns (2019)), although it is too extreme, as women's frequency never breaks the 50% threshold in economics (although it does in other areas, such as psychology). This result is also in stark contrast with the case of meritocracy that is illustrated in Proposition 3(b).

In that case, θ^* prevails in the limit which generates a symmetric distribution of M and F researchers across characteristics.

Under the interpretation that different research characteristics are more prevalent in different fields, the model then also implies that the typical research of M-researchers becomes "mainstream" compared to the typical research of F-researchers. Under self-image bias in refereeing, in the limit, the probability that a researcher of type θ^m (resp. θ^f) publishes successfully is $\overline{\gamma} \cdot \bar{\lambda}^{\theta^m}$ (resp. $\overline{\gamma} \cdot \bar{\lambda}^{\theta^m}$), where $\overline{\gamma} = \gamma^{\theta^m} = \gamma^{\theta f}$. We then obtain the following corollary:

Corollary 5 *In the limit, the mass of published work in the field characterized by θ^m is larger than the one in the field characterized by θ^f. That is:*

$$\overline{\gamma} \cdot \bar{\lambda}^{\theta^m} > \overline{\gamma} \cdot \bar{\lambda}^{\theta f}$$

This asymmetry arises in the limit notwithstanding the ex-ante symmetry of the model.

2.4.4. Publication Success and F-Underrepresentation

Our model is also consistent with the evidence that more research-intensive universities have lower female representation, as shown in the bottom panel of Figure 1. In particular, suppose that the condition in part (a) of Proposition 4 holds, so that, in the limit, only two types of researchers survive, namely θ^m and θ^f. We analyze the resulting limit economy; see Section 3. for numerical results with calibrated parameter values and a finite time horizon.

Consider an institution with an arbitrary fraction $y \in [0, 1]$ of θ^m-researchers and a complementary fraction $(1 - y)$ of θ^f-researchers; since no other types survive in the limit, these are the only researcher types that an institution can employ.[10]

Under self-image bias in refereeing, the probability that a researcher of type θ^m (resp. θ^f) publishes successfully is $\overline{\gamma} \cdot \bar{\lambda}^{\theta^m}$ (resp. $\overline{\gamma} \cdot \bar{\lambda}^{\theta^m}$), where $\overline{\gamma} = \gamma^{\theta^m} = \gamma^{\theta f}$. Hence, the *average publication frequency* of the institution is

$$P(y) = \overline{\gamma}\,(y\overline{\lambda}^{\theta^m} + (1-y)\overline{\lambda}^{\theta f})$$

Since $\overline{\lambda}^{\theta^m} > \overline{\lambda}^{\theta f}$, $P(y)$ increases in y.

[10] The total mass of researchers in the institution under consideration is irrelevant to the analysis, and can be considered small. In Section 3., we consider a different parameterization in which the entire population is divided into a given, fixed number of institutions.

We assume that the type-θ^m and type-θ^f researchers at the institution under consideration belong to the F and M groups in proportions analogous to those in the population: that is, a fraction $y\bar{\lambda}^{\theta^m,f}$ are of type θ^m and group F, a fraction $y\bar{\lambda}^{\theta^f,f}$ are of type θ^f and group F, etc. Then, the fraction of F-researchers in an institution parameterized by y is given by:

$$F(y) = \frac{(y\overline{\lambda}^{\theta^m,f} + (1-y)\overline{\lambda}^{\theta^f,f})}{(y\overline{\lambda}^{\theta^m} + (1-y)\overline{\lambda}^{\theta^f})}$$

Proposition A.4 in the Online Appendix immediately implies that $F(y)$ decreases in y. We thus have the following corollary:

Corollary 6 *In part (a) of Proposition 4, in the limit, institutions with higher exogenous fraction y of θ^m-researcher, and $(1-y)$ of θ^f researchers, have higher publication frequency and lower percentage of F-researchers.*

Thus, institutions with higher research intensity – i.e., higher publication frequency – also have a lower share of F-researchers, consistently with Figure 1. Intuitively, the result follows from the fact that the limit mass of θ^m-researchers in the population is higher than that of θ^f-researchers: $\overline{\lambda}^{\theta^m} > 0.5 > \overline{\lambda}^{\theta^f}$. Self-image bias implies that types θ^m have a higher chance of publication success, because the probability they are matched with referees of their own type is higher. Consequently, on average, institutions with a higher fraction y of type-θ^m scholars are more likely on average to achieve successful publications. However, types θ^m are also more likely to come from the M group: this generates a negative relation between research success and F-representation. In the limit, F-researchers are least represented in "top institutions," where "top" is defined as in terms of publication record. This inverse relation is surprising as the referees in the model have no group bias, only self-image bias.

2.4.5. Perceived Trade-off Between Quality and Diversity

Self-image bias also explains why the current population of referees may incorrectly perceive that there is a trade-off between "quality" and diversity—that increasing diversity implies compromising on quality. The intuition is simple: by definition, self-image bias implies that the closer another researcher is to one's own type, the higher their subjectively perceived quality. Hence, in particular, M-researchers will subjectively perceive other M-researchers to be of higher quality on average than F-researchers. Therefore, if the population of established scholars consists mainly of M-researchers, a random sample of established scholars will misperceive the average quality of M-researchers to be higher than that of F-researchers. Hence the mistaken impression that, in order to increase F-representation, one has to "accept" a loss of quality.

This conclusion is in stark contrast with the fact that, in our model, the average *objective* quality of each cohort M- and F-researchers, which is given by the average probability of producing quality research, is exactly the same, by construction. Furthermore, as shown in Proposition 6, for $N = 2$, the average quality of accepted F researchers is actually *higher* than that of accepted M researchers; the same is true for the calibrated model of Section 3., with multiple characteristics (see Fig. 5). Thus, in fact, increasing diversity can potentially *increase* average objective quality.

We now formally derive this conclusion from Proposition 5. Recall that, under self-image bias, a referee r of type θ^r accepts a researcher of type θ only if $\theta = \theta^r$. We can interpret this by saying that referee r believes researcher θ has a quality of γ^θ if $\theta = \theta^r$, and 0 otherwise. Therefore, for this referee, the perceived quality of a randomly drawn M-researcher is $Q(M|\theta^r) = \gamma^{\theta^r} p_m^{\theta^r}$, while that of a randomly drawn F-researcher is $Q(F|\theta^r) = \gamma^{\theta^r} p_f^{\theta^r}$. Finally, given the distribution λ_t of established researchers, the average perceived quality of young M- and F-researchers are, respectively,

$$Q(M|\lambda_t) = \sum_{\theta \in \Theta} p_m^\theta \gamma^\theta \lambda_t^\theta \quad \text{and} \quad Q(F|\lambda_t) = \sum_{\theta \in \Theta} p_f^\theta \gamma^\theta \lambda_t^\theta$$

Proposition 5 then yields

Corollary 7 *Assume that $\lambda_0 = p^m$. Then, the population of referees λ_t (mis)perceives the quality of a random F-researcher to be lower than the quality of a random M-researcher. That is,*

$$Q(F|\lambda_t) < Q(M|\lambda_t) \tag{15}$$

2.4.6. Talent Loss and Welfare Loss

The loss of research characteristics more prevalent among F researchers can reduce societal welfare. To flesh out this intuition, in this section we consider a simple extension of our model, inspired by the "science of science" literature.

Assume that, in every period, society confronts a new real-world problem, and a randomly drawn established researcher is selected to solve it. Identify each problem with the set of characteristics that make it more likely for a researcher to find a solution. Thus, we identify the set of real-world problems with the set Θ of types, and assume that problem ϑ will be solved with higher probability by a researcher whose type θ satisfies $\theta_n = 1$ for every

$n = 1, \ldots, N$ such that $\vartheta_n = 1$. Thus, type θ^* can solve any problem with high probability, whereas type θ^m (resp. θ^f) is only likely to find solutions to problems that require M-prevalent (resp. F-prevalent) research characteristics. Specifically, we assume that type θ solves problem ϑ with probability

$$Pr\left(\text{solve problem } \vartheta|\text{type } \theta\right) = \alpha\beta^{\frac{1}{N}\sum_{n=1}^{N}\max(1-\vartheta_n,\theta_n)} \tag{16}$$

with $\alpha > 0$ and $\beta > 1$, and $\alpha < \beta^{-1}$.

Finally, we assume that real-world problems are uniformly drawn from Θ. This implies that no characteristic—whether it be more prevalent among M or F researchers—is intrinsically more useful to solve real-world problems.

The key feature of this environment is that the type distribution of established researchers at each time t, or in the limit, determines society's ability to solve real-world problems. This allows us to derive sharp conclusions about welfare. Self-image bias in this environment has the natural interpretation as a referee's perception of which real-world problems are relevant: A referee of type θ believes that only problems that require his/her research characteristics are worth solving.

Furthermore, in this environment, "objective quality" has also a natural interpretation: it is the ability to solve a uniformly drawn real-world problem. The following result ensures that our specification of the function γ^θ in equation (2) is, in fact, consistent with this interpretation—it is the "reduced form" of the probability of solving a real-world problem in the extended model just described:

Proposition 7 *The probability that type $\theta \in \Theta$ solves a uniformly drawn problem is*

$$\frac{1}{2^N}\left(\alpha\sum_{\vartheta\in\Theta}\beta^{\frac{1}{N}\sum_{n=1}^{N}\max(1-\vartheta_n,\theta_n)}\right) = \gamma_0\,\rho^{\frac{1}{N}\sum_n\theta_n} = \gamma^\theta$$

where $\beta = \left(\frac{1}{2-\rho^{1/N}}\right)^N\rho$, $\alpha = \gamma_0(2-\rho^{1/N})^N$, *and* $1 < \rho < 2^N$.

Given this interpretation of our basic model, we can now draw welfare conclusions. In particular, consider two symmetric problems ϑ and ϑ' where the former depends mostly on M-characteristics, and the latter, symmetrically, mostly on F-characteristics. Specifically, let $n_m(\vartheta) = \sum_{n=1}^{N/2}\vartheta_n$ and $n_f(\vartheta) = \sum_{n=N/2+1}^{N}\vartheta_n$ denote the number of 1's of problem ϑ in the M-group and the F-group, respectively. We assume that problems ϑ and ϑ' are symmetric in the sense that $n_m(\vartheta) = n_f(\vartheta')$ and $n_f(\vartheta) = n_m(\vartheta')$; furthermore, we assume that $n_m(\vartheta) > n_f(\vartheta)$, and hence, symmetrically, $n_m(\vartheta') < n_f(\vartheta')$.

The probability of solving ϑ or ϑ' can be obtained by averaging the expression in Eq. (16) over all researcher types θ, weighted by the distribution of these types among established researchers. Consider the limiting type distribution as an example.

In the first best solution in Section 2.1., the share of each type θ is given in Proposition 1, and is readily seen to be symmetric, in the sense that, if $\theta, \theta' \in \Theta$ satisfy $n_m(\theta) = n_f(\theta')$ and $n_f(\theta) = n_m(\theta')$, then θ and θ' have the same limiting shares. This immediately implies that $Pr(\text{society solves } \vartheta) = Pr(\text{society solves } \vartheta')$.

The economy with self-image bias is instead more likely, in the limit, to solve problems that require characteristics typical in the M-group, even though characteristics more prevalent in the F-group are also well represented in the M-group if ϕ is close to 0.5, as in the calibration of Section 3..

Corollary 8 *Consider two symmetric real-world problems ϑ and ϑ' with $n_m(\vartheta) = n_f(\vartheta') > n_f(\vartheta) = n_m(\vartheta')$. Then, under the conditions of part (a) of Proposition 4, in the limit,*

$$Pr(\textit{society solves } \vartheta) > Pr(\textit{society solves } \vartheta'). \tag{17}$$

This result readily follows from the fact that the only surviving types in part (a) of Proposition 4 are θ^m and θ^f, with limiting shares $\bar{\lambda}^{\theta^m} > \bar{\lambda}^{\theta^f}$.

Thus, the loss of F-talent in an economy subject to self-image bias will result in diminished ability to solve certain real-world problems. This is in line with evidence from Bell et al. (2019) on the innovation potential intrinsic in greater participation across different groups.

3. Calibration with Many Characteristics

The previous section provided the main propositions along with a simple illustrative numerical example. The parameter ϕ can be easily related to Cohen's d statistic for an individual characteristic: for $n = 1, \ldots, \frac{N}{2}$,

$$d = \frac{\mathrm{E}[\theta_n^i | i \in M] - \mathrm{E}[\theta_n^i | i \in F]}{\sigma_{\text{pooled}}(\theta_n^i)} = \frac{2\phi - 1}{\sqrt{\phi(1-\phi)}}. \tag{18}$$

For $n = \frac{N}{2} + 1, \ldots, N$, the d statistic is the negative of the above expression. Cohen (2013) suggests that values of d around 0.2 should be considered "small," values around

0.5 “medium,” and values around or above 0.8 “large.” In the example in the previous section, Cohen’s d statistic for each characteristic then equals

$$d = \frac{2\phi - 1}{\sqrt{\phi(1-\phi)}} = \frac{0.6}{\sqrt{0.16}} = 1.5,$$

which, as we noted above, is excessively large for most characteristics likely to be relevant to research activity. However, Proposition 3 shows that, if the number of characteristics is sufficiently large, such extreme across-group differences are not required for our conclusions to hold.

This section considers a more realistic parametrization of our model. The first issue is the number of characteristics that lead to quality research and are taken into account by referees when they evaluate a candidate. We suggest that the number of characteristics is actually large. The following is but a partial list: (i) Economic motivation; (ii) “Nose” for good questions; (iii) Institutional knowledge; (iv) Ability to find new data sources; (v) Solid identification strategy; (vi) Sophisticated empirical analysis; (vii) Clever experimental design; (viii) Skilful theoretical modelling; (ix) Ability to highlight insights, strategic effects, etc. (x) Mathematical sophistication, proof techniques, etc. (xi) Ability to position within the literature; (xii) Ability to highlight policy implications; (xiii) Presentation skills; (xiv) Ability to address questions from audience; (xv) Honesty;[11] and so on. Likely, there are many others. Perhaps some of these research traits are more important than others, but as a first pass, it is indeed plausible that the positive or negative result of a review depends on a combination of research characteristics, and not just a small number. In light of these considerations, and to be conservative, we assume that $N = 10$.

The second issue is the magnitude of between-group differences, which depends on the parameter ϕ. We set $\phi = 0.5742$, so the implied Cohen’s d is

$$d = \frac{2 \times 0.5742 - 1}{\sqrt{0.5742 \times (1 - 0.5742)}} = 0.3,$$

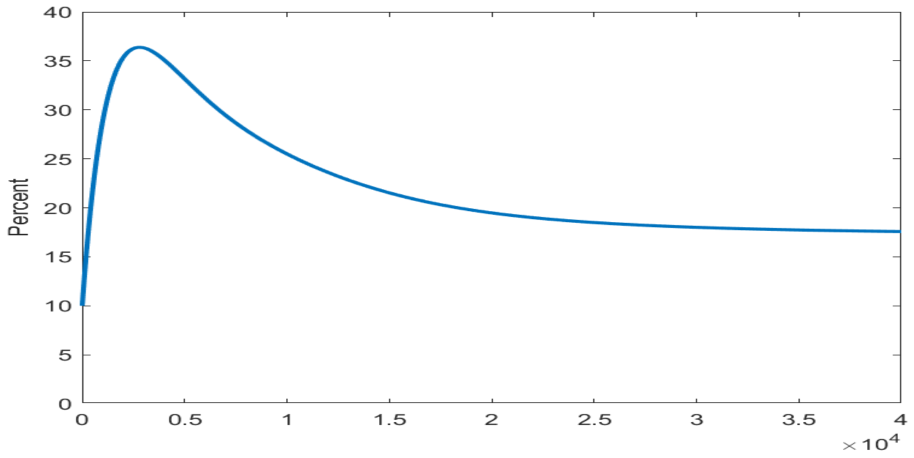
This value is considered “small” and in line with the estimated group differences of the various traits discussed in the introduction. With $\phi = 0.5742$, between-group heterogeneity in each characteristic is far smaller than within-group heterogeneity.[12]

As for the parameterization of $\gamma^\theta = \gamma_0 \, \rho^{\frac{1}{N}\sum_{n=1}^{N}\theta_n}$, we proceed as follows: First, we assume the best researchers $\theta^* = (1, 1, ..., 1)$ has 100% probability of producing quality

[11] For instance, some researchers may be more keen to “torture” the data than others, or search for variables that lead to statistical significance. See e.g. discussion in Mayer (2009) and, on the impact of conflict of interests on economic research, Fabo, Jancokova, Kempf, and Pastor (2020).

[12] We focus on effect size for a single characteristic as its magnitude has been widely documented in the psychology literature (see introduction). Unfortunately, we were unable to identify experimental studies measuring multidimensional effect sizes between genders for us to use in our calibration.

Figure 4: Percent of F-Researchers in Calibrated Model



Percent of F- researchers in calibrated model. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9\, p^m + 0.1\, p^f$.

research, ie. $\gamma^{\theta^*} = 1$. Second, we calibrate γ_0 to match the rate at which economics PhD students succeed in getting an academic job. We compute the latter from the NSF Survey of Doctoral Recipients. We take the ratio of economics PhD recipients who are employed in 4-year educational institutions over the total of economics PhD recipients, both inside and outside the US.[13] That ratio is 0.462. Choosing $\gamma_0 = 0.2$ yields an objective success rate $\sum_\theta \gamma^\theta(p^{\theta,f} + p^{\theta,m})/2 = 0.462$. Interestingly, the implied $\rho = \gamma^{\theta^*}/\gamma_0 = 5$ entails that researcher N is objectively five times as productive as researcher 0, which is roughly in line with the evidence on research productivity reported in Conley and Önder (2014).[14]

Finally, we assume that initially F-researchers represent 10% of the total mass, which is roughly consistent with the percentage of women faculty in 1975, and be consistent with the distribution of annual inflows of young researchers, i.e. $\lambda_0 = 0.9p^m + 0.1p^f$.

The results are in Figures 4 through 7. Figure 4 shows that the system converges to a large imbalance between M- and F-researchers, with F-researchers representing around 20%

[13] The 2017 survey is the latest as of the time of this writing and it is available at `https://ncsesdata.nsf.gov/doctoratework/2017/index.html`. The total number of economics PhD recipients is 32,000 in US and 12,750 outside the US. The total number of them working in a 4-year educational institution are 12,750 in the US and 7,900 outside the US. The ratio of economics PhDs who undertake an academic career is (12,750+7,900)/(32,000+12,750) = 0.462.

[14] These parallels with the data should be taken with a grain of salt, given that the data would reflect the outcome of the model with self-image bias, and not just objective refereeing. On the other hand, we have more degrees of freedom: recall that we normalized that mass of reviewers to 1, but we can choose another mass K to match the failure rate from the data. See footnote 5.

of the population.[15] This large imbalance obtains despite the fact that the distribution of characteristics is now very similar across M and F types.

Panel (a) of Figure 5 plots the average quality of F- and M-researchers conditional on being accepted, and shows that the average quality of F-researchers is uniformly higher than $M-$ researchers, although both eventually converge to $N/2 = 5$. This plot is consistent with Proposition 6 and our conjecture that the result should hold for every N.[16] Panel (b) reports Figure 4(a) of Card et al. (2020) which shows the analogous result in the data, except that there is no dynamics and the graphs refers to different referees' recommendations. The figure shows that "[a]t nearly each referee recommendation, female-authored papers have higher citations than male-authored papers, with a 20 log point average difference. This suggests that papers by all-female authors are held to a higher bar by the referees." (Card et al. (2020), page 296)

Next, we build on Section 2.4.4. and examine the publication success of researchers in institutions that differ in their type composition. Specifically, consider J institutions, and assume that each institution $j = 1, \ldots, J$ employs an exogenously specified fraction x_j^θ of all type-θ researchers, with $\sum_{j=1}^J x_j^\theta = 1$. The mass of θ researchers in institution j at time t is thus $x_j^\theta \lambda_t^\theta$. We make no assumptions about the distribution of *groups* in institutions.

Since, at each time t, institution j employs $x_j^\theta \lambda_t^\theta$ researchers of type θ, each such researcher publishes with probability $\gamma^\theta \lambda_t^\theta$, and the total mass of researchers employed by institution j is $\sum_\theta x_j^\theta \lambda_{j,t}^\theta$, the weighted-average probability of publications of (established) researchers in institution j is

$$P_{j,t} = \frac{\sum_{\theta\in\Theta} \gamma^\theta \left(\lambda_t^\theta\right)^2 x_j^\theta}{\sum_{\theta\in\Theta} \lambda_t^\theta x_j^\theta}.$$

On the other hand, the fraction of F-researchers in institution j is

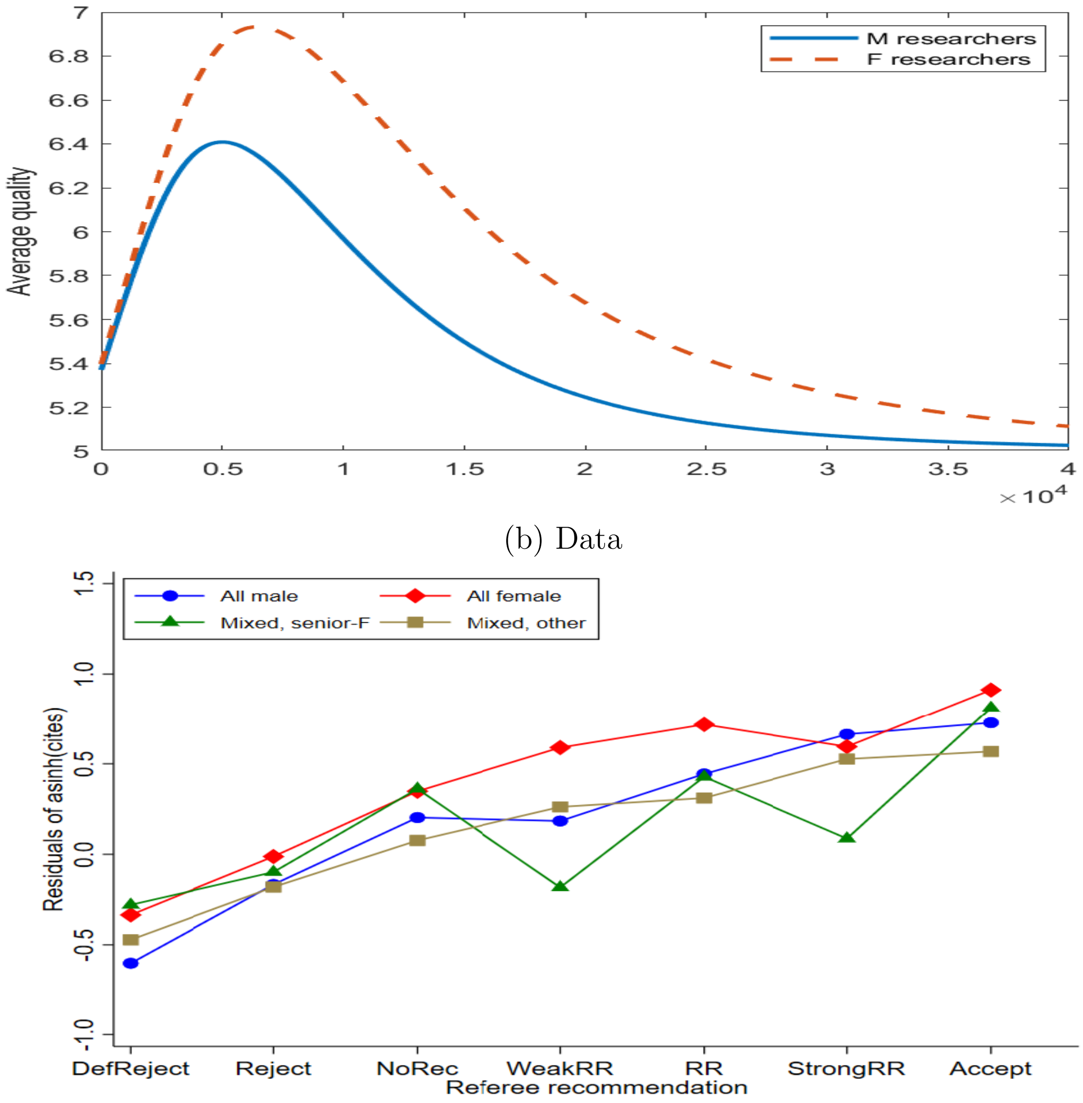
$$F_{j,t} = \frac{\sum_{\theta\in\Theta} \lambda_t^{\theta,f} x_j^\theta}{\sum_{\theta\in\Theta} \lambda_t^\theta x_j^\theta}$$

Panel (a) of Figure 6 shows the scatterplot of $P_{j,t}$ and $F_{j,t}$ from the model simulation for a random draw of x_j's for $J = 100$ institutions, for large t (i.e., "in the limit"). As the plot shows, there is a negative relation between an institution publishing intensity (x-axis) and its F-representation ($y-$axis). In Panel (b) we first rank the J institution in terms of publishing intensity $P_{j,t}$, and then take the average of the fraction of F-researchers across

[15] By comparison, setting $\phi = 0.5742$ and $N = 10$ in equation (11), the limiting fraction of M researchers is $\bar{\Lambda}^m \approx 91\%$. The difference is due to the fact that Eq. (11) was derived assuming that $\lambda_0 = p^m$.

[16] Again, Proposition 6 assumes that $\lambda_0 = p^m$; the results in Panel (a) of Figure 5 thus suggest that the conclusions of the Proposition are robust to small changes the initial population.

Figure 5: Quality of $M-$ and F-Researchers: Model vs. Data


(a) Model

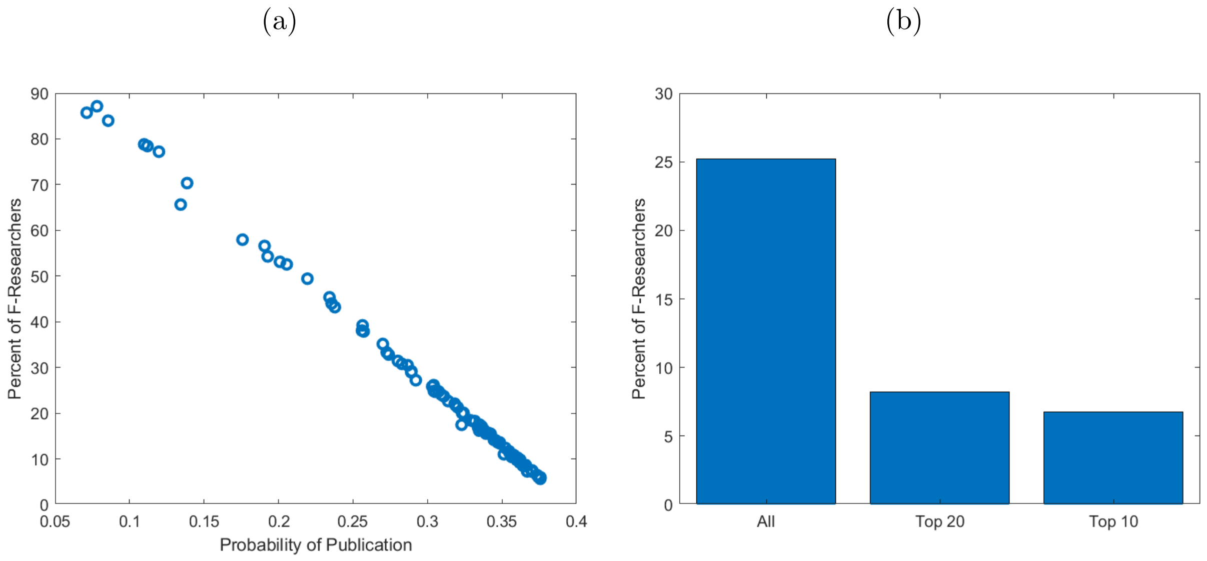
(b) Data

Panel (a) plots the average quality of accepted M and F researchers in the model. The quality is measured as $\sum_\theta L^\theta w_t^{\theta,g}$ where $L^\theta = \sum_{n=1}^N \theta_n$ and $w_t^{\theta,g} = a_t^{\theta,g} / \sum_{\theta'} a_t^{\theta',g}$, $g = f, m$. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9p^m + 0.1p^f$. Panel (b) reports Figure 4 (a) of Card, Della Vigna, Funk, and Iriberri, "Are referees and editors in economics gender neutral?" *Quarterly Journal of Economics*, 135, 2020, which plots the average (residualized) citation rate of non-desk rejected papers across types of referee recommendation.

the top-10, top-20, and, respectively, all institutions. This plot should be compared with the bottom panel in Figure 1. While the absolute levels are smaller in our model, the close match is surprising, given that our model has no group bias at all.

Our model has additional implication for the time-series dynamics of the fraction of F-

Figure 6: Publishing Intensity and F-researchers underrepresentation


(a)

(b)

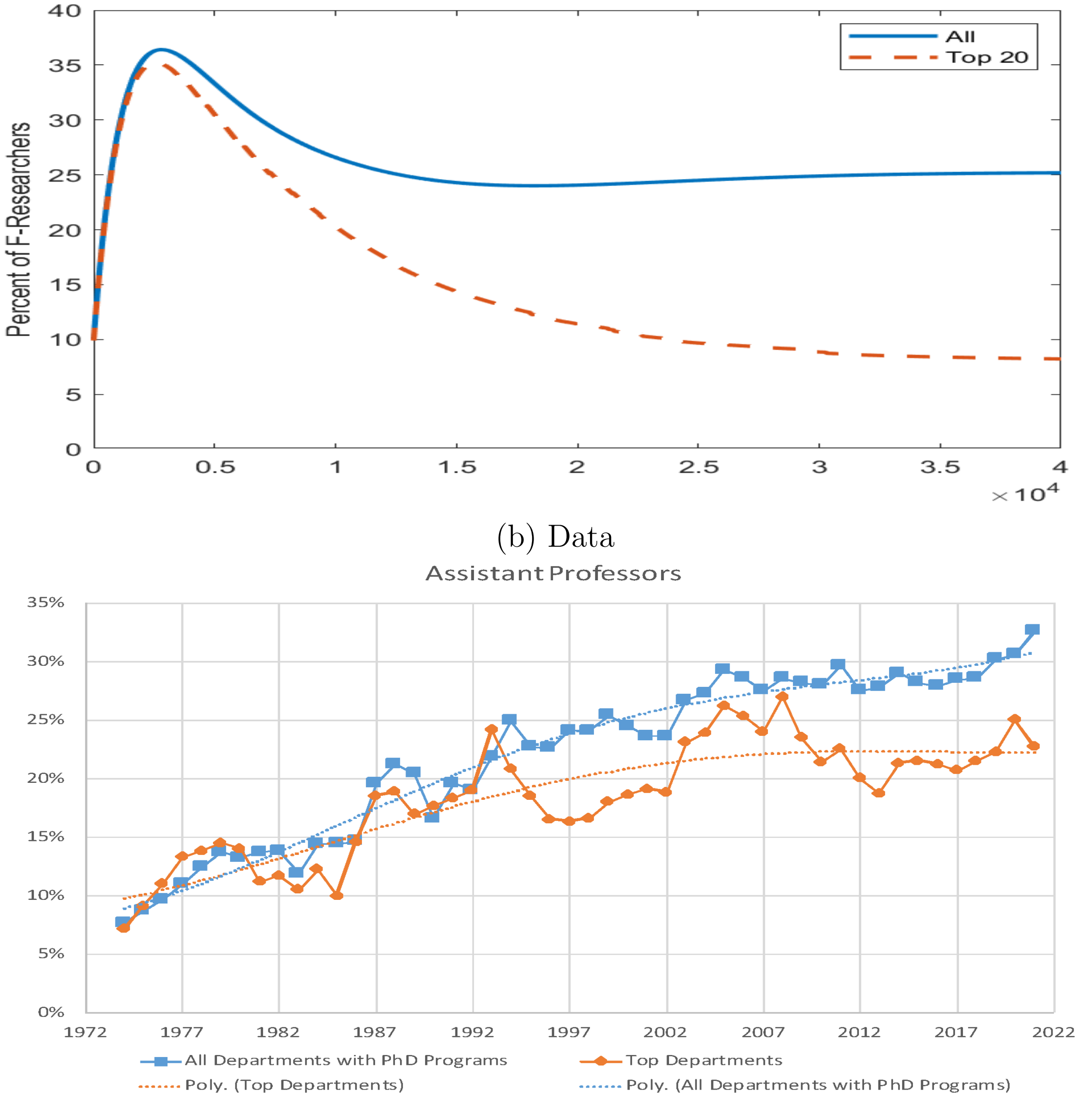
Panel (a) presents the scatterplot of 100 simulated institution publishing probabilities ($x-$axis) versus the F-researcher representation in the same institution. Panel (b) ranks the top 10 and top 20 institution by publishing probability, and compare the average F-representation against the overall population of institutions. Both panels are from the simulation of the model and plotted at the last period. Initially $\lambda_0 = 0.9p^m + 0.1p^f$. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$.

researchers over time in top institutions. This is shown in panel (a) of Figure 7. In our model, the gap between the fraction of F-researchers in top institutions vs. all institutions increases over time. Intuitively, sorting institutions by their publishing intensity implicitly defines "top institutions" as those whose types are increasingly similar to the types of the majority of referees—that is, in the limit, $\overline{\lambda}^{\theta m}$. By construction, the remaining institutions will have a larger fraction of researchers that are less represented in the refereeing population.

Panel (b) of Figure 7 indeed shows that the percentage of women assistant professors in top universities used to be similar to those across all PhD granting universities in the 1970s and 1980s, albeit small for all institutions. However, over time, the gap between the top universities and all PhD-granting institutions has increased, as predicted by our model. The plot also shows the 4^{th}-order polynomial trends of the two lines, denoting the different convergence endpoints. We caveat these results, however, by noticing that from 1974 to 1993, CSWEP defines "top institution" as those above the median according to the National Research Council rankings, while in the 1994 to 2021 data, CSWEP defines "top institutions" as the top 20 schools. Still, the gap is visibly increasing also just in the latter sample.

We conclude this section by returning to the prediction that, in the limit, we should

Figure 7: The Dynamics of F Researchers in Top Research Institutions: Model vs. Data



(a) Model

(b) Data

Panel (a) reports the fraction F-researchers in top research institutions and across all institutions as simulated from the model. Top research institutions are the top 20 out 100 with the highest frequency of publication. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9p^m + 0.1p^f$. Panel (b) reports the fraction women assistant professors in top institution vs. all institutions with a PhD program from 1974 to 2021. Data from 1974 to 1993 were extracted from Figures 2 and 3 of the 1994 CSWEP report available at `https://www.aeaweb.org/content/file?id=682`, while data from 1994 onward are from the 2021 CSWEP report. For the former sample, CSWEP defined "top" as "above median department" according to the National Research Council rankings, while for the latter sample, CSWEP defined "top" as the "top 20" schools. The figure also reports 4^{th}-order polynomial trend lines.

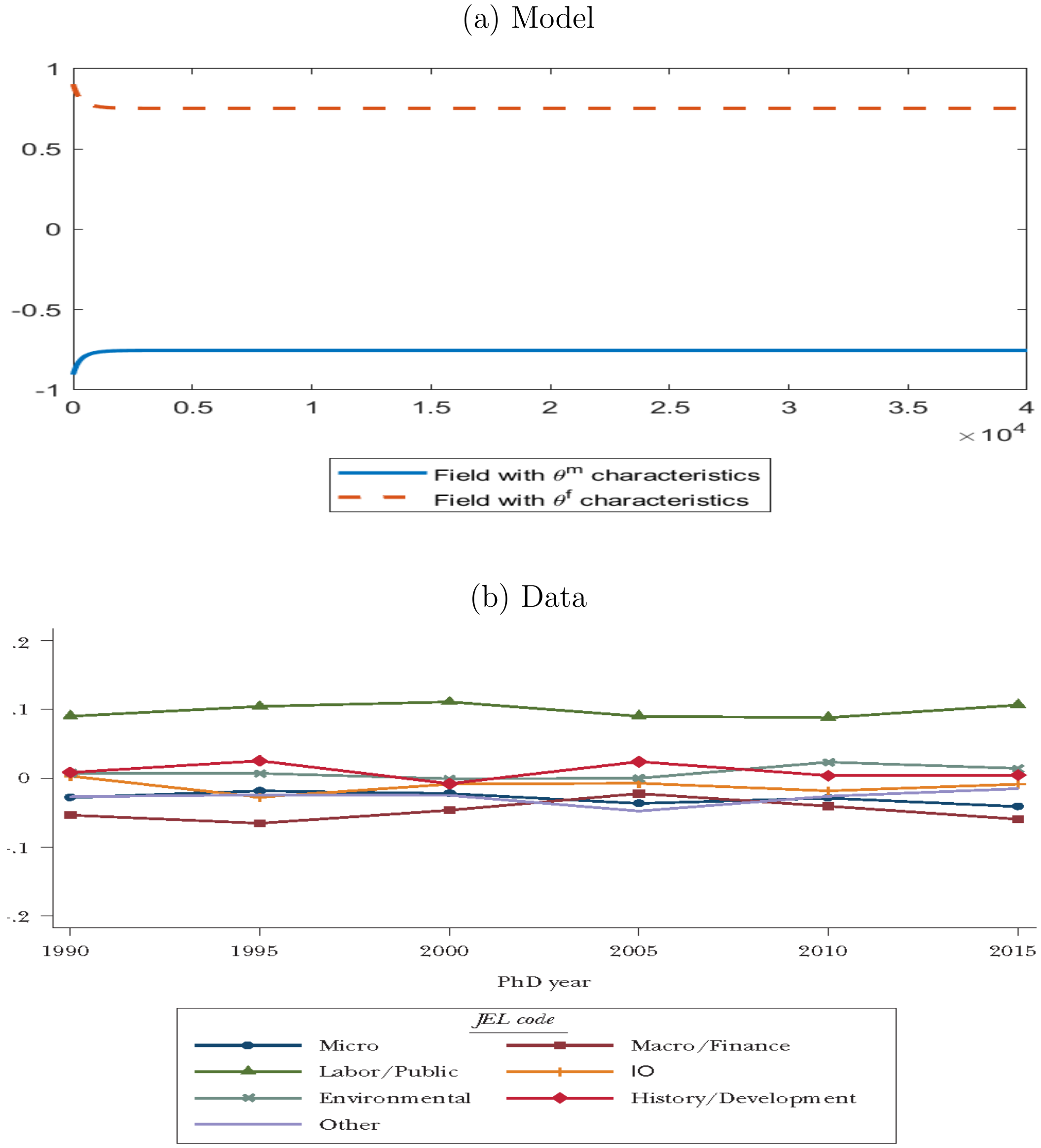
see clustering of "fields," if we assume that at least some characteristics are more valuable and hence prevalent in certain fields. Specifically, consider types θ^m and θ^f (see Corollary 4). Panel (a) of Figure 8 plots the difference between the share of F- and M-researchers across the θ^f-field and the θ^m-field. As also shown in Corollary 4, F-researchers are relatively more represented in the former (top line), and M-researcher relatively more the latter (bottom line). Panel (b) reproduces the same statistics, but in the data. Specifically, Panel (b) reproduces Figure 5 in Lundberg and Stearns (2019) which reports the difference between share of women and share of men in particular fields of economics. The data used are from the annual list of Doctoral Dissertations in Economics, from 1991–2017. As it can be seen, there is nearly no change over time in gender representation in different fields: Women concentrate the most in Labor/Public economics and men concentrate the most in Micro and Macro/Finance, with nearly no variation over almost three decades. While our model with only two fields in the limit (θ^f and θ^m) shows a more extreme behavior, the pattern in the data, with its lack of dynamics, is consistent with our model's predictions.

In conclusion, our model based on self-image bias can explain several of the dynamic patterns observed in the data: The apparent convergence to a steady state with a strong gender bias towards men, the fact that top research institutions display a stronger bias that increases through time, that women are held to a higher standard, and that the choice of fields of study between women and men differ and do not change over time. Moreover, F-underrepresentation comes at a welfare loss: In the language of Section 2.4.6. and consistently with Proposition 8, in the calibrated model society resolves 57% of problems that require more M-characteristics but only 34% of problems that require more F-characteristics.[17] The strong asymmetry in the solution of real-world problems may lead society to suffer a loss in innovation, consistently with the evidence of Bell et al. (2019).

Not only our model predicts these results, but it is not clear how taste-based discrimination and statistical discrimination would predict the same collection of dynamic patterns. Do economists in macro/finance discriminate more than those in labor and public economics? Do economists in top institutions discriminate more than the others? The American Economic Association has started several new initiatives in the last 20 years (e.g. several mentoring programs, such as CeMENT, mentoring breakfasts, professional development initiative, AEA summer economic fellows program) but, while welcome, they do not appear to have generated major changes to the economics profession. The alternative simple explanation is that self-image bias is the main underlying reason of the current state of the economic profession, and therefore a different set of policies should be considered to decrease the talent loss, as

[17] Problems that require more M-characteristics are defined as those with $n_m(\vartheta) > n_f(\vartheta)$, and viceversa.

Figure 8: Difference between Share of F- and M-researchers across Fields: Model vs. Data



(a) Model

(b) Data

Panel (a) reports the difference in share of F-researchers and M-researcher in the field with characteristics θ^m and θ^f, that is, respectively, $\frac{\lambda^{\theta^m,f}}{(\lambda^{\theta^m,f}+\lambda^{\theta^f,f})} - \frac{\lambda^{\theta^m,m}}{(\lambda^{\theta^m,m}+\lambda^{\theta^f,m})}$ and $\frac{\lambda^{\theta^f,f}}{(\lambda^{\theta^f,f}+\lambda^{\theta^m,f})} - \frac{\lambda^{\theta^f,m}}{(\lambda^{\theta^f,m}+\lambda^{\theta^m,m})}$. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9p^m + 0.1p^f$. Panel (b) reports Figure 5 from "Women in Economics: Stalled Progress" by Lundberg and Stearns, *Journal of Economic Perspectives*, 33, 1, 2019. This figure reports the difference between share of women and share of men in particular fields of economics. Data from the annual list of Doctoral Dissertations in Economics, 1991–2017, was collapsed into five-year bins for smoothness. The 1990 bin contains data from 1991 to 1994 and the 2015 bin contains data from 2015 to 2017; all other bins contain five years of data.

discussed in Section 4.

4. The Impact of Policy Actions

In this section we discuss the impact of policy actions that have been proposed to address gender imbalance. We consider (*i*) the impact of mentoring (section 4.1.); and (*ii*) the impact of affirmative action (section 4.2.).

4.1. Mentoring: Group Balance versus Talent Loss

The adoption of mentoring to improve the prospects of female economists is one of the most popular proposals. Indeed, there is evidence that mentoring does help increase the success rate of female economists (Ginther, Currie, Blau, and Croson (2020)). We now investigate the implications of mentoring in our model.

We assume that at the beginning of each period t every young researcher of type θ is randomly matched with an advisor a of type θ^a drawn from the established group, whose mass is $\lambda_{t-1}^{\theta^a}$. Upon matching, the researcher of type θ can choose to pay a cost $C(\theta, \theta^a)$ to "become" the same type of the advisor. Assume that P is the payoff from being hired and U is the utility from an outside option. Researcher θ will then pay the cost if and only if

$$\gamma^{\theta^a}\lambda_{t-1}^{\theta^a}\left(P - C(\theta, \theta^a)\right) + \left(1 - \gamma^{\theta^a}\lambda_{t-1}^{\theta^a}\right)\left(U - C(\theta, \theta^a)\right) > \gamma^{\theta}\lambda_{t-1}^{\theta}P + \left(1 - \gamma^{\theta}\lambda_{t-1}^{\theta}\right)U$$

That is, a young researcher θ pays the cost if and only if

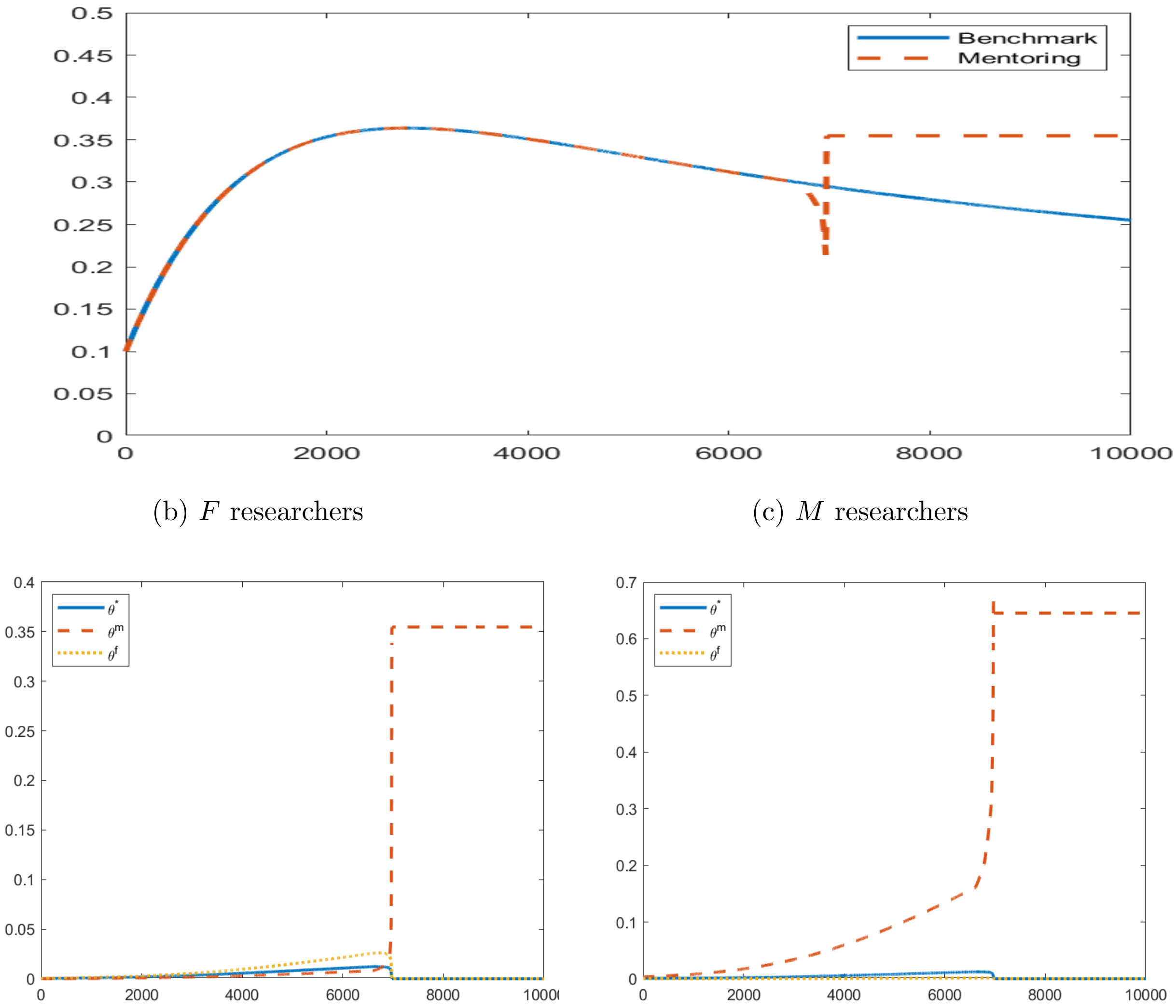
$$\widetilde{C}(\theta, \theta^a) = \frac{C(\theta, \theta^a)}{P - U} < \gamma^{\theta^a}\lambda_{t-1}^{\theta^a} - \gamma^{\theta}\lambda_{t-1}^{\theta}$$

In words, the increase in the probability of getting hired must be sufficiently high relative to the cost of undergoing mentoring. For instance, if the right-hand-side was negative (type θ is already likely to succeed), nobody of that type would pay such a cost.

We assume that the cost itself depends on the distance between the young researcher's type θ and the type of the advisor θ^a: The larger the distance and the higher the cost, indicating that it will take a higher effort to "learn" to become a type that is likely to be hired. Note that such distance may be high as the young researcher θ may have some characteristics that are desirable from an objective standpoint, but that are not viewed as important or relevant by the majority of established researchers. The cost, in that case, is to "unlearn" what is deemed "irrelevant."

The Online Appendix contains the details of the system dynamics. For brevity, we only provide the intuition here. Panel (a) of Figure 9 illustrates the dynamics resulting

Figure 9: Costly Mentoring



(a) Fraction of F-Researchers

(b) F researchers

(c) M researchers

Fraction of M and F researchers (panel (a)), and mass of established $F-$researchers (panel (b)) and of M-researchers (panel (c)) under costly mentoring. Initial $\lambda_0 = 0.9p^m + 0.1p^f$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$, cost function $C(\theta, \theta') = 0.0750 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

from Eq.(A.31), under the same parameters as in Section 3. and a cost function $C(\theta, \theta') = \beta \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$, with $\beta = 0.075$. We choose this cost so that not all of the young researchers want to pay the switching cost to become like their advisors, which seems plausible. The resulting steady state is roughly consistent with the percentage of female participation in economics.

Initially, the dynamics are as in the base case, as all θ_t^θ are small and thus no young researcher wants to pay the cost of mentoring. In this dynamics, as we know, $\theta_t^{\theta^m}$ and $\theta_t^{\theta^f}$ increase, with the former increasing faster, as shown in panel (c) of Figure 9. At some point, the mass of $\lambda_t^{\theta^m}$ becomes large enough to induce many young researchers, both M and F, to pay the mentoring cost, and the system (nearly) jumps. The reason is that many young researchers now expect that their advisor will likely be of type θ^m, which is also the type of established researchers who will evaluate their research. They are thus happy to pay the cost and become like their advisors.

The bottom panels of Figure 9 show, however, that the mass of young M-researchers jumps by more than the mass of F-researchers. The reason is that even though the cost function is the same for M- and F-researchers, young M-researchers are on average closer to θ^m and thus have have systematically lower cost to switch than F-researchers. For this reason, group imbalance persists forever.[18] Moreover, only type θ^m survives and therefore the research characteristics mildly more common in the F-population, but also very common in the M-population, disappear, thus yielding talent loss and loss of knowledge.

4.2. Affirmative Action

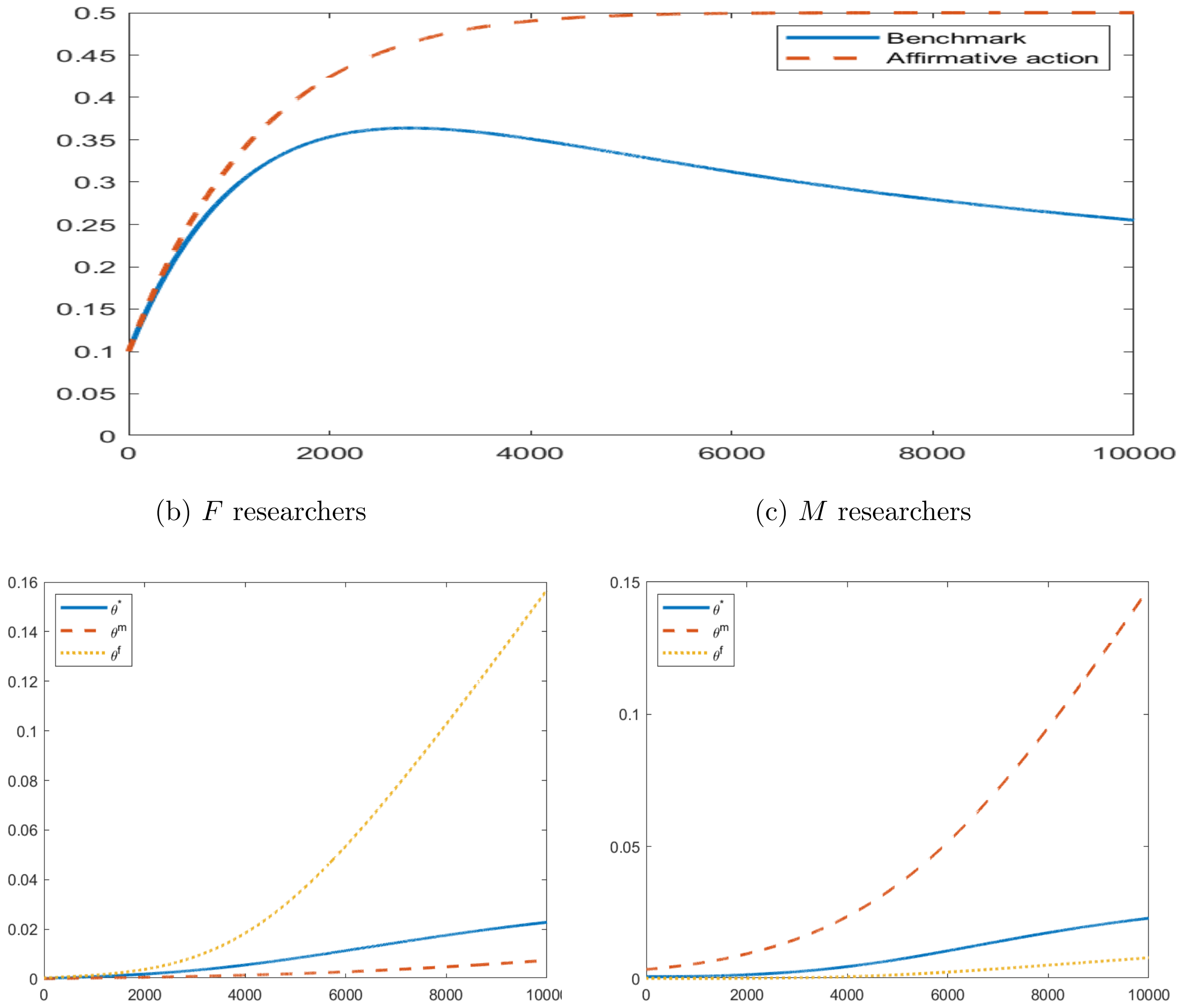
A common policy to increase diversity is "affirmative action", which effectively increases the representation of specified groups by mandate. We consider a simple rule in this section: in each round, it is mandated that evaluators must hire the same number of M and F researchers. We change just one assumption to the dynamics in the benchmark case: namely, we require that

$$a_t^{\theta,m} = k_t \; \gamma^\theta \; \lambda_{t-1}^\theta \; p^{\theta,m} \qquad \text{where} \qquad k_t = \frac{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',f}}{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',m}}. \tag{19}$$

The scaling factor k_t ensures that $\sum_\theta a_t^{\theta,f} = \sum_\theta a_t^{\theta,m}$. Figure 10 provide the dynamics for this case. The affirmative action policy reaches group balance, which is not surprising, given the definition of k_t. However, it also attains diversity in research characteristics: in the limit, M researchers are of type θ^m and F researchers are of type θ^f. Assuming that maximizing the representation of research characteristics is beneficial to society (see Section 2.4.6.), this policy appears superior to mentoring, as it does not skew the distribution of such characteristics towards θ^m even when reaching group balance.

[18] If the cost function was lower, however, then *all* young researchers, M and F, would pay the cost and the system would jump to group balance. This extreme case is illustrated in Figure A.12 in the Online Appendix.

Figure 10: Affirmative Action



(a) Fraction of F-Researchers

(b) F researchers

(c) M researchers

Fraction of F researchers (panel (a)), mass of established F-researchers (panel (b)) and of M-researches (panel (c)) when an affirmative action policy requires to accept the same number of M and F researchers. Parameters: $\phi = 0.5742$ $(d = 0.3)$, $\gamma_0 = 0.2$, $\rho = 4$, and $N = 10$. Initial $\lambda_0 = 0.9p^m + 0.1p^f$

Intuitively, by expanding the set of referee characteristics, affirmative action makes it possible to reward the research of *talented* F researchers—those who are more likely to produce quality research. It is still the case that F researchers who are not (objectively) as productive will not survive in the limit and will be weeded out from the system.

5. Extensions

In this section we summarize the results of numerous extensions of the model; detailed results are in the Online Appendix. In particular, we study the impact of candidates' decisions to apply to positions, or institution decisions to hire (Section 5.1.); a two-layer hierarchy of junior and senior established researchers (Section 5.2.); attribution of co-authored work (Section 5.3.); more general distance functions between established researchers and juniors (Section 5.4.).

5.1. Endogenous Entry

Section A1.1. endogenizes the choice of candidates to apply for a position (Section A1.1.1.), and (symmetrically) the hiring decision of institutions (Section A1.1.4.).

5.1.1. Endogenous Application Decision

We assume that candidates may choose a career in academia, in which case their success depends on the judgement of the established group of researchers as described in Section 2., or can opt for a different career. If they choose the academic career, candidates pay a utility cost C but receive a payoff P if successful. The outside options gives a benchmark utility of zero. We obtain several results:

1. When the relative cost C/P is small and below a cutoff, the same equilibrium and results as in Section 2.3. obtain;
2. When the relative cost C/P is intermediate, between two cutoffs, the equilibrium changes, and only type θ^m survives in the limit. In this case, only characteristics that are mildly more common in the M-population survive.
3. The pool of applicants skew towards the M-population. That is, imbalance occurs even in the "pipeline."

Thus, there are parameter values that would lead to group balance in our main model, for which costly entry generates imbalance. The cost of entry into the profession tends to keep F-researchers out from the beginning, which may explain the low percentage of women applications to PhD program, for instance.

5.1.2. The Endogenous Choice of Hiring Institutions

The choice of hiring institutions can be modelled in a symmetric fashion. In particular, suppose an institution receives a reward P (e.g. grant money, better students etc.) from hiring a candidate who turns out to be successful, and bears an cost C for all hires, independently of their publication success. We show that the equilibrium is identical to the one in the previous subsection. In particular, for intermediate values of C/P, only type θ^m survives. This result emphasize the role of hiring institution that base their decisions only on publication potential (which, in our model, is endogenously determined by the prevailing distribution of types among established researchers) rather than objective quality. It suggests an additional (endogenous) reason why gender imbalance is especially strong in research-oriented institutions.

5.2. Seniors and Juniors

Next, we enrich the model to mimic the tenure-track process. In each period, young researchers apply for positions, and are evaluated by all established researchers. If they are hired, they become "junior" established researchers. In the following period, these junior researchers are put "up for tenure," and must be further evaluated by senior researchers only. At each step of the tenure process, the evaluation mechanism is as in our basic model, but the evaluator is drawn from a different population—all established researchers vs. only seniors.

The headline finding in this model is that, under suitable parameter values, a *leaky pipeline* can arise: that is, the representation of F researchers is higher among juniors than among seniors. As we noted in the Intoduction, the CSWEP report (Chevalier, 2020) highlighted such a pattern in the data.

5.3. Co-Authorship

We briefly examine co-authorship between a M author and an F author. We assume that, if coauthors a, b with types θ^a and θ^b write a paper, the quality of the finished product will be the component-wise maximum of θ^a and θ^b. We can then compute the expected objective quality (number of research characteristics) of θ^a and θ^b, conditional on the paper being accepted in an economy in which gender imbalance prevails—hence, conditional on the component-wise maximum being either θ^m or θ^f.

The main finding is that, if a is drawn from the M population and b is from the F population, then the expected quality of a is higher than that of b. The intuition is that, since referees are most likely to be of type θ^m, if the paper was accepted, its quality is likely to be θ^m; and since M researchers are more likely to possess characteristics $n = 1, \ldots, N/2$, they are more likely to have contributed such characteristics. This conclusion is in line with findings in Sarsons et al. (2021).

5.4. General Distance Functions from Referees

Finally, we explore less extreme forms of self-image bias. One approach that is closest to the model in Section 2. is to assume that a referee of type θ^r accepts researchers whose type θ is no more than some distance $\eta \geq 0$ away from her own, where distance is defined as

$$D(\theta^r, \theta) = \sum_{n=1}^{N} (\theta_n^r - \theta_n)^2 = \#\{n = 1, \ldots, N : \theta_n^r = \theta_n\}.$$

The model in the main text corresponds to $\eta = 0$.

For $\eta > 0$, we have three main results.

- If the set Θ of types is "connected," in the sense that one can reach any type θ' starting from any other type θ by a sequence of steps of "size" no more than η, then group balance obtains in the limit. This is in particular the case if $\Theta = \{0, 1\}^N$, as in the main body of the paper.
- If instead we consider a strict subset $\Theta \subsetneq \{0, 1\}^N$ that can be separated into two or more subsets whose elements are more than η apart, then imbalance can obtain.
- Finally, with endogenous costly entry, depending on parameter values, group imbalance can obtain for *any* specification of the type set Θ.

6. Literature Review

There is a considerable body of research on the underlying reason of under-representation of women in the economics profession. We do not attempt an exhaustive survey here, but refer the reader to Bayer and Rouse (2016), who review the literature on both "supply-side" and "demand-side" factors. Among supply-side factors, these authors argue that prior exposure to economics, as well as the performance in introductory courses, and the lack of role models all have documented effects on the gender imbalance in applications to Economics Ph.D. programs. On the other hand, the evidence suggests that differences in math preparation do not explain a significant fraction of the imbalance. On the demand side, Bayer and Rouse (2016) suggest that policy changes in most academic institutions have diminished, if not completely removed, the impact of explicit or statistical discrimination in recruiting Ph.D. students. At the same time, these authors argue that the literature suggests that an important role is played by *implicit bias* and *stereotyping*. Our model with self-image bias is consistent with the persistence of gender bias even when all structural sources of gender-biases have been removed.

In a more recent contribution, Sarsons et al. (2021)'s work on recognition for coauthored papers shows that, for men, an additional coauthored paper has the same effect on the likelihood of tenure as a solo-authored paper; however, for women, coauthorship entails a significant "discount factor," especially if the coauthor(s) are men. The large body of research on the gender pay gap and on the "glass ceiling" in other labor markets is also indirectly relevant in our context: see e.g. Blau and Kahn (2017); Goldin and Rouse (2000); Goldin (2014); Weber and Zulehner (2014); Aigner and Cain (1977); Lazear and Rosen (1990).

On the theoretical side, our model is related to the literature on statistical discrimination: a relative recent survey is Fang and Moro (2011). One strand within that literature, originating from Phelps (1972), posits the existence of exogenous differences between groups, either in the distribution of productivity ("Case 1"), or in the quality of signals about it ("Case 2"). In Case 2, the employer does not observe the productivity of individual applicants, but receives a signal about it. Differential average treatment of the two groups can emerge either through risk aversion of the employer (Aigner and Cain, 1977), investment in human capital (Lundberg and Startz, 1983), or if hiring occurs in a tournament (Cornell and Welch, 1996). In Conde-Ruiz, Ganuza, and Profeta (2020), the difference in signal quality leads members of the group in the minority of a hiring committee to underinvest in human capital; this perpetuates the imbalance. A recent contribution, Bardhi, Guo, and Strulovici (2019), revisits Phelp's Case 1, but assume that success or failure is observed over time and is informative about the worker's type. This can lead to large differences in ex-post treatment of the two groups, even if ex-ante productivity differences are small. Differently from this literature, in our model the ex-ante distributions of productivity are the same in the M and F group, because all characteristics are equally valuable. Furthermore, productivity is observed. In our model, standard statistical discrimination does not lead to gender imbalance.

Becker (1957)'s model of taste-based discrimination instead posits that employers may have a preference for hiring members of one specific group. This is not the case in our model:

while referees only accept applicants whose research characteristics match their own, they do not take group membership into consideration at all.

Heidhues, Kőszegi, and Strack (2019) proposes a model in which an agent's ability is unobserved, both by herself and by others. Agents belong to different groups, each potentially subject to "discrimination," and are "stubbornly overconfident" about their own ability. Overconfidence leads agents to have a more favorable view of individuals in their own social group, ascribing poor performance to discrimination against them. In our model, ability is observed, and there is no exogenously imposed discrimination on either group. Incorporating (possibly biased) learning (cf. e.g. Bohren, Imas, and Rosenberg, 2019) about young researchers' characteristics is an interesting direction for future work.

7. Conclusions and Policy Implications

Our model highlights a novel mechanism that endogenously perpetuates specific research characteristics over time without relying on implicit or explicit gender bias. This occurs due to self-image bias, grounded in the psychology literature, and its application to the reviewing process: established researchers use their own personal research characteristics as a guidance to judge others' output. Findings in psychology and experimental economics point to mild between-group heterogeneity; yet, in our model, such mild differences are enough to lead the initially prevalent group to dominate forever. It is *as if* the initially dominant group decided for society what are the important research characteristics and topic.

Our results are consistent with several empirical regularities, in addition to the trends illustrated in Figure 1. First, gender imbalance can persist long after steps are taken to eliminate outright, or structural, gender bias (see Bayer and Rouse, 2016): if evaluators are predominantly male due to past discrimination against women, our model predicts that self-image bias will perpetuate this imbalance forward. Second, our model implies that women are held to higher standards (Card et al., 2020; Dupas et al., 2021) and receive less credit for joint work with co-authors (Sarsons, 2017; Sarsons et al., 2021). Third, it is consistent with a different representation of women across fields (Chari and Goldsmith-Pinkham, 2018). Fourth, it predicts that the under-representation of women should be especially severe in research-oriented institutions, and that the gap in female under-representation between top institutions and the rest should increase through time (Chevalier, 2020, Figure 1, and Figure 6). Finally, it can generate a "leaky pipeline," with women applying less to Economics PhD programs and their representation being lower the higher the rank (Chevalier, 2020).

Standard solutions to the gender bias problem may not be very effective in our model. For instance, outreach programs to encourage members of a given group to apply to PhD programs may prove ineffective. Such outreach program are akin to lowering the cost of doing research, which we explore in an extension of the model in Section 5.1. While lowering the cost may indeed switch the path towards convergence for some parameter configurations, our basic model in Section 2. assumes zero costs and yet, under the conditions of Proposition 3 (2.a), and, in fact, under our calibrated parameters, gender bias persists.

Similarly, mentorship programs for female researchers will only be effective to increase female representation in the profession insofar as they induce female researchers to adopt those characteristics that are prevalent in the reviewer population (see Section 4.1.). While this may improve female participation (as it has: see e.g. Ginther et al., 2020), it still propagates the bias towards male research characteristics. This leads to under-representation of valuable research characteristics relative to the efficient benchmark.

Because the problem is self-image bias, the best policy intervention must involve limiting the ability of reviewers to use their own research style as a yardstick while judging others' research. One solution is to provide strict guidelines in the refereeing process. Indeed, in light of Proposition 1 and 2, editors should guide referees to limit the number of aspects of the submitted research paper they should focus on. For instance, a journal may provide questionnaires with precise, pointed questions and explicitly ask referees to leave aside other judgemental elements that are most susceptible to self-image bias. Dunning, Meyerowitz, and Holzberg (1989) provides suggestive evidence in support of this approach.

Another solution is instead to change the reviewing process to include input from the full distribution of researchers, as opposed to just the established ones. While radical as a proposal, it would be reasonable to consider an editorial policy that requires young researchers to participate in the evaluation process, or in fact, "oversample" young female researchers.

Our model suggests a novel rationale for affirmative-action policies: diversifying the pool of reviewers. In our model, scientific progress requires a combination of all research characteristics, regardless of whether they are more prevalent among males or females—because all such characteristics are equally productive. If males are initially dominant, they will remain so, and research characteristics more prevalent among females will be under-represented. Facilitating the promotion of female researchers counteracts this force, and leads to a more balanced representation of research characteristics in the steady-state population. While facilitating the promotion of female researchers may be perceived by the (dominant) male population as lowering the standards of the profession, we show that this trade-off is a mis-perception that is due itself to self-image bias.

Finally, in this paper we emphasize gender discrimination in academia. However, a similar force may help explain discrimination against other groups and in other settings. Even if evaluators are group-neutral in their reviews, self-image bias may lead majority evaluators to unconsciously fail to promote socially valuable characteristics that are (possibly slightly) more prevalent in an underrepresented group. We leave this investigation to future research.

References

Dennis J. Aigner and Glen G. Cain. Statistical theories of discrimination in labor markets. *ILR Review*, 30(2):175–187, 1977.

Steffen Andersen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics*, 95(4):1438–1443, 2013.

Peter Andre and Martin Falk. What's worth knowing? economists' opinions about economics. ECONtribute Discussion Paper 102, Bonn and Cologne, 2021. URL `http://hdl.handle.net/10419/237347`.

Emmanuelle Auriol, Guido Friebel, Alisa Weinberger, and Sascha Wilhem. Women in economics: Europe and the world. mimeo, Toulose School of Economics, January 2022.

Arjada Bardhi, Yingni Guo, and Bruno Strulovici. Spiraling or self-correcting discrimination: A multi-armed bandit approach. Technical report, Technical report, Northwestern University, 2019.

Amanda Bayer and Cecilia Elena Rouse. Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives*, 30(4):221–42, 2016.

Gary S Becker. *The economics of discrimination.* University of Chicago press, 1957.

Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly Journal of Economics*, 134(2):647–713, 2019.

Michael Betz, Lenahan O'Connell, and Jon M Shepard. Gender differences in proclivity for unethical behavior. *Journal of Business Ethics*, 8(5):321–324, 1989.

Francine D. Blau and Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, 2017.

J Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019.

Lex Borghans, Bart H.H. Golsteyn, James J. Heckman, and Huub Meijers. Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3):649–658, 2009.

David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Are referees and editors in economics gender neutral? *Quarterly Journal of Economics*, 135:269–327, February 2020.

Magnus Carlsson, Henning Finseraas, Arnfinn H Midtbøen, and Gudbjörg Linda Rafnsdóttir. Gender bias in academic recruitment? evidence from a survey experiment in the nordic region. *European Sociological Review*, 37(3):399–410, 2021.

Christoph Carnehl and Johannes Schneider. A quest for knowledge. *arXiv preprint arXiv:2102.13434*, 2021.

Anusha Chari and Paul Goldsmith-Pinkham. Gender representation in economics across topics and time: Evidence from the nber summer institute. Technical report, Working Paper, Yale University, 2018.

Judy Chevalier. Report: committee on the status of women in the economics profession. Technical report, American Economic Association, 2020.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

J. Ignacio Conde-Ruiz, Juan José Ganuza, and Paola Profeta. Statistical discrimination and committees. mimeo, Universitat Pompeu Fabra, December 2020.

John P. Conley and Ali Sina Önder. The research productivity of new phds in economics: The surprisingly high non-success of the successful. *Journal of the Economic Perspectives*, 28(3):205–216, 2014.

Bradford Cornell and Ivo Welch. Culture, information, and screening discrimination. *Journal of Political Economy*, 104(3):542–571, 1996.

Paul T Costa, Antonio Terracciano, and Robert R McCrae. Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2):322, 2001.

Faye J Crosby, Aarti Iyer, Susan Clayton, and Roberta A Downing. Affirmative action: Psychological data and the policy debates. *American Psychologist*, 58(2):93, 2003.

Rachel Croson and Uri Gneezy. Gender differences in preferences. *Journal of Economic literature*, 47(2):448–74, 2009.

Marcus Dittrich and Kristina Leipold. Gender differences in time preferences. *Economics Letters*, 122(3):413–415, 2014.

Anna Dreber and Magnus Johannesson. Gender differences in deception. *Economics Letters*, 99(1):197–199, 2008.

David Dunning, Judith A Meyerowitz, and Amy D Holzberg. Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 57(6):1082, 1989.

David Dunning, Marianne Perie, and Amber L Story. Self-serving prototypes of social categories. *Journal of Personality and Social Psychology*, 61(6):957, 1991.

Pascaline Dupas, A Modestino, Muriel Niederle, and Justin Wolfers. Gender and the dynamics of economics seminars. mimeo, February 2021.

Brian Fabo, Martina Jancokova, Elisabeth Kempf, and Lubos Pastor. Fifty shades of qe: Conflicts of interest in economic research. Technical report, University of Chicago, 2020.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692, 2018.

Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, volume 1, pages 133–200. Elsevier, 2011.

First Round Review, 2022. Eight ways to make your d&i efforts less talk and more walk. `https://review.firstround.com/eight-ways-to-make-your-dandi-efforts-less-talk-and-more-walk`, 2022. Accessed: 2022-05-24.

Donna K Ginther, Janet Currie, Francine D Blau, and Rachel Croson. Can mentoring help female assistant professors in economics? an evaluation by randomized trial. Technical report, NBER, March 2020. Working Paper 26864.

Claudia Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119, 2014.

Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

Luigi Guiso, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. Culture, gender, and math. *Science*, 320(5880):1164, 2008.

Paul Heidhues, Botond Kőszegi, and Philipp Strack. Overconfidence and prejudice. *arXiv preprint arXiv:1909.08497*, 2019.

Thomas Hill, Nancy D Smith, and Hunter Hoffman. Self-image bias and the perception of other persons' skills. *European Journal of Social Psychology*, 18(3):293–298, 1988.

Janet Shibley Hyde. Gender similarities and differences. *Annual Review of Psychology*, 65: 373–398, 2014.

Janet Shibley Hyde and Marcia C. Linn. Gender similarities in mathematics and science. *Science*, 314(5799):599–600, 2006.

Edward P. Lazear and Sherwin Rosen. Male-female wage differentials in job ladders. *Journal of Labor Economics*, 8(1, Part 2):S106–S123, 1990.

Pawel Lewicki. Self-image bias in person perception. *Journal of Personality and Social Psychology*, 45(2):384, 1983.

Shelly J Lundberg and Richard Startz. Private discrimination and social intervention in competitive labor market. *The American Economic Review*, 73(3):340–347, 1983.

Shelly J Lundberg and Jenna Stearns. Women in economics: Stalled progress. *The Journal of Economic Perspectives*, 33(1):3–22, 2019.

Thomas Mayer. Honesty and integrity in economics. Technical report, University of California at Davis, 2009. Working Paper 09-2.

Edmund S Phelps. The statistical theory of discrimination. *American Economic Review*, 62 (4):659–661, 1972.

Heather Sarsons. Recognition for group work: Gender differences in academia. *American Economics Review: Papers and Proceedings*, 107:141–45, 2017.

Heather Sarsons, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129, 2021.

Amber L Story and David Dunning. The more rational side of self-serving prototypes: The effects of success and failure performance feedback. *Journal of Experimental Social Psychology*, 34(6):513–529, 1998.

Andrea Weber and Christine Zulehner. Competition and gender prejudice: Are discriminatory employers doomed to fail? *Journal of the European Economic Association*, 12(2): 492–521, 2014.