NBER WORKING PAPER SERIES

SELF-IMAGE BIAS AND LOST TALENT

Marciano Siniscalchi
Pietro Veronesi

Working Paper 28308
http://www.nber.org/papers/w28308

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2020, Revised March 2021

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, Nicola Gennaioli, Lars P. Hansen, Alex Imas, Jessica Jaffers, Seema Jayachandran, Emir Kamenica, Elisabeth Kempf, Lubos Pastor, Jane Risen, Antoinette Schoar, Oleg Urminsky, Adrien Verdhelan, participants at Behavioral Science workshop and the Finance workshop at the University of Chicago, and seminar participants at the Chicago Fed, Bocconi University, and MIT. 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Self-image Bias and Lost Talent
Marciano Siniscalchi and Pietro Veronesi
NBER Working Paper No. 28308
December 2020, Revised March 2021
JEL No. A11,J16,J7

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 with characteristics similar to theirs. Promoted Fresearchers are few and similar to M-researchers, perpetuating imbalance. Candidates' career concerns and institutions' hiring practices exacerbate talent loss. Mentorship reduces group imbalance, but increases F-group talent loss. Affirmative action reduces both. Our mechanism explains existing evidence and suggests different policies.

Marciano Siniscalchi
Department of Economics
Northwestern University
2211 Campus Drive, 3rd Floor
Evanston, IL 60208
USA
marciano@northwestern.edu

Pietro Veronesi
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
pietro.veronesi@chicagobooth.edu

1. Introduction

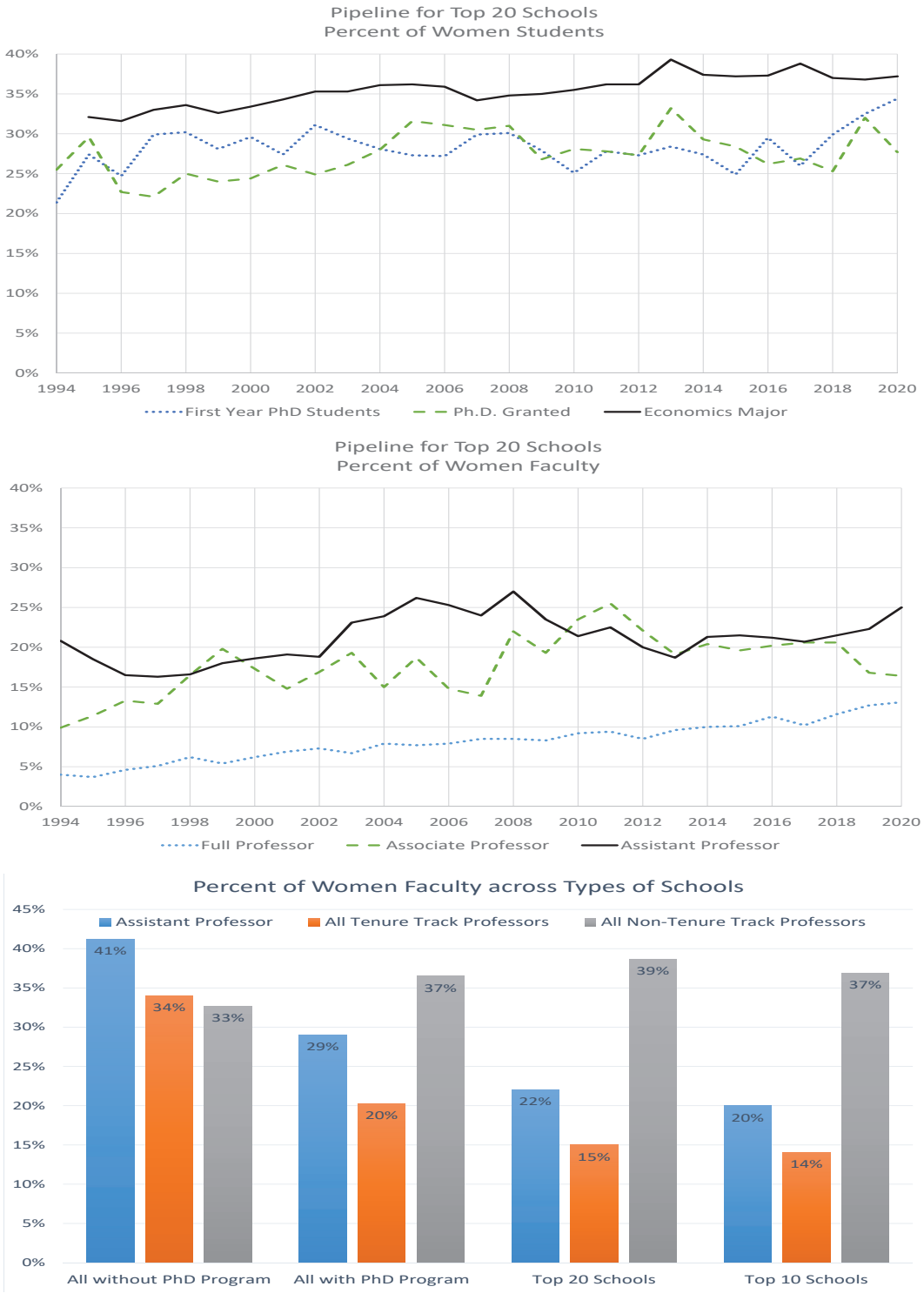
The economics profession has long been male-dominated. The Committee on the Status of Women in the Economics Profession (CSWEP), a standing committee of the AEA since 1971,[1] has been regularly documenting the progress of female economists (or lack thereof): see Chevalier (2020). This phenomenon has 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 almost 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%.

This 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 7. 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), which would then suggest that gender-bias is more prominent in some economic fields than others.

The present paper provides a different perspective on these empirical findings, which suggests alternative policy approaches to addressing gender imbalance. 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

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

Figure 1: Percentage of Women in Academia



Source: CSWEP Report, 2020. The data in the bottom panel are averages over the 2016-2020 sample.

to place more weight on their own positive attributes when judging others, and 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, explains why female under-representation persists in the long run—even when reviewers apply gender-neutral criteria to evaluate others' work—and why this phenomenon is especially concentrated in research-oriented institutions, and differs across fields. In addition, our mechanism predicts that women are held to higher standards and their contributions are valued less than men in co-authored work, notwithstanding gender-neutral evaluation criteria. From a normative standpoint, our model suggests different interventions to promote gender balance and reduce talent loss, as discussed in the concluding section.

To be clear, we do not suggest that outright gender discrimination does not exist in academia. However, our paper identifies a subtle, yet powerful, mechanism that can perpetuate a bias against an initially underrepresented group even if other forms of bias or discrimination are eliminated. This calls for policies that are not just "gender-blind," but instead aim at achieving gender balance and preserving a diversity of talents in the profession.

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.

We first show 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.

Our model thus 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] Similarly, 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). In the online appendix we show that when coauthored work reflects the characteristics of both M and F coauthors, but referees are unaware of each coauthor's characteristics, then conditional on their joint work being accepted, the *expected* objective quality of the M-coauthor increases more than the one of the F-coauthor. 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. Finally, self-image bias implies that M and F researchers cluster around types whose characteristics are (mildly) more common in their own groups, explaining differences in women representation across fields (e.g. Chari and Goldsmith-Pinkham, 2018).

[2]The evidence in Card et al. (2020) is more nuanced and we discuss it in Section 5. 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).

Gender imbalance and loss of talent are exacerbated by candidates' career concerns. We extend the model to allow young agents from both groups F and M to choose whether to pay a cost to become researchers, or enjoy an outside option. Anticipating a bias against their research characteristics, the mass of F-agents who pays the cost 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.

Gender imbalance is also exacerbated by institutions' hiring practices. In a second extension, we assume that 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. Such payoff may be in terms of visibility, recognition, grant money, and so forth. Crucially, institutions anticipate that new hires' research will be reviewed by established scholars who are affected by self-image bias. For this reason, hiring institutions skew the distribution of their hires towards characteristics more prevalent in the M-group and thus exacerbates the loss of talent. This result may explain why "the share female [sic] falls as the research intensity of the department increases (e.g. from top 20 to top 10)" (Chevalier, 2020, p. 15) as shown in the bottom panel of Figure 1. Indeed, consistently with this interpretation, we see little difference in the female share of teaching faculty.

In a further extension, in the online appendix, 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.

We finally investigate the impact of some policy actions. We first investigate the impact of mentorship. 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; young researchers can do so by paying a cost that increases in the distance between them and their advisors. We show that, while mentorship may help reduce (but not necessarily eliminate) gender imbalance, it also accelerates the loss of $F-$group characteristics. Intuitively, mentors are drawn from the dominant population, which over-represents $M-$group characteristics. Thus, young M-researchers have lower costs of switching than F-researchers, on average. Moreover, since referees are drawn from the same dominant population, young F researchers may find it profitable to adopt their mentors' characteristics,

but they also give up their own characteristics. This exacerbates the loss of talent.

We then consider 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.

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 ($d = 0.49$)[3]. 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.

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 dif-

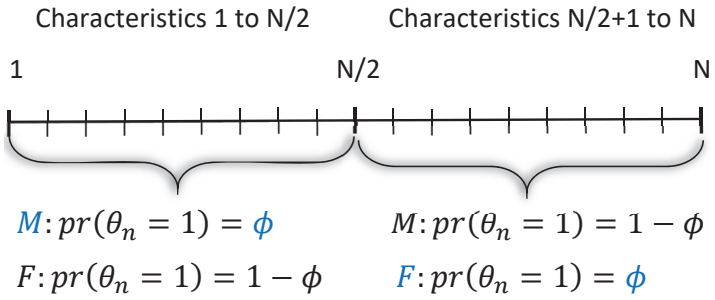
[3]Cohen (2013)'s d measures the standardized mean difference between two populations. $d \approx 0.2$ is considered "small" and $d \approx 0.5$ is considered "medium."

ferences 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 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.

Figure 2: Symmetric Distribution of Research Characteristics



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 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 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 $n = 1, \ldots, \frac{N}{2}$, $Pr(\theta^i_n = 1) = \phi$ for M-researchers and $Pr(\theta^i_n = 1) = (1-\phi)$ 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,

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

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, $\gamma^{(0,\ldots,0)} = \gamma_0$, 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 $\gamma^{(1,\ldots,1)} = \gamma_0 \, \rho \in [\gamma_0, 1]$. 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 $(0,\ldots,0)$ 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 $(1,\ldots,1)$ is four times more likely to produce quality research than the worst researcher, $(0,\ldots,0)$.

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 with 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 $\lambda_t^{\theta,g}$ 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: $\sum_\theta \sum_g \lambda_0^{\theta,g} = 1$.[4] In order

[4] The fact that the total mass of established scholars (a stock) equals the mass of young M and F

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.[5] The resulting dynamic is then described by the following equation:

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

The limiting behavior of this system is readily characterized. First, *initial conditions have no long-run effect*. Eq. (3) shows that $a_t^{\theta,g}$ is time-invariant for $g \in \{f, m\}$; hence, so is a_t^{θ}, and therefore a_t. Then, dropping time indices, for $g \in \{f, m\}$,

$$\lambda_t^{\theta,g} = (1 - a)\lambda_{t-1}^{\theta,g} + a^{\theta,g} = (1 - a)^t \lambda_0^{\theta,g} + a^{\theta,g} \frac{1 - (1 - a)^t}{a} \to \frac{a^{\theta,g}}{a} \tag{5}$$

so the limiting fraction of M- to F-researchers is

$$\frac{\sum_\theta a^{\theta,m}}{\sum_\theta a^{\theta,g}} = \frac{\sum_\theta \gamma^\theta p^{\theta,m}}{\sum_\theta \gamma^\theta p^{\theta,f}}.$$

Second, in our symmetric model, for every type $\theta = (\theta_1, \ldots, \theta_N)$, there is a corresponding type $\bar{\theta} = (\theta_{N/2+1}, \ldots, \theta_N, \theta_1, \ldots, \theta_{N/2})$ such that $p^{\theta,m} = p^{\bar{\theta},f}$ and $\gamma^\theta = \gamma^{\bar{\theta}}$; hence, the above fraction equals 1. This establishes the main result of this section: regardless of initial conditions, the system 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.

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*

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

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

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.

[5] 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.

(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, 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

$$a_t^{\theta,g} = \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g}. \tag{6}$$

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

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g} (1 - a_t) + a_t^{\theta,g} \tag{7}$$

where as above $a_t = \sum_\theta \sum_g a_t^{\theta,g}$. Equations (6) and (7) 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,

$$\theta^m = (1, \ldots, 1, 0, \ldots, 0) \quad \textit{and} \quad \theta^f = (0, \ldots, 0, 1, \ldots, 1); \textit{ or} \tag{8}$$

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

$$\theta^* = (1, \ldots, 1). \tag{9}$$

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.

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

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*

$$\bar{\rho}(\phi, N) = \frac{1}{4}\left(\left(\frac{1-\phi}{\phi}\right)^{N/2} + \left(\frac{\phi}{1-\phi}\right)^{N/2}\right)^2. \tag{10}$$

(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

$$\bar{\lambda}^{\theta^m} = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1}{2}; \quad \bar{\lambda}^{\theta^f} = 1 - \bar{\lambda}^{\theta^m}. \tag{11}$$

(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 that, 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.

2.3.1. 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 4 *Assume that initially $\lambda_0 = p^m$. 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.

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.[6] Indeed, the following Proposition shows 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.

[6] Card et al. (2020) also show that, unconditionally, men- and women-authored papers are equally likely 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 4, one readily sees that young M researchers are more likely to be accepted on average. The model with endogenous choice in Section 5. yields more uniform unconditional acceptance across genders, and fewer female acceptance overall due to self-selection.

Proposition 5 *(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.

The intuition builds upon Proposition 4. 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. 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.[7]

2.3.2. Group Imbalance in the Limit

Proposition 3 mostly concerns the distribution of researcher types irrespective of their group. In the Online Appendix we analyze in detail how the mass of each type θ evolves among M- and F-researchers separately, and also characterize group (im)balance in the limit. Here we report the main result about group imbalance.

Proposition 6 *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 mass of M and F researchers are

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} > 0.5. \tag{13}$$

[7] 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 5 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.

(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 6 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.

2.3.3. 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.

Corollary 3 *In part (a) of Proposition 6, in the limit,*

$$0.5 = \frac{\bar{\lambda}^{\theta^f, f}}{\bar{\lambda}^{\theta^m, f} + \bar{\lambda}^{\theta^f, f}} = \frac{\bar{\lambda}^{\theta^m, f}}{\bar{\lambda}^{\theta^m, f} + \bar{\lambda}^{\theta^f, f}} \tag{14}$$

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.

Self-image bias also implies clustering of characteristics and groups. In particular, M-researchers will be mostly of type θ^m; in contrast, F-researchers, while a minority, will tend to be mostly of type θ^f. Thus, if at least some of the characteristics correspond to research topics, we conclude that different groups will be relatively more prevalent in different "fields:"

Table 1: Type Frequencies in a Simple Example

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

Corollary 4 *In part (a) of Proposition 6, in the limit, M-researchers are relatively more frequent as type θ^m and F-researchers are relatively more frequent as type θ^f:*

$$\frac{\bar{\lambda}^{\theta^m,m}}{\bar{\lambda}^{\theta^m,m} + \bar{\lambda}^{\theta^m,f}} = \frac{\bar{\lambda}^{\theta^f,f}}{\bar{\lambda}^{\theta^f,m} + \bar{\lambda}^{\theta^f,f}} = \frac{1}{1 + \left(\frac{1-\phi}{\phi}\right)^N} > 0.5; \tag{15}$$

This results is qualitatively consistent with the evidence documenting large gender differences across economics topics (see e.g. Chari and Goldsmith-Pinkham (2018)), 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.

3. A Simple Numerical Example

To illustrate the intuition of our model, we first provide a simple example. Consider the case in which agents have only two characteristics, so $N = 2$. Thus, we have a set of four types:

$$\Theta = \{(0,0), (1,0), (0,1), (1,1)\}.$$

In the notation of the preceding subsections, $\theta^m = (1,0)$, $\theta^f = (0,1)$, and $\theta^* = (1,1)$.

To characterize the population of young researchers, we choose $\phi = 0.8$. That is, 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 of types θ are in Table 1. The between-group heterogeneity in this example is large and not realistic. Our objective in this section is simply to illustrate the patterns that our model can generate. Section 4. provides a numerical analysis of a more realistic case, with between-group heterogeneity in line with the data.

We first consider parameters γ_0 and ρ for which self-image bias prevails. Specifically, we let $\gamma_0 = 0.2$ 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, it turns out that, with $\phi = 0.8$, by Proposition 3 self-image bias prevails:

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

Part (a) of Proposition 3 states that, in the limit, only the two intermediate types have positive mass. So, in particular, the "best" researcher type $(1,1)$ disappears in the limit. Furthermore, if $\lambda_0 = p^m$, then eventually $\theta^m = (1,0)$ becomes the majority type; specifically,

$$\bar{\lambda}^{(1,0)} = \bar{\lambda}^{\theta^m} = \frac{0.8^2}{0.8^2 + 0.2^2} \approx 94\%.$$

As may be expected, correspondingly, established researchers are predominantly M-type in the limit: from Eq. (13) in Proposition 6, the fraction of M-researchers in the limit is

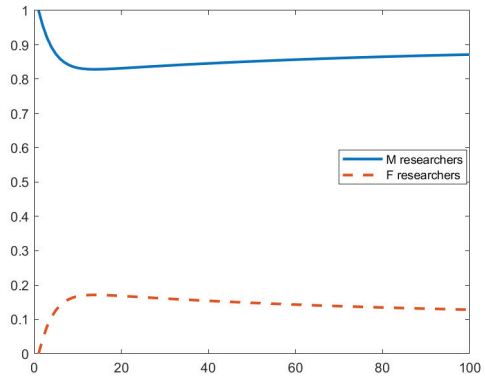
$$\bar{\Lambda}^m = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} = \frac{1 + \left(\frac{.8}{.2}\right)^4}{1 + \left(\frac{.8}{.2}\right)^4 + 2\left(\frac{.8}{.2}\right)^2} \approx 89\%.$$

This is the case despite the fact that an equal mass of young M- and F-researchers appear in every period, and also despite the absence of any explicitly group-biased evaluation of young researchers. The result is driven solely by the initial condition and the referees' self-image bias. To give a sense of the dynamics of the system at finite times, panel (a) of Figure 3 displays the evolution of the fraction of M- and F-researchers in the population (that is, Λ_t^m and Λ_t^f) 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.

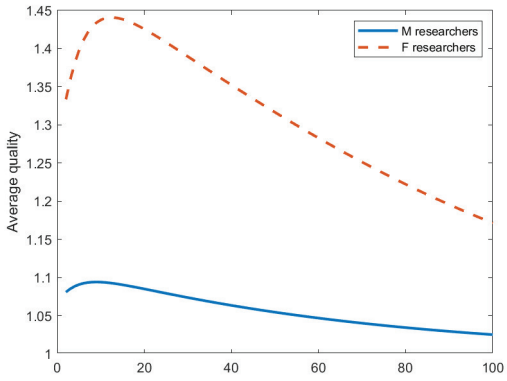
Panel (b) of Figure 3 shows the average objective quality of accepted F and M researchers, with the former being uniformly higher than the latter. This is consistent with Proposition 5 and the intuition provided there. Graphically, the two panels of Figure 4 show the percentage of established F-researchers (left) and M-researchers (right). Concentrating on M-researchers first, type $\theta^m = (1,0)$ prevails from the beginning, compared to other types. As discussed in Proposition 5, this is due to referees being from M-group initially, and thus oversampling the characteristics (1,0). The opposite is true for the F-researchers. In this case, the population of successful F-researchers are initially mostly of type $\theta^* = (1,1)$ (blue line), with smaller—and equal—masses of types $\theta^m = (1,0)$ and $\theta^f = (0,1)$. (The masses are small in all three cases.) Hence, *conditional* on being accepted, F-researchers have a large representation of the best characteristics (1,1) initially, as explained in Proposition 5.

Figure 3: Fraction of M and F Researchers and Acceptance Rates

(a) M and F researchers



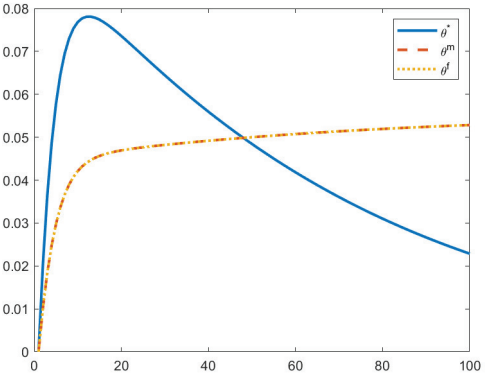
(b) Average Quality



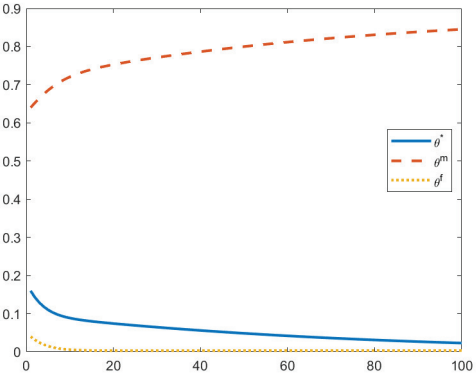
Fraction of M and F researchers (Panel a) and average acceptance rates of M and F researchers, i.e. $\sum_\theta L^\theta w_t^{\theta,g}$ where $L^\theta = \sum_{n=1}^N \theta_n$ and $w_t^{\theta,g} = a_t^{\theta,g} / \sum_{\theta'} a_t^{\theta',g}$, $g = f, m$ (Panel b). Initially $\lambda_0 = p^m$.Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

Figure 4: Types of Established F and M Researchers

(a) F researchers



(b) M researchers



Types of established F (left) and M (right) researchers. We show types $\theta^* = (1,1)$, $\theta^m = (1,0)$, and $\theta^f = (0,1)$. Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

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.

In summary, the system "weeds" out the least productive types $\theta = (0,0)$, but also the efficient type $\theta^* = (1,1)$, despite its higher objective quality. Moreover, because the M-population dominates, and, in this population, characteristic $n = 2$ is under-represented, we end up with a self-perpetuating state in which the dominant M-characteristic $n = 1$ is over-sampled at the expense of the dominant F-characteristic $n = 2$.

3.1. Convergence to Efficiency

We now demonstrate how group balance may arise even with an unbalanced initial population. We continue to assume that the initial population is M-dominated: $\lambda_0 = p^m$, and that $N = 2$ and $\phi = 0.8$. However, we now take $\gamma_0 = 0.1$ and $\rho = 9$. Compared with our previous parameterization, research characteristics now have a greater impact on the likelihood of producing quality research. For instance, type θ^* is 3 times as likely to produce quality research as types θ^f and θ^m, who are themselves 3 times as likely to do so as type $(0,0)$. Thus, the system is now more "meritocratic." Now

$$\rho = 9 > 4.25 = \frac{1}{4}\left(\frac{0.2}{0.8} + \frac{0.8}{0.2}\right)^2 = \bar{\rho}(0.8, 2),$$

so Proposition 3 part (b) implies that type θ^* will dominate in the limit. Figures 5 and 6 illustrate the dynamics. Now the percentage of F-researchers indeed converges to 50%. Moreover, the system weeds out those researchers that do not possess both characteristics. Panel (b) of Figure 5 shows that accepted F researchers are of higher quality than accepted M researchers, as in Proposition 5, until convergence to quality $L = 2$.

4. Many Characteristics and a Calibration

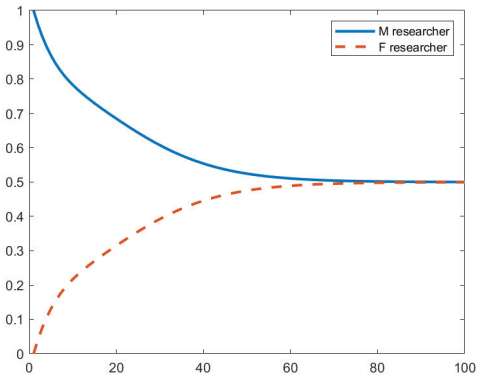
The previous section illustrated the dynamics and the implicit bias that arises from the case with only two research characteristics. The bias was evident and extreme when we considered a large difference in the distribution of each characteristic in the population—we assumed $\phi = 0.8$, so 80% of young M-researchers and 20% of young F-researchers were endowed with characteristics 1, while the opposite was true for characteristics 2. The parameter ϕ can be easily related to Cohen's d statistic for an individual characteristic: for $n = 1, \ldots, \frac{N}{2}$,

$$d = \frac{\mathrm{E}[\theta_n^i | i \in M] - \mathrm{E}[\theta_n^i | i \in F]}{\sigma_{\text{pooled}}(\theta_n^i)} = \frac{2\phi - 1}{\sqrt{\phi(1-\phi)}}. \tag{16}$$

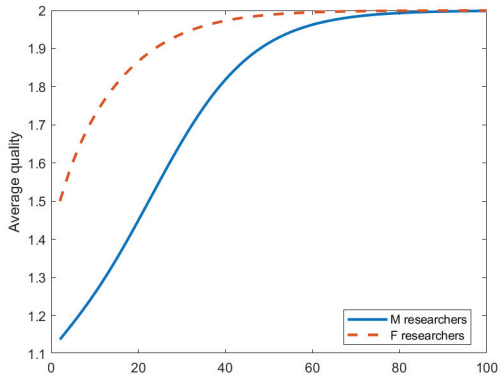
For $n = \frac{N}{2} + 1, \ldots, N$, the d statistic is the negative of the above expression. Cohen (2013) suggests that values of d around 0.2 should be considered "small," values around

Figure 5: Fraction of M and F Researchers and Acceptance Rates with More Meritocracy

(a) M and F researchers



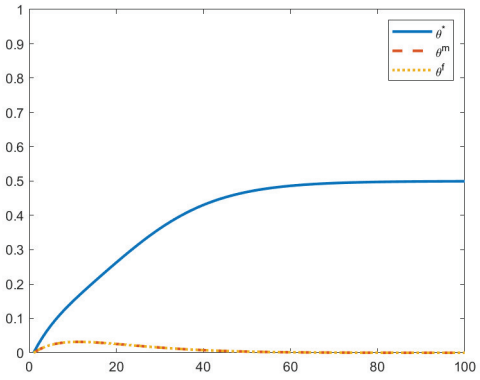
(b) Average Quality



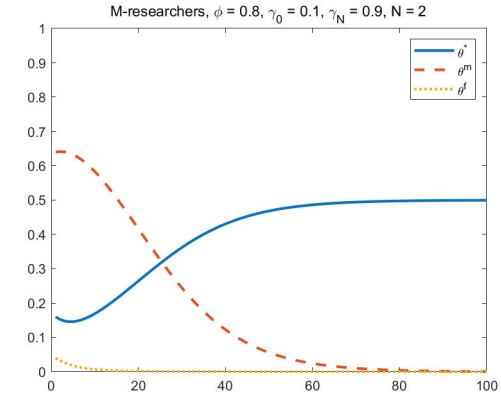
Fraction of M and F researchers (panel a) and average acceptance rates of M and F researchers, i.e. $\sum_\theta L^\theta w_t^{\theta,g}$ where $L^\theta = \sum_{n=1}^N \theta_n$ and $w_t^{\theta,g} = a_t^{\theta,g} / \sum_{\theta'} a_t^{\theta',g}$, $g = f, m$ (Panel b). Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.1$, $\rho = 9$, $N = 2$.

Figure 6: Types of Established Female and Male Researchers with More Meritocracy

(a) F researchers



(b) M researchers



Types of established F (left) and M (right) researchers. We show types $\theta^* = (1,1)$, $\theta^m = (1,0)$, and $\theta^f = (0,1)$. Initially $\lambda_0 = p_m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.1$, $\rho = 9$, $N = 2$.

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 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;[8] 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.

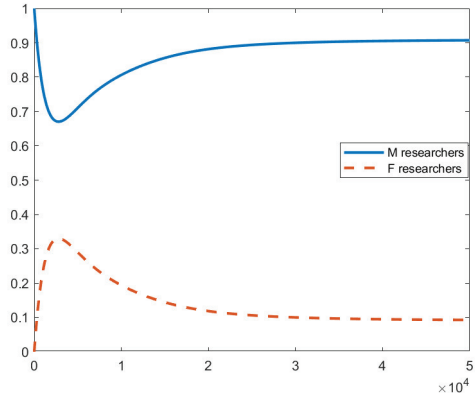
As for the parametrization of $\gamma^\theta = \gamma_0 \, \rho^{\frac{1}{N}\sum_{n=1}^{N}\theta_n}$, we proceed as follows: First, we assume the best researchers $\theta^* = (1, 1, ..., 1)$ has 100% probability of producing quality research, ie. $\gamma^{\theta^*} = 1$. Second, we calibrate γ_0 to match 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.[9] 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

[8]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).

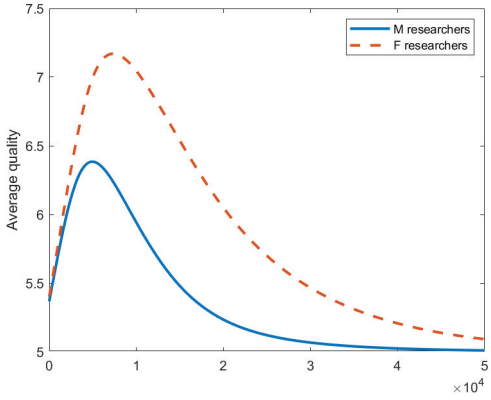
[9]The 2017 survey is the latest as of the time of this writing and it is available at `https://ncsesdata.nsf.gov/doctoratework/2017/index.html`. The total number of economics PhD recipients is 32,000 in US and 12,750 outside the US. The total number of them working in a 4-year educational institution 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.

Figure 7: Fraction of M and F Researchers and Acceptance Rates with Ten Characteristics

(a) M and F researchers



(b) Average Quality



Fraction of M and F researchers (Panel a) and average quality of M and F researchers, 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 b). Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$.

on research productivity reported in Conley and Önder (2014).[10]

The result is in Figure 7. Panel (a) shows that the system converges to a large imbalance between M- and F-researchers, with F-researchers representing less than 10% of the population.[11] This large imbalance obtains despite the fact that the distribution of characteristics is now very similar across M and F types. Panel (b) plots the average quality of accepted F- and M-researchers, and shows the average quality of the former group is uniformly higher, although both eventually converge to $N/2 = 5$. This plot is consistent with Proposition 5 and our conjecture that the result should hold for every N.

5. Endogenous Entry

In this section we extend the model to consider the optimal choice of young researchers on whether to undertake a research career (Section 5.1.) and the optimal choice of hiring institutions on whether to hire young researchers (Section 5.2.).

[10]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 4.

[11]Indeed, inserting $\phi = 0.5742$ and $N = 10$ in equation (13), the limiting fraction of M researchers is $\bar{\Lambda}^m \approx 91\%$, which is where the system converges in Figure 7.

5.1. Endogenous Choice of Young Researchers

Consider a potential researcher choosing between an academic career and an outside option. The prospective researcher knows her type θ, and is aware of both the likelihood of producing quality research, and the evaluation criteria used by the referees. Attempting to pursue research entails a cost C, which is identical across agents. If the potential researcher is hired (accepted), he or she receives a payoff of P; finally, the outside option is normalized to 0. Thus, the total payoff is $P - C$ if the researcher is hired, and $-C$ otherwise. What types of agents decide to pay the cost C and thus take their chance with the academic career?

Assume that the entry decision, research activity, and hiring decision all occur at time t. Then, given the time-t distribution $\lambda_t = (\lambda_t^\theta)_{\theta\in\Theta}$ of referees' types, a prospective researcher of type θ pursues an academic career—"applies"—if and only if

$$\gamma^\theta \lambda_t^\theta (P - C) + (1 - \gamma^\theta \lambda_t^\theta)(-C) > 0. \tag{17}$$

Consequently, the accepted mass of researchers is as follows: for $g = f, m$,

$$a_t^{\theta,g} = \begin{cases} \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g} & \text{if } \gamma^\theta \lambda_{t-1}^\theta \geq \frac{C}{P} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}(1 - a_t) + a_t^{\theta,g} \tag{19}$$

Expression (18) shows that if the mass of type-θ reviewers drops below $\frac{C}{\gamma^\theta P}$ at time $t-1$, both M and F young type-θ researchers will not apply at date t. From Eq. (19), this implies that the total mass of such types will decrease, at least weakly, because some type-θ established researchers will have to retire in order to make room for researchers of other types who are accepted. In fact, the mass of such types will decrease strictly, except in case no young researcher wants to apply.

While the dynamics with endogenous entry is considerably more complicated than in the benchmark case, we prove the following Proposition:

Proposition 7 *Assume that at time 0, all referees are from the M-group with $\lambda_0 = p^m$.*

(a.1) If $\rho < \bar{\rho}(\phi, N)$ and $\frac{C}{P} \leq (1-\phi)^N \gamma_0 \sqrt{\rho}$, then the steady state is as in Proposition 3(a).

(a.2) If $\rho < \bar{\rho}(\phi, N)$ and $(1-\phi)^N \gamma_0 \sqrt{\rho} < \frac{C}{P} \leq \phi^N \gamma_0 \sqrt{\rho}$, then only type θ^m survives in the limit, i.e. $\bar{\lambda}^{\theta^m} = 1$. The limiting mass of M researchers is strictly larger than in (a.1):

$$\bar{\Lambda}^m = \lim_{t\to\infty} \sum_\theta \lambda_t^{m,\theta} = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^N}. \tag{20}$$

(b) If $\rho > \bar{\rho}(\phi, N)$ and $[\phi(1-\phi)]^{N/2} \geq \frac{C}{\gamma_0 \rho P}$, then the steady state is as in Proposition 3(b).

In each of the above cases, if $\bar{\lambda}^\theta = 0$, then there is $t^\theta \geq 0$ such that $\lambda_t^\theta = 0$ for all $t \geq t^\theta$.

Part (a.1) and (b) of this proposition shows that if the cost C is low enough, then the the steady state is the same as in the basic model in Section 2. for the same two conditions about ρ, respectively. This is intuitive. The only difference is that all types other than surviving ones drop out in finite time, rather than only in the limit.

The interesting new part is (a.2). In this case, the only type that survives in the long-run is θ^m, the most prevalent type in the $M-$population. In particular, θ^f now disappears. Thus, the characteristics that are mildly more frequent in the $F-$population, but also common in the M-population, eventually disappear. In this case, endogenous entry greatly exacerbates the loss of talent compared to the base case. Indeed, the total mass of M researchers, $\bar{\Lambda}^m$, is now even larger than in its counterpart without endogenous entry, whose expression is in Eq. (11) in Proposition 3. Thus, if the conditions in part (a.2) are satisfied, the distribution of established researchers will be even more skewed towards the M group.

Parts (a.1)–(b) do not exhaust all possible cases; for instance, they do not analyze the possibility that the first condition in part (b) holds, but the second does not—that is, θ^* is not willing to apply. The following section illustrates a stark instance of one such possibility. The proof of the above Proposition in the Appendix provides a general characterization that can be used to further explore different parametric choices.

5.1.1. Example of Group Imbalance due to Endogenous Entry

We first illustrate how endogenous entry can exacerbate group imbalance, provided the cost of entry is not too small. Consider the parameterization in Section 4. In our basic model, M-researchers represent 91% of the overall population in the limit. If we add endogenous entry, Proposition 7 shows that the steady state either remains the same, if the cost C is sufficiently low, as in case (a.1), or it becomes even more skewed towards the M group, as in case (a.2). In the latter case, the limiting fraction of M-researchers is $\bar{\Lambda}^m = \phi^N/(\phi^N + (1-\phi)^N) = 95\%$.

We now illustrate how endogenous choice may prevent convergence to group balance even when group balance would in fact attain in the basic model. We use the same parameterization as in Section 4., except that the number of characteristics is $N = 8$ instead of $N = 10$. With these parameter values, Proposition 3 part (b) implies that the system will converge to an equal mass of M and F researchers, because $\rho = 5 > 3.61 = \bar{\rho}(\phi, N)$. The solid and

dashed lines in Figure 8 confirm this.

However, assume now that entry is endogenous; the payoff if a researcher is hired is $P = 1,000$, and the cost of entry is $C = 4$ (i.e., 0.4% of the payoff of becoming a researcher over the outside option). Note that these parameters apply equally to M and F researchers. The key point is that now the efficient type θ^* (M or F) does not want to apply at date 0:

$$\lambda_0^{\theta^*} = p^{\theta^*,m} = \phi^{N/2}(1-\phi)^{N/2} = 0.3574\% < 0.4\% = \frac{C}{\gamma^{\theta^*}P}.$$

Moreover, type θ^f (M or F) does not want to apply either:

$$\lambda_0^{\theta^f} = p^{\theta^f,m} = (1-\phi)^N = 0.1081\% < 0.8944\% = \frac{C}{\gamma^{\theta^f}P}.$$

On the other hand, type θ^m (M or F) does:

$$\lambda_0^{\theta^m} = p^{\theta^m,m} = \phi^N = 1.18\% > 0.8944\% = \frac{C}{\gamma^{\theta^m}P}.$$

Therefore, while other types are also willing to apply, type θ^m will prevail, which will lead to a severe imbalance between M and F researchers in the limit, as shown in Figure 8. Indeed, in this case the talent loss is rather severe, as the only surviving type $\theta^m = (1, ..., 1, 0, ...0)$ has none of the research characteristics that are (mildly) more common in the $F-$population. Figure 9 shows that both F and M researchers are of type θ^m in the long run.

To sum up, even if the basic environment is meritocratic, in the sense that differences in talents γ^θ across types are sufficient to lead to group balance, endogenous entry introduces a bias in favor of M-researchers which leads to an imbalance steady state. In this case, policies aimed at lowering the cost C can lead to group balance in the long run.

5.1.2. Characterization of the Applicant Pool

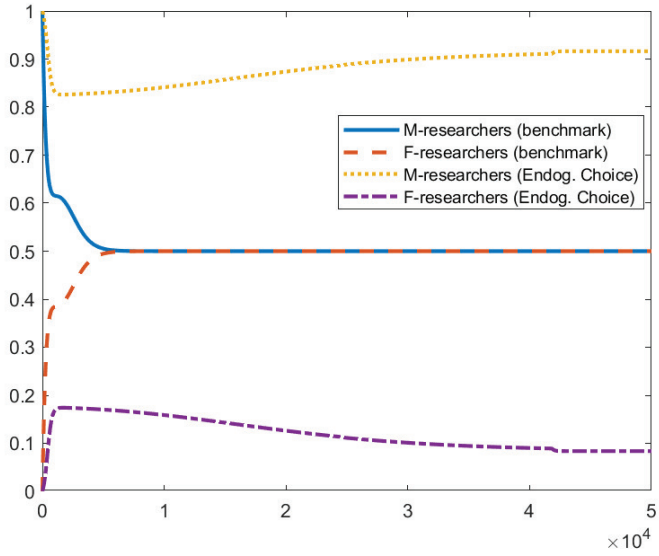
Due to variation in the distribution of characteristics, Proposition 7 also has implications for the mass of young M and F researchers who decide to apply for an academic job:

Proposition 8 *For every t, let*

$$A_t^m = \sum_{\theta:\lambda_t^\theta \geq \frac{C}{\gamma^\theta P}} p^{\theta,m} \quad \text{and} \quad A_t^f = \sum_{\theta:\lambda_t^\theta \geq \frac{C}{\gamma^\theta P}} p^{\theta.,f}$$

Then $A_t^m \geq A_t^f$. Moreover, if $\lambda_0^{\theta^m} > \frac{C}{\gamma_0\sqrt{\rho}P} > \lambda_0^{\theta^f}$, then $A_t^f \to 1 - \bar{\Lambda}^m$, where $\bar{\Lambda}^m$ is as in part (a.2) of Proposition 7.

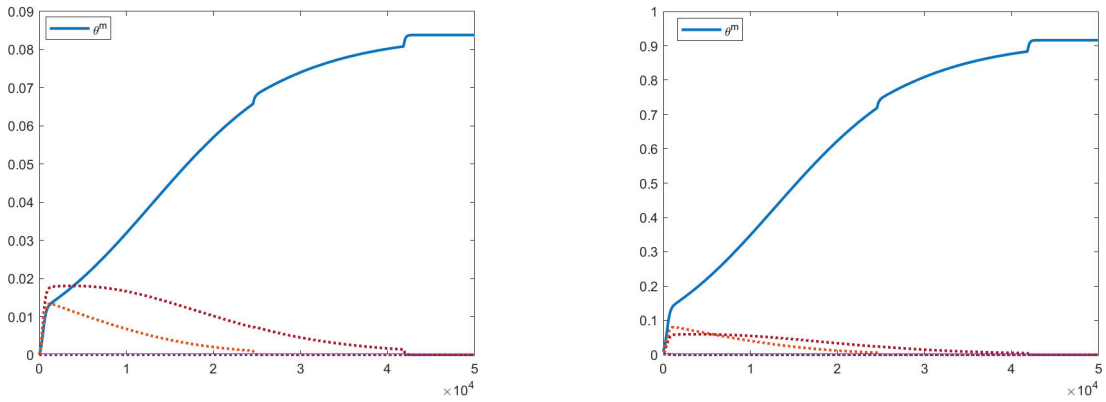
Figure 8: Fraction of M and F Researchers with Endogenous Entry



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 8$, $P = 1000$, and $C = 4$.

Figure 9: Types of Established F and M Researchers with Endogenous Entry

(a) F researchers (b) M researchers



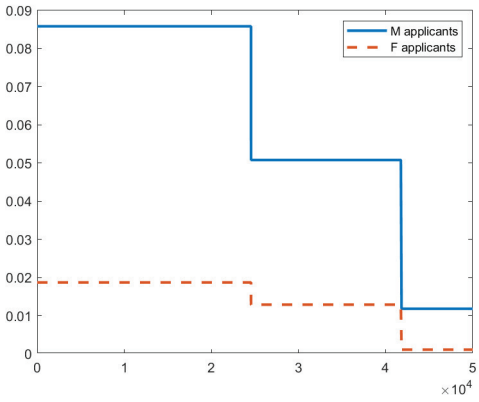
Types of established F (left) and M (right) researchers with endogenous entry. $\theta^m = (1, ..., 1, 0, ..., 0)$ dominates; all other types eventually vanish. Parameters: $\phi = 0.0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 8$, $P = 1000$, and $C = 4$.

The intuition stems from the fact that when the majority of referees is from the M-group, it is more likely for an M-researchers to be accepted than for a F-researcher, on average. Thus, mass of applicants from the M-group is higher than from the F-group.

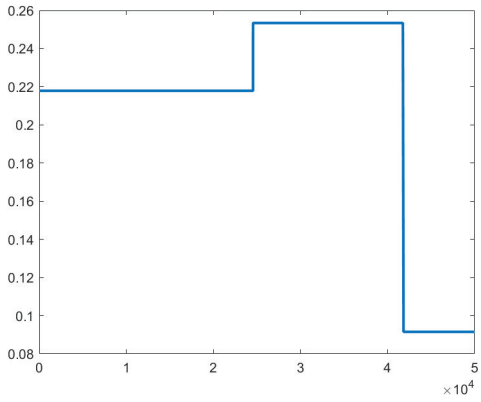
Figures 10a and 10b show the total masses of M and F applicants and, respectively, the

Figure 10: Endogenous entry: applicants

(a) M and F Applicants



(b) Fraction of F Applicants



Total mass of M and F applicants (left) and fraction of F applicants (right). Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 8$, $P = 1000$, and $C = 4$.

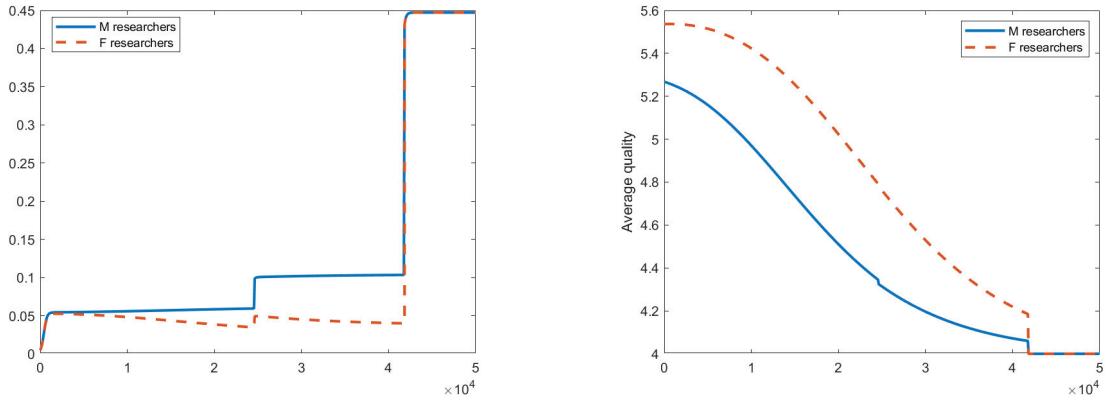
percentage of F applicants over the total application pool. The parameter values are the same as for Figure 8. Consistently with Corollary 8, the mass of M applicants is always greater than that of F applicants; furthermore, the latter declines over time. The discrete jumps in these masses occur whenever, for some type θ, the population fraction λ_t^θ falls below the cutoff $C/(\gamma^\theta P)$. In the limit, the fraction of F applicants equals the fraction of F researchers of the only surviving type θ^m over the total:

$$\lim_{t\to\infty} \frac{A_t^f}{A_t^m + A_t^f} = \frac{p^{\theta^m,f}}{p^{\theta^m,f} + p^{\theta^m,m}} = \frac{(1-\phi)^N}{\phi^N + (1-\phi)^N} = \frac{0.4258^8}{.4258^8 + .5742^8} = 0.0838$$

Finally, the left panel of Figure 11 shows the total acceptance rates of M and F applicants. In the initial period, the acceptance rates of M and F applicants are similar. They though diverge in the intermediate period, in which M applicants are accepted more often than the (fewer) F applicants, and then they finally converge, when only type θ^m survives. Interestingly, the right panel shows that the average quality of F researchers is uniformly higher until the time of convergence. This implies that in the initial period our model predicts similar acceptance rates of M and F researchers, even if the latter have higher objective quality. This result is reminiscent of Card et al. (2020), who show that unconditionally, acceptance rates of men- and women-authored papers are similar, but that the average quality of accepted women-authored papers, proxied by their future citations, is higher.

Figure 11: Endogenous entry: Acceptance Rates

(a) Acceptance rate of applicants (b) Average quality



Acceptance rate of M and F applicants (left) and average quality of accepted ones (right). Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 8$, $P = 1000$, and $C = 4$.

5.2. Endogenous Selection by Hiring Institutions

The previous section demonstrates that endogenizing the choice of entry into academia may shrink the supply of talent. We now show that the a similar mechanism operates on the demand side: when hiring decisions are based on the expectation of academic success, the anticipation of self-image bias in the refereeing process (Section 2.2.) induces institutions to hire only those types θ that can produce research that is more likely to be "accepted" by the established refereeing population.

Consider the following alternative interpretation of our model. When a hiring institution evaluates a candidate, it takes into account whether or not the candidate will produce quality work that the profession recognizes, or—in the language of Section 2.2.—"accepts." A candidate who is accepted by the profession yields a payoff P to the institution; this reflects e.g. visibility, grant money, or increased ability to attract top students. Hiring a candidate involves a cost C, which may be monetary but may also reflect mentoring resources and/or opportunity cost. This cost is borne by the institution whether or not the candidate is eventually accepted, and it is the same for M and F researchers. If the candidate is eventually not accepted or if the institution does not hire any candidate, the institution's payoff is zero. As above, a candidate of type θ produces quality work with probability γ^θ. To analyze demand effects, we reinterpret the key assumption of Section 2.2. as follows: the hiring institution anticipates that referees are subject to self-image bias, so that a type-θ researcher will be accepted by the profession with probability $\gamma^\theta \lambda_t^\theta$ at the end of time t.

Under these conditions, the institution hires an agent of type θ if and only if

$$\gamma^{\theta}\lambda_t^{\theta}\left(P-C\right)+\left(1-\lambda_t^{\theta}\gamma^{\theta}\right)\left(-C\right)>0 \tag{21}$$

This is the same condition as in Equation (17) in the previous section. Thus, the mass of established researchers λ_t^{θ} follows the system dynamics described by Equations (18) - (19). Proposition 7 then applies and group imbalance and loss of talent obtains.

Moreover, under the conditions of case (a.2) of Proposition 7, the system converges, in finite time, to a steady state in which only type θ^m survives. That is, if institutions *only* take acceptance by the profession into account at the hiring stage, type θ^f eventually disappears, even when such type would survive without endogenous selection. Again, this implies talent loss: research characteristics that are (mildly) more common in the F-population disappear.

We can also re-interpret the example in subsection 5.1.1. as a consequence of the hiring practices of hiring institutions. In the absence of endogenous selection, the parametric choices in that example lead to group balance, with both types θ^m and θ^f being represented in the limit. However, if institutions wish to hire only young researchers who are sufficiently likely to be accepted by the *current* population of referees, then group imbalance emerges, as in Figure 8. Again, in this example type θ^f then disappears completely, as in Figure 9.

This mechanism may explain the patterns in Figure 1. From the top panel, the female representation of undergraduate students with economics major in the top-20 schools has been rising over the past 25 years, reaching almost 40% by the late 2010s. This shows interests in economics among female undergraduates. Yet, in the same period, the percentage of female PhD students has been flat at around 30%, and that of assistant professors has been flat at around 22%. The bottom panel shows a striking difference between schools with and without PhD programs: In the latter group, the share of assistant professors is over 40%, while in the former is below 30%, with the top 10 schools at 20%.[12] These differences do not apply to the female share of teaching faculty, which are around 37% across all schools. This is consistent with our model: when a school has research as the guiding principle in hiring, it tends to skew towards the characteristics of established researchers, i.e. θ^m in our model.

6. The Impact of Policy Action

In this section we discuss the impact of policy actions that have been proposed to address gender imbalance. Our discussion of endogenous entry, as in e.g. subsection 5.1.1., already

[12]We use the "top-X schools" terminology as in Chevalier (2020). School names are not reported.

suggests that, if the cost of pursuing an academic career is the main cause of group imbalance, then reducing this cost is the appropriate policy response. This corresponds to *outreach*, which we discuss in Section 8.. Here we focus on situations in which the cost of entry is *not* the main cause of imbalance; instead, self-image bias is—formally, entry is costless, and $\rho < \overline{\rho}(\phi, N)$ in point (a) of Proposition 6. We consider (i) the impact of mentoring (section 6.1.); and (ii) the impact of affirmative action (section 6.2.).

6.1. The Impact of Mentoring

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. Assuming again that P is the payoff from being hired and U is the utility from an outside option, researcher θ will 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. Figure 12 illustrates the dynamics resulting from Eq. (A.29),

under the same parameters as in Section 4. 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 the in the right panel of Figure 13. 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.

Figure 13 shows, 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.[13] 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.

6.2. The Impact of Affirmative Action

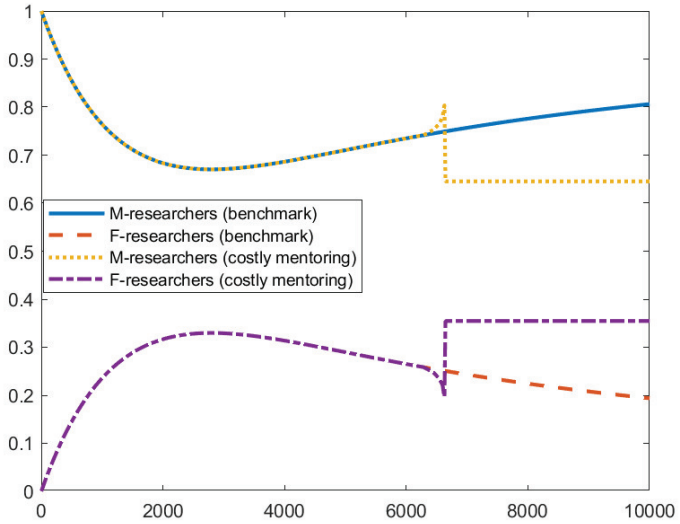
A common policy to increase diversity is "affirmative action", that is, the policy to increase the representation of under-represented groups by mandate. We consider a simple rule in this section: it each round, it is mandated that reviewers must accept in their group of established researchers the same number of M and F researchers. We change just one assumption to the dynamics in the benchmark case: namely, we assume

$$a_t^{\theta,m} = k_t \; \gamma^\theta \; \lambda_{t-1}^\theta \; p^{\theta,m} \qquad \text{where} \qquad k_t = \frac{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',f}}{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',m}}. \tag{22}$$

The scaling factor k_t ensures that $\sum_\theta a_t^{\theta,f} = \sum_\theta a_t^{\theta,m}$. Figures 14 and 15 provide the dynamics for this case. The affirmative action policy reaches group balance (and this is not surprising, given the definition of k_t) as well as diversity in research characteristics, as in the limit M researchers are of type θ^m and F researchers are of type θ^f. Assuming that maximizing the

[13] 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.8 in the online appendix.

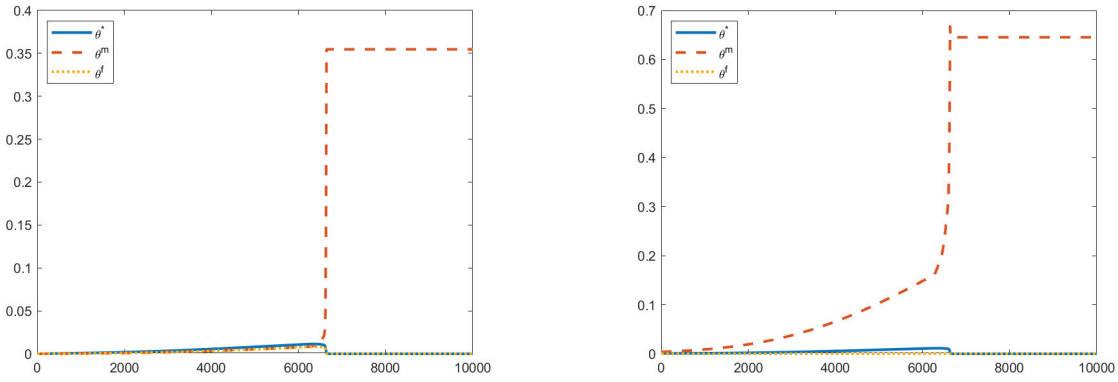
Figure 12: Fraction of F and M Researchers with Costly Mentoring



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$, cost function $C(\theta, \theta') = 0.0750 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

Figure 13: Types of Established F and M Researchers with Costly Mentoring .

(a) F researchers (b) M researchers

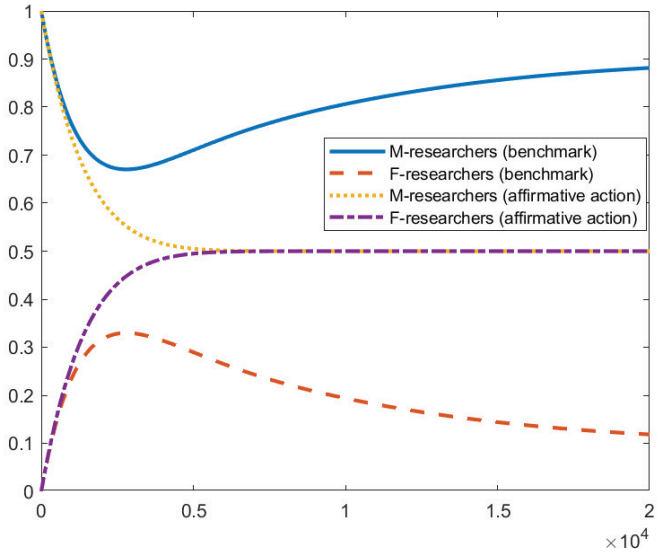


Types of established F (left) and M (right) researchers with costly mentoring. We show the masses of types $\theta^* = (1, 1, ..., 1)$, $\theta^m = (1, ..., 1, 0, ..., 0)$, and $\theta^f = (0, ..., 0, 1, ..., 1)$. Initial reviewers: $\lambda_0 = p^m$. Parameters: $\phi = 0.0.5742$ ($d = 0.3$) , $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$; cost function: $C(\theta, \theta') = 0.0750 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

representation of research characteristics is beneficial to society, this policy appears superior to mentoring, as it does not skew the distribution onto θ^m even when reaching group balance.

Intuitively, by expanding the set of referee characteristics, affirmative action makes it

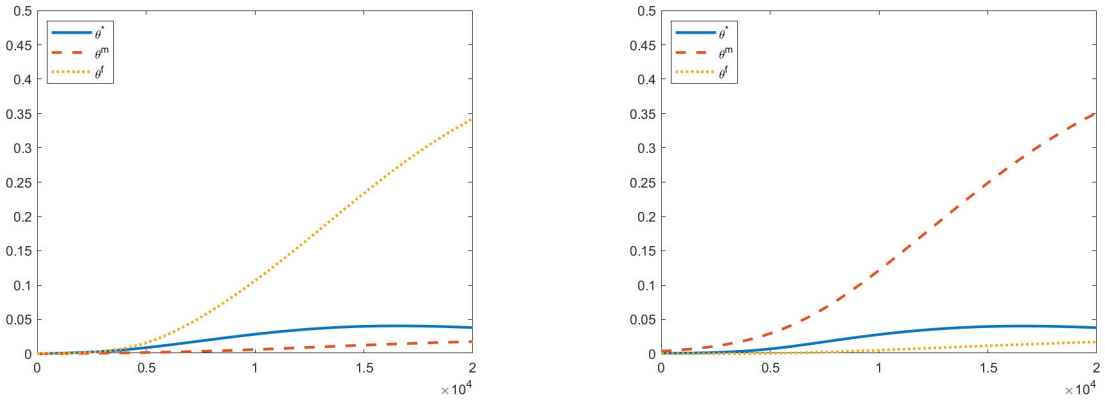
Figure 14: Fraction of F and M Researchers with Affirmative Action



Fraction of M and F researchers when $\lambda_0 = p^m$ and there is an affirmative action policy that 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 $N = 10$.

Figure 15: Types of Established F and M Researchers with Affirmative Action

(a) F researchers (b) M researchers



Types of established F (left) and M (right) researchers when affirmative action requires accepting the same number of M and F researchers. We show types $\theta^* = (1, 1, ..., 1)$, $\theta^m = (1, ..., 1, 0, ..., 0)$, and $\theta^f = (0, ..., 0, 1, ..., 1)$. Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, and $N = 10$

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.

7. 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.

8. 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 empirical evidence and the trends 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 (Chevalier, 2020 and Figure 1). 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 (see Section 5.1.). While lowering the cost may indeed switch the path towards convergence for some parameter configurations, as shown in Section 5.1.1., our basic model in Section 2.2. assumes zero costs and yet, under the conditions of Proposition 3, (2.a), gender bias persists. In particular, if reviewers evaluate others' research on a multitude of research characteristics, gender imbalance would persist.

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 6.1.). 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 underrepresented. Facilitating the promotion of female researchers counteracts this force, and leads to a more balanced representation of research characteristics in the steady-state population.

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.

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.

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.

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.

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.

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.

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*, 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.

Self-Image Bias and Talent Loss

On-Line Appendix

Marciano Siniscalchi

Northwestern University

Pietro Veronesi

University of Chicago

This on-line appendix contains additional analysis and the proofs of our propositions.

A1. Additional Analysis and Results

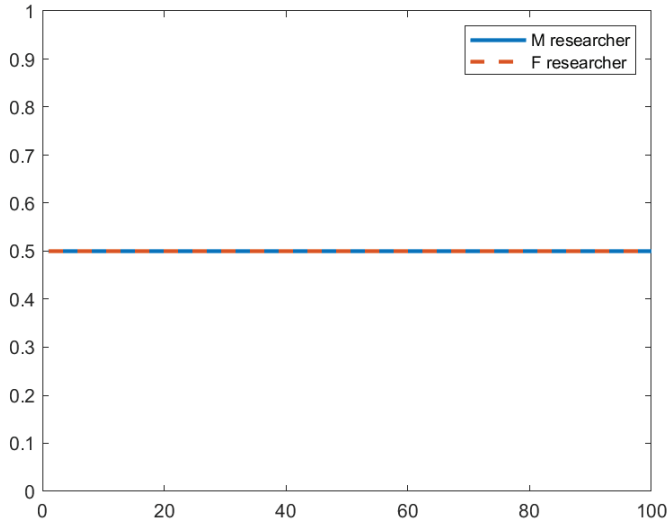
A1.1. Balanced Steady State

In Section 3. we considered a simple numerical example with only two characteristics ($N = 2$), which led to types $\Theta = \{(0,0),(0,1),(1,0),(1,1)\}$. In that section, we showed that when $\rho < \overline{\rho}(\phi, N)$ and the initial population of referees is only from the M-group, $\lambda_0^{\theta,m} = p^{\theta,m}$, then the dynamics never converges. Here we now consider a different initial condition.

Indeed, the dynamics of the mass of each type depends upon their frequencies in the population of young researchers, p_m and p_f, as well as the initial conditions λ_0. In particular, suppose that the initial mass of referees is composed of M- and F-researchers in equal proportions: $\lambda_0 = \frac{1}{2}p_m + \frac{1}{2}p_f$. One implication is that then the two M-prevalent and F-prevalent types $\theta^m = (1,0)$ and $\theta^f = (0,1)$ both represent 34% of the initial mass of referees, whereas the other two types $(0,0)$ and $(1,1)$ each represent 16% of the initial population. While we can no longer invoke the results in Sections 2.2.-2.3.3., we can plot the dynamics of the fractions of established M- and F-researchers, as well as those of established M-and F-researcher types. (Theorem A.1 in the Appendix characterizes the limiting behavior of the system for arbitrary initial conditions and type distributions.)

Figures A.1 and A.2 display the results. The figures are self explanatory: an equal proportion of M- and F-researchers is maintained throughout. However, importantly, type θ^f (resp. θ^m) will eventually become prevalent among F-researchers (resp. M-researchers), which means that established F- (resp. M-) economists are oversampled from those who

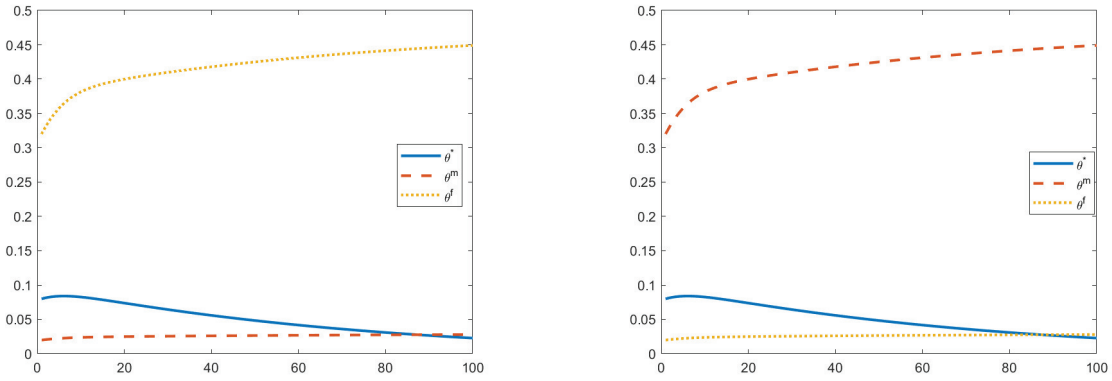
Figure A.1: Fraction of M and F researchers with Start from Equal Proportions



Fraction of M and F researchers when $\lambda_0 = \frac{1}{2}p_m + \frac{1}{2}p_f$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

Figure A.2: Types of Established F and M Researchers with Start from Equal Proportions

(a) F researchers (b) M researchers



Types of established F (left) and M (right) researchers. We show types $\theta^* = (1, 1, ..., 1)$, $\theta^m = (1, ..., 1, 0, ..., 0)$, and $\theta^f = (0, ..., 0, 1, ..., 1)$. Initially $\lambda_0 = \frac{1}{2}p_m + \frac{1}{2}p_f$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

possess characteristic 2 (resp. 1). Furthermore, the efficient type θ^* will disappear in the limit.

A1.2. Seniors and Juniors

We now extend the basic model (without endogenous entry) in a different direction, namely, to the case in which there are different levels of seniority in the population of established researchers, with the seniors judging the research of the juniors, before accepting them onto their group. For instance, junior assistant professors may judge candidates from the rookie market and senior professors judge both assistant professors and rookies.

To avoid introducing new symbols, we add a subscript "1" to denote the mass of junior established researchers, and a subscript "2" for the senior established researchers. The difference from the previous case is mainly the mass of candidates of each type θ at each time t. For simplicity, we assume that, at time 0 and thereafter, the mass of seniors is fixed at σ and the mass of juniors is $1-\sigma$, so that the overall population of established researchers has mass 1, as in previous sections. That is, for all t, we must have

$$\sum_{\theta} \lambda_{1,t}^{\theta} = 1 - \sigma, \quad \sum_{\theta} \lambda_{2,t}^{\theta} = \sigma.$$

The flows are similar to before: young researchers are evaluated by all, and juniors are evaluated by seniors only. For each group $g \in \{f, m\}$ and type $\theta \in \Theta$, the flows of juniors $a_{1,t}^{\theta,g}$ and seniors $a_{2,t}^{\theta,g}$ evolve according to

$$a_{1,t}^{\theta,g} = \gamma^{\theta} \cdot p^{\theta,g} \cdot (\lambda_{1,t-1}^{\theta} + \lambda_{2,t-1}^{\theta}) \tag{A.23}$$

$$a_{2,t}^{\theta,g} = \gamma^{\theta} \cdot \lambda_{1,t-1}^{\theta,m} \cdot \lambda_{2,t-1}^{\theta}. \tag{A.24}$$

Again, we assume that current seniors are randomly replaced by newly promoted juniors, and current juniors are randomly replaced by newly accepted young researchers. However, we now must take into account the fact that juniors promoted to seniors leave the junior pool. We thus obtain the dynamics

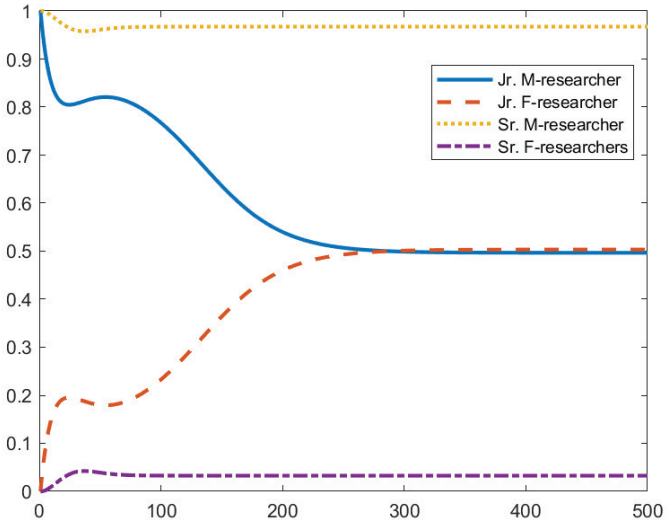
$$\lambda_{1,t}^{\theta,g} = \lambda_{1,t-1}^{\theta,m} \left(1 - \frac{1}{1-\sigma}(a_{1,t} - a_{2,t})\right) + a_{1,t}^{\theta,g} - a_{2,t}^{\theta,g} \tag{A.25}$$

$$\lambda_{2,t}^{\theta,g} = \lambda_{2,t-1}^{\theta,g} \left(1 - \frac{1}{\sigma} a_{2,t}\right) + a_{2,t}^{\theta,g} \tag{A.26}$$

for $g \in \{f, m\}$, where $a_{j,t} = \sum_{\theta}(a_{j,t}^{\theta,f} + a_{j,t}^{\theta,m})$ for $j = 1, 2$.

The dynamics are far more complex than in the base case, and we rely on numerical simulations.

Figure A.3: Leaky pipeline



Fraction of senior and junior M and F researchers, relative to σ (seniors) and $1-\sigma$ (juniors), when $\lambda_0 = p^m$. Parameters: $\phi = 0.7$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 4$ and $\sigma = 0.5$.

A1.2.1. Leaky Pipeline

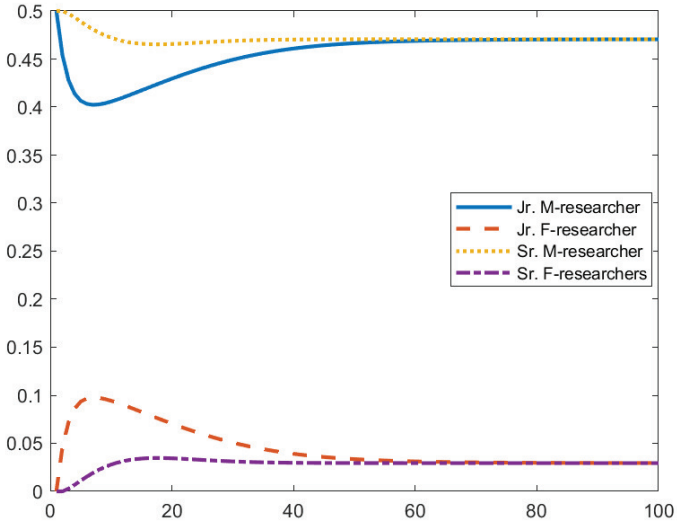
Here we focus on the most interesting case, namely, the fact that this extension can also account for the "leaky pipeline" pattern highlighted in the CSWEP report (Chevalier, 2020). Figure A.3 provides a stark illustration: under the given parametric assumptions, group balance attains among juniors, but not among seniors. A rough intuition is that the self-image bias may not be strong enough to result in a prevalence of θ^m types among juniors, given the constant influx of new researchers with a more balanced distribution of types. However, it may be strong enough if the candidates' types are themselves more biased towards the M researchers' distribution—as is the case for junior up for promotion to the senior rank.

A1.2.2. Other Patterns

We now consider other cases, for illustration. All the simulations in this section assume equal fractions of juniors and seniors ($\sigma = 0.5$).

First, the presence of a second screening—and hence a second opportunity for self-image bias to exert its influence—can exacerbate group imbalance in the senior rank, at least in the short run. Figure A.4 demonstrates this. Model parameters are as in Figure 3, so in a single-cohort environment significant group imbalance emerges. The same is true with two ranks;

Figure A.4: More extreme imbalance for senior rank



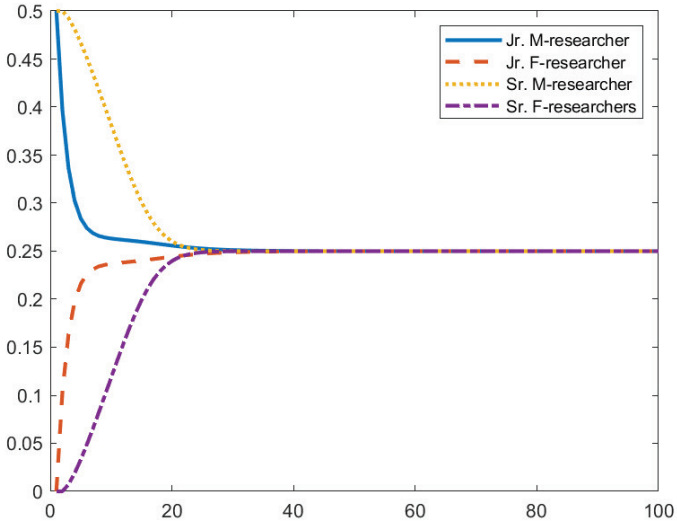
Fraction of senior and junior M and F researchers when $\lambda_0 = p_m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

however, in the short run, the imbalance is more pronounced in the senior rank. The reason is that, in order to be promoted to the senior rank, a researcher must match with a referee of the same type *twice*. Initially, both junior and senior referees have the same type distribution, which by assumption coincides with that of M researchers. Hence, whatever effect is present at the junior rank is compounded at the senior rank.[1] The difference between the two ranks vanishes in the long run because, as type θ^m becomes prevalent among established juniors and seniors, promotion eventually is driven solely by objective research quality—matching with a senior reviewer of the junior candidate's own type is virtually guaranteed.

A more pronounced group imbalance can also arise, in the short / medium run, for parameter values for which convergence is eventually attained. This is demonstrated in Figure A.5, where we take $\phi = 0.6$ rather than $\phi = 0.8$. Again, the need to match with a like type twice, coupled with the assumption that the initial population consists entirely of M-researchers, leads to a lower representation of F researchers at the senior rank. However, over time, type θ^* prevails among juniors and seniors, so matching with like types is virtually guaranteed; and since convergence is attained amongst juniors, it must obtain among seniors as well.

[1] In fact, the bias becomes stronger over time at the senior rank. The reason is that the initial population of junior candidates up for promotion is characterized by types distributed as among male researchers, whereas the initial population of young researchers applying for a junior position is balanced.

Figure A.5: Convergence, but greater short-run imbalance among seniors



Fraction of senior and junior M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.6$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

A1.3. Similarity in Research Characteristics

In this section we extend the model to investigate the case in which referees accept researchers who have characteristics close but not necessarily identical to their own. In particular, we assume that referee r of type θ^r accepts the research of young researcher θ if

$$D(\theta^r, \theta) = \sum_n (\theta^r_n - \theta_n)^2 \leq \eta \tag{A.27}$$

where η is a non-negative integer. That is, referee θ^r treats candidate θ as "close enough" if it differs from his or her own type in no more than η characteristics.

Our models so far correspond to $\eta = 0$. If instead $\eta > 0$, the dynamics for λ^θ_t are still as in Eq. (7), but the mass $a^{\theta,g}_t$ of accepted researchers of type θ in group $g \in \{f, m\}$ is given by

$$a^{\theta,g}_t \;=\; \gamma^\theta \sum_{\theta^r : D(\theta^r,\theta) \leq \eta} \lambda^{\theta^r}_{t-1}\, p^{\theta,g} \tag{A.28}$$

Unfortunately, obtaining general analytical results in this case seems difficult. Therefore, we consider illustrative special cases.

A1.3.1. Connected Set of Types

The set Θ of types we have considered so far enjoys a special structure that is relevant to the relaxed definition of "acceptance" in Eq. (A.27). For every $\eta \geq 1$, and every pair $\theta, \theta' \in \Theta$, there is a finite ordered list $\theta_1, \ldots, \theta_K \in \Theta$ such that $\theta_1 = \theta$, $\theta_K = \theta'$, and $D(\theta_k, \theta_{k+1}) \leq \eta$ for all $k = 1, \ldots, K-1$. In this sense, we say that $\Theta = \{0,1\}^N$ is η-*connected* for every $\eta \geq 1$. Of course, being 1-connected implies being η-connected for $\eta > 1$; we shall see in the next subsection that a subset of $\{0,1\}^N$ may be η-connected for some $\eta > 1$, but for any smaller integer η' (including $\eta' = 1$).

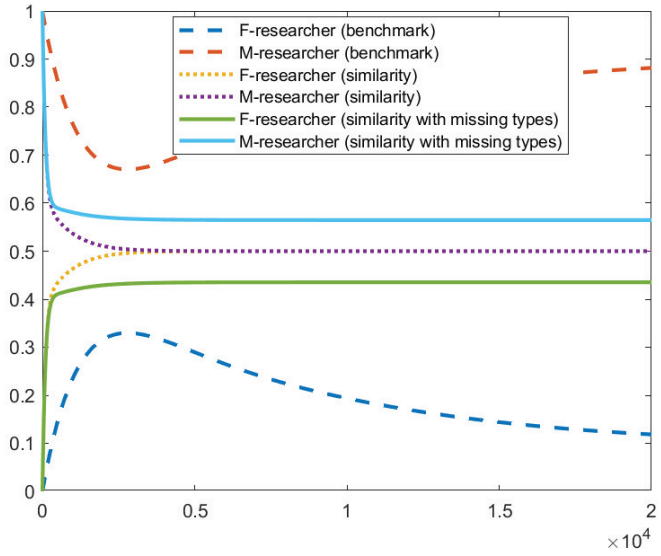
With $\Theta = \{0,1\}^N$, and for the parameter values used in the examples of Sections 3. and 4., the relaxed acceptance criterion in Eq. (A.27) leads to convergence. For instance, Figure A.6 illustrates the parameterization used in Section 4.. The dashed lines represent the benchmark case $\eta = 0$, where there is no convergence. The dotted lines reflect the assumption that referees accept young researchers that are closely similar to them: specifically, taking $\eta = 1$. Notably, group balance obtains. (The solid lines are discussed in the next section.) Moreover, we have not been able to find parameterizations for which convergence did *not* occur. We conjecture that this is a general property of the special structure of the type space $\Theta = \{0,1\}^N$. Intuitively, a referee of type θ accepts a positive mass of young researchers of similar, but not identical type θ'; these become referees in the following period, and accept a positive mass of young researchers of type θ'' that type-θ referees would reject; and so on. A contagion argument suggests that, in the limit, the impact of self-image bias should vanish, so that group balance should emerge.

A1.3.2. Disconnected Set of Types

A subset of $\{0,1\}^N$ may well be η-disconnected for some η. For a trivial example, $\{\theta^m, \theta^f\}$ is $(N-1)$-disconnected, because each of the N coordinates of θ^f is different from the corresponding coordinate of θ^f. A fortiori, it is η-disconnected for every $\eta \leq N-1$.

Intuition suggests that the contagion argument given above breaks down with a disconnected set of types. We now verify this intuition. The solid lines in Figure A.6 represent the same parameterization as in the previous subsection, with $\eta = 1$, but applied to a state space Θ obtained by randomly removing 20% of the elements of $\{0,1\}^N$ and suitably renormalizing probabilities. As expected, the system does not attain group balance in the limit.

Figure A.6: Fraction of M and F Researchers under the Research Similarity Assumption



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$, which implied $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$, and, under research similarity, $\eta = 1$.

A1.3.3. Endogenous Entry

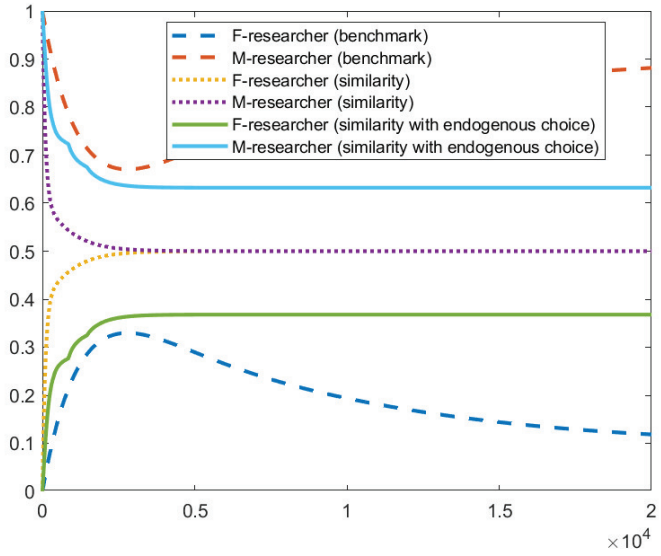
Finally, return to the case in which $\Theta = \{0,1\}^N$ (a connected set of types) but consider endogenous entry, as in Section 5.. In this case, even if the connected set of types would lead to convergence (see subsection A1.3.1.), the endogenous entry prevents such convergence, as shown in Section 5.1.1.. This is shown in Figure A.7. Again, the dashed lines and the dotted lines show the total fraction of M- and F-researchers in the benchmark case ($\eta = 0$) and, respectively, the research similarity case ($\eta = 1$). The solid lines now show the the fraction of M- and F-researchers under research similarity ($\eta = 1$) but with endogenous entry. The intuition is the same as the one given in Section 5..

In sum, this section suggests that the main results of the paper are robust to a weaker assumption about the referees' selection mechanism.

A2. Mentoring

In this section we provide additional intuition on the dynamics of the system under mentoring, and further illustration. The mass of young researchers from group $g \in \{f, m\}$ of type

Figure A.7: Fraction of M and F Researchers under Research Similarity and Endogenous Entry



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$, which implied $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$, and, under research similarity, $\eta = 1$. Endogenous choice assume $P = 1000$ and $C = 6$.

θ accepted at time t is then

$$a_t^{\theta,g} = \gamma^{\theta} \lambda_{t-1}^{\theta} \left[\left(p^{\theta,g} \sum_{\theta^a : \widetilde{C}(\theta,\theta^a) \geq \gamma^{\theta^a} \lambda_{t-1}^{\theta^a} - \gamma^{\theta} \lambda_{t-1}^{\theta}} \lambda_{t-1}^{\theta^a} \right) + \left(\lambda_{t-1}^{\theta} \sum_{\theta' : \theta' \neq \theta, \widetilde{C}(\theta',\theta) < \gamma^{\theta} \lambda_{t-1}^{\theta} - \gamma^{\theta'} \lambda_{t-1}^{\theta'}} p^{\theta',g} \right) \right] \tag{A.29}$$

The first term in brackets captures all of the young researchers of type θ from group g who are matched with mentors of types θ^a (with probability $\lambda_{t-1}^{\theta^a}$) and choose not to be advised as the cost is too large; these young researchers thus remain of type θ. The inequality is weak to reflect the fact that type $\theta^a = \theta$ will also not want to pay the cost to "acquire" his or her own current type. The second term in the bracket captures young g-researchers of type $\theta' \neq \theta$ who are matched with a mentor of type θ (whose mass is λ_{t-1}^{θ}) and decide to be advised by them. The remaining dynamics for $\lambda_t^{\theta,m}$ and $\lambda_t^{\theta,f}$ are the same as in the main model. Note that if $\widetilde{C}(\theta, \theta^a) \to \infty$ for all types (e.g. $P \to U$) then the first term in the bracket converges to $p^{\theta,m}$ and the second to 0, returning to the original dynamics.

To gauge the type of dynamics that emerges from Eq. (A.29), note that initially all λ_{t-1}^{θ} are likely small, and thus for a given cost function, both conditions $\widetilde{C}(\theta, \theta^a) \geq \gamma^{\theta^a} \lambda_{t-1}^{\theta^a} - \gamma^{\theta} \lambda_{t-1}^{\theta}$ and $\widetilde{C}(\theta', \theta) < \gamma^{\theta} \lambda_{t-1}^{\theta} - \gamma^{\theta'} \lambda_{t-1}^{\theta'}$ are likely to hold. That is, in this case, the system

runs as in the benchmark case in Section 2.2.. However, as we know from our preceding analysis, λ_t^θ converges to zero for all $\theta \notin \{\theta^*, \theta^m, \theta^f\}$. Specifically, consider the case in which eventually only θ^m and θ^f survive, so $\lambda_t^{\theta^m}$ and $\lambda_t^{\theta^f}$ increase, and the former does so at a faster rate. Intuitively, suppose t is large enough so the mass of established researchers satisfies $\lambda_{t-1}^{\theta^m} + \lambda_{t-1}^{\theta^f} \approx 1$. By symmetry, recall also that $\gamma^{\theta^m} = \gamma^{\theta^f} = \overline{\gamma}$ and the distance between θ^f and θ^m is just $\widetilde{C}(\theta^m, \theta^f) = \widetilde{C}(\theta^f, \theta^m) = \widetilde{C}$. Consistently with the assumption of a large M-group mass of referees initially, let $\left(\lambda_{t-1}^{\theta^m} - \lambda_{t-1}^{\theta^f}\right) > 0$ with $0 < \widetilde{C} < \overline{\gamma}\left(\lambda_{t-1}^{\theta^m} - \lambda_{t-1}^{\theta^f}\right)$. The dynamics then specializes to

$$a_t^{\theta^m,g} = \overline{\gamma}\ \lambda_{t-1}^{\theta^m}\left[p^{\theta^m,g}\ +\ \lambda_{t-1}^{\theta^m}\left(p^{\theta^f,g} + \sum_{\theta':\theta'\neq\theta^m,\theta^f,\widetilde{C}(\theta',\theta^m)<\overline{\gamma}\lambda_{t-1}^{\theta^m}} p^{\theta',g}\right)\right] \tag{A.30}$$

$$a_t^{\theta^m,f} = \overline{\gamma}\ \lambda_{t-1}^{\theta^m}\left[p^{\theta^m,f}\ +\ \lambda_{t-1}^{\theta^m}\left(p^{\theta^f,f} + \sum_{\theta':\theta'\neq\theta^m,\theta^f,\widetilde{C}(\theta',\theta^m)<\overline{\gamma}\lambda_{t-1}^{\theta^m}} p^{\theta',f}\right)\right] \tag{A.31}$$

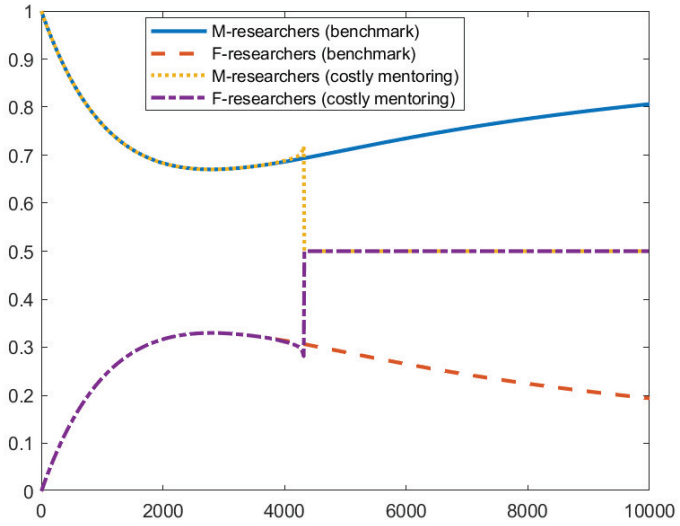
$$a_t^{\theta^f,g} = \overline{\gamma}\ \left(\lambda_{t-1}^{\theta^f}\right)^2\left(\sum_{\theta':\theta'\neq\theta^m,\theta^f,\widetilde{C}(\theta',\theta^f)<\overline{\gamma}\lambda_{t-1}^{\theta^f}} p^{\theta',g}\right) \tag{A.32}$$

$$a_t^{\theta^f,f} = \overline{\gamma}\ \left(\lambda_{t-1}^{\theta^f}\right)^2\left(\sum_{\theta':\theta'\neq\theta^m,\theta^f,\widetilde{C}(\theta',\theta^f)<\overline{\gamma}\lambda_{t-1}^{\theta^f}} p^{\theta',f}\right) \tag{A.33}$$

for $g \in \{f, m\}$. Comparing these expressions with the benchmark case, we see that each $a_t^{\theta^m g}$ is weakly larger than then in the benchmark case, and hence $\lambda_t^{\theta^m}$ increases further over time. If $\lambda_t^{\theta^m}$ becomes sufficiently large, for many young researchers of type θ' the condition $\widetilde{C}(\theta', \theta^m) < \overline{\gamma}\lambda_{t-1}^{\theta^m}$ will hold, but the condition $\widetilde{C}(\theta', \theta^f) < \overline{\gamma}\lambda_{t-1}^{\theta^f}$ will not hold (the details depend on the cost structure). Indeed, if $\max_{\theta'}\left\{\widetilde{C}(\theta', \theta^m)\right\} = \overline{C} < \overline{\gamma}$ and $\lambda_{t-1}^{\theta^m} > \overline{C}/\overline{\gamma}$, then *all* young researchers will be willing to pay a cost to become type θ^m and none will be willing to pay to became any other type. The system then quickly converges to $\lambda_t^{\theta^m} = 1$.

Figure A.8 illustrates the dynamics resulting from Eq. (A.29), under the same parameters as in Section 4. and a cost function $C(\theta, \theta') = \beta\sum_{n=1}^{N}(\theta_n - \theta'_n)^2$, with $\beta = 0.025$. 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 the in the right panel of Figure A.9. At some point, the mass of $\lambda_t^{\theta^m}$ is sufficiently large to induce all young researchers, M and F, decide to pay the cost and the system (nearly) jumps. The reason is that *all* 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

Figure A.8: Fraction of F and M Researchers with Costly Mentoring (low costs)



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$, cost function $C(\theta, \theta') = 0.0250 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

their advisors. Moreover, we reach group balance, as all young M- and F-researchers decide to become θ^m, and there are equal masses of them. However, the downside is that group balance is achieved at the expense of weeding out valuable research characteristics that are more prevalent among young F-researchers—there is, again, loss of talent.

A3. Co-authorship

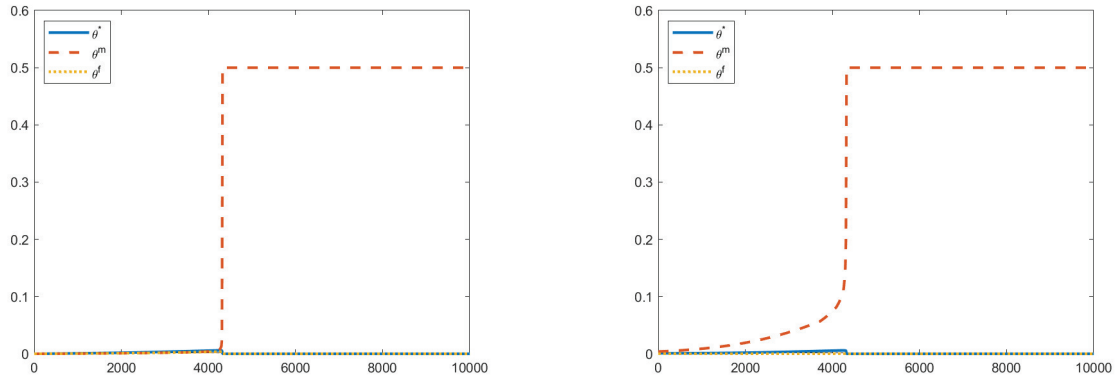
This section briefly explores the implications of our model's dynamics for inferences about the relative (objective) quality of coauthors in a joint project.

We show that, consistently with the findings in Sarsons et al. (2021), if research co-authored by a young M-researcher and a young F-researcher is accepted, then the expected quality of the M-researcher is higher. For simplicity, we consider an economy that has reached its steady state, and such that only types θ^m and θ^f are represented in the population of established scholars. Hence, a joint research project is accepted if and only if its vector of characteristics is θ^m or θ^f.

Proposition A.1 *Let the economy be at its steady state with only types θ^f and θ^m surviving. For each researcher of type θ, define $L(\theta) = \sum_{n=1}^{N} \theta_n$ its objective quality. Let a*

Figure A.9: Types of Established F and M Researchers with Costly Mentoring (low cost) .

(a) F researchers (b) M researchers



Types of established F (left) and M (right) researchers with costly mentoring. We show the masses of types $\theta^* = (1, 1, ..., 1)$, $\theta^m = (1, ..., 1, 0, ..., 0)$, and $\theta^f = (0, ..., 0, 1, ..., 1)$. Initial reviewers: $\lambda_0 = p^m$. Parameters: $\phi = 0.0.5742$ $(d = 0.3)$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$; cost function: $C(\theta, \theta') = 0.0250 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

research that is coauthored by type θ^a and θ^b be of type $\theta = \theta^a \vee \theta^b$, where $\vee$ denotes the component-wise maximum. Let researcher $a \in M$ and $b \in F$. Then, conditional on acceptance of the joint work, i.e. $\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}$, we have

$$\mathrm{E}[L(\theta^a)|\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}] > \mathrm{E}[L(\theta^b)|\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}]$$

The intuition of the result is that referees are more frequently of type θ^m, and, in addition, θ^m is more frequent in the M population than in the F population. It follows that conditional on the joint work being accepted, it is then more likely it is due for the M characteristics than the F characteristics.

Proof of Proposition A.1 Let θ^a and θ^b be the types of the two young researchers. We assume that the type of the joint project is the elementwise maximum of θ^a and θ^b: that is, the project displays characteristics i if and only if at least one of the researchers displays it.

For $g = m, f$, let $\Theta^g = \{(\theta, \theta') : \theta \vee \theta' = \theta^g\}$, where $\vee$ denotes the component-wise maximum. Note that, if $(\theta, \theta') \in \Theta^m$, then $\theta_i = \theta'_i = 0$ for $i = N/2 + 1, \ldots, N$; similarly, if $(\theta, \theta') \in \Theta^f$, then $\theta_i = \theta'_i = 0$ for $i = 1, \ldots, N/2$. Moreover, $(\theta, \theta') \in \Theta^g$ iff $(\theta', \theta) \in \Theta^g$ for $g = m, f$. Finally, $(\theta, \theta') \in \Theta^m$ if and only if $(\bar{\theta}, \bar{\theta}') \in \Theta^f$, where $\bar{\theta}, \bar{\theta}'$ are defined by $\bar{\theta}_{i+N/2} = \theta_i$, $\bar{\theta}'_{i+N/2} = \theta'_i$ and $\bar{\theta}_i = \bar{\theta}'_i = 0$ for $i = 1, \ldots, N/2$; furthermore, these types satisfy

$$p^{\theta,m} = p^{\bar{\theta},f} \qquad \text{and} \quad p^{\theta',f} = p^{\bar{\theta}',m}. \tag{A.34}$$

Then, invoking the above properties, the probability that the joint project is accepted—that is, the probability that $\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}$—is

$$\begin{aligned}
&\gamma^{\theta^m}\bar{\lambda}^{\theta^m} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f} \sum_{(\theta,\theta')\in\Theta^f} p^{\theta,m}\cdot p^{\theta',f} \\
=&\gamma^{\theta^m}\bar{\lambda}^{\theta^m} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f} \sum_{(\theta,\theta')\in\Theta^m} p^{\bar{\theta},m}\cdot p^{\bar{\theta}',f} \\
=&\gamma^{\theta^m}\bar{\lambda}^{\theta^m} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f} \sum_{(\theta',\theta)\in\Theta^m} p^{\bar{\theta}',m}\cdot p^{\bar{\theta},f} \\
=&\gamma^{\theta^m}\bar{\lambda}^{\theta^m} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f} \sum_{(\theta',\theta)\in\Theta^m} p^{\theta',f}\cdot p^{\theta,m} \\
=&\gamma^{\theta^m}\bar{\lambda}^{\theta^m} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,f}\cdot p^{\theta',m} \\
=&(\gamma^{\theta^m}\bar{\lambda}^{\theta^m} + \gamma^{\theta^f}\bar{\lambda}^{\theta^f}) \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f} \\
=&\gamma_0\rho^{N/2} \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}p^{\theta',f} \equiv \gamma_0\rho^{N/2}\Pi,
\end{aligned}$$

where the last equality follows from the definition of γ^θ and the fact that θ^m, θ^f are the only surviving types.

Now let $L(\theta) = \sum_i \theta_i$. We claim that the expectation of $L(\theta^a) - L(\theta^b)$ conditional on $\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}$ is strictly positive—that is, the expected quality of a, the young M coauthor, is strictly higher than the expected quality of that of the young F coauthor b.

First,

$$\begin{aligned}
\Delta &\equiv \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m}\cdot p^{\theta',f}[L(\theta) - L(\theta')] \\
&= \sum_{(\theta,\theta')\in\Theta^m:L(\theta)>L(\theta')} p^{\theta,m}\cdot p^{\theta',f}[L(\theta) - L(\theta')] + \sum_{(\theta,\theta')\in\Theta^m:L(\theta)<L(\theta')} p^{\theta,m}\cdot p^{\theta',f}[L(\theta) - L(\theta')] \\
&= \sum_{(\theta,\theta')\in\Theta^m:L(\theta)>L(\theta')} [p^{\theta,m}\cdot p^{\theta',f} - p^{\theta',m}\cdot p^{\theta,f}][L(\theta) - L(\theta')] > 0.
\end{aligned}$$

The last equality follows because $(\theta, \theta') \in \Theta^m$ if and only if $(\theta', \theta) \in \Theta^m$, and of course $L(\theta) > L(\theta')$ iff $L(\theta') < L(\theta)$. The inequality follows because, if $L(\theta) > L(\theta')$, then by assumption $p^{\theta,m} > p^{\theta',m}$ and $p^{\theta',f} > p^{\theta,f}$.

Repeating the calculations for Θ^f and again appealing to the properties of pairs $(\theta, \theta') \in$

Θ^m and the corresponding types $(\bar{\theta}, \bar{\theta}') \in \Theta^f$,

$$\sum_{(\theta,\theta')\in\Theta^f} p^{\theta,m} \cdot p^{\theta',f}[L(\theta) - L(\theta')] = \sum_{(\theta,\theta')\in\Theta^f : L(\theta)>L(\theta')} [p^{\theta,m} \cdot p^{\theta',f} - p^{\theta',m} \cdot p^{\theta,f}][L(\theta) - L(\theta')]$$
$$= \sum_{(\theta,\theta')\in\Theta^m : L(\theta)>L(\theta')} [p^{\bar{\theta},m} \cdot p^{\bar{\theta}',f} - p^{\bar{\theta}',m} \cdot p^{\bar{\theta},f}][L(\bar{\theta}) - L(\bar{\theta}')]$$
$$= \sum_{(\theta,\theta')\in\Theta^m : L(\theta)>L(\theta')} [p^{\theta,f} \cdot p^{\theta',m} - p^{\theta',f} \cdot p^{\theta,m}][L(\theta) - L(\theta')] =$$
$$= - \sum_{(\theta,\theta')\in\Theta^m} p^{\theta,m} \cdot p^{\theta',f}[L(\theta) - L(\theta')] = -\Delta.$$

Finally, the expected difference in the number of characteristics of θ^a and θ^b is

$$\mathrm{E}[L(\theta^a) - L(\theta^b) | \theta^a \vee \theta^b \in \{\theta^m, \theta^f\}] = \frac{\gamma^{\theta^m}\bar{\lambda}^{\theta^m}\Delta - \gamma^{\theta^f}\bar{\lambda}^{\theta^f}\Delta}{\gamma_0\rho^{N/2}\Pi} = \frac{\rho^{N/2}\Delta}{\Pi}(\bar{\lambda}^{\theta^m} - \bar{\lambda}^{\theta^f}) > 0,$$

as asserted.

Q.E.D

A4. Proofs

We first characterize key features of the population dynamics for an arbitrary, finite set Θ of types, with initial distribution $\lambda_0 \in \Delta(\Theta)$, such that $\lambda_0 = \lambda_0^m + \lambda_0^f$ for $\lambda_0^m, \lambda_0^f \in \mathbb{R}_+^\Theta$, and per-period inflows $q^g = (q^{\theta,g})_{\theta\in\Theta} \in \mathbb{R}_+^\Theta \setminus \{0\}$, for $g \in \{f, m\}$. It is also convenient to define $q = q^m + q^f$. Then, for $g \in \{f, m\}$, the dynamics are given by

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}\left(1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}\right) + \lambda_{t-1}^{\theta} q^{\theta,g} \tag{A.35}$$
$$\lambda_t^{\theta} = \lambda_t^{\theta,m} + \lambda_t^{\theta,f}. \tag{A.36}$$

The body of the paper focuses on the special case $q^{\theta,m} = \gamma^\theta p^{\theta,m}$, $q^{\theta,f} = \gamma^\theta p^{\theta,f}$.

Theorem A.1 *Assume that $q^\theta \leq 1$ for all $\theta \in \Theta$. Then, for all $t \geq 0$, $\lambda_t \in \Delta(\Theta)$, and $\lambda_t^m, \lambda_t^f \in \mathbb{R}_+^\Theta$. Moreover:*

1. *if $\lambda_0^\theta = 0$, then $\lambda_t^\theta = 0$ for all $t \geq 0$;*
2. *if $\lambda_0^\theta > 0$, then $\lambda_t^\theta > 0$ for all $t \geq 0$;*

3. *for* $\theta, \tilde{\theta} \in \Theta$ *with* $\lambda_0^\theta \cdot \lambda_0^{\tilde{\theta}} > 0$*:*

   (a) $\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} - \frac{\lambda_t^{\tilde{\theta}}}{\lambda_{t-1}^{\tilde{\theta}}} = q^\theta - q^{\tilde{\theta}}$ *for all* $t \geq 1$*, and*

   (b) $q^\theta > q^{\tilde{\theta}}$ *implies* $\frac{\lambda_t^\theta}{\lambda_t^{\tilde{\theta}}} \to \infty$*, and* $q^\theta = q^{\tilde{\theta}}$ *implies* $\frac{\lambda_t^\theta}{\lambda_t^{\tilde{\theta}}} = \frac{\lambda_o^\theta}{\lambda_o^{\tilde{\theta}}}$ *for all* $t \geq 0$*;*

4. *define the set*

$$\Theta^{\max} = \{\theta \in \Theta \,:\, \lambda_0^\theta > 0,\, \theta \in \arg\max_{\theta' \in \Theta} q^{\theta'}\} \tag{A.37}$$

   and let $\bar{\lambda} \in \Delta(\Theta)$ *be such that*

$$\bar{\lambda}^{\tilde{\theta}} = \begin{cases} \frac{\lambda_0^{\tilde{\theta}}}{\sum_{\theta \in \Theta^{\max}} \lambda_0^\theta} & \tilde{\theta} \in \Theta^{\max} \\ 0 & \tilde{\theta} \notin \Theta^{\max} \end{cases}: \tag{A.38}$$

   then $\lim_{t\to\infty} \lambda_t = \bar{\lambda}$*;*

5. *define*

$$\bar{\lambda}^{\tilde{\theta},f} = \begin{cases} \frac{\lambda_0^{\tilde{\theta}} q^{\tilde{\theta},f}}{\sum_{\theta \in \Theta^{\max}} \lambda_0^\theta q^\theta} & \tilde{\theta} \in \Theta^{\max} \\ 0 & \tilde{\theta} \notin \Theta^{\max} \end{cases} \quad \textit{and} \quad \bar{\lambda}^{\tilde{\theta},m} = \begin{cases} \frac{\lambda_0^{\tilde{\theta}} q^{\tilde{\theta},m}}{\sum_{\theta \in \Theta^{\max}} \lambda_0^\theta q^\theta} & \tilde{\theta} \in \Theta^{\max} \\ 0 & \tilde{\theta} \notin \Theta^{\max} \end{cases}: \tag{A.39}$$

   then $\lim_{t\to\infty} \lambda_t^f = \bar{\lambda}^f$ *and* $\lim_{t\to\infty} \lambda_t^m = \bar{\lambda}^m$*.*

Proof: Eqs. (A.35) and (A.36) imply that

$$\lambda_t^\theta = \left(1 - \sum_{\theta' \in \Theta} \lambda_{t-1}^{\theta'} q^{\theta'}\right) \lambda_{t-1}^\theta + \lambda_{t-1}^\theta q^\theta. \tag{A.40}$$

By assumption $\lambda_0 \in \Delta(\Theta)$. Inductively, suppose $\lambda_{t-1} \in \Delta(\Theta)$ and $\lambda_{t-1}^m, \lambda_{t-1}^f \in \mathbb{R}_+^\Theta$. Summing over Θ on both sides of Eq. (A.40) yields $\sum_\theta \lambda_t^\theta = (1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'})(\sum_\theta \lambda_{t-1}^\theta) + \sum_\theta \lambda_{t-1}^\theta q^\theta = (1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}) + \sum_\theta \lambda_{t-1}^\theta q^\theta = 1$. Furthermore, since $\lambda_{t-1} \in \Delta(\Theta)$, $\sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} \in [\min_{\theta'} q^{\theta'}, \max_{\theta'} q^{\theta'}] \subseteq [0,1]$; moreover, $q^\theta \geq 0$ and $\lambda_{t-1}^\theta \geq 0$, so Eq. (A.40) implies that $\lambda_t^\theta \geq 0$ as well. By the same argument, $q^\theta \geq 0$ and $\lambda_{t-1}^{\theta,g} \geq 0$ for $g \in \{f, m\}$ imply $\lambda_t^{\theta,g} \geq 0$ for $g \in \{f, m\}$ as well by Eq. (A.35). Thus, $\lambda_t \in \Delta(\Theta)$, and $\lambda_t^g \in \mathbb{R}_+^\Theta$ for each g.

Claim 1 is immediate. For Claim 2, again we argue by induction. For $t = 0$, the claim is trivially true. Inductively, assume $\lambda_{t-1}^\theta > 0$. By Eq. (A.40), since as was just shown $1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} \geq 0$, and the inductive hypothesis implies that $\lambda_{t-1}^\theta > 0$, if $q^\theta > 0$ then $\lambda_t^\theta \geq \lambda_{t-1}^\theta q^\theta > 0$. Suppose instead $q^\theta = 0$. If $\sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} = 1$, then, since $q^{\theta'} \leq 1$ for all θ' by assumption, and $\lambda_{t-1} \in \Delta(\Theta)$, it must be that $\lambda_{t-1}^{\theta'} > 0$ implies $q^{\theta'} = 1$: but then $\lambda_{t-1}^\theta = 0$, which contradicts the inductive hypothesis. Thus, $0 \leq \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} < 1$, so Eq. (A.40) implies that $\lambda_t^\theta = \left(1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}\right) \lambda_{t-1}^\theta > 0$.

For Claim 3, divide both sides of Eq. (A.40) for type θ by λ^θ_{t-1}, which is assumed to be positive; this yields

$$\frac{\lambda^\theta_t}{\lambda^\theta_{t-1}} = 1 + q^\theta - \sum_{\theta'} \lambda^{\theta'}_{t-1} q^{\theta'}. \tag{A.41}$$

A similar equation holds for $\tilde{\theta}$. This immediately yields 3(a). To derive 3(b), since $\lambda^{\theta'}_t = \lambda^{\theta'}_0 \cdot \prod_{s=1}^t \frac{\lambda^{\theta'}_s}{\lambda^{\theta'}_{s-1}}$ for $\theta' = \theta, \tilde{\theta}$,

$$\frac{\lambda^\theta_t}{\lambda^{\tilde{\theta}'}_t} = \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \frac{\prod_{s=1}^t \frac{\lambda^\theta_s}{\lambda^\theta_{s-1}}}{\prod_{s=1}^t \frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}}} = \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \prod_{s=1}^t \frac{\frac{\lambda^\theta_s}{\lambda^\theta_{s-1}}}{\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}}} = \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \prod_{s=1}^t \frac{\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}} + q^\theta - q^{\tilde{\theta}}}{\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}}} = \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \prod_{s=1}^t \left(1 + \frac{q^\theta - q^{\tilde{\theta}}}{\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}}}\right).$$

If $q^\theta = q^{\tilde{\theta}}$, then every term in parentheses equals 1, and the claim follows. If instead $q^\theta > q^{\tilde{\theta}}$, recall that, by Eq. (A.41), for all $s \geq 1$, since $\lambda_{s-1} \in \Delta(\Theta)$ and $q \in [0,1]^{|\Theta|}$, $\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}} \leq 1 + q^{\tilde{\theta}}$. Therefore, each term in parentheses is not smaller than $1 + \frac{q^\theta - q^{\tilde{\theta}}}{1+q^{\tilde{\theta}}} > 1$. It follows that

$$\frac{\lambda^\theta_t}{\lambda^{\tilde{\theta}'}_t} = \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \prod_{s=1}^t \left(1 + \frac{q^\theta - q^{\tilde{\theta}}}{\frac{\lambda^{\tilde{\theta}}_s}{\lambda^{\tilde{\theta}}_{s-1}}}\right) \geq \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0} \cdot \left(1 + \frac{q^\theta - q^{\tilde{\theta}}}{1 + q^{\tilde{\theta}}}\right)^t \to \infty.$$

For Claim 4, consider first $\tilde{\theta} \notin \Theta^{\max}$, and fix $\theta \in \Theta^{\max}$ arbitrarily. Then $\frac{\lambda^\theta_t}{\lambda^{\tilde{\theta}}_t} \to \infty$ by Claim 3(b). Suppose that there is a subsequence $(\lambda_{t(\ell)})_{\ell \geq 0}$ such that $\lambda^{\tilde{\theta}}_{t(\ell)} \geq \epsilon$ for some $\epsilon > 0$ and all $\ell \geq 0$. Since $\frac{\lambda^\theta_{t(\ell)}}{\lambda^{\tilde{\theta}}_{t(\ell)}} \to \infty$ as well, there is ℓ large enough such that $\frac{\lambda^\theta_{t(\ell)}}{\lambda^{\tilde{\theta}}_{t(\ell)}} > \frac{1}{\epsilon}$: but then $\Lambda^\theta_{t(\ell)} > 1$ for such ℓ: contradiction. Thus, for every $\epsilon > 0$, eventually $\lambda^{\tilde{\theta}}_t < \epsilon$: that is, $\lambda^{\tilde{\theta}}_t \to 0$.

Next, consider $\tilde{\theta} \in \Theta^{\max}$. By Claim 2, $\lambda^{\tilde{\theta}}_t > 0$ and $\sum_{\theta \in \Theta^{\max}} \lambda^\theta_t > 0$, and

$$\frac{\lambda^{\tilde{\theta}}_t}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_t} = \frac{1}{\sum_{\theta \in \Theta^{\max}} \frac{\lambda^\theta_t}{\lambda^{\tilde{\theta}}_t}} = \frac{1}{\sum_{\theta \in \Theta^{\max}} \frac{\lambda^\theta_0}{\lambda^{\tilde{\theta}}_0}} = \frac{\lambda^{\tilde{\theta}}_0}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_0} = \bar{\lambda}^{\tilde{\theta}},$$

where the third inequality follows from Claim 3(b). Therefore,

$$\lambda^{\tilde{\theta}}_t = \frac{\lambda^{\tilde{\theta}}_t}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_t} \cdot \left(\sum_{\theta \in \Theta^{\max}} \lambda^\theta_t\right) = \bar{\lambda}^{\tilde{\theta}} \cdot \left(1 - \sum_{\theta \notin \Theta^{\max}} \lambda^\theta_t\right) \to \bar{\lambda}^{\tilde{\theta}},$$

because, as was just shown above, $\lambda^\theta_t \to 0$ for $\theta \notin \Theta^{\max}$.

Finally, consider Claim 5. Fix $g \in \{f, m\}$. First, since $0 \leq \lambda^{\theta,g}_t \leq \lambda^\theta_t$ for all $t \geq 0$, if $\theta \notin \Theta^{\max}$ then by Claim 4 $\lambda^\theta_t \to \bar{\lambda}^\theta = 0$, and so $\lambda^{\theta,g}_t \to 0 = \bar{\lambda}^{\theta,g}$ as well. Thus, focus on the case $\theta \in \Theta^{\max}$, so that by Claim 4 $\bar{\lambda}^\theta > 0$.

If $\sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} = 1$, then Eq. (A.35) and the fact that $\sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} \in [0,1]$ and $0 \leq \lambda_{t-1}^{\theta,g} \leq \lambda_{t-1}^{\theta} \leq 1$ for all θ imply that

$$\lambda_t^{\theta,g} = \left(1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}\right) \lambda_{t-1}^{\theta,g} + \lambda_{t-1}^{\theta} q^{\theta,g} \in \left[\lambda_{t-1}^{\theta} q^{\theta,g}, 1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} + \lambda_{t-1}^{\theta} q^{\theta,g}\right]$$

and both endpoints of the interval in the r.h.s. converge to $\bar{\lambda}^{\theta} q^{\theta,g}$ by Claim 4 if $\sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} = 1$. Furthermore, the same assumption implies that $\bar{\lambda}^{\theta} q^{\theta,g} = \bar{\lambda}^{\theta,g}$, so $\lambda_t^{\theta,g} \to \bar{\lambda}^{\theta,g}$.

Now consider the case $0 < \sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} < 1$. (The set $\Theta^{\max}$ is non-empty, and since $q \in \mathbb{R}_+^{\Theta} \setminus \{0\}$, there is $\theta^+ \in \Theta^{\max}$ with $q^{\theta^+} > 0$; by Claim 4, $\bar{\lambda}^{\theta'} > 0$ for $\theta' \in \Theta^{\max}$, so in particular $\bar{\lambda}^{\theta^+} > 0$; but then $\sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} \geq \bar{\lambda}^{\theta^+} q^{\theta^+} > 0$.) It is convenient to let $q_t = \sum_{\theta'} \lambda_t^{\theta'} q^{\theta'}$ and $\bar{q} = \sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} = \lim_{t\to\infty} q_t$, where the second equality follows from Claim 4. Thus, Eq. (A.35) can be written as

$$\lambda_t^{\theta,g} = (1 - q_{t-1}) \lambda_{t-1}^{\theta,g} + \lambda_{t-1}^{\theta} q^{\theta,g}. \tag{A.42}$$

In addition, $\bar{q} \in (0,1)$.

We claim that, for all $T \geq 0$ and $t > T$,

$$\lambda_t^{\theta,g} = \lambda_T^{\theta,g} \prod_{s=T}^{t-1} (1 - q_s) + q^{\theta,g} \sum_{s=T}^{t-1} \lambda_s^{\theta} \prod_{r=s+1}^{t-1} (1 - q_r). \tag{A.43}$$

For $t = T + 1$, this follows from Eq. (A.42). Inductively, assume it holds for $t - 1 > T$. Then, by Eq. (A.42) and the inductive hypothesis,

$$\begin{aligned}
\lambda_t^{\theta,g} &= (1 - q_{t-1}) \left[\lambda_T^{\theta,g} \prod_{s=T}^{t-2} (1 - q_s) + q^{\theta,g} \sum_{s=T}^{t-2} \lambda_s^{\theta} \prod_{r=s+1}^{t-2} (1 - q_r)\right] + \lambda_{t-1}^{\theta} q^{\theta,g} = \\
&= \lambda_T^{\theta,g} \prod_{s=T}^{t-1} (1 - q_s) + q^{\theta,g} \sum_{s=T}^{t-1} \lambda_s^{\theta} \prod_{r=s+1}^{t-1} (1 - q_r),
\end{aligned}$$

as claimed.

Fix $\epsilon > 0$ such that $\bar{\lambda}^{\theta} - \epsilon > 0$, $\bar{q} - \epsilon > 0$, $1 - \bar{q} + \epsilon < 1$, and $1 - \bar{q} - \epsilon > 0$. This is possible because $\bar{\lambda}^{\theta} > 0$ and $\bar{q} \in (0,1)$, hence $1 - \bar{q} \in (0,1)$.

Since $\lambda_t^{\theta} \to \bar{\lambda}^{\theta}$ and $q_t \to \bar{q}$, there is $T \geq 0$ such that, for all $t > T$, $\lambda_t^{\theta} < \bar{\lambda}^{\theta} + \epsilon$ and

$q_t > \bar{q} - \epsilon$. Hence, for such $t > T$, Eq. (A.43) implies that

$$\begin{aligned}
\lambda_t^{\theta,g} \leq & \lambda_T^{\theta,g} \prod_{s=T}^{t-1}(1-\bar{q}+\epsilon) + q^{\theta,g} \sum_{s=T}^{t-1}(\bar{\lambda}^\theta + \epsilon) \prod_{r=s+1}^{t-1}(1-\bar{q}+\epsilon) = \\
= & \lambda_T^{\theta,g}(1-\bar{q}+\epsilon)^{t-T} + q^{\theta,g}(\bar{\lambda}^\theta+\epsilon) \sum_{s=T}^{t-1}(1-\bar{q}+\epsilon)^{t-1-s} = \\
= & \lambda_T^{\theta,g}(1-\bar{q}+\epsilon)^{t-T} + q^{\theta,g}(\bar{\lambda}^\theta+\epsilon) \sum_{s=0}^{t-1-T}(1-\bar{q}+\epsilon)^{s} = \\
= & \lambda_T^{\theta,g}(1-\bar{q}+\epsilon)^{t-T} + q^{\theta,g}(\bar{\lambda}^\theta+\epsilon) \frac{1-(1-\bar{q}+\epsilon)^{t-T}}{\bar{q}-\epsilon} \to \frac{q^{\theta,g}(\bar{\lambda}^\theta+\epsilon)}{\bar{q}-\epsilon}.
\end{aligned}$$

This implies that $\limsup_t \lambda_t^{\theta,g} \leq \frac{q^{\theta,g}(\bar{\lambda}^\theta+\epsilon)}{\bar{q}-\epsilon}$. Since this must hold for all $\epsilon > 0$, it must be that $\limsup_t \lambda_t^{\theta,g} \leq \frac{q^{\theta,g}\bar{\lambda}^\theta}{\bar{q}} = \bar{\lambda}^{\theta,g}$.

Similarly, $\lambda_t^\theta \to \bar{\lambda}^\theta$ and $q_t \to \bar{q}$ imply that there is $T \geq 0$ such that, for all $t > T$, $\lambda_t^\theta > \bar{\lambda}^\theta - \epsilon > 0$ and $q_t < \bar{q} + \epsilon < 1$. Then

$$\begin{aligned}
\lambda_t^{\theta,g} \geq & \lambda_T^{\theta,g} \prod_{s=T}^{t-1}(1-\bar{q}-\epsilon) + q^{\theta,g} \sum_{s=T}^{t-1}(\bar{\lambda}^\theta - \epsilon) \prod_{r=s+1}^{t-1}(1-\bar{q}-\epsilon) = \\
= & \lambda_T^{\theta,g}(1-\bar{q}-\epsilon)^{t-T} + q^{\theta,g}(\bar{\lambda}^\theta-\epsilon) \frac{1-(1-\bar{q}-\epsilon)^{t-T}}{\bar{q}+\epsilon} \to \frac{q^{\theta,g}(\bar{\lambda}^\theta-\epsilon)}{\bar{q}+\epsilon},
\end{aligned}$$

so $\liminf_t \lambda_t^{\theta,g} \geq \frac{q^{\theta,g}(\bar{\lambda}^\theta-\epsilon)}{\bar{q}+\epsilon}$. Again, since this must hold for all $\epsilon > 0$, $\liminf_t \lambda_T^{\theta,g} \geq \frac{q^{\theta,g}\bar{\lambda}^\theta}{\bar{q}} = \bar{\lambda}^{\theta,g}$. Hence, $\lambda_t^{\theta,g} \to \bar{\lambda}^{\theta,g}$. *Q.E.D.*

Next, we establish certain basic properties of the symmetric model considered in the paper. Claims 1 and 3 characterize the set $\Theta^{\max}$ for this specification. Claim 2 ensures that the parameterization satisfies the conditions in Theorem A.1.

Lemma A.1 *Assume that, for every $\theta \in \Theta$, γ^θ, $p^{\theta,m}$ and $p^{\theta,f}$ are as defined in Section 2.. Then, for every $\phi \in (\frac{1}{2}, 1)$, N even, $\gamma_0 \in (0,1)$, and $\rho \in (1, \frac{1}{\gamma_0})$:*

1. *the set of maximizers of $\gamma^\theta \cdot (p^{\theta,m} + p^{\theta,f})$ is $\{\theta^m, \theta^f\}$ if $\rho < \bar{\rho}(\phi, N)$ and $\{\theta^*\}$ if $\rho > \bar{\rho}(\phi, N)$.*
2. *$0 < \gamma^\theta \cdot [p^{\theta,m} + p^{\theta,f}] \leq 1$.*
3. *there is $\bar{N} > 0$ such that, for all even $N \geq \bar{N}$, the maximizers of $\gamma^\theta \cdot (p^{\theta,m} + p^{\theta,f})$ are θ^m and θ^f.*

Recall that $\bar{\rho}(\cdot)$ is defined in Eq. (10).

Proof: Write

$$\begin{aligned}p^{\theta,m} &= \phi^{\sum_{n=1}^{N/2}\theta_n}(1-\phi)^{N/2-\sum_{n=1}^{N/2}\theta_n}\cdot(1-\phi)^{\sum_{n=N/2+1}^{N}\theta_n}\phi^{N/2-\sum_{n=N/2+1}^{N}\theta_n} = \\ &= \phi^{N/2+\sum_{n=1}^{N/2}\theta_n-\sum_{n=N/2+1}^{N}\theta_n}(1-\phi)^{N/2+\sum_{n=N/2+1}^{N}\theta_n-\sum_{n=1}^{N/2}\theta_n} = \\ &= \phi^{N/2}(1-\phi)^{N/2}\left(\frac{\phi}{1-\phi}\right)^{\sum_{n=1}^{N/2}\theta_n-\sum_{n=N/2+1}^{N}\theta_n}.\end{aligned}$$

Similarly

$$p^{\theta,f} = \phi^{N/2}(1-\phi)^{N/2}\left(\frac{\phi}{1-\phi}\right)^{\sum_{n=N/2+1}^{N}\theta_n-\sum_{n=1}^{N/2}\theta_n}.$$

Then $F(\theta)\equiv\gamma^{\theta}(p^{\theta,m}+p^{\theta,f})$ equals

$$\gamma_0\,\rho^{\sum_n\theta_n/N}\cdot\phi^{N/2}(1-\phi)^{N/2}\left[\left(\frac{\phi}{1-\phi}\right)^{\sum_{n=1}^{N/2}\theta_n-\sum_{n=N/2+1}^{N}\theta_n}+\left(\frac{\phi}{1-\phi}\right)^{-\sum_{n=1}^{N/2}\theta_n+\sum_{n=N/2+1}^{N}\theta_n}\right].$$

Since Θ is finite, there exists at least one maximizer θ of $F(\cdot)$. We claim that, if θ satisfies $\theta_n=\theta_m=0$ for some $n\in\{1,\ldots,N/2\}$ and $m\in\{N/2+1,\ldots,N\}$, then it is not a maximizer. To see this, define θ' by $\theta'_\ell=\theta_\ell$ for $\ell\in\{1,\ldots,N\}\setminus\{n,m\}$ and $\theta'_n=\theta'_m=1$. Then $\sum_n\theta'_n>\sum_n\theta_n$, so for $\rho>1$, $\gamma^{\theta'}>\gamma^{\theta}$. On the other hand, the term in square brackets is the same for θ and θ' (and it is strictly positive). Hence, θ is not a maximizer of $F(\cdot)$. It follows that the only candidate maximizers of $F(\cdot)$ have either $\theta_n=1$ for all $n=1,\ldots,N/2$, or $\theta_n=1$ for all $n=N/2,\ldots,N$, or both.

If $\theta_n=1$ for $n=1,\ldots,N/2$, then $F(\theta)=F(\theta')$, where $\theta'_n=1$ for $n=N/2+1,\ldots,N$ and $\theta'_n=\theta_{n+N/2}$ for $n=1,\ldots,N/2$. Hence, it is enough to consider θ such that $\theta_n=1$ for $n=N/2+1,\ldots,N$. Let Θ^f be the collection of such types, and notice that it contains both θ^f (for which $\theta^f_n=0$ for $n=1,\ldots,N/2$) and $\theta^*=(1,\ldots,1)$. We show that the maximizer of $F(\cdot)$ on Θ^f is either θ^f or θ^*.

For each $\theta\in\Theta^f$, factoring out all terms not involving $\sum_{n=1}^{N/2}\theta_n$, $F(\theta)$ is proportional to

$$\rho^{\sum_{n=1}^{N/2}\theta_n/N}\cdot\left[\left(\frac{\phi}{1-\phi}\right)^{\sum_{n=1}^{N/2}\theta_n}+\left(\frac{1-\phi}{\phi}\right)^{\sum_{n=1}^{N/2}\theta_n}\right].$$

Hence, $F(\theta)$ is proportional to $\tilde{F}(\sum_{n=1}^{N/2}\theta_n)$, where $\tilde{F}:[0,\frac{1}{2}]\to\mathbb{R}_+$ is defined by

$$\tilde{F}(x)=\rho^x\left[\left(\frac{\phi}{1-\phi}\right)^x+\left(\frac{1-\phi}{\phi}\right)^x\right].$$

The functions $x \mapsto \rho^{\frac{x}{N}}\Phi^x = \left(\rho^{\frac{1}{N}}\right)^x \Phi^x = \left(\rho^{\frac{1}{N}} \cdot \Phi\right)^x$, for $\Phi = \frac{\phi}{1-\phi} \neq 1$ and $\Phi = \frac{1-\phi}{\phi} \neq 1$ respectively, are non-constant and exponential, hence strictly convex on $[0, \frac{1}{2}]$. Hence, $\tilde{F}(\cdot)$ is also strictly convex on $[0, \frac{1}{2}]$, so its maximum is either at 0 or at $\frac{1}{2}$. Correspondingly, $F(\cdot)$ attains a maximum either at θ^f or at θ^* on the set Θ^f.

To conclude the proof of Claim 1, we calculate the values attained by $F(\cdot)$ at these two extremes:

$$F(\theta^f) = \gamma_0\sqrt{\rho} \cdot [(1-\phi)^N + \phi^N]$$
$$F(\theta^*) = \gamma_0\rho \cdot 2\phi^{N/2}(1-\phi)^{N/2}.$$

Dividing $F(\theta^*)$ and $F(\theta^f)$ by $\gamma_0\sqrt{\rho}\phi^{N/2}(1-\phi)^{N/2}$ and comparing the resulting quantities, we conclude that θ^* is (uniquely) optimal iff

$$2\sqrt{\rho} > \left[\left(\frac{\phi}{1-\phi}\right)^{-\frac{N}{2}} + \left(\frac{1-\phi}{\phi}\right)^{-\frac{N}{2}}\right]$$

or equivalently

$$\rho > \frac{1}{4}\left(\left(\frac{1-\phi}{\phi}\right)^{\frac{N}{2}} + \left(\frac{\phi}{1-\phi}\right)^{\frac{N}{2}}\right)^2 = \bar{\rho}(\phi, N), \tag{A.44}$$

which is Claim 1.

For Claim 2, we show that $(1-\phi)^N + \phi^N \leq 1$ and $\phi^{N/2}(1-\phi)^{N/2} \leq \frac{1}{2}$; this is sufficient, because $\gamma_0 \in (0,1)$ and $\rho \in (1, \frac{1}{\gamma_0})$ by assumption, so also $\gamma_0\sqrt{\rho} \leq \gamma_0\rho < 1$.

The function $N \mapsto (1-\phi)^N + \phi^N$ is strictly decreasing in N, so it is enough to prove the claim for $N = 2$. In this case, $(1-\phi)^2 + \phi^2 = 1 - 2\phi + \phi^2 + \phi^2 = 1 + 2\phi(\phi - 1) < 1$, because $\phi < 1$. Similarly, $N \mapsto [\phi(1-\phi)]^{N/2}$ is decreasing in N, and for $N = 2$ it reduces to $\phi(1-\phi) = \phi - \phi^2$; this is concave and maximized at $\phi = \frac{1}{2}$, where it takes the value $\frac{1}{4} < \frac{1}{2}$.

Finally, for Claim 3, as $N \to \infty$, the first term in the rhs of Eq. (A.44) converges to zero, but the second diverges to infinity. Thus, for N large, only θ^m and θ^f maximize $F(\cdot)$. *Q.E.D.*

We now turn to the proofs of the main Propositions and Corollaries in the text.

Proof of Proposition 3 and Corollary 1: convergence of $(\lambda_t)_{t\geq 0}$, $(\lambda_t^m)_{t\geq 0}$ and $(\lambda_t^f)_{t\geq 0}$ follows from Theorem A.1 and Claim 2 of Lemma A.1. Parts (a) and (b) follow from Claim 1 in Lemma A.1 and Claim 4 in Theorem A.1. Corollary 1 follows from Claim 3 in Lemma A.1. *Q.E.D.*

Proposition 2 follows from Proposition 3.

Proof of Proposition 4: Fix $\theta \in \Theta$, and define θ^{sym} by $\theta^{\text{sym}}_n = \theta_{N+1-n}$ for all $n = 1, \ldots, N$. (Notice that, for some θ, it may be the case that $\theta^{\text{sym}} = \theta$.) We first claim that

$$a_t^{\theta,m} + a_t^{\theta^{\text{sym}},m} \geq a_t^{\theta,f} + a_t^{\theta^{\text{sym}},f}. \tag{A.45}$$

Notice that, if $\theta^{\text{sym}} = \theta$, the above inequality just says that $a_t^{\theta,m} \geq a_t^{\theta,f}$.

Let $m_0 = \sum_{n=1}^{N/2} \theta$ and $m_1 = \sum_{n=N/2+1}^{N} \theta_n$. By definition, $p^{\theta,m} = \phi^{m_0}(1-\phi)^{N/2-m_0}\phi^{N/2-m_1}(1-\phi)^{m_1} = \phi^{(m_0-m_1)+N/2}(1-\phi)^{N/2-(m_0-m_1)} = [\phi(1-\phi)]^{N/2}\left(\frac{\phi}{1-\phi}\right)^{m_0-m_1}$, and similarly $p^{\theta^{sym},m} = [\phi(1-\phi)]^{N/2}\left(\frac{1-\phi}{\phi}\right)^{m_0-m_1}$. Moreover, since p_f is defined with the roles of ϕ and $1-\phi$ reversed, $p^{\theta,f} = p^{\theta^{\text{sym}},m}$ and $p^{\theta,m} = p^{\theta^{\text{sym}},f}$, so $p^{\theta,m} + p^{\theta,f} = p^{\theta^{\text{sym}},m} + p^{\theta^{\text{sym}},f}$. Finally, by construction $\gamma^{\theta} = \gamma^{\theta^{sym}}$.

Suppose that $m_0 \geq m_1$. Since $\phi > \frac{1}{2}$, $p^{\theta,m} \geq p^{\theta^{\text{sym}},m}$. At time 0 we thus have $\lambda_0^{\theta} = p^{\theta,m} \geq p^{\theta^{\text{sym}},m} = \lambda_0^{\theta^{\text{sym}}} > 0$. Then, since $q^{\theta} = \gamma^{\theta}(p^{\theta,m} + p^{\theta,f}) + \gamma^{\theta^{sym}}(p^{\theta^{sym},m} + p^{\theta^{sym},f}) = q^{\theta^{sym}}$, by part 3(a) of Theorem A.1, for every $t > 0$, $\frac{\lambda_t^{\theta}}{\lambda_{t-1}^{\theta}} = \frac{\lambda_t^{\theta^{\text{sym}}}}{\lambda_{t-1}^{\theta^{\text{sym}}}}$, and hence $\frac{\lambda_t^{\theta}}{\lambda_t^{\theta^{\text{sym}}}} = \frac{\lambda_{t-1}^{\theta}}{\lambda_{t-1}^{\theta^{\text{sym}}}} = \frac{\lambda_0^{\theta}}{\lambda_0^{\theta^{\text{sym}}}} \geq 1$. Thus, $\lambda_t^{\theta} \geq \lambda_t^{\theta^{\text{sym}}}$ for all $t > 0$ as well. Finally, letting $\bar{\gamma} \equiv \gamma^{\theta^{\text{sym}}} = \gamma^{\theta}$, for every $t \geq 1$,

$$a_t^{\theta} = a_t^{\theta,m} + a_t^{\theta,f} = \bar{\gamma}\lambda_{t-1}^{\theta}(p^{\theta,m} + p^{\theta,f}) \geq \bar{\gamma}\lambda_{t-1}^{\theta^{\text{sym}}}(p^{\theta^{\text{sym}},m} + p^{\theta^{\text{sym}},f}) = a_t^{\theta^{\text{sym}},m} + a_t^{\theta^{\text{sym}},f} = a_t^{\theta^{\text{sym}}}.$$

All the inequalities in the above paragraph are strict if $m_0 > m_1$; they are reversed if $m_0 \leq m_1$; and hold as equalities if $m_0 = m_1$.

Now, regardless of the values of m_0 and m_1,

$$\begin{aligned}
& a_t^{\theta,m} + a_t^{\theta^{\text{sym}},m} \geq a_t^{\theta,f} + a_t^{\theta^{\text{sym}},f} \\
\Leftrightarrow \quad & \bar{\gamma}(\lambda_{t-1}^{\theta} p^{\theta,m} + \lambda_{t-1}^{\theta^{sym}} p^{\theta^{sym},m}) \geq \bar{\gamma}(\lambda_{t-1}^{\theta} p^{\theta,f} + \lambda_{t-1}^{\theta^{sym}} p^{\theta^{\text{sym}},f}) \\
\Leftrightarrow \quad & \lambda_{t-1}^{\theta}[p^{\theta,m} - p^{\theta,f}] \geq \lambda_{t-1}^{\theta^{\text{sym}}}[p^{\theta^{\text{sym}},f} - p^{\theta^{\text{sym}},m}] \\
\Leftrightarrow \quad & [\lambda_{t-1}^{\theta} - \lambda_{t-1}^{\theta^{\text{sym}}}] \cdot [p^{\theta,m} - p^{\theta,f}] \geq 0,
\end{aligned}$$

where the last step follows from $p^{\theta,m} = p^{\theta^{\text{sym}},f}$ and $p^{\theta,f} = p^{\theta^{\text{sym}},m}$.

If $m_0 = m_1$, then both terms in square brackets equal zero, so equality obtains; in particular, this is true if $\theta = \theta^{\text{sym}}$. If $m_0 > m_1$, then both terms are positive, if $m_0 < m_1$, then both terms are negative. Thus, in any event, the last inequality, and hence Eq. (A.45), holds; furthermore, if $\theta = \theta^{\text{sym}}$, then $a_t^{\theta,m} = a_t^{\theta,f}$.

Now fix $L \in \{0, \ldots, N\}$. Then

$$\begin{aligned}
\sum_{\theta:\sum_n \theta_n = L} a_t^{\theta,m} &= \sum_{\theta:\sum_n \theta_n = L, \theta = \theta^{\text{sym}}} a_t^{\theta,m} + \sum_{\theta:\sum_n \theta_n = L, \theta \neq \theta^{\text{sym}}} a_t^{\theta,m} = \\
&= \sum_{\theta:\sum_n \theta_n = L, \theta = \theta^{\text{sym}}} a_t^{\theta,m} + \frac{1}{2} \sum_{\theta:\sum_n \theta_n = L, \theta \neq \theta^{\text{sym}}} [a_t^{\theta,m} + a_t^{\theta^{\text{sym}},m}] \geq \\
&\geq \sum_{\theta:\sum_n \theta_n = L, \theta = \theta^{\text{sym}}} a_t^{\theta,f} + \frac{1}{2} \sum_{\theta:\sum_n \theta_n = L, \theta \neq \theta^{\text{sym}}} [a_t^{\theta,f} + a_t^{\theta^{\text{sym}},f}] = \\
&= \sum_{\theta:\sum_n \theta_n = L} a_t^{\theta,f}.
\end{aligned}$$

The second equality follows from the observation that, restricting attention to types θ with $\sum_n \theta_n = L$, also $\sum_n \theta_n^{\text{sym}} = L$, so that adding $a_t^{\theta,m} + a_t^{\theta^{\text{sym}},m}$ over all θ with $\theta \neq \theta^{\text{sym}}$ counts each type twice. The inequality follows from Eq. (A.45), which in particular implies that $a_t^{\theta,m} = a_t^{\theta,f}$ if $\theta = \theta^{\text{sym}}$. This inequality is strict if the second summation is non-empty, i.e., if there is θ with $\sum_n \theta_n = L$ and $\theta_n \neq \theta_{N+1-n}$ for some n, because the latter condition implies $\theta \neq \theta^{\text{sym}}$. Finally, the last equality follows by repeating the first two steps backwards, for F-group researchers. *Q.E.D*

Proof of Proposition 5: We begin with a preliminary result.

Lemma A.2 *For all parameter values and initial conditions, and for all $\theta \in \Theta$ and $t \geq 1$,*

$$\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} = (1 - a_t) + \gamma^\theta (p^{\theta,m} + p^{\theta,f});$$

and for $t \geq 2$,

$$\frac{a_t^\theta}{a_{t-1}^\theta} = \frac{a_t^{\theta,m}}{a_{t-1}^{\theta,m}} = \frac{a_t^{\theta,f}}{a_{t-1}^{\theta,f}} = \frac{\lambda_{t-1}^\theta}{\lambda_{t-2}^\theta}.$$

Proof: From Eq. (7), $\lambda_t^\theta = \lambda_t^{\theta,m} + \lambda_t^{\theta,f} = (\lambda_{t-1}^{\theta,m} + \lambda_{t-1}^{\theta,f})(1 - a_t) + \gamma^\theta (p^{\theta,m} + p^{\theta,f})$, which yields the first equation because $\lambda_\tau^\theta > 0$ for all θ and τ.

From Eq. (6), for $t \geq 2$,

$$\frac{a_t^{\theta,g}}{a_{t-1}^{\theta,g}} = \frac{\lambda_{t-1}^\theta \gamma^\theta p^{\theta,g}}{\lambda_{t-2}^\theta \gamma^\theta p^{\theta,g}} = \frac{\lambda_{t-1}^\theta}{\lambda_{t-2}^\theta};$$

similarly,

$$\frac{a_t^\theta}{a_{t-1}^\theta} = \frac{\lambda_{t-1}^\theta \gamma^\theta (p^{\theta,m} + p^{\theta,f})}{\lambda_{t-2}^\theta \gamma^\theta (p^{\theta,m} + p^{\theta,f})} = \frac{\lambda_{t-1}^\theta}{\lambda_{t-2}^\theta}.$$

Q.E.D.

We now prove Proposition 5. For $N = 2$ we only have 4 types, $\Theta = \{(0,0),(1,0),(0,1),(1,1)\}$. Let $a^{L,g} = \sum_{sum_{n=1}^2 \theta_n = L} a^{\theta,g}$ and $a_t^g = \sum_{\ell=0}^2 a_t^{\ell,g}$. From Proposition 4, for all t, $a_t^{1,m} > a_t^{1,f}$, $a_t^{2,m} = a_t^{2,f}$, and $a_t^{0,m} = a_t^{0,f}$. Therefore, $a_t^m > a_t^f$, which implies that the weight on $L = 1$ for accepted M researchers is

$$\frac{a_t^{1,m}}{a_t^m} = 1 - \frac{a_t^{2,m} + a_t^{0,m}}{a_t^m} = 1 - \frac{a_t^{2,f} + a_t^{0,f}}{a_t^m} > 1 - \frac{a_t^{2,f} + a_t^{0,f}}{a_t^f} = \frac{a_t^{1,f}}{a_t^f}.$$

Similarly, $a_t^m > a_t^f$ and $a_t^{0,m} = a_t^{0,f}$, $a_t^{2,m} = a_t^{2,f}$ imply

$$\frac{a_t^{0,m}}{a_t^m} < \frac{a_t^{0,f}}{a_t^f}, \qquad \frac{a_t^{2,m}}{a_t^m} < \frac{a_t^{2,f}}{a_t^f}.$$

Moreover, we claim that, $a_t^{2,g} > a_t^{0,g}$. For $t = 0$, $a_0^{2,g} = a_0^{(1,1),g} = p^{(1,1),m}\gamma^{(1,1)}p^{(1,1),g} > p^{(0,0),m}\gamma^{(0,0)}p^{(0,0),g} = a_0^{(0,0),g} = a_0^{0,g}$, because $p^{(0,0),g} = p^{(1,1),g}$ but $\gamma^{(1,1)} > \gamma^{(0,0)}$. Inductively, from Lemma A.2,

$$\begin{aligned}
a_t^{2,g} = a_t^{(1,1),g} = a_{t-1}^{(1,1),g} \cdot \frac{a_t^{(1,1),g}}{a_{t-1}^{(1,1),g}} = a_{t-1}^{(1,1,g)} \left(1 - a_{t-1} + \gamma^{(1,1)}(p^{(1,1),m} + p^{(1,1),f}\right) > \\
> a_{t-1}^{(1,1,g)} \left(1 - a_{t-1} + \gamma^{(0,0)}(p^{(0,0),m} + p^{(0,0),f}\right) > a_{t-1}^{(0,0,g)} \left(1 - a_{t-1} + \gamma^{(0,0)}(p^{(0,0),m} + p^{(0,0),f}\right) = \\
= a_{t-1}^{(0,0,g)} \frac{a_t^{(0,0),g}}{a_{t-1}^{(0,0,g)}} = a_t^{(0,0),g} = a_t^{0,g}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
0 < \frac{a_t^{0,f}}{a_t^{1,f} + a_t^{2,f} + a_t^{0,f}} - \frac{a_t^{0,m}}{a_t^{1,m} + a_t^{2,m} + a_t^{0,m}} = \frac{a_t^{0,f}}{a_t^{1,f} + a_t^{2,f} + a_t^{0,f}} - \frac{a_t^{0,f}}{a_t^{1,m} + a_t^{2,m} + a_t^{0,m}} < \\
< \left(\frac{a_t^{2,f}}{a_t^{0,f}}\right) \cdot \left(\frac{a_t^{0,f}}{a_t^{1,f} + a_t^{2,f} + a_t^{0,f}} - \frac{a_t^{0,f}}{a_t^{1,m} + a_t^{2,m} + a_t^{0,m}}\right) = \frac{a_t^{2,f}}{a_t^{1,f} + a_t^{2,f} + a_t^{0,f}} - \frac{a_t^{2,f}}{a_t^{1,m} + a_t^{2,m} + a_t^{0,m}} = \\
= \frac{a_t^{2,f}}{a_t^{1,f} + a_t^{2,f} + a_t^{0,f}} - \frac{a_t^{2,m}}{a_t^{1,m} + a_t^{2,m} + a_t^{0,m}};
\end{aligned}$$

the first inequality follows from $a_t^{1,f} < a_t^{1,m}$ and $a_t^{0,f} = a_t^{0,m}$ and $a_t^{2,f} = a_t^{2,m}$, the next equality from $a_t^{0,m} = a_t^{0,f}$, the second inequality from $a_t^{2,f} > a_t^{0,f} > 0$ and the fact that the difference of fractions is positive, and the last equality from $a_t^{2,m} = a_t^{2,f}$.

The result then follows from a symmetry argument.

$$\begin{aligned}
E[L|F] &= \frac{0 \times a_t^{0,f} + a_t^{1,f} + 2a_t^{2,f}}{a_t^f} \\
E[L|M] &= \frac{0 \times a_t^{0,m} + a_t^{1,m} + 2a_t^{2,m}}{a_t^m}
\end{aligned}$$

which, since $a_t^{1,g} = 1 - a_t^{0,g} - a_t^{2,g}$, implies

$$E[L|F] = -1\frac{a_t^{0,f}}{a_t^f} + 1 + \frac{a_t^{2,f}}{a_t^f}$$
$$E[L|M] = -1\frac{a_t^{0,m}}{a_t^m} + 1 + \frac{a_t^{2,m}}{a_t^m}$$

It follows that

$$E[L|F] - E[L|M] = -\left(\frac{a_t^{0,f}}{a_t^f} - \frac{a_t^{0,m}}{a_t^m}\right) + \left(\frac{a_t^{2,f}}{a_t^f} - \frac{a_t^{2,m}}{a_t^m}\right) > 0$$

Q.E.D

Detailed dynamics of the mass of M and F accepted agents. It is useful to rewrite Equation (7) for each group $g = m, f$ as follows:

$$\lambda_t^{\theta,m} - \lambda_{t-1}^{\theta,m} = -\lambda_{t-1}^{\theta,m}\, a_t + \lambda_{t-1}^{\theta,m}\gamma^\theta p^{\theta,m} + \lambda_{t-1}^{\theta,f}\gamma^\theta p^{\theta,m} \quad \text{(A.46)}$$
$$\lambda_t^{\theta,f} - \lambda_{t-1}^{\theta,f} = -\lambda_{t-1}^{\theta,f}\, a_t + \lambda_{t-1}^{\theta,f}\gamma^\theta p^{\theta,f} + \lambda_{t-1}^{\theta,m}\gamma^\theta p^{\theta,f} \quad \text{(A.47)}$$

Consider the dynamics of F-researchers in (A.47), for instance. The change in the mass of F-researchers of type θ decreases due to replacement at the rate a_t, and it then increases due to the young F-researchers who produce quality research and are matched with referees from the F group who share their type and hence view them positively ($\lambda_{t-1}^{\theta,f}\gamma^\theta p^{\theta,f}$), plus the young F-researchers who produce quality research and are matched with M-referees of their own type ($\lambda_{t-1}^{\theta,m}\gamma^\theta p^{\theta,f}$). The asymmetry between the two dynamics (A.46) and (A.47) is apparent in the last two terms of each. If θ is a type that is more prevalent among M-researchers—for instance, $\theta = \theta^m$—then $p^{\theta,f}$ will be small while $p^{\theta,m}$ will be large. If the current mass of M-researchers of type θ is large, then $\lambda_{t-1}^{\theta}\gamma^\theta p^{\theta,m}$ will act to further increase the mass of M-researchers, while the respective term $\lambda_{t-1}^{\theta}\gamma^\theta p^{\theta,f}$ in the F-group dynamics will lead to a smaller increase in the mass of type-θ F-researchers. In particular, if we start from a situation in which *all* referees of type θ are in M-group, then, while they will accept some F-researchers of type θ, they will accept a much larger mass of M-researchers.

This force is at play regardless of the parameter values, and for all types. However, its implications for the limiting group (im)balance in the population depend upon whether or not we are in a "meritocratic" scenario. If research characteristics have a limited effect on the probability of quality research, as in Part (a) of Proposition 3, then θ^m and θ^f are the only types that survive in the limit. These are also the types for which the difference in proportions among young M- and F-researchers is greatest. Thus, in the scenario of Part

(a), the force thus described has the greatest effect, which is further reinforced if initially *all* referees are in M-group. The result is that, in the limit, despite the fact that the mass of young M- and F-researchers appearing at each time t is the same, the referees' self-image bias leads to a limiting population in which the majority of scholars are in M group.

By way of contrast, in the meritocratic scenario of Part (b) in Proposition 3, the type that prevails in the limit is the efficient one, namely θ^*. In our symmetric model, the *same* fraction of young M- and F-researchers are of type θ^*. Therefore, the effect described above becomes more and more muted over time. Consequently, in the limit, the mass of M- and F-scholars is the same.

The following Proposition formalizes the above discussion. We denote by $\Lambda_t^m \equiv \sum_\theta \lambda_t^{\theta,m}$ and $\Lambda_t^f \equiv \sum_\theta \lambda_t^{\theta,f}$ the total mass of M- and F-scholars at date t; $\bar{\Lambda}^m$ and $\bar{\Lambda}^f$ are the corresponding limiting quantities.

Proposition A.2 *Assume that all referees are initially from the M-group, i.e., $\lambda_0 = p^m$.*

(a) If $\rho < \bar{\rho}(\phi, N)$, then the limiting masses are

$$(M\text{-researchers of type } \theta^m)\text{: } \bar{\lambda}^{\theta^m,m} = \frac{(\phi^N)^2}{(\phi^N + (1-\phi)^N)^2}; \tag{A.48}$$

$$(F\text{-researchers of type } \theta^m)\text{: } \bar{\lambda}^{\theta^m,f} = \frac{\phi^N\ (1-\phi)^N}{(\phi^N + (1-\phi)^N)^2}; \tag{A.49}$$

$$(M\text{-researchers of type } \theta^f)\text{: } \bar{\lambda}^{\theta^f,m} = \frac{((1-\phi)^N)^2}{(\phi^N + (1-\phi)^N)^2}; \tag{A.50}$$

$$(F\text{-researchers of type } \theta^f)\text{: } \bar{\lambda}^{\theta^f,f} = \frac{(1-\phi)^N\ \phi^N}{(\phi^N + (1-\phi)^N)^2}; \tag{A.51}$$

with

$$\bar{\lambda}^{\theta^m,m} > \bar{\lambda}^{\theta^m,f} = \bar{\lambda}^{\theta^f,f} > \bar{\lambda}^{\theta^f,m} \tag{A.52}$$

In addition, the total mass of M and F researchers are

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} > 0.5. \tag{A.53}$$

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

Proof of Proposition 6, A.2 and Corollary 2. For Part (a), since $\gamma^{\theta^m} = \gamma^{\theta^f} = \gamma_0 (\rho)^{N/2}$ and, by Proposition 3, $\Theta^{\max} = \{\theta^m, \theta^f\}$, $\bar{\lambda}^{\tilde{\theta},m} = \frac{\lambda_0^{\tilde{\theta}} p^{\tilde{\theta},m}}{\lambda_0^{\theta^m} p^{\theta^m,m} + \lambda_0^{\theta^f} p^{\theta^f,m}}$ for $\tilde{\theta} \in \Theta^{\max}$, and $\bar{\lambda}^{\tilde{\theta},m} = 0$ otherwise; a similar expression holds for $\bar{\lambda}^{\tilde{\theta},f}$. Equations (A.48) through (A.51) then follow from the specification of p^m and p^f. Eq. (13) follows from $\bar{\Lambda}^g = \bar{\lambda}^{\theta^m,g} + \bar{\lambda}^{\theta^f,g}$.

Part (b) follows from the fact that, by Proposition 3 part (b), $\Theta^{\max} = \{\theta^*\}$ in this scenario. Corollary 2 follows from Lemma A.1 Claim (3).

Proposition 6 consists of (b) and the last claim in (a) of Proposition A.2. *Q.E.D.*

Proof of Proposition 7: let $\Theta_{-1} = \Theta$ and $t(-1) = 0$. Also let $\lambda_{0,0}^m = \lambda_{1,0}^m = \lambda_0^m$, $\lambda_{0,0}^f = \lambda_{1,0}^f = \lambda_0^f$, and $\lambda_{0,0} = \lambda_{1,0} = \lambda_{1,0}^m + \lambda_{1,0}^f$. Finally, let $\Theta_0 = \left\{\theta \in \Theta : \lambda_{1,0}^\theta \geq \frac{C}{\gamma^\theta P}\right\}$.

For $j \geq 0$, say that *Conditions* $C(j)$ *hold* if there is a set $\Theta_j \subseteq \Theta_{j-1}$, a period $t(j) > t(j-1)$, and for $\tau = 0, \ldots, t(j) - t(j-1)$, vectors $\lambda_{\tau,j}^m, \lambda_{\tau,j}^f, \lambda_{\tau,j} \in \mathbb{R}_+^\Theta$ such that

(i) for $0 \leq \tau \leq t(j) - t(j-1)$, $\lambda_{\tau,j}^m = \lambda_{t(j-1)+\tau}^m$, $\lambda_{\tau,j}^f = \lambda_{t(j-1)+\tau}^f$, and $\lambda_{\tau,j} = \lambda_{\tau,j}^m + \lambda_{\tau,j}^f$;

(ii) for $0 \leq \tau < t(j) - t(j-1)$, $\lambda_{\tau,j}^\theta \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_j$;

(iii) $\lambda_{\tau,j}^\theta < \frac{C}{\gamma^\theta (P-U)}$ for $0 \leq \tau \leq t(j) - t(j-1)$ and all $\theta \in \Theta \backslash \Theta_j$, and $\lambda_{t(j)-t(j-1),j}^{\theta_0} < \frac{C}{\gamma^{\theta_0}(P-U)}$ for some $\theta_0 \in \Theta_j$.

We claim that, for every $k \geq 0$, if either $k = 0$ or $k > 0$ and Conditions $C(k-1)$ hold, then either Conditions $C(k)$ hold as well, with $\Theta_k \subsetneq \Theta_{k-1}$ in case $k > 0$, or else there exist vectors $\lambda_{\tau,k}^m, \lambda_{\tau,k}^f, \lambda_{\tau,k} \in \mathbb{R}_+^\Theta$ for all $\tau \geq 1$ such that (i) holds for $j = k$, and $\lambda_{\tau,j}^\theta \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_k$. In the latter case, if the sequences of such vectors converge, then $\lim_{\tau\to\infty} \lambda_{\tau,k}^m = \lim_{t\to\infty} \lambda_t^m$ and similarly for $\lambda_{\tau,k}^f$ and $\lambda_{\tau,k}$.

Let $\lambda_{0,k}^{\theta,g} = \lambda_{t(k-1)}^{\theta,g}$ for $g = f, m$; also let $\lambda_{0,k} = \lambda_{0,k}^m + \lambda_{0,k}^f$. Let $\Theta_k = \left\{\theta \in \Theta : \lambda_{0,k}^\theta \geq \frac{C}{\gamma^\theta P}\right\}$. If $k = 0$, then $\Theta_0 \subseteq \Theta = \Theta_{-1}$. Otherwise, $C(k-1)$ must hold, so $\lambda_{0,k} = \lambda_{t(k-1)} = \lambda_{t(k-1)-t(k-2),k-1}$. By (iii), if $\theta \notin \Theta_{k-1}$ then $\lambda_{0,k}^\theta = \lambda_{t(k-1)-t(k-2),k-1}^\theta < \frac{C}{\gamma^\theta P}$, so $\theta \notin \Theta_k$ as well; firthermore, there exists $\theta_0 \in \Theta_{k-1}$ such that $\lambda_{0,k}^{\theta_0} = \lambda_{t(k-1)-t(k-2),k-1}^{\theta_0} < \frac{C}{\gamma^\theta P}$. Therefore, if $k > 0$, then $\Theta_k \subsetneq \Theta_{k-1}$.

Define $q_k^g \in \mathbb{R}_+^\Theta \setminus \{0\}$ for $g = f, m$ by $q_k^{\theta,g} = \gamma^\theta p^{\theta,g}$ if $\theta \in \Theta_k$, and $q_k^{\theta,g} = 0$ otherwise. Then $q_k^{\theta,m} + q_k^{\theta,f} \leq 1$ for all θ. Consider the sequences $(\lambda_{\tau,k}^{\theta,g})_{\tau\geq 0}$ for $g = f, m$ and $(\lambda_{\tau,k}^\theta)_{\tau\geq 0}$ defined by Eqs. (A.35)–(A.36) for the vectors q_k^f, q_k^m .

Suppose first that there are $\bar{\tau} > 0$ and $\theta_0 \in \Theta_k$ such that $\lambda_{\bar{\tau},k}^{\theta_0} < \frac{C}{\gamma^{\theta_0}(P-U)}$. Let $t(k) = t(k-1) + \bar{\tau}$. Then, for each group $g = f, m$, the dynamics in Eqs. (A.35)–(A.36) induced

by the vectors q_k^f, q_k^m for the subsequence $(\lambda^g_{\tau,k})_{\tau=0,\ldots,\bar\tau}$ coincide with those in Eq. (19) for the subsequences $(\lambda^g_t)_{t=t(k-1),\ldots,t(k)}$; thus, (i) holds for $j=k$. Furthermore, (ii) and the second part of (iii) hold for $j=k$ by the definition of $\bar\tau$. For the first part of (iii) with $j=k$, recall that by definition $q_k^{\theta,m}+q_k^{\theta,f}=0$ for $\theta\in\Theta\setminus\Theta_k$; hence, for all $\theta'\in\Theta$ and all $\theta\in\Theta\setminus\Theta_k$, $q_k^{\theta,m}+q_k^{\theta,f}\le q_{m,k}^{\theta'}+q_{f,k}^{\theta'}$. By part 3(a) in Theorem A.1, it must be the case that $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k}\le 1$: otherwise, $\sum_{\theta'\in\Theta}\lambda^{\theta'}_{\tau+1,k}>\sum_{\theta'\in\Theta}\lambda^{\theta'}_{\tau,k}=1$, which contradicts the fact that $\lambda_{\tau+1,k}\in\Delta(\Theta)$ per Theorem A.1. Since by definition $\lambda^\theta_{0,k}<\frac{C}{\gamma^\theta P}$ for $\theta\notin\Theta_k$, it follows that also $\lambda^\theta_{\tau,k}<\frac{C}{\gamma^\theta P}$ for $\tau=0,\ldots,\bar\tau$ and for any such θ. Thus, in this case Conditions $C(k)$ hold.

If instead $\lambda^\theta_{\bar\tau,k}\ge\frac{C}{\gamma^\theta(P-U)}$ for all $\theta\in\Theta_k$, then for each group $g=f,m$, the dynamics in Eqs. (A.35)–(A.36) induced by the vectors $q_{m,k}, q_{f,k}$ for the subsequence $(\lambda^g_{\tau,k})_{\tau\ge0}$ coincide with those in Eq. (19) for the subsequence $(\lambda^g_t)_{t\ge t(k-1)}$. Again, in this case (i) holds for $j=k$. This completes the proof of the claim.

Since the set Θ is finite, there exists $K\ge0$ such that the induction stops—that is, $\lambda^\theta_{\bar\tau,K}\ge\frac{C}{\gamma^\theta(P-U)}$ for all $\theta\in\Theta_K$. Let $\Theta_k^{\max}=\arg\max\{q_k^{\theta,m}+q_k^{\theta,f}:\theta\in\Theta\}$. Since $\Theta_0\supsetneq\Theta_1\supsetneq\ldots\supsetneq\Theta_K$, by the definition of the vectors q_k^g for $g=f,m$, also $\Theta_0^{\max}\supseteq\Theta_1^{\max}\supseteq\ldots\supseteq\Theta_K^{\max}$. Moreover, for every $k=0,\ldots,K-1$, and every $\theta\in\Theta_k^{\max}$, $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k}\ge1$ for $0\le\tau<t(k)-t(k)$; otherwise, by part 3(a) in Theorem A.1, $\sum_{\theta\in\Theta}\lambda^\theta_{\tau+1,k}<\sum_{\theta\in\Theta}\lambda^\theta_{\tau,k}=1$, which contradicts the fact that $\lambda_{\tau+1}\in\Delta(\Theta)$ per Theorem A.1.

Now assume that $\Theta_0^{\max}\subseteq\Theta_0$. Then, for every $\theta\in\Theta_0^{\max}$,

$$\frac{C}{\gamma^\theta P}\le\lambda^\theta_{0,0}\le\lambda^\theta_{t(1)-t(0),0}=\lambda^\theta_{0,1}\le\lambda^\theta_{t(2)-t(1),1}\cdots\le\lambda^\theta_{0,K},$$

so $\theta\in\Theta_k$ for all $k=0,\ldots,K$, and thus $\Theta_0^{\max}=\Theta_1^{\max}=\ldots=\Theta_K^{\max}\equiv\Theta^{\max}$. In addition, again by part 3(a) of Theorem A.1, if $\theta,\theta'\in\Theta^{\max}$, then $\frac{\lambda^\theta_{\tau+1,k}}{\lambda^\theta_{\tau,k}}=\frac{\lambda^{\theta'}_{\tau+1,k}}{\lambda^{\theta'}_{\tau,k}}$ for all $k=0,\ldots,K-1$ and $\tau=0,\ldots,t(k)-t(k-1)$, and for $k=K$ and all $\tau\ge0$. Rearranging terms, $\frac{\lambda^\theta_{\tau+1,k}}{\lambda^{\theta'}_{\tau+1,k}}=\frac{\lambda^\theta_{\tau,k}}{\lambda^{\theta'}_{\tau,k}}$ for such k and τ. Therefore, (i) in Conditions $C(0)...C(K)$ imply that

$$\frac{\lambda^\theta_{0,K}}{\lambda^{\theta'}_{0,K}}=\frac{\lambda^\theta_{t(K-1)}}{\lambda^{\theta'}_{t(K-1)}}=\frac{\lambda^\theta_{t(K-1)-t(K-2),K-1}}{\lambda^{\theta'}_{t(K-1)-t(K-2),K-1}}=\frac{\lambda^\theta_{0,K-1}}{\lambda^{\theta'}_{0,K-1}}=\ldots=\frac{\lambda^\theta_{t(0)-t(-1),0}}{\lambda^{\theta'}_{t(0)-t(-1),0}}=\frac{\lambda^\theta_{0,0}}{\lambda^{\theta'}_{0,0}}=\frac{\lambda^\theta_0}{\lambda^{\theta'}_0}.$$

Therefore, for $\theta\in\Theta^{\max}=\Theta_K^{\max}$, from Theorem A.1 part (4),

$$\bar\lambda^\theta=\bar\lambda^\theta_K=\frac{\lambda^\theta_{0,K}}{\sum_{\theta'\in\Theta^{\max}}\lambda^{\theta'}_{0,K}}=\frac{1}{\sum_{\theta'\in\Theta^{\max}}\frac{\lambda^{\theta'}_{0,K}}{\lambda^\theta_{0,K}}}=\frac{1}{\sum_{\theta'\in\Theta^{\max}}\frac{\lambda^{\theta'}_0}{\lambda^\theta_0}}=\frac{\lambda^\theta_0}{\sum_{\theta'\in\Theta^{\max}}\lambda^{\theta'}_0}.\qquad(A.54)$$

Similarly, for $\theta \in \Theta^{\max}$, part (5) in the same Theorem implies that

$$\bar{\lambda}^{\theta,m} = \bar{\lambda}_K^{\theta,m} = \frac{\lambda_{0,K}^{\theta} q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_{0,K}^{\theta'} q_K^{\theta'}} = \frac{q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_{0,K}^{\theta'}}{\lambda_{0,K}^{\theta}} q_K^{\theta'}} = \frac{q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_0^{\theta'}}{\lambda_0^{\theta}} q_K^{\theta'}} = \frac{\lambda_0^{\theta} q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} q_K^{\theta'}}, \tag{A.55}$$

and analogously for $\bar{\lambda}^{\theta,f}$.

Statements (a.1)–(b) now follow. Recall that $\lambda_0 = p^m$. In (a.1), by assumption $\Theta^{\max} = \Theta_0^{\max} = \{\theta^m, \theta^f\} \subseteq \Theta_0$. Substituting $\lambda_0^{\theta^m} = \phi^N$ and $\lambda_0^{\theta^f} = (1-\phi)^N$ in Eq. (A.54) yields $\bar{\lambda}^{\theta^m} = \frac{\phi^N}{\phi^N + (1-\phi)^N}$. Similarly, substituting for q_K^g, $g = f, m$, and $q_K = q_K^f + q_K^m$ in Eq. (A.55) yields the same expression for $\bar{\lambda}^{\theta^m,m}$ as in Proposition 3, because $\theta \in \Theta^{\max}$ implies that $q_K^{\theta,g} = \gamma^{\theta} p^{\theta,g}$; ditto for $\bar{\lambda}^{\theta^m,f}$, $\bar{\lambda}^{\theta^f,m}$ and $\bar{\lambda}^{\theta^f,f}$, and hence for $\bar{\Lambda}^m$.

For (a.2), $\Theta^{\max} = \Theta_0^{\max} = \{\theta^m\}$. This immediately implies that $\bar{\lambda}^{\theta^m} = \bar{\lambda}_K^{\theta^m} = 1$. Furthermore, from Eq. (A.55), $\bar{\Lambda}^m = \bar{\lambda}^{m,\theta^m} = \bar{\lambda}_K^{m,\theta^m} = \frac{\gamma^{\theta^m} p^{\theta^m,m}}{\gamma^{\theta^m}(p^{\theta^m,m} + p^{\theta^m,f})} = \frac{p^{\theta^m,m}}{p^{\theta^m,m} + p^{\theta^m,f}} = \frac{\phi^N}{\phi^N + (1-\phi)^N}$, as asserted. Finally, we compare this quantity with its counterpart in Eq. (13):

$$\begin{aligned} &\frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} = \frac{(1-\phi)^{2N} + \phi^{2N}}{\left[(1-\phi)^N + \phi^N\right]^2} < \\ &< \frac{(1-\phi)^N \phi^N + \phi^{2N}}{\left[(1-\phi)^N + \phi^N\right]^2} = \frac{(1-\phi)^N + \phi^N}{(1-\phi)^N + \phi^N} \cdot \frac{\phi^N}{(1-\phi)^N + \phi^N} = \frac{\phi^N}{(1-\phi)^N + \phi^N} = \bar{\Lambda}^m, \end{aligned}$$

where the inequality follows from the assumption that $\phi > 0.5$.

The analysis of (b) is analogous to that of (a.2), with θ^* in lieu of θ^m; in this case, $p^{\theta^*,m} = p^{\theta^*,f} = \phi^{N/2}(1-\phi)^{N/2}$, so $\bar{\Lambda}^m = \bar{\lambda}^{\theta^*,m} = \frac{1}{2}$.

The statements about t^{θ} for $\theta \notin \Theta^{\max}$ follow from the construction of $t(0), \ldots, t(K)$. *Q.E.D.*

Proof of Proposition 8. For part 1, the key step is analogous to the proof of Proposition 4, modified to allow for endogenous entry. Let $m_0 = \sum_{n=1}^{N/2} \theta$ and $m_1 = \sum_{n=N/2+1}^{N} \theta_n$. By assumption, $m_0 > m_1$. By definition, $p^{\theta,m} = \phi^{m_0}(1-\phi)^{N/2-m_0}\phi^{N/2-m_1}(1-\phi)^{m_1} = \phi^{(m_0-m_1)+N/2}(1-\phi)^{N/2-(m_0-m_1)} = [\phi(1-\phi)]^{N/2}\left(\frac{\phi}{1-\phi}\right)^{m_0-m_1}$, and similarly $p^{\theta^{sym},m} = [\phi(1-\phi)]^{N/2}\left(\frac{1-\phi}{\phi}\right)^{m_0-m_1}$; since $\phi > \frac{1}{2}$, $p^{\theta,m} > p^{\theta^{\mathrm{sym}},m}$. At time 0 we thus have $\lambda_0^{\theta} = p^{\theta,m} > p^{\theta^{\mathrm{sym}},m} = \lambda_0^{\theta^{\mathrm{sym}}}$. Moreover, since p_f is defined with the roles of ϕ and $1-\phi$ reversed, $p^{\theta,f} = p^{\theta^{\mathrm{sym}},m} < p^{\theta,m} = p^{\theta^{\mathrm{sym}},f}$.

Since $\gamma^{\theta^{\mathrm{sym}}} = \gamma^{\theta}$, it follows that at time 0, if $\lambda_0^{\theta^{\mathrm{sym}}} > \frac{C}{\gamma^{\theta^{\mathrm{sym}}} P}$, then also $\lambda_0^{\theta} > \frac{C}{\gamma^{\theta} P}$. In addition, $p_m^{\theta} + p_f^{\theta} = p_m^{\theta^{\mathrm{sym}}} + p_f^{\theta^{\mathrm{sym}}}$. Thus, in the notation of Proposition 7, for $t < \min(t^{\theta}, t^{\theta^{\mathrm{sym}}})$,

both θ and θ^{sym} apply, and applying part 3(a) of Theorem A.1 to the relevant subsequence of $(\lambda_t)_{t\geq 0}$ as in the proof of Proposition 7, $\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} = \frac{\lambda_t^{\theta^{\text{sym}}}}{\lambda_{t-1}^{\theta^{\text{sym}}}}$, and hence $\frac{\lambda_t^\theta}{\lambda_t^{\theta^{\text{sym}}}} = \frac{\lambda_{t-1}^\theta}{\lambda_{t-1}^{\theta^{\text{sym}}}} = \frac{\lambda_0^\theta}{\lambda_0^{\theta^{\text{sym}}}} > 1$. Thus, $\lambda_t^\theta > \lambda_t^{\theta^{\text{sym}}}$, so again, if $\lambda_t^{\theta^{\text{sym}}} > \frac{C}{\gamma^{\theta^{\text{sym}}}P}$, then also $\lambda_t^\theta > \frac{C}{\gamma^\theta P}$, i.e., $t^\theta \geq t^{\theta^{\text{sym}}}$. In particular, if the inequality is strict and $t^{\theta^{\text{sym}}} < t < t^\theta$, then researchers of type θ will apply at time t, but those of type θ^{sym} will not.

For part 2, We have

$$
\begin{aligned}
A_t^m - A_t^f &= \sum_{\theta:\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} p^{\theta,m} - \sum_{\theta:\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} p^{\theta,f} = \\
&= \sum_\theta p^{\theta,m} 1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - \sum_\theta p^{\theta,f} 1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} = \\
&= \sum_\theta p^{\theta,m} 1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - \sum_\theta p^{\theta^{\text{sym}},f} 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} = \\
&= \sum_\theta p^{\theta,m} \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) = \\
&= \sum_{\theta:\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n} p^{\theta,m} \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) + \\
&+ \sum_{\theta:\sum_{n=1}^{N/2}\theta_n = \sum_{n=N/2+1}^{N}\theta_n} p^{\theta,m} \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) + \\
&+ \sum_{\theta:\sum_{n=1}^{N/2}\theta_n < \sum_{n=N/2+1}^{N}\theta_n} p^{\theta,m} \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) = \\
&= \sum_{\theta:\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n} p^{\theta,m} \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) + \\
&+ \sum_{\theta:\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n} p^{\theta^{\text{sym}},m} \left(1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} - 1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} \right) = \\
&= \sum_{\theta:\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n} (p^{\theta - p_m^{\theta^{\text{sym}}},m}) \left(1_{\lambda_t^\theta \geq \frac{C}{\gamma^\theta}P} - 1_{\lambda_t^{\theta^{\text{sym}}} \geq \frac{C}{\gamma^{\theta^{\text{sym}}}}P} \right) \geq 0.
\end{aligned}
$$

The third equality follows from the fact that $\theta \mapsto (1-\theta_n)_{n=1}^N$ is a bijection. The fourth follows from the fact that $p^{\theta^{\text{sym}},f} = p^{\theta,f}$. To obtain the fifth, we break up the sum into types θ with more (resp. as many, resp. fewer) characteristics between 1 and $N/2$ than between $N/2+1$ and N. For the sixth, observe that if a type θ has the same number of features between 1 and $N/2$ and between $N/2+1$ and N, then $p^{\theta,m} = p^{\theta^{\text{sym}},m}$ and so $\lambda_0^\theta = \lambda_0^{\theta^{\text{sym}}}$; arguing as in Proposition 8, $\lambda_t^\theta = \lambda_t^{\theta^{\text{sym}}}$ for all $t \geq 0$ (note that as soon as one type stops applying, so does the other); but then, since also $\gamma^\theta = \gamma^{\theta^{\text{sym}}}$, the term in parentheses for such types is

identially zero. In addition, we express the sum over θ's for which $\sum_{n=1}^{N/2}\theta_n < \sum_{n=N/2+1}^{N}\theta_n$ iterating over types θ for which $\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n$, but adding up terms corresponding to the associated symmetric types θ^{sym}. The seventh equality is immediate. Finally, the inequality follows because, for θ such that $\sum_{n=1}^{N/2}\theta_n > \sum_{n=N/2+1}^{N}\theta_n$, the term in parentheses is non-negative by Proposition 8, and in addition $p^{\theta > p_m^{\theta^{\text{sym}}},m}$. *Q.E.D.*