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

Veronesi acknowledges financial support from the Fama-Miller Center for Research in Finance and 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, Hui Chen, Alexia Delfino, Nicola Gennaioli, Lars P. Hansen, Emir Kamenica, Elisabeth Kempf, Jessica Jaffers, Seema Jayachandran, Lubos Pastor, Jane Risen, Antoinette Schoar, Adrien Verdhelan, and seminar participants at the University of Chicago, the Chicago Fed, Bocconi University, and MIT. Errors are our own. 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
JEL No. A11,J16,J7

ABSTRACT

We propose an overlapping-generations model in which established researchers evaluate the research of new researchers. All researchers are differentially endowed with equally desirable research characteristics and belong to two groups, M or F, which have identical ex-ante productivity distributions. Evaluations are group-blind. Yet, when research is evaluated on many characteristics, evaluators' self-image bias and mild between-group heterogeneity lead the initially larger group, say M, to dominate indefinitely. M-researchers only accept the research of young scholars with characteristics close to theirs. Promoted F-researchers are thus few and "similar" to M-researchers, perpetuating the asymmetry. This talent loss is exacerbated by candidates' career concerns and institutions' focus on hiring faculty whose research will be approved by established researchers. Mentorship reduces group imbalance, but it increases the F-group talent loss. Affirmative action reduces both. Our model's predictions are consistent with existing empirical evidence on female participation in academic economics.

Marciano Siniscalchi
Department of Economics
Northwestern University
2211 Campus Drive, 3rd Floor
Evanston, IL 60208
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, though differences across fields do exist. 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 (2019). However, this phenomenon has recently received renewed attention, possibly due to the very slow progress attained in the last 15 years. Indeed, the top panel of Figure 1 shows that in this time span, while the fraction of women in undergraduate economics majors increased to almost 40%, the fraction of women PhD students and assistant professors was flat at roughly 30%. The bottom panel shows that in the "top–10" schools, the fraction of women assistant professors even declined to 19.8% by 2019. Gender imbalance is strong at every stage, from the applicant pool to PhD programs, to their graduation rates, to differential promotion rates through the ranks. Perhaps the most striking fact is that "[w]omen have been less likely to transition to tenured associate or full professors, creating a leaky pipeline" (Chevalier, 2019, p. 14).[2]

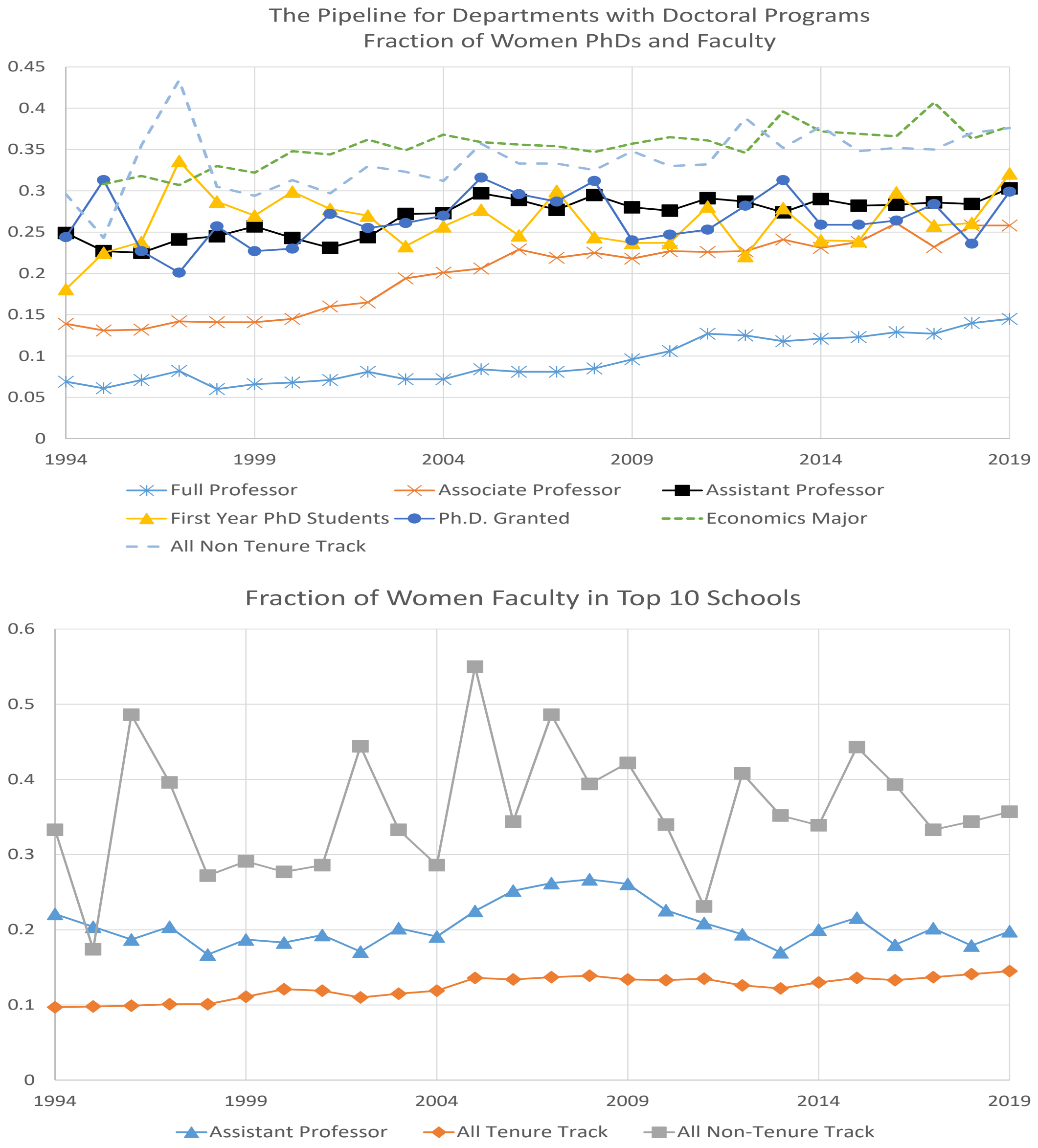
There is a substantial literature that explores different factors that contribute to the lack of female representation in economics; we review this literature in Section 8. In this paper we provide a new model to highlight an additional and more subtle, but still powerful, source of implicit bias that does not depend on stereotypes or discrimination, whether taste-based or statistical. This bias is due to the combination of mild population heterogeneity in research characteristics, and the tendency of scholars to use their personal research style to evaluate others' research output. Both assumptions find strong support in the data, as we discuss below. Our model thus explains why female under-representation may persist in the long run, even when reviewers apply gender-neutral criteria to evaluate others' work.

More specifically, we consider an overlapping-generations model in which agents belong to one of two groups: the M-group or the F-group. A new cohort of young M- and F-researchers appear in every period, in equal proportions. Each researcher is characterized by a set of characteristics. These 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. All researchers, old and young, are endowed with a subset of such desirable research characteristics. Such research characteristics are randomly and symmetrically distributed in the population of young researchers, with some of them slightly more common in the F-group and some others

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

[2] This mirrors the well-documented fact that the gender wage gap is higher at the high end of the distribution: see Blau and Kahn (2017), §2.2.

Figure 1: Percentage of Women in Academia



Source: CSWEP Report, 2019

slightly more common in the M-group. These slight differences are also symmetric: that is, for every characteristic that is slightly more common in the M-group, there is another that is slightly more common, by the same measure, in the F-group, and conversely. 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 accomplishes its objectives). This implies that 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) determines whether the young researcher can be made part of the established population and thus become him- or her-self a referee. Each referee's perceptions of young researchers' research reflect a self-image bias (Lewicki, 1983): that is, they tend to use their own research characteristics as yardstick to evaluate others' research. At the same time, the referees' evaluation is group-neutral: each given referee uses the same set of research characteristics for 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, even mild between-group heterogeneity and self-image bias generate a persistent bias that favors the research of 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, thereby perpetuating the bias forward.

Intuitively, because the M-group is larger initially and referees use their own research style to judge others, the M-group effectively "decides" on behalf of the whole society which research characteristics are important and worthy of reward, and which are not. This is despite the fact that, in our model, all research characteristics are equally conducive to quality research, and therefore both groups are ex-ante symmetric in terms of the likelihood to advance knowledge. 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 a sense, the "bar" for F-researchers is higher. These results are consistent with the evidence in Card, DellaVigna, Funk, and Iriberri (2020) who show that, on average, there is no apparent gender difference in referees' valuation of men- and women-authored papers, although conditional on quality (proxied by citations post-publication) women-authored papers tend to be accepted less frequently than men's.

Gender imbalance and loss of talent is exacerbated by candidates' career concerns, as well as institutions' focus on hiring faculty based on the likelihood their research will meet with the approval of the profession. To demonstrate the first point, we allow young agents from both groups F and M to choose whether to pay a cost to become researchers, or enjoy an outside option. We show that this endogenous choice tends to skew the distribution further towards the M-group. Intuitively, anticipating a bias against their research characteristics, the mass of F-agents who decide to pay the cost of entry 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: Chevalier (2019) reports that the female share of the entering cohort of PhD students in 2018 was 33.2%, much higher than the 7.6% share in 1971, but actually slightly lower than the share in 1994.

The second extension of our model assumes 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 will skew the distribution of their hires towards characteristics more prevalent in the M-group. The steady state in this extension of our model is the same as in the case of career concerns. In other words, endogenous selection affects both the supply and the demand for talent. In both cases, it skews the steady state towards the M group, and exacerbates the loss of talent. This result may explain why "the share female falls as the research intensity of the department increases (e.g. from top 20 to top 10)" (Chevalier, 2019, p. 14): compare the top and bottom panels in Figure 1. Consistently with this interpretation, the two panels show essentially no difference in the female share of teaching faculty.

In a further extension, we allow for different levels of seniority for established researchers. We assume that 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, 2019): 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. While other sources of gender bias may exist in practice (but see Card et al., 2020), the fact that self-image bias can also lead to gender bias and especially talent loss may warrant different interventions. We first investigate the impact of mentorship. We assume that each young researcher is matched with a random advisor from the set of established researchers. Given self-image bias, such advisor will advise the young researcher to become like him or her; the young researcher can do so by paying a cost. We show that, while mentorship may achieve gender balance, it also accelerates the convergence to a steady state with loss of $F-$group characteristics. Intuitively, mentors are drawn from the dominant population, which—by our results—over-represents $M-$group characteristics. Since referees are drawn from the same dominant population, it can be profitable for young researchers to adopt their mentors' characteristics. On one hand, this makes it easier for young F–group researchers to achieve success. On the other hand, such young researchers give up their own characteristics to acquire those more common in the $M-$group. This exacerbates the loss of talent.

We then consider the impact of "affirmative action" policies. In particular, we study a mandated requirement to accept the same number of F researchers as M researchers. Clearly, such policy action mechanically brings about gender balance in the long run. However, we also find that gender balance is reached while having all characteristics in the population represented in the limit. Intuitively, increasing the $F-$group representation in hiring by mandate also increases heterogeneity in the future pool of referees. This makes it more likely that research characteristics (mildly) more prevalent across F researchers will be accepted. The resulting steady state thus represents a wider array of research styles than the one obtained from e.g. mentorship, and is thus beneficial to society.

Our results depend on two main assumptions: mild heterogeneity in research characteristics between M researchers and F researcher, and 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 literature.

First, there is a considerable body of research studying gender differences in cognitive traits, preferences, and attitudes. Regarding cognitive traits, in general these differences are small, in the sense that within-group differences are far larger than the between-group differences. Hyde and Linn (2006) review the evidence on gender difference in mathematics and science, and conclude that, even when statistically significant, such differences as measured by Cohen's d (e.g. Cohen, 2013, §2.2), are small (e.g. $d = 0.11$ for high-schoolers).[3]

[3] For any attribute of interest, the quantity d equals the difference in means between the two samples or populations (here, males and females), divided by the pooled standard deviation. In the cited reference, Cohen suggests that values of d around 0.2 should be considered "small," values around 0.5 "medium," and values around or above 0.8 "large."

Similarly, based on a meta-analysis of research on mathematics performance, males outperform females in "complex problem solving;" while more pronounced ($d = 0.29$) the effect size is still relatively small. Larger differences exist in personality traits and preferences. For instance, medium-sized effects are found for aggression (d between 0.40 and 0.60) and activity level in the classroom ($d = 0.49$). 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). Niederle and Vesterlund (2010) review evidence showing that, relative to the general population, the mathematics gender gap widens considerably when restricting attention to students at the highest levels of math performance.[4] They relate this to differential attitudes toward competitiveness. Goldin (2014) reviews studies finding that the gender pay gap is largest in professions where "working long hours" is especially rewarded; this 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 differences in math scores across countries, as measured by the PISA assessment, are largely explained by broad measures of gender equality in those countries. Falk, Becker, Dohmen, Enke, Huffman, and Sunde (2018) document variation in preference traits across 76 countries and find that women are more risk-averse than men in most countries; however, for trust and patience, the correlation with gender is only significant for a subset of countries. This suggests that cultural factors may partly account for gender differences in preference traits. Andersen, Ertac, Gneezy, List, and Maximiano (2013) provide experimental evidence indicating that the gender gap in competitiveness does not arise in a matriarchal society.

[4] Hyde (2014, p. 391) also reviews work documenting that, for traits such as mathematical ability, spatial reasoning, and verbal ability, there is slightly greater variance among males than females; this might help explain the finding that there are more males at the very top of the ability distribution—though, conceivably, also at the very bottom.

The second important assumption of our model is that each referee uses his/her own research style characteristics to form his/her opinions of what constitutes important or worthy research. The psychological literature on the "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). While this is often just a useful heuristic, it can also reflect a "hidden agenda of self-affirmation" (Dunning and Beauregard, 2000). On the other hand, 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.

A second possible interpretation of our assumption is that referees have preferences over characteristics; in particular, they prefer candidates who share their own characteristics (e.g. theorists like theorists, and empiricists like empiricists). Importantly, this preference does *not* take group membership into account at all. Preferences are an intrinsic trait of a referee: they do not arise out of the belief that one's own characteristics are the ones that make a good economist. This interpretation is consistent with our analysis. However, it implies that referees do not value heterogeneity (e.g., theorists derive no benefit from interacting with empirical researchers, and conversely). It also implies that referees do not take the candidate's objective productivity into account, and hence disregard the benefits that would accrue to a department—or, in fact, from the profession as a whole—from hiring and advancing a productive young researcher who however does not share their own characteristics.

Organization. Section 2. introduces our basic model. Section 3. provides a simple numerical illustration, and Section 4. a more elaborate one that is closer to the data. Section 5. endogenizes entry (§5.1.) and hiring (§5.2.) decisions. Section 6. studies the case of junior and senior researchers. We then turn to evaluate policy actions: Section 7.1. studies the impact of mentorship and Section 7.2. discusses the impact of affirmative action policies. Section 8. reviews the literature and Section 9. concludes.

2. The Basic Model

We consider an overlapping-generations model in which unit masses of two groups of young researchers, 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 particularly stark, symmetric environment in which each type corresponds to a vector of N characteristics which can only take two values, 0 and 1: that is, $\Theta \equiv \{0, 1\}^N$. (We consider a more general specification in the Appendix.) For $\theta \in \Theta$, $1 \leq n \leq N$, and $i \in M \cup F$, we denote by θ_n^i the value of the n-th characteristic for agent i. We assume that the number N of characteristics is even, and that their distribution among M and F researchers is determined by a single parameter $\phi \in (\frac{1}{2}, 1)$. Specifically:

- characteristics are mutually independent;
- for $n = 1, \ldots, \frac{N}{2}$, the probability that $\theta_n^i = 1$ is ϕ for M-researchers and $1 - \phi$ for F-researchers; and
- for $n = \frac{N}{2} + 1, \ldots, N$, the probability that $\theta_n^i = 1$ is $1 - \phi$ for M-researchers and ϕ for F-researchers.

For every $\theta \in \Theta$, let $p^{\theta,f}$ (resp. $p^{\theta,f}$) denote the fraction of types in the F (resp. M) population of young researchers. Let $p^g = (p^{\theta,g})_{\theta \in \Theta} \in \Delta(\Theta)$ for $g = f, m$. Thus,

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

The parameter ϕ can also be related to Cohen's d statistic for an individual characteristic: for $n = 1, \ldots, \frac{N}{2}$,

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

For $n = \frac{N}{2} + 1, \ldots, N$, the d statistic is the negative of the above expression. We will focus our numerical exercises to ϕ close to 0.5, which implies that between-group heterogeneity is small compared to within-group heterogeneity.

We model each characteristic as a desirable research attribute, which makes it more likely for the researcher to produce a quality paper. A "quality" paper 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 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

$$\gamma^\theta \equiv \gamma_0\, \rho^{\frac{1}{N}\sum_n \theta_n}. \tag{3}$$

Thus, $\gamma^{(0,\ldots,0)} = \gamma_0$, and the probability of writing a quality paper depends solely on the number of 1's in $\sum_n \theta_n$, with the maximum attained for $\gamma^{(1,\ldots,1)} = \gamma_0\, \rho \in [\gamma_0, 1]$. The parameter ρ reflects the relative abilities of researchers with different characteristics to write a "quality" paper. If $\rho = 1$, for instance, then all types write a quality paper with probability γ_0, which entails that other factors will need to be used by referees in promoting researchers. If $\rho = 4$, instead, it means that the best researcher $\gamma^{(1,\ldots,1)}$ is four times more likely to produce quality research than the worse researcher, with $\gamma^{(0,\ldots,0)}$.

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 depends on whether the paper is of sufficient quality or not, as described in previous section. Since each young scholar with type θ produces quality research with probability γ^θ, given in (3), this is also the probability with which the research is "accepted" by referees. This probability increases with the number of desirable characteristics $\sum_n \theta_n$. This assumption captures the fact that a young scholar with many desirable characteristics is more likely to produce quality research than another scholar with fewer desirable characteristics. Still, even a scholar θ' with $\sum_n \theta'_n = 0$ has probability $\gamma_0 > 0$ to produce quality research, perhaps by sheer luck. To sum up, in this setting, the referee is only certifying that the research is of sufficient quality, that is, it reaches its goals.

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:

$$a_t^{\theta,g} = \gamma^\theta \cdot p^{\theta,g}, \quad g \in \{f, m\}. \tag{4}$$

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

$$\lambda_t^{\theta,g} = (1 - a_t)\lambda_{t-1}^{\theta,g} + a_t^{\theta,g}, \quad g \in \{f, m\}. \tag{5}$$

The limiting behavior of this system is readily characterized. First, *initial conditions have no long-run effect.* Eq. (4) 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 $g \in \{f, m\}$,

$$\lambda_t^{\theta,g} = (1 - a)\lambda_{t-1}^{\theta,g} + a^{\theta,g} = (1 - a)^t\lambda_0^{\theta,g} + a^{\theta,g}\frac{1 - (1 - a)^t}{a} \to \frac{a^{\theta,g}}{a} \tag{6}$$

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 $\theta = (\theta_1, \dots, \theta_N)$, there is a corresponding type $\bar\theta = (\theta_{N/2+1}, \dots, \theta_N, \theta_1, \dots, \theta_{N/2})$ 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.

[5] The fact that the total mass of established scholars (a stock) equals the mass of young M and F researchers (flows) is of course not realistic, but immaterial for our analysis. Normalizing the stock of established researchers to any positive number L yields the same predictions. Furthermore, L could be calibrated, for instance, by matching the fraction of young researchers who are hired to data on the academic job market (e.g., see Conley and Önder, 2014)

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

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 (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 and the adoption of self-serving prototypes, 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. 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

$$a_t^{\theta,g} = \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g}. \tag{7}$$

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 (5), the total mass of established researchers of type θ and group g is given by

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}\,(1-a_t) + a_t^{\theta,g} \tag{8}$$

where as above $a_t = \sum_\theta \sum_g a_t^{\theta,g}$. Equations (7) and (8) 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 establishes the types that can potentially survive (i.e. have positive mass) in the limit. All other types vanish over time.

Proposition 2 *For every even $N > 0$, $\phi \in (\frac{1}{2}, 1)$, $\gamma_0 \in (0,1)$, and $\rho \in (1, \frac{1}{\gamma_0})$, the sequences $(\lambda_t)_{t\geq 0}$, $(\lambda_t^m)_{t\geq 0}$, and $(\lambda_t^f)_{t\geq 0}$, admit limits. Furthermore, only three types can potentially survive in the limit: either*

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

$$\theta^m = (1,\dots,1,0,\dots,0) \quad \textit{and} \quad \theta^f = (0,\dots,0,1,\dots,1); \;\textit{or} \tag{9}$$

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

$$\theta^* = (1,\dots,1). \tag{10}$$

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.

The fact that these three types are the only ones that can potentially survive reflects the observation that both the distribution of entrants and the relative chances of quality research determine the evolution of the system. Furthermore, not all three types can survive. 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 eventually prevails.

In the next proposition, recall that the parameter ρ measures agents' heterogeneity in producing quality research (see equation (3)).

Proposition 3 *Under the assumptions of Proposition 2, let $\bar{\lambda} = \lim_{t\to\infty} \lambda_t$ and*

$$\bar{\rho}(\phi, N) = \frac{1}{4}\left(\left(\frac{1-\phi}{\phi}\right)^{N/2} + \left(\frac{\phi}{1-\phi}\right)^{N/2}\right)^2. \tag{11}$$

(a) If $\rho < \bar{\rho}(\phi, N)$, then only types θ^m and θ^f survive in the limit. In particular, if at time 0, all referees are in the M-group, so that $\lambda_0 = p^m$, then[7]

$$\bar{\lambda}^{\theta^m} = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1}{2}; \quad \bar{\lambda}^{\theta^f} = 1 - \bar{\lambda}^{\theta^m}. \tag{12}$$

(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 $N/2$. The interpretation is simple once we remember 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 consider his or her traits as the important ones, and discounts the other ones.

By way of contrast, part (b) characterizes a more "meritocratic" scenario in which research characteristics significantly improve the chances of producing quality research. In this case, regardless of the initial conditions, the system converges to an efficient steady state in

[7] What matters for the result to hold is that $\lambda_0 = p^m$; in principle, this may hold even if not all referees are initially from the M group, but in practice, this is the case of interest here.

which all researchers possess every research characteristics—regardless of their group. The self-image bias is still at work in this scenario: referees still evaluate research according to their own characteristics. However, in this scenario 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 self-image bias, as well as meritocratic and unbiased 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, a main take-away message of our model is that, if the number of research characteristics is large, if 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. The F-researchers have no chance to grow to equality, even without any explicit bias against them.

In fact, 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. 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 all even N, $\phi \in (\frac{1}{2}, 1)$, $\gamma_0 \in (0, 1)$ and $\rho \in (1, \frac{1}{\gamma_0})$, and for every $L \in \{0, \ldots, N\}$,*

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

and the inequality is strict if there is $\theta \in \Theta$ with $\sum_n \theta_n$ 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, namely, the number of desirable characteristics $\sum_n \theta_n = L$; this is in line with the cited evidence in Card et al. (2020).

2.4. M- and F-Researcher Types in the Limit

Proposition 3 mostly concerns the distribution of researcher types irrespective of their group. We now analyze how the mass of each type θ evolves among M- and F-researchers separately, and also characterize group (im)balance in the limit.

To do so, it is useful to rewrite Equation (8) for each group $g = m, f$ as follows:

$$\lambda_t^{\theta,m} - \lambda_{t-1}^{\theta,m} = -\lambda_{t-1}^{\theta,m}\, a_t \; + \lambda_{t-1}^{\theta,m}\gamma^\theta p^{\theta,m} + \lambda_{t-1}^{\theta,f}\gamma^\theta p^{\theta,m} \tag{13}$$

$$\lambda_t^{\theta,f} - \lambda_{t-1}^{\theta,f} = -\lambda_{t-1}^{\theta,f}\, a_t \; + \lambda_{t-1}^{\theta,f}\gamma^\theta p^{\theta,f} + \lambda_{t-1}^{\theta,m}\gamma^\theta p^{\theta,f} \tag{14}$$

Consider the dynamics of F-researchers in (14), 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 (13) and (14) 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 5 *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

$$\text{(M-researchers of type θ^m): } \bar{\lambda}^{\theta^m,m} = \frac{(\phi^N)^2}{(\phi^N + (1-\phi)^N)^2}; \tag{15}$$

$$\text{(F-researchers of type θ^m): } \bar{\lambda}^{\theta^m,f} = \frac{\phi^N\,(1-\phi)^N}{(\phi^N + (1-\phi)^N)^2}; \tag{16}$$

$$\text{(M-researchers of type θ^f): } \bar{\lambda}^{\theta^f,m} = \frac{((1-\phi)^N)^2}{(\phi^N + (1-\phi)^N)^2}; \tag{17}$$

$$\text{(F-researchers of type θ^f): } \bar{\lambda}^{\theta^f,f} = \frac{(1-\phi)^N\,\phi^N}{(\phi^N + (1-\phi)^N)^2}; \tag{18}$$

with

$$\bar{\lambda}^{\theta^m,m} > \bar{\lambda}^{\theta^m,f} = \bar{\lambda}^{\theta^f,f} > \bar{\lambda}^{\theta^f,m} \tag{19}$$

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

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} > 0.5. \tag{20}$$

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

The next proposition 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 5 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.

2.5. Research Characteristics of Established F-Researchers

One further implication of Proposition 3 (see Eqs. 15 and 18) is that, in an environment in which self-image bias prevails, the *same* limiting fraction of established F-researcher exhibit types θ^m and θ^f. In other words, the research characteristics of *established* F-researchers are "biased," i.e., they are slanted towards the characteristics prevalent among M-researchers.

Corollary 3 *In part (a) of Proposition 5,*

$$0.5 = \frac{\bar{\lambda}^{\theta^f, f}}{\bar{\lambda}^{\theta^m, f} + \bar{\lambda}^{\theta^f, f}} = \frac{\bar{\lambda}^{\theta^m, f}}{\bar{\lambda}^{\theta^m, f} + \bar{\lambda}^{\theta^f, f}} \tag{21}$$

This result is in stark contrast with the assumption that θ^f is the prevalent type in each cohort of young F-researchers. In other words, 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 had $(1 - \phi)^N$ frequency in each population of young F researchers.

The intuition is as follows. The fraction of M agents of type θ^f is small, and in the limit most established agents are from the M group. This means that, although types θ^f are the most prevalent in the F group, there are few type-θ^f referees that will accept their research. On the other hand, while the fraction of types θ^m in the F group is small, there are many potentially positive type-θ^m referees. The symmetry of our model implies that these effects exactly balance out in the limit.

Table 1: Type Frequencies in a Simple Example

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

3. A Simple Numerical Example

To illustrate the results in the previous subsections, 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 $\theta^* = (1,1)$. We will consider different choices for the parameters γ_0 and ρ, but since γ^θ only depends upon $\sum_n \theta_n$, θ^m and θ^f have the same probability of producing quality research, and θ^* is the most likely type to do so.

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 between-group heterogeneity in this example is large and not realistic, and of course restricting attention to only two research characteristics is just for simplicity. 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 much smaller between-group heterogeneity and a larger number of research characteristics. M- and F-research output is characterized by the frequencies in Table 1.

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. (20) in Proposition 5, 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, the left panel of Figure 2 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.

Despite the fact that both characteristics are important to produce quality research, this dynamics tends to weed out those researchers who indeed possess both such characteristics (their mass vanishes). Moreover, because the M-population dominates, and in such population the second characteristic is under-represented, we end up with a self-perpetuating state in which the dominant M-characteristic 1 is over-sampled at the expense of the dominant F-characteristic 2.

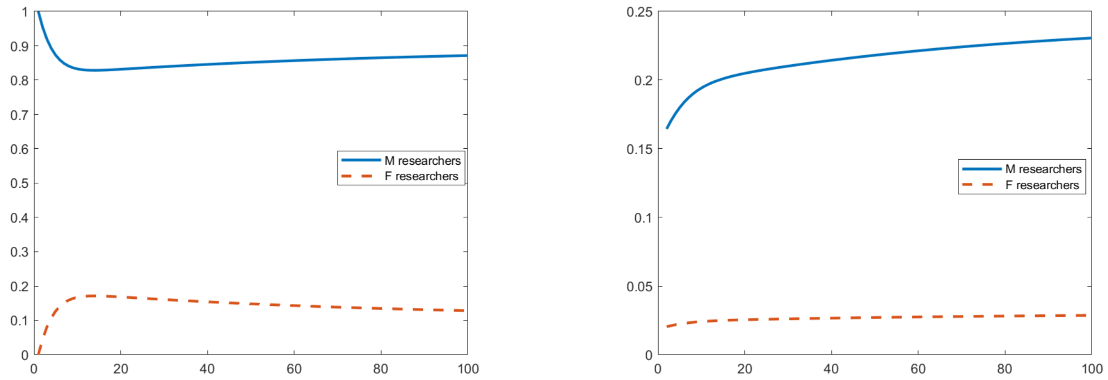
Panel (b) of Figure 2 shows the total acceptance rates of symmetric types θ^m and θ^f for M and F researchers: $a_t^{\theta^m,m} + a_t^{\theta^f,m}$ and $a_t^{\theta^m,f} + a_t^{\theta^f,f}$ (see Proposition 4). Despite the identical quality of these researchers, the acceptance rate of M-researchers is far higher than the acceptance rate of F-researchers.

Of interest is that established F-researchers are overly frequently of the type $\theta^m = (1,0)$ that is most prevalent among the M-group, consistently with Corollary 3. And, because established F-researchers carry the characteristics of M-researchers, they will judge other F-researchers just like M-referees do: they will exhibit an implicit bias against types prevalent among F-researchers, downplaying characteristic 2 and instead putting excessive weight on characteristic 1.

The left panel of Figure 3 shows the percentage of established F-researchers over time. Recall that, by assumption, there are no F-researchers at time 0. Initially, intrinsic research

Figure 2: Fraction of M and F Researchers and Acceptance Rates

(a) M and F researchers (b) Acceptance Rates



Fraction of M and F researchers (Panel a) and sum of acceptance rates of symmetric types θ^m and θ^f for M and F researchers, i.e., $a_t^{\theta^m,m} + a_t^{\theta^f,m}$ and $a_t^{\theta^m,f} + a_t^{\theta^f,f}$ (Panel b). Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

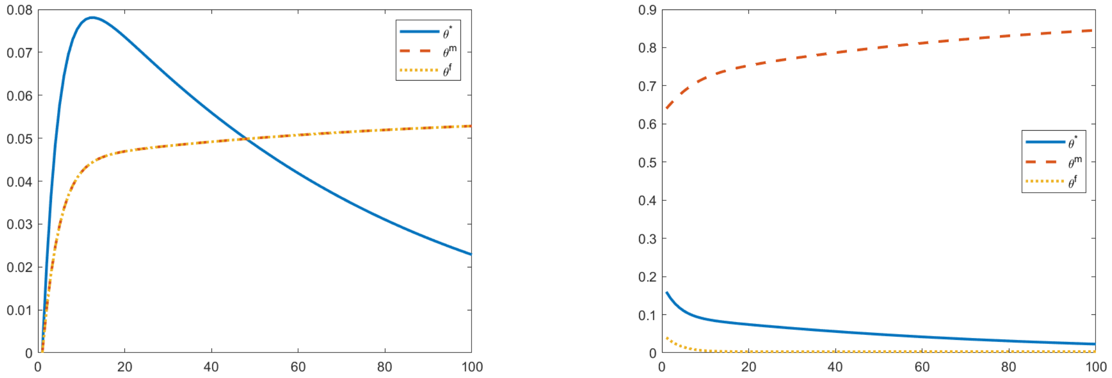
ability prevails, and type θ^* is most prevalent among F-researchers that become established. However, over time, types θ^m and θ^f dominate. In particular, even though $\frac{\phi^2}{(1-\phi)^2} = \frac{0.64}{0.04} = 16$ times as many θ^f types as θ^m types appear among F-researchers in *every* period, this is compensated 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, and the fraction of established θ^m and θ^f types is the same among F-researchers at each point in time. This is in stark contrast with the asymmetry between θ^m and θ^f types in each young F-cohort. On the other hand, the right panel of Figure 3 shows that the percentage of M-researchers of type θ^m increases to one. Thus, the system "weeds" out the least productive types $\theta = (0, 0)$, but it also weeds out the efficient type θ^*.

3.1. Convergence to Efficiency

Can the dynamics lead to convergence to equal shares of established M- and F-researchers, even if one starts from an unbalanced initial population? Proposition 3 shows that this is the case if agents' differences in probability to produce quality research, ρ, is sufficiently high.

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 $\rho = 9$. Thus, now 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" than in our

Figure 3: 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 $\theta^* = (1,1,....,1)$, $\theta^m = (1,...,1,0,...,0)$, and $\theta^f = (0,...,0,1,...,1)$. Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

previous numerical examples. Now

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

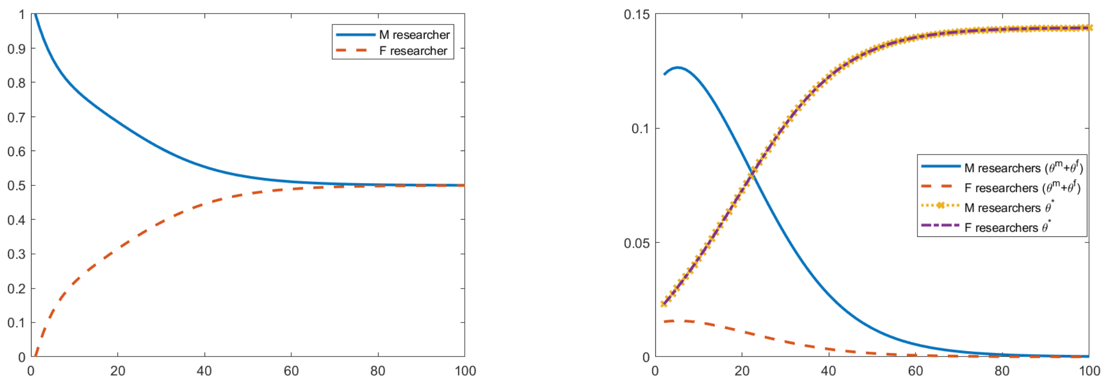
so Proposition 3 part (b) implies that type θ^* will dominate in the limit. Figures 4 and 5 illustrate the dynamics. In this case, the percentage of F-researchers indeed converges to the 50-50 symmetric configuration, and eventually, the same fraction of M- and F-researchers is accepted. Moreover, the system is able to weed out those researchers that do not possess both characteristics—i.e., the system "works." Yet, Panel (b) in Figure 4 shows that, in the short run, a greater mass of M researchers whose type is either θ^m or θ^f is accepted relative to F researchers of the same types.

4. Many Characteristics

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. This implies that Cohen's d statistic for each characteristic equals

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

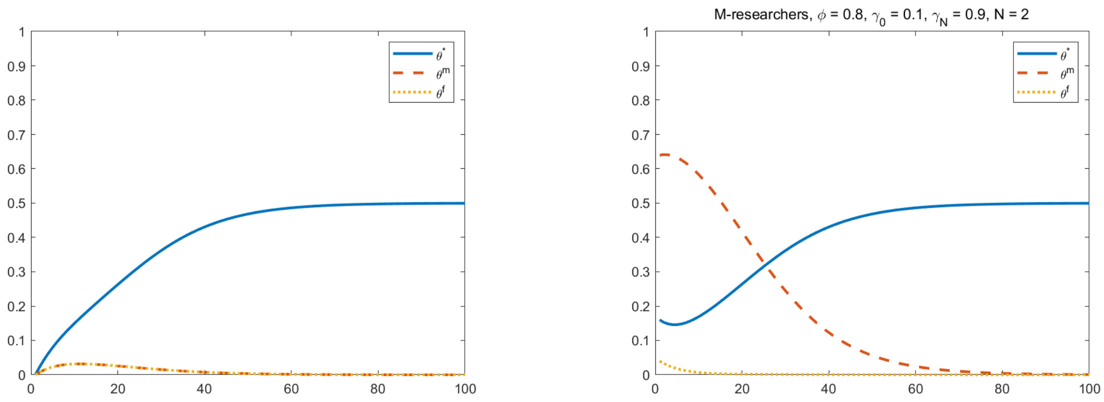
Figure 4: Fraction of M and F Researchers and Acceptance Rates with More Meritocracy


(a) M and F researchers

(b) Acceptance Rates

Fraction of M and F researchers (panel a) and sum of acceptance rates of symmetric types, θ^m and θ^f, and of type θ^* for M and F researchers, i.e., $a_t^{\theta^m,m} + a_t^{\theta^f,m}$, $a_t^{\theta^m,f} + a_t^{\theta^f,f}$, $a_t^{\theta^*,m}$ and $a_t^{\theta^*,f}$ (panel b). Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.1$, $\rho = 9$, $N = 2$.

Figure 5: 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 $\theta^* = (1, 1, ..., 1)$, $\theta^m = (1, ..., 1, 0, ..., 0)$, and $\theta^f = (0, ..., 0, 1, ..., 1)$. Initially $\lambda_0 = \frac{1}{2}p_m + \frac{1}{2}p_f$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.1$, $\rho = 9$, $N = 2$.

which, as discussed in the Introduction, 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.

The relevant question is then how many research characteristics 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.

Second, here we only consider $\{0, 1\}$–valued characteristics for simplicity: either a researcher possesses a trait, or he/she does not. In most cases, each characteristics has different degrees; this provide further scope for self-image bias, and hence amplify its impact. For instance, a referee might like a style of research that combines theory and empirical evidence, but not work that is either "hard-core theory," or that, on the contrary, lacks any theoretical underpinning.

The next example displays the dynamics of the fraction of established M- and F-researchers in an environment with $N = 10$ characteristics. We set $\phi = 0.5742$, so the implied Cohen's d is

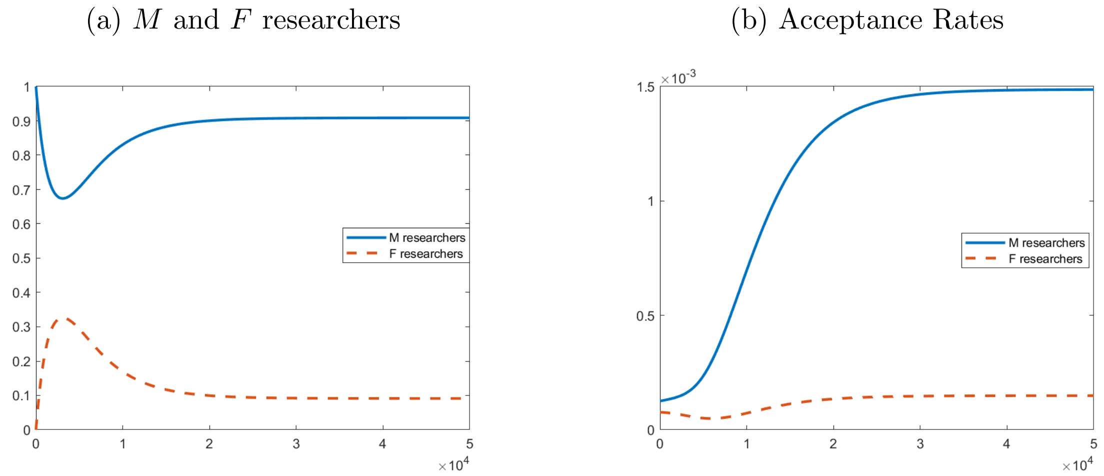
$$d = \frac{2 \times 0.5742 - 1}{\sqrt{0.5742 \times (1 - 0.5742)}} = 0.3,$$

which is considered "small" and in line with the estimated group differences of the various traits discussed in the introduction. As for γ^{θ}, we assume $\gamma_0 = 0.2$ and $\rho = 4$. This implies an ex-ante objective failure rate $\sum_{\theta}(1 - \gamma^{\theta})(p^{\theta,f} + p^{\theta,m})/2 = 59\%$, which seems plausible. In addition, the choice of ρ implies that researchers N is objectively four times as productive as researcher 0, which is roughly in line with the evidence reported in Conley and Önder (2014).[9] The result is in Figure 6. As can be seen in Panel (a), eventually, the system converges again to large disparity between M- and F-researchers. Eventually, the percentage of F-researchers is below 10%, even if the distribution of characteristics is much more similar across M and F types. Panel (b) plots the acceptance rates of all young M-

[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] 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 L to match the failure rate from the data. See footnote 5.

Figure 6: Fraction of M and F Researchers and Acceptance Rates with Ten Characteristics


(a) M and F researchers

(b) Acceptance Rates

Fraction of M and F researchers (Panel a) and sum of acceptance rates for M- and F researchers of all types θ such that $\sum_n \theta_n = N/2$ i.e., $\sum_{\{\theta:\sum_n \theta_n = N/2\}} a_t^{\theta,m}$ and $\sum_{\{\theta:\sum_n \theta_n = N/2\}} a_t^{\theta,f}$ (Panel b). Initially $\lambda_0 = p^m$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

and F-researchers with types θ's such that $\sum_n \theta_n = N/2$, i.e., of the same quality as θ^m and θ^f. As can be seen, controlling for objective quality, F-researchers are accepted far less than M-researchers.

From Corollary 3, the limiting fraction of male researchers is

$$\bar{\Lambda}^m = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} = \frac{1 + \left(\frac{0.5742}{0.4258}\right)^{20}}{1 + \left(\frac{0.5742}{0.4258}\right)^{20} + 2\left(\frac{0.5742}{0.4258}\right)^{10}} \approx 91\%,$$

which is where the system converges in Figure 6.

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.).

5.1. Endogenous Choice of Young Researchers

We now extend the model to consider the optimal choice of a potential researcher who can choose between a career in research or an outside option, which we normalize to zero. We assume that a prospective researcher must pay a utility cost C, which is identical across all agents, in order to undertake a career in academia. Each agent knows his/her type θ and also knows that the screening criteria are the ones illustrated in the previous section. We assume that a young researcher who is hired obtains a payoff of P. If he/she is not hired, the researcher's overall payoff is thus $-C$ as the researcher has to go back to the outside option, but paid the utility cost C. What types of agents decide to pay the cost C and thus take their chance with the academic career?

Given a distribution λ_t^θ of referees across all types $\theta \in \Theta$ at time t, assume that each young researcher decides whether or not to pay the cost C, then (upon deciding to pursue an academic career) produces research, and is then evaluated—all at time t. An agent of type θ then pays the cost to become a researcher if and only if

$$\gamma^\theta \lambda_t^\theta (P - C) + (1 - \gamma^\theta \lambda_t^\theta)(-C) > 0. \tag{22}$$

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

$$a_t^{\theta,g} = \begin{cases} \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g} & \text{if } \gamma^\theta \lambda_{t-1}^\theta \geq \frac{C}{P} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g} (1 - a_t) + a_t^{\theta,g} \tag{24}$$

Expression (23) 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—they will choose not to pay the cost to enter the profession. From Eq. (24), 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 presence of cutoffs makes the analysis slightly different from that of the basic model in Section 2., Proposition 3 still suggests what the dynamics will look like. Set aside the trivial case in which no type wants to apply (so that the limit distribution of types is just λ_0). Suppose first the condition in part (a) of Proposition 3 holds, so the only types that survive in the limit of the basic dynamics are θ^m and θ^f. It turns out that the population shares of these types increase monotonically. Hence, if, say, $\lambda_0^{\theta^m} \geq \frac{C}{\gamma^{\theta^m} P}$ at time $t = 0$, then this will also be the case at all subsequent times. So, this type will continue to apply and its dynamic will be similar to the one in Section 2.. On the other hand, all other types $\theta \notin \{\theta^m, \theta^f\}$ eventually vanish in the basic model. This suggests that, for each such type θ, there is a time $t(\theta)$ such that $\lambda^{\theta}_{t(\theta)} < \frac{C}{\gamma^{\theta} P}$. Thus, young researchers of type θ stop applying at time $t(\theta)$. This implies that the masses of type θ^m and θ^f increase faster relative to the basic model, because fewer "retirements" are needed to make room for those types θ who no longer apply but would have been hired if they had applied.

In this case, if both types θ^m and θ^f apply at date 0 (and thereafter), their limiting masses will be the same as in Section 2.. However, since $\gamma^{\theta^m} = \gamma^{\theta^f} = \gamma_0\sqrt{\rho}$ and we assume that the initial population consists solely of male researchers, $\lambda^{\theta^m}_0 = \phi^N > (1-\phi)^N = \lambda^{\theta^f}_0$, if

$$\phi^N > \frac{C}{\gamma_0\sqrt{\rho}P} > (1-\phi)^N, \tag{25}$$

young researchers of type θ^m apply at date 0 and thereafter, whereas those of type θ^f *never* apply. In the limit, relative to the basic model, this leads to both a more pronounced imbalance between M and F researchers and further talent loss, as the characteristics of type θ^f are not represented at all.

On the other hand, if the condition in part (b) of Proposition 3 holds, one can again obtain a balanced long-run population of established researchers. However, this requires that type θ^* be willing to apply. This is an additional constraint on efficiency that arises under endogenous entry.

The following Proposition formalizes this discussion.

Proposition 6 *For every even $N > 0$, $\phi \in (\frac{1}{2}, 1)$, $\gamma_0 \in (0,1)$, and $\rho \in (1, \frac{1}{\gamma_0})$, the sequences $(\lambda_t)_{t\geq 0}$, $(\lambda^m_t)_{t\geq 0}$, and $(\lambda^f_t)_{t\geq 0}$ in Eqs. (23)–(24), admit limits. Furthermore, assume that at time 0, all referees are from M-group, i.e., $\lambda_0 = p^m$, and let $\bar{\lambda} = \lim_{t\to\infty} \lambda_t$ and $\bar{\Lambda}^m = \lim_{t\to\infty} \sum_{\theta} \lambda^{m,\theta}_t$. Then:*

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

(a.2) If $\rho < \bar{\rho}(\phi, N)$ and $\phi^N \geq \frac{C}{\gamma_0\sqrt{\rho}P} > (1-\phi)^N$, then only one type, θ^m, survives in the limit, i.e. $\bar{\lambda}^{\theta^m} = 1$. In addition, the total mass of M researchers is

$$\bar{\Lambda}^m = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}}. \tag{26}$$

(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.

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 more frequent in the $F-$population, but also common in the M-population, disappear altogether in the limit. 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. (12) 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 6 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 omit the figure for brevity).

Interestingly, endogenous choice may prevent convergence to group balance even when group balance would in fact attain in the basic model. This situation is illustrated in Figure 7, which shows the fraction of M- and F-researchers with endogenous choice. 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

$$\rho = 4 > 3.61 = \bar{\rho}(\phi, N).$$

The solid and dashed lines in Figure 7 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 $C = 3$ (i.e. 0.3% 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:

$$\lambda_0^{\theta^*} = p^{\theta^*,m} = \phi^{N/2}(1-\phi)^{N/2} = 0.3574\% < 0.3750\% = \frac{C}{\gamma^{\theta^*}P}.$$

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

$$\lambda_0^{\theta f} = p^{\theta f,m} = (1-\phi)^N = 0.1081\% < 0.75\% = \frac{C}{\gamma^{\theta f}P}.$$

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

$$\lambda_0^{\theta^m} = p^{\theta^m,m} = \phi^N = 1.18\% > 0.75\% = \frac{C}{\gamma^{\theta^m}P}.$$

Therefore, while other types are also willing to apply, type θ^m will prevail (cf. Lemma 1 in the Appendix), which will lead to a severe imbalance between M and F researchers in the limit, as shown in Figure 7. 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 8 shows that both F and M researchers are of type θ^m in the long run.

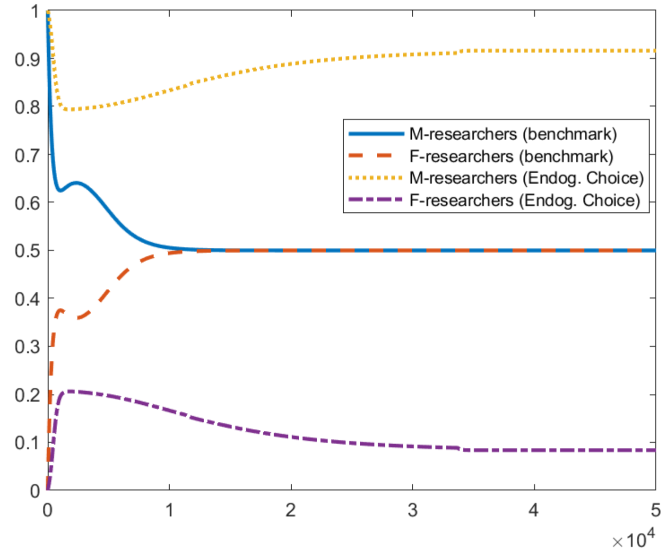
To sum up, even if the basic environment is meritocratic, in the sense that even with self-image bias, the differences in talents γ^θ across types are sufficient to eventually lead to group balance, the introduction of endogenous entry introduces a bias in favor of M-researchers which leads to an imbalance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

5.1.2. Characterization of the Applicant Pool

Because of the variation in the population of research characteristics, Proposition 6 also has implications on the mass of M- and F-applicants—that is, young researchers who decide to apply for a job. These masses are, respectively,

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

Figure 7: Fraction of M and F Researchers with Endogenous Entry

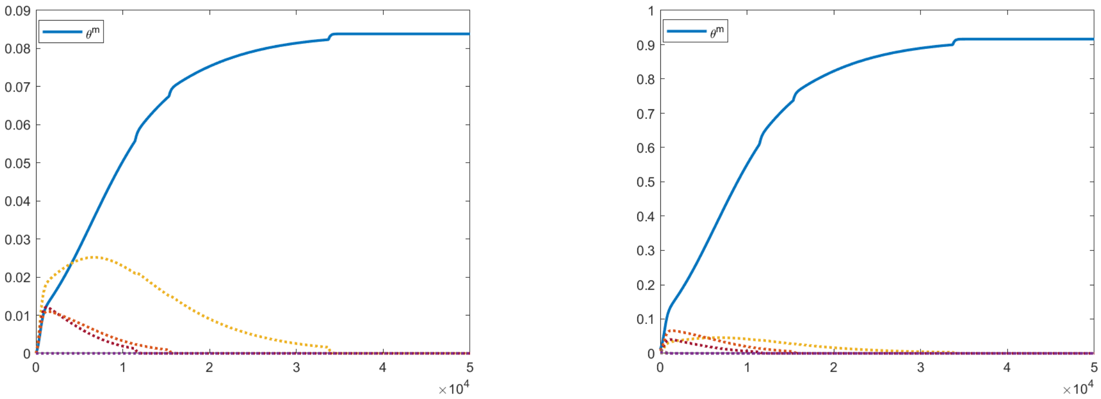


Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, $N = 8$, $P = 1000$, and $C = 3$.

Figure 8: Types of Established F and M Researchers with Endogenous Entry

(a) F researchers

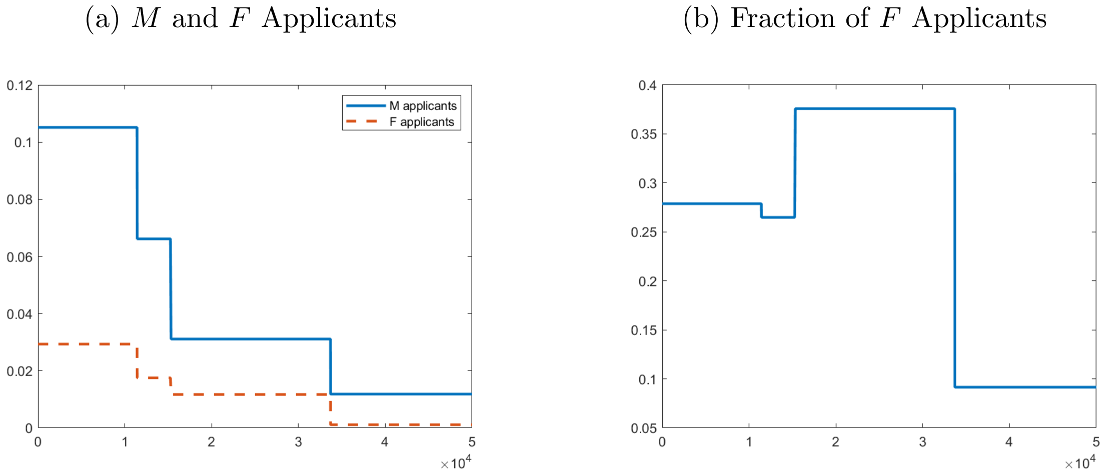
(b) M researchers



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: $\phi = 0.0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, $N = 8$, $P = 1000$, and $C = 3$.

We obtain the following result: consider a type θ who has m_0 characteristics numbered 1 through $N/2$ (i.e., "M-prevalent" features) and m_1 characteristics numbered $N/2+1$ through N ("F-prevalent" features), with $m_0 > m_1$. Associate with it a "symmetric" type θ^{sym} who has m_0 characteristics numbered $n \in \{N/2, \ldots, N\}$ and m_1 characteristics numbered

Figure 9: 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: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, $N = 8$, $P = 1000$, and $C = 3$.

$n \in \{1, \ldots, N/2\}$. For definiteness, define θ^{sym} so that $\theta^{\text{sym}}_n = \theta_{N+1-n}$ for all n. Then, θ applies whenever θ^{sym} does, but θ may apply when θ^{sym} does not. The following proposition formalizes the argument and provides the limiting result:

Proposition 7 1. *For each type $\theta \in \Theta = \{0,1\}^N$, consider the symmetric type θ^{sym} with $\theta^{\text{sym}}_n = \theta_{N+1-n}$. Suppose that*

$$\sum_{n=1}^{N/2} \theta_n > \sum_{n=N/2+1}^{N} \theta_n.$$

Then, at any time $t \geq 0$, if researchers of type θ^{sym} apply, so do researchers of type θ. However, the reverse need not hold.

2. *For every t, $A^m_t \geq A^f_t$. Moreover, if $\lambda_0^{\theta m} > \frac{C}{\gamma_0\sqrt{\rho P}} > \lambda_0^{\theta f}$,then $A^f_t \downarrow 1 - \bar{\Lambda}^m$, where $\bar{\Lambda}^m$ is as in part (a.2) of Proposition 6.*

Proposition 7 shows that M researchers in aggregate are more likely to apply than F researchers. Moreover, self-selection pushes out F-researchers from the research pool and eventually they do not apply.

Figures 9a and 9b show the total masses of M and F applicants and, respectively, the percentage of F applicants over the total application pool. The parameter values are the same as for Figure 7. Consistently with Corollary 7, the mass of M applicants is always greater than that of F applicants; furthermore, the latter declines over time. The decline is not steady, but in steps, as there are sudden moves of types that decide not to apply anymore. Furthermore, in the limit, the fraction of F applicants is of course given by the fraction of F researchers of the only surviving type θ^m over the total:

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

5.2. The Endogenous Selection of 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—that is, from the perspective of hiring institutions. 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. This leads to the same conclusions as in Section 5.1..

Specifically, consider the following alternative interpretation of the basic model in Section 2.2.. 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 to its programs. At the same time, 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. Consistently with Sections 2.1. and 2.2., a candidate of type θ produces quality work with probability γ^θ. The key assumption concerns the probability that the candidate's quality work is accepted by the profession. Here, we reinterpret the key assumption of Section 2.2.: the hiring institution anticipates that referees are subject to self-image bias, so that a type-θ researcher will be accepted 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

$$\gamma^\theta \lambda_t^\theta \, (P - C) + \left(1 - \lambda_t^\theta \gamma^\theta\right)(-C) > 0 \tag{27}$$

This is the same condition as in Equation (22) in the previous section. Thus, the mass of established researchers λ_t^θ follows the system dynamics described by Equations (23) - (24). Proposition 6 then applies and group imbalance and loss of talent obtains.

Moreover, under the conditions of case (a.2) of Proposition 6, the system converges, in finite time, to a steady state in which only type θ^m survives: $\lambda_t^{\theta^m} \to 1$. The implication of this result is that, under these conditions, the hiring practice of institutions to *only* take publication potential into account leads to a steady state in which type θ^f disappears, even when such type would survive without endogenous selection. Again, this implies talent loss: the 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 population of researchers. However, if we account for institutions' desire to hire only young researchers who are sufficiently likely to be accepted by the *current* population of referees, then group imbalance (Figure 7) emerges. Again, in this example type θ^f disappears completely (Figure 8).

This mechanism may explain the dynamics observed in the top and bottom panels of Figure 1. First, from the top panel, the female representation of undergraduate students with economics major has been rising over the past 15 years, reaching almost 40% by the late 2010s. The same trend applies to non-tenure track faculty, for whom research promise is not a primary consideration in the hiring decision. In contrast, not only has the percentage of female faculty at the entry-level (assistant professor) rank been flat at 30% in the last 10 years (top panel), but, in so-called “top-10” schools, it has actually declined to 19.8% (bottom panel).[10] Yet, female teaching faculty in “top-10” schools hover around 40%, as in the aggregate. These patterns are consistent with our model: when a hiring institution has research as the guiding principle in hiring, it may tend to skew towards the characteristics of the existing established profession, i.e. θ^m in our model.

6. 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

[10] We use the “top-10” schools terminology as per Chevalier (2019). The school names are not reported.

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

$$a_{1,t}^{\theta,g} = \gamma^{\theta} \cdot p^{\theta,g} \cdot (\lambda_{1,t-1}^{\theta} + \lambda_{2,t-1}^{\theta}) \tag{28}$$

$$a_{2,t}^{\theta,g} = \gamma^{\theta} \cdot \lambda_{1,t-1}^{\theta,m} \cdot \lambda_{2,t-1}^{\theta}. \tag{29}$$

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

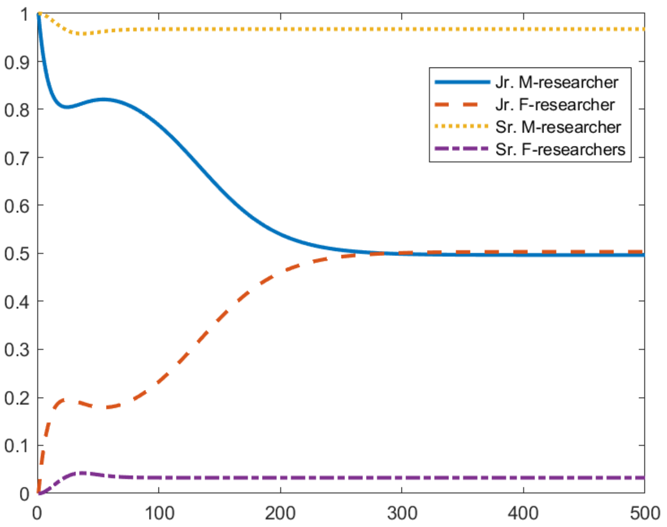
$$\lambda_{1,t}^{\theta,g} = \lambda_{1,t-1}^{\theta,m}\left(1 - \frac{1}{1-\sigma}(a_{1,t} - a_{2,t})\right) + a_{1,t}^{\theta,g} - a_{2,t}^{\theta,g} \tag{30}$$

$$\lambda_{2,t}^{\theta,g} = \lambda_{2,t-1}^{\theta,g}\left(1 - \frac{1}{\sigma}a_{2,t}\right) + a_{2,t}^{\theta,g} \tag{31}$$

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. The online appendix highlights some basic insights via numerical simulations. 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, 2019) . Figure 10 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.

Figure 10: 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 $\sigma = 0.5$.

7. The Impact of Policy Action

In this section we discuss the impact of some policy actions that have been proposed to address gender imbalance. In particular, we consider (i) the impact of mentoring (section 7.1.); and (ii) the impact of affirmative action (section 7.2.). In this section, we take it as a running assumption that endogenous entry is not the cause of group imbalance, as in e.g. subsection 5.1.1.. Indeed, in those cases, lowering the cost of entry C would fully solve the problem. What can we do instead when the underlying reason is just self-image bias, i.e. $\rho < \overline{\rho}(\phi, N)$ in point (a) of Proposition 5?

7.1. 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. First, a clarification: by mentoring we mean policies aimed at helping young researchers (e.g., Ph.D. students) who have already decided to pursue a career in economics. This is distinct from *outreach*, which aims at increasing entry into the profession. We discuss outreach in Section 9..

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" of 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.

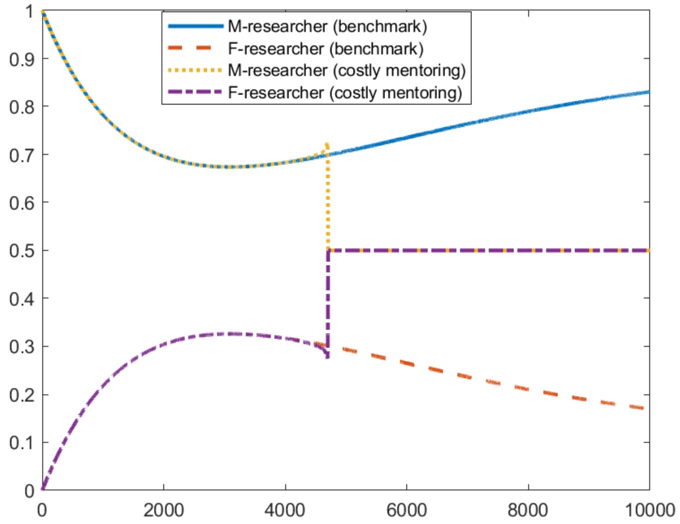
We assume that the cost itself depends on the distance between the young researcher's type θ and the type of the advisor θ^a: The larger the distance and the higher the cost, indicating that it will take a higher effort to "learn" to become a type that is likely to be hired. Note that such distance may 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." For instance, agent $\theta^* = (1, 1, ..., 1)$ is far away from $\theta^m = (1, ..., 1, 0, ..., 0)$ but in an environment such as the benchmark case in Section 2.2., researchers of type θ^* disappear, and so it is in the interest of such researcher to rather become of type θ^m.

The mass of young researchers from group $g \in \{f, m\}$ of type θ who is accepted at time t is then

$$a_t^{\theta,g} = \gamma^{\theta} \lambda_{t-1}^{\theta} \Bigg[\Bigg(p^{\theta,g} \sum_{\theta^a : \widetilde{C}(\theta,\theta^a) \geq \gamma^{\theta^a} \lambda_{t-1}^{\theta^a} - \gamma^{\theta} \lambda_{t-1}^{\theta}} \lambda_{t-1}^{\theta^a} \Bigg) + \Bigg(\lambda_{t-1}^{\theta} \sum_{\theta' : \theta' \neq \theta, \widetilde{C}(\theta',\theta) < \gamma^{\theta} \lambda_{t-1}^{\theta} - \gamma^{\theta'} \lambda_{t-1}^{\theta'}} p^{\theta',g} \Bigg) \Bigg) \Bigg] \tag{32}$$

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

Figure 11: Fraction of F and M Researchers with Costly Mentoring



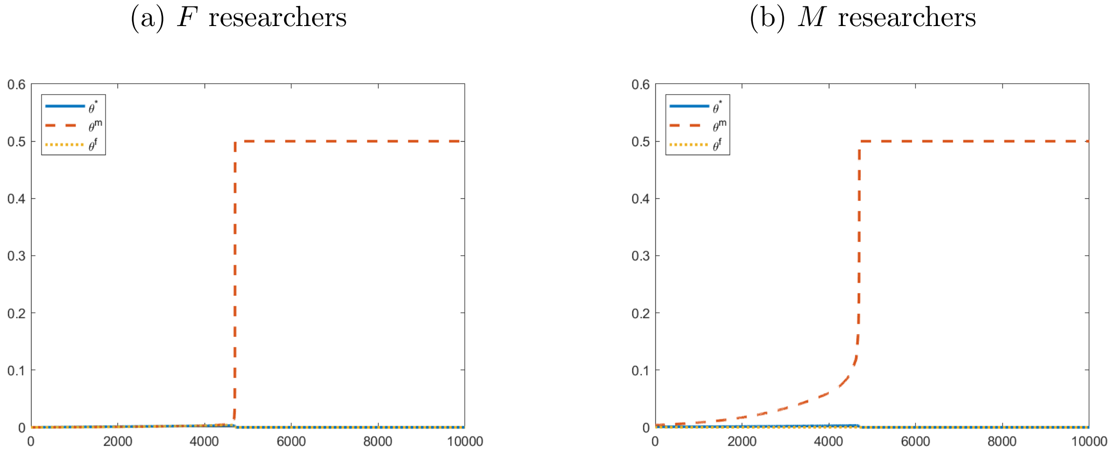
Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, $N = 10$, cost function $C(\theta, \theta') = 0.0250 \sum_{n=1}^{N}(\theta_n - \theta'_n)^2$.

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.

Figure 11 illustrates the dynamics resulting from Eq. (32), 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 12. 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 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.

If the cost function is such that $\max_{\theta'} \widetilde{C}(\theta', \theta^m)$ is large enough, however, then not all young researchers want to switch to θ^m; only those who are sufficiently close to their ran-

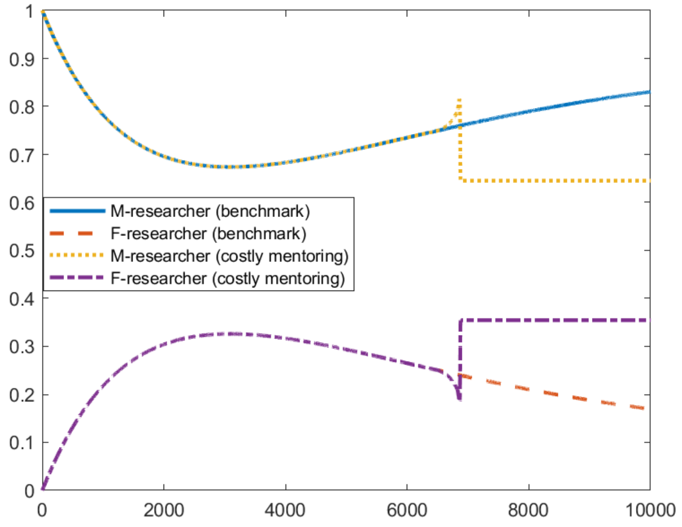
Figure 12: Types of Established F and M Researchers with Costly Mentoring .


(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: $\phi = 0.0.5742$ $(d = 0.3)$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 10$; cost function: $C(\theta, \theta') = 0.0250 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

Figure 13: Fraction of F and M Researchers with High-Cost Mentoring



Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ $(d = 0.3)$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 10$. Cost function: $C(\theta, \theta') = 0.0750 \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$.

domly assigned advisor do. Thus, even though the cost function is the same for M- and F-researchers, young M-researchers have systematically lower cost to switch to θ^m than for F-researchers. In this case, the system still jumps, but the group imbalance persist forever, as illustrated in the next example in Figure 13.

In conclusion, this section shows in a world in which the M-population is initially large, their most prevalent research characteristics θ^m become the standard for society. It follows that it is on average less costly for M-researchers to transform themselves into θ^m-types than for a F-researchers, as M-researchers are already "closer" to θ^m, the most prevalent type of established researcher. So, mentoring helps compared to a world with no mentoring, as shown in Figure 12, but it does not solve the basic problem that research characteristics are dominated by M-types, which involves both loss of knowledge and need not even ensure converge to group balance.

7.2. 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

$$a_t^{\theta,m} = k_t \; \gamma^\theta \; \lambda_{t-1}^\theta \; p^{\theta,m} \tag{33}$$

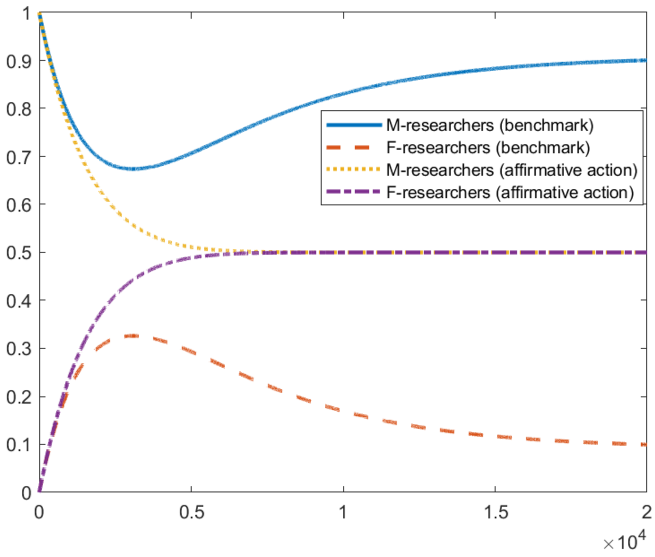
where k_t is a scaling factor

$$k_t = \frac{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',f}}{\sum_{\theta'} \gamma^{\theta'} \; \lambda_{t-1}^{\theta'} \; p^{\theta',m}}$$

This scaling factor 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 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.

What is the intuition of the result? By allowing a larger representation of F researchers, the refereeing population becomes more diverse without requesting the same researcher to "change their type" to be more likely to be accepted by the profession. Note that refereeing is still taking place, in the sense that a researcher of type θ must still be matched with a reviewer of the same type to be accepted. Indeed, the main impact of affirmative action is to affect the set of referees, which makes it possible to reward the research of *talented* F researchers—those who possess a large number of characteristics, and are thus 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.

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

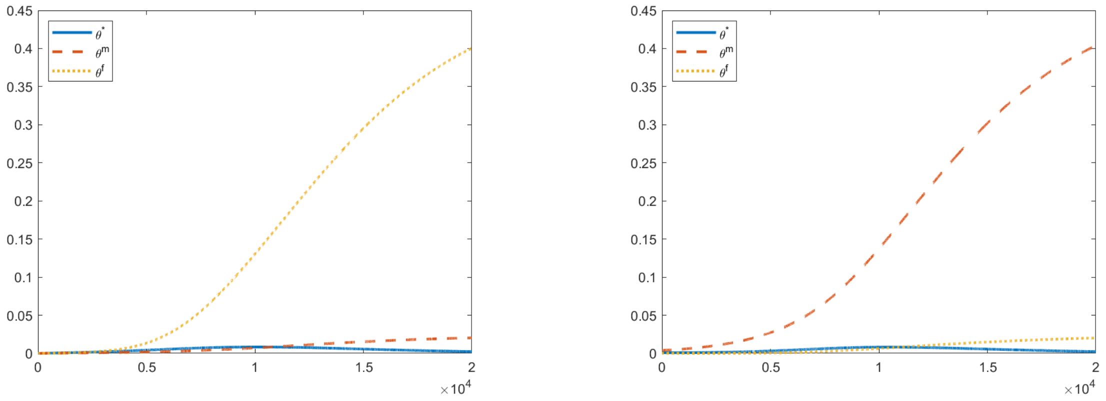


Fraction of M and F researchers when $\lambda_0 = p^m$ and there is an affirmative action policy that requires to accept the same number of M and F researchers. Parameters: $\phi = 0.5742$ ($d = 0.3$), $\gamma_0 = 0.2$, $\rho = 4$, and $N = 10$.

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 $N = 10$

8. Literature Review

There is a considerable body of research on the underlying reason of under-representation of women in the economics profession. In their recent survey paper, Bayer and Rouse (2016) 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 make a compelling argument that the literature suggests that an important role is played by *implicit bias* and *stereotyping.* Among the studies they cite, Milkman, Akinola, and Chugh (2015) fictional prospective students contacted 6,500 professors in 89 disciplines, at 259 institutions, inquiring about research opportunities prior to applying to a Ph.D. program. Analogously to the pioneering study by Bertrand and Mullainathan (2004) (which was in a labor-market setting), student names were randomly assigned to signal gender and race.[11] Bayer and Rouse (2016) also review evidence from the natural and life sciences documenting biases in hiring and gender stereotyping in recommendation letters. Finally, Bayer and Rouse (2016) point to instances in which institutional policies may cause unintended biases in hiring. Two of the examples they provide are especially notable. First, hiring only candidates who have completed their Ph.D. in six years or less may discriminate against minority or female candidates, who on average take longer to complete their doctoral studies. Second, extending the tenure clock for new parents turns out to have a positive effect on male tenure rates, but a negative effect (!) on female tenure rates.

All these findings point to a complex web of interrelated factors that contribute to creating and/or exacerbating the gender bias in the economics profession. In a more recent contribution, Sarsons (2019)'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. We have already noted the nuanced evidence on biases in the refereeing process documented by Card et al. (2020). 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).

[11] Table 1 in Milkman et al. (2015) shows that, across almost all disciplines (with the exception of Fine Arts), a higher proportion of emails from Caucasian males received a reply relative to emails from other students, including women; the difference is largest in "business" (25% additional emails replied) and smaller in the "social sciences," (7%) which comprises economics. Unfortunately, the paper does not provide the results for economics specifically.

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 setting (Cornell and Welch, 1996). 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. They show that, depending on the nature of the information, this may lead to large differences in ex-post treatment of the two groups, even if ex-ante productivity differences are small. Differently from this literature, we assume that, even though the distribution of characteristics is slightly different in the M and F groups, the associated ex-ante distributions of productivity are the same. Furthermore, productivity is observed. In our model, the standard statistical discrimination mechanism would not lead to gender imbalance.

Becker (2010)'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.

9. Conclusions and Policy Implications

Our model highlights a mechanism that endogenously perpetuates certain specific research characteristics over time. This occurs through the reviewing process, as a consequence of reviewers' self-image bias, i.e., the fact that reviewers use their own personal characteristics as a guidance to judge others' research output. Because we are agnostic about the characteristics of male and female researchers (besides what the empirical literature suggests), the policy implications we discuss cannot depend on specific interpretations of the characteristics.

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), the 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 by the reviewer population (see Section 7.1.). 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 (while still maintaining anonymity). 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 (e.g. is the research paper correct? is the research topic relevant? why or why not?) 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. We leave such investigation to future research.

Finally, our model suggests a rationale for affirmative-action policies aimed at 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. We have shown that if males are initially dominant, they will remain so, and research characteristics more prevalent among females will be under-represented. Facilitating the promotion of young female researchers directly counteracts this force, and can lead to a more balanced representation of research characteristics in the steady-state population.

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, 2010.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

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.

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.

Judy Chevalier. Report: committee on the status of women in the economics profession. Technical report, American Economic Association, 2019.

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

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 and Keith S Beauregard. Regulating impressions of others to affirm images of the self. *Social Cognition*, 18(2):198–222, 2000.

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.

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.

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.

Katherine L Milkman, Modupe Akinola, and Dolly Chugh. What happens before? a field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, 100(6):1678, 2015.

Muriel Niederle and Lise Vesterlund. Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–44, 2010.

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

Heather Sarsons. Gender differences in recognition for group work. *Journal of Political Economy*, 2019. conditionally accepted.

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

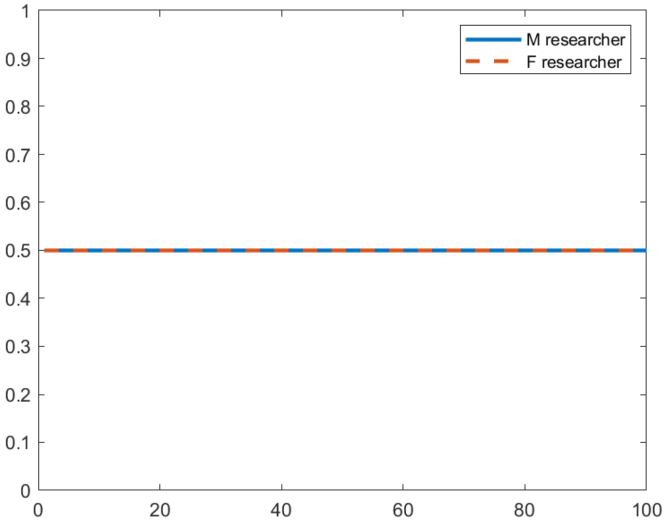
A1.1. Balanced Steady State

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.-2.5., 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 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

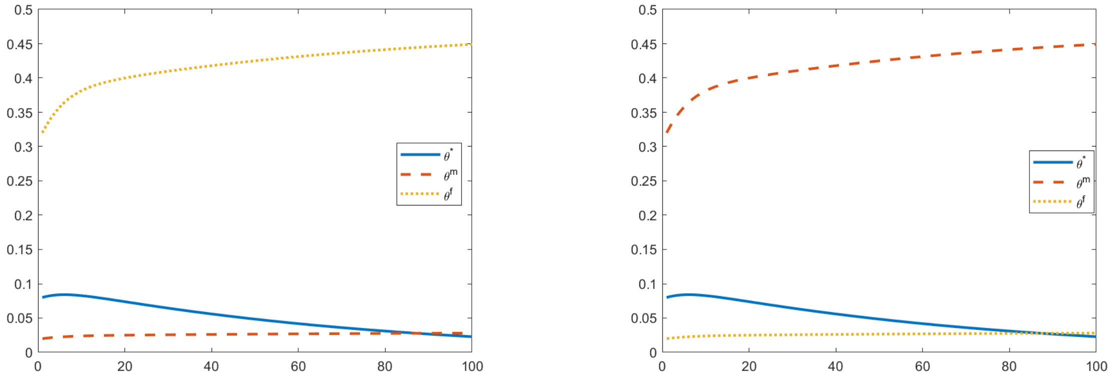


Fraction of M and F researchers when $\lambda_0 = \frac{1}{2}p_m + \frac{1}{2}p_f$. Parameters: $\phi = 0.8$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 2$.

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

In Section 6. we extended the basic model to include different levels of seniorities in the established set of researchers, with seniors evaluating juniors before accepting them into their group, and both seniors and juniors evaluating the young researchers. The analysis is substantially more complex in this case, and we only rely on numerical simulations. The following cases add up to the one discussed in the body of the paper.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.3 demonstrates this. Model parameters are as in Figure 2, so in a single-cohort environment significant group imbalance emerges. The same is true with two ranks; 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.

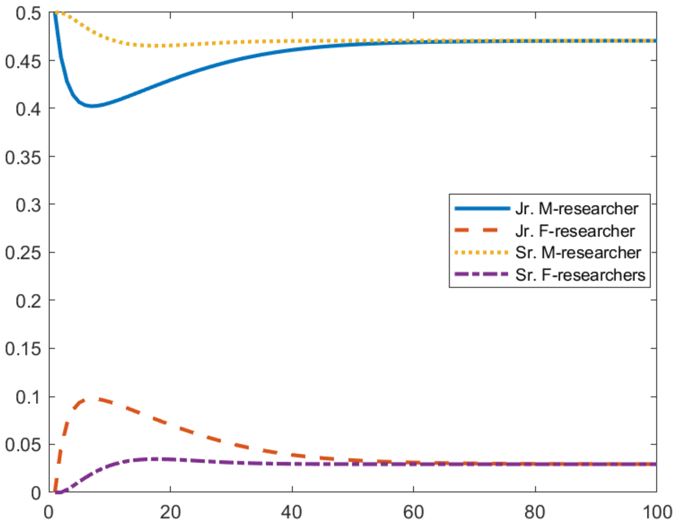
A more pronounced group imbalance can also arise, in the short / medium run, for parameter values for which convergence is eventually attained. This is demonstrated in Figure A.4, where we take $\phi = 0.6$ rather than $\phi = 0.8$. Again, the need to match with a like type twice, coupled with the assumption 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.

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

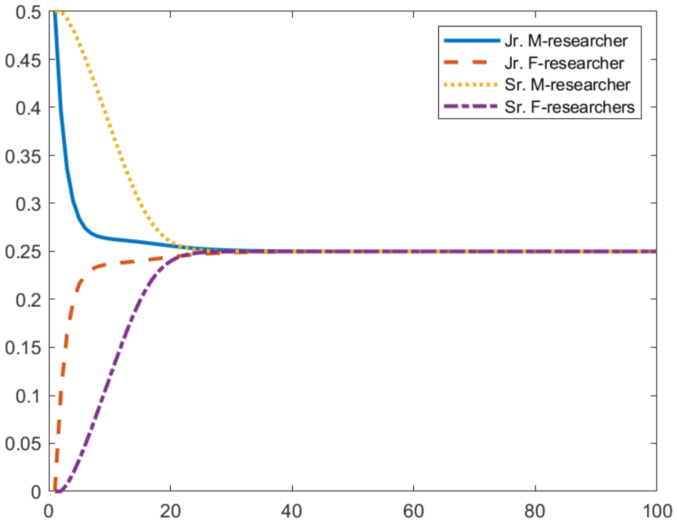
[1] In fact, the bias becomes stronger over time at the senior rank. The reason is that the initial population of junior candidates up for promotion is characterized by types distributed as among male researchers, whereas the initial population of young researchers applying for a junior position is balanced.

Figure A.3: More extreme imbalance for senior rank



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

Figure A.4: Convergence, but greater short-run imbalance among seniors



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

assume that referee r of type θ^r accepts the research of young researcher θ if

$$D(\theta^r, \theta) = \sum_n (\theta^r_n - \theta_n)^2 \le \eta \tag{A.34}$$

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. (8), but the mass $a^{\theta,g}_t$ of accepted researchers of type θ in group $g \in \{f, m\}$ is given by

$$a^{\theta,g}_t \;\; = \;\; \gamma^\theta \sum_{\theta^r : D(\theta^r,\theta) \le \eta} \lambda^{\theta^r}_{t-1} \, p^{\theta,g} \tag{A.35}$$

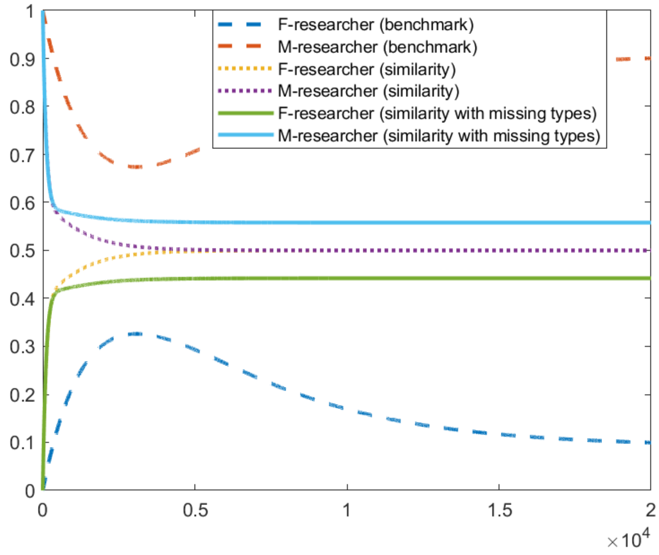
Unfortunately, obtaining general analytical results in this case seems difficult. Therefore, we consider illustrative special cases.

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.34). For every $\eta \ge 1$, 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 $D(\theta_k, \theta_{k+1}) \le \eta$ for all $k = 1, \ldots, K-1$. In this sense, we say that $\Theta = \{0, 1\}^N$ is *η-connected* for every $\eta \ge 1$. 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.34) leads to convergence. For instance, Figure A.5 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.

Figure A.5: 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 $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 10$, and, under research similarity, $\eta = 1$.

A1.3.2. Disconnected Set of Types

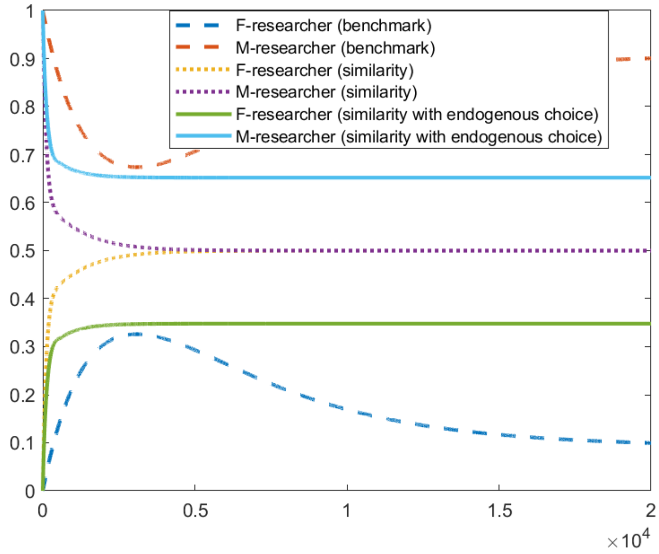
A subset of $\{0,1\}^N$ may well be η-disconnected for some η. For a trivial example, $\{\theta^m, \theta^f\}$ is $(N-1)$-disconnected, because each of the N coordinates of θ^f is different from the corresponding coordinate of θ^f. A fortiori, it is η-disconnected for every $\eta \leq N-1$.

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.5 represent the same parameterization as in the previous subsection, with $\eta = 1$, but applied to a state space Θ obtained by randomly removing 15% of the elements of $\{0,1\}^N$ and suitably renormalizing probabilities. As expected, the system does not attain group balance in the limit.

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.6. Again, the dashed lines and the dotted lines show the total fraction of M- and F-researchers in the benchmark case ($\eta = 0$) and,

Figure A.6: 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 $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 4$, $N = 10$, and, under research similarity, $\eta = 1$.

respectively, the research similarity case ($\eta = 1$). The solid lines now show the the fraction of M- and F-researchers under research simularity ($\eta = 1$) 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. 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

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

$$\lambda_t^\theta = \lambda_t^{\theta,m} + \lambda_t^{\theta,f}. \tag{A.37}$$

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 1 *Assume that $q^\theta \le 1$ for all $\theta \in \Theta$. Then, for all $t \ge 0$, $\lambda_t \in \Delta(\Theta)$, and $\lambda^m_t, \lambda^f_t \in \mathbb{R}^\Theta_+$. Moreover:*

1. *if $\lambda^\theta_0 = 0$, then $\lambda^\theta_t = 0$ for all $t \ge 0$;*
2. *if $\lambda^\theta_0 > 0$, then $\lambda^\theta_t > 0$ for all $t \ge 0$;*
3. *for $\theta, \tilde\theta \in \Theta$ with $\lambda^\theta_0 \cdot \lambda^{\tilde\theta}_0 > 0$:*
   (a) *$\frac{\lambda^\theta_t}{\lambda^\theta_{t-1}} - \frac{\lambda^{\tilde\theta}_t}{\lambda^{\tilde\theta}_{t-1}} = q^\theta - q^{\tilde\theta}$ for all $t \ge 1$, and*
   (b) *$q^\theta > q^{\tilde\theta}$ implies $\frac{\lambda^\theta_t}{\lambda^{\tilde\theta}_t} \to \infty$, and $q^\theta = q^{\tilde\theta}$ implies $\frac{\lambda^\theta_t}{\lambda^{\tilde\theta}_t} = \frac{\bar\lambda^\theta_o}{\bar\lambda^{\tilde\theta}_o}$ for all $t \ge 0$;*
4. *define the set*
$$\Theta^{\max} = \{\theta \in \Theta \,:\, \lambda^\theta_0 > 0,\ \theta \in \arg\max_{\theta' \in \Theta} q^{\theta'}\} \tag{A.38}$$
   and let $\bar\lambda \in \Delta(\Theta)$ be such that
$$\bar\lambda^{\tilde\theta} = \begin{cases} \frac{\lambda^{\tilde\theta}_0}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_0} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} : \end{cases} \tag{A.39}$$
   then $\lim_{t\to\infty} \lambda_t = \bar\lambda$;
5. *define*
$$\bar\lambda^{\tilde\theta,f} = \begin{cases} \frac{\lambda^{\tilde\theta}_0 q^{\tilde\theta,f}}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_0 q^\theta} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} \end{cases} \quad \text{and} \quad \bar\lambda^{\tilde\theta,m} = \begin{cases} \frac{\lambda^{\tilde\theta}_0 q^{\tilde\theta,m}}{\sum_{\theta \in \Theta^{\max}} \lambda^\theta_0 q^\theta} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} : \end{cases} \tag{A.40}$$
   then $\lim_{t\to\infty} \lambda^f_t = \bar\lambda^f$ and $\lim_{t\to\infty} \lambda^m_t = \bar\lambda^m$.

Proof: Eqs. (A.36) and (A.37) imply that
$$\lambda^\theta_t = \left(1 - \sum_{\theta' \in \Theta} \lambda^{\theta'}_{t-1} q^{\theta'}\right) \lambda^\theta_{t-1} + \lambda^\theta_{t-1} q^\theta. \tag{A.41}$$

By assumption $\lambda_0 \in \Delta(\Theta)$. Inductively, suppose $\lambda_{t-1} \in \Delta(\Theta)$ and $\lambda^m_{t-1}, \lambda^f_{t-1} \in \mathbb{R}^\Theta_+$. Summing over Θ on both sides of Eq. (A.41) yields $\sum_\theta \lambda^\theta_t = (1 - \sum_{\theta'} \lambda^{\theta'}_{t-1} q^{\theta'})(\sum_\theta \lambda^\theta_{t-1}) + \sum_\theta \lambda^\theta_{t-1} q^\theta = (1 - \sum_{\theta'} \lambda^{\theta'}_{t-1} q^{\theta'}) + \sum_\theta \lambda^\theta_{t-1} q^\theta = 1$. Furthermore, since $\lambda_{t-1} \in \Delta(\Theta)$, $\sum_{\theta'} \lambda^{\theta'}_{t-1} q^{\theta'} \in [\min_{\theta'} q^{\theta'}, \max_{\theta'} q^{\theta'}] \subseteq [0,1]$; moreover, $q^\theta \ge 0$ and $\lambda^\theta_{t-1} \ge 0$, so Eq. (A.41) implies that $\lambda^\theta_t \ge 0$ as well. By the same argument, $q^\theta \ge 0$ and $\lambda^{\theta,g}_{t-1} \ge 0$ for $g \in \{f,m\}$ imply $\lambda^{\theta,g}_t \ge 0$ for $g \in \{f,m\}$ as well by Eq. (A.36). Thus, $\lambda_t \in \Delta(\Theta)$, and $\lambda^g_t \in \mathbb{R}^\Theta_+$ 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.41), 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.41) 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.41) for type θ by λ_{t-1}^{θ}, which is assumed to be positive; this yields

$$\frac{\lambda_t^{\theta}}{\lambda_{t-1}^{\theta}}=1+q^{\theta}-\sum_{\theta'}\lambda_{t-1}^{\theta'}q^{\theta'}. \tag{A.42}$$

A similar equation holds for $\tilde{\theta}$. This immediately yields 3(a). To derive 3(b), since $\lambda_t^{\theta'}=\lambda_0^{\theta'}\cdot\prod_{s=1}^{t}\frac{\lambda_s^{\theta'}}{\lambda_{s-1}^{\theta'}}$ for $\theta'=\theta,\tilde{\theta}$,

$$\frac{\lambda_t^{\theta}}{\lambda_t^{\tilde{\theta}'}}=\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\frac{\prod_{s=1}^{t}\frac{\lambda_s^{\theta}}{\lambda_{s-1}^{\theta}}}{\prod_{s=1}^{t}\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}}=\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\prod_{s=1}^{t}\frac{\frac{\lambda_s^{\theta}}{\lambda_{s-1}^{\theta}}}{\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}}=\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\prod_{s=1}^{t}\frac{\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}+q^{\theta}-q^{\tilde{\theta}}}{\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}}=\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\prod_{s=1}^{t}\left(1+\frac{q^{\theta}-q^{\tilde{\theta}}}{\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}}\right).$$

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.42), for all $s\geq 1$, since $\lambda_{s-1}\in\Delta(\Theta)$ and $q\in[0,1]^{|\Theta|}$, $\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}\leq 1+q^{\tilde{\theta}}$. Therefore, each term in parentheses is not smaller than $1+\frac{q^{\theta}-q^{\tilde{\theta}}}{1+q^{\tilde{\theta}}}>1$. It follows that

$$\frac{\lambda_t^{\theta}}{\lambda_t^{\tilde{\theta}'}}=\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\prod_{s=1}^{t}\left(1+\frac{q^{\theta}-q^{\tilde{\theta}}}{\frac{\lambda_s^{\tilde{\theta}}}{\lambda_{s-1}^{\tilde{\theta}}}}\right)\geq\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}\cdot\left(1+\frac{q^{\theta}-q^{\tilde{\theta}}}{1+q^{\tilde{\theta}}}\right)^{t}\to\infty.$$

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

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

$$\frac{\lambda_t^{\tilde{\theta}}}{\sum_{\theta\in\Theta^{\max}}\lambda_t^{\theta}}=\frac{1}{\sum_{\theta\in\Theta^{\max}}\frac{\lambda_t^{\theta}}{\lambda_t^{\tilde{\theta}}}}=\frac{1}{\sum_{\theta\in\Theta^{\max}}\frac{\lambda_0^{\theta}}{\lambda_0^{\tilde{\theta}}}}=\frac{\lambda_0^{\tilde{\theta}}}{\sum_{\theta\in\Theta^{\max}}\lambda_0^{\theta}}=\bar{\lambda}^{\tilde{\theta}},$$

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

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

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

Finally, consider Claim 5. Fix $g \in \{f, m\}$. First, since $0 \leq \lambda_t^{\theta,g} \leq \lambda_t^\theta$ for all $t \geq 0$, if $\theta \notin \Theta^{\max}$ then by Claim 4 $\lambda_t^\theta \to \bar{\lambda}^\theta = 0$, and so $\lambda_t^{\theta,g} \to 0 = \bar{\lambda}^{\theta,g}$ 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.36) 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.36) can be written as

$$\lambda_t^{\theta,g} = (1 - q_{t-1})\lambda_{t-1}^{\theta,g} + \lambda_{t-1}^\theta q^{\theta,g}. \tag{A.43}$$

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

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

$$\lambda_t^{\theta,g} = \lambda_T^{\theta,g} \prod_{s=T}^{t-1} (1 - q_s) + q^{\theta,g} \sum_{s=T}^{t-1} \lambda_s^\theta \prod_{r=s+1}^{t-1} (1 - q_r). \tag{A.44}$$

For $t = T + 1$, this follows from Eq. (A.43). Inductively, assume it holds for $t - 1 > T$. Then, by Eq. (A.43) and the inductive hypothesis,

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

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.44) 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 1.

Lemma 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. (11).

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:

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

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

$$\rho > \frac{1}{4} \left(\left(\frac{1-\phi}{\phi}\right)^{\frac{N}{2}} + \left(\frac{\phi}{1-\phi}\right)^{\frac{N}{2}} \right)^2 = \bar{\rho}(\phi, N), \tag{A.45}$$

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.45) 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 1 and Claim 2 of Lemma 1. Parts (a) and (b) follow from Claim 1 in Lemma 1 and Claim 4 in Theorem 1. Corollary 1 follows from Claim 3 in Lemma 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

$$a^{\theta,m}_t + a^{\theta^{\text{sym}},m}_t \geq a^{\theta,f}_t + a^{\theta^{\text{sym}},f}_t. \tag{A.46}$$

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

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 $p^{\theta,m} + p^{\theta,f} = p^{\theta^{\text{sym}},m} + p^{\theta^{\text{sym}},f}$.

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 $\lambda^\theta_0 = p^{\theta,m} \geq p^{\theta^{\text{sym}},m} = \lambda^{\theta^{\text{sym}}}_0 > 0$. Then, by part 3(a) of Theorem 1, for every $t > 0$, $\frac{\lambda^\theta_t}{\lambda^\theta_{t-1}} = \frac{\lambda^{\theta^{\text{sym}}}_t}{\lambda^{\theta^{\text{sym}}}_{t-1}}$, and hence $\frac{\lambda^\theta_t}{\lambda^{\theta^{\text{sym}}}_t} = \frac{\lambda^\theta_{t-1}}{\lambda^{\theta^{\text{sym}}}_{t-1}} = \frac{\lambda^\theta_0}{\lambda^{\theta^{\text{sym}}}_0} \geq 1$. Thus, $\lambda^\theta_t \geq \lambda^{\theta^{\text{sym}}}_t$ for all $t > 0$ as well. Finally, $\gamma^{\theta^{\text{sym}}} = \gamma^\theta \equiv \bar{\gamma}$. Therefore, for every $t \geq 1$,

$$a^\theta_t = a^{\theta,m}_t + a^{\theta,f}_t = \bar{\gamma}\lambda^\theta_{t-1}(p^{\theta,m} + p^{\theta,f}) \geq \bar{\gamma}\lambda^{\theta^{\text{sym}}}_{t-1}(p^{\theta^{\text{sym}},m} + p^{\theta^{\text{sym}},f}) = a^{\theta^{\text{sym}},m}_t + a^{\theta^{\text{sym}},f}_t = a^{\theta^{\text{sym}}}_t.$$

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,

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

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.46), holds; furthermore, if $\theta = \theta^{\text{sym}}$, then $a^{\theta,m}_t = a^{\theta,f}_t$.

Now fix $L \in \{0, \dots, N\}$. 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.46), 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 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 (15) through (18) then follow from the specification of p^m and p^f. Eq. (20) 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 1 Claim (3). *Q.E.D.*

Proof of Proposition 6: 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 $\tau = 0, \dots, t(j) - t(j-1)$, 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 $\lambda^m_{\tau,k}, \lambda^f_{\tau,k}, \lambda_{\tau,k} \in \mathbb{R}^\Theta_+$ for all $\tau \geq 1$ such that (i) holds for $j = k$, and $\lambda^\theta_{\tau,j} \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_k$. In the latter case, if the sequences of such vectors converge, then $\lim_{\tau\to\infty} \lambda^m_{\tau,k} = \lim_{t\to\infty} \lambda^m_t$ and similarly for $\lambda^f_{\tau,k}$ and $\lambda_{\tau,k}$.

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

Define $q^g_k \in \mathbb{R}^\Theta_+ \setminus \{0\}$ for $g = f, m$ by $q^{\theta,g}_k = \gamma^\theta p^{\theta,g}$ if $\theta \in \Theta_k$, and $q^{\theta,g}_k = 0$ otherwise. Then $q^{\theta,m}_k + q^{\theta,f}_k \leq 1$ for all θ. Consider the sequences $(\lambda^{\theta,g}_{\tau,k})_{\tau\geq 0}$ for $g = f, m$ and $(\lambda^\theta_{\tau,k})_{\tau\geq 0}$ defined by Eqs. (A.36)–(A.37) for the vectors q^f_k, q^m_k .

Suppose first that there are $\bar\tau > 0$ and $\theta_0 \in \Theta_k$ such that $\lambda^{\theta_0}_{\bar\tau,k} < \frac{C}{\gamma^{\theta_0}(P-U)}$. Let $t(k) = t(k-1) + \bar\tau$. Then, for each group $g = f, m$, the dynamics in Eqs. (A.36)–(A.37) induced by the vectors q^f_k, q^m_k for the subsequence $(\lambda^g_{\tau,k})_{\tau=0,\ldots,\bar\tau}$ coincide with those in Eq. (24) 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 $q^{\theta,m}_k + q^{\theta,f}_k = 0$ for $\theta \in \Theta \setminus \Theta_k$; hence, for all $\theta' \in \Theta$ and all $\theta \in \Theta \setminus \Theta_k$, $q^{\theta,m}_k + q^{\theta,f}_k \leq q^{\theta'}_{m,k} + q^{\theta'}_{f,k}$. By part 3(a) in Theorem 1, it must be the case that $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k} \leq 1$: 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 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 $\lambda^\theta_{\bar\tau,k} \geq \frac{C}{\gamma^\theta(P-U)}$ for all $\theta \in \Theta_k$, then for each group $g = f, m$, the dynamics in Eqs. (A.36)–(A.37) induced by the vectors $q_{m,k}, q_{f,k}$ for the subsequence $(\lambda^g_{\tau,k})_{\tau\geq 0}$ coincide with those in Eq. (24) for the subsequence $(\lambda^g_t)_{t\geq t(k-1)}$. Again, in this case (i) holds for $j = k$. This completes the proof of the claim.

Since the set Θ is finite, there exists $K \geq 0$ such that the induction stops—that is, $\lambda^\theta_{\bar\tau,K} \geq \frac{C}{\gamma^\theta(P-U)}$ for all $\theta \in \Theta_K$. Let $\Theta^{\max}_k = \arg\max\{q^{\theta,m}_k + q^{\theta,f}_k : \theta \in \Theta\}$. Since $\Theta_0 \supsetneq \Theta_1 \supsetneq \ldots \supsetneq \Theta_K$, by the definition of the vectors q^g_k for $g = f, m$, also $\Theta^{\max}_0 \supseteq \Theta^{\max}_1 \supseteq \ldots \supseteq \Theta^{\max}_K$. Moreover, for every $k = 0, \ldots, K-1$, and every $\theta \in \Theta^{\max}_k$, $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k} \geq 1$ for $0 \leq \tau < t(k) - t(k)$; otherwise, by part 3(a) in Theorem 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 1.

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

$$\frac{C}{\gamma^\theta P} \leq \lambda_{0,0}^\theta \leq \lambda_{t(1)-t(0),0}^\theta = \lambda_{0,1}^\theta \leq \lambda_{t(2)-t(1),1}^\theta \dots \leq \lambda_{0,K}^\theta,$$

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

$$\frac{\lambda_{0,K}^\theta}{\lambda_{0,K}^{\theta'}} = \frac{\lambda_{t(K-1)}^\theta}{\lambda_{t(K-1)}^{\theta'}} = \frac{\lambda_{t(K-1)-t(K-2),K-1}^\theta}{\lambda_{t(K-1)-t(K-2),K-1}^{\theta'}} = \frac{\lambda_{0,K-1}^\theta}{\lambda_{0,K-1}^{\theta'}} = \dots = \frac{\lambda_{t(0)-t(-1),0}^\theta}{\lambda_{t(0)-t(-1),0}^{\theta'}} = \frac{\lambda_{0,0}^\theta}{\lambda_{0,0}^{\theta'}} = \frac{\lambda_0^\theta}{\lambda_0^{\theta'}}.$$

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

$$\bar\lambda^\theta = \bar\lambda_K^\theta = \frac{\lambda_{0,K}^\theta}{\sum_{\theta' \in \Theta^{\max}} \lambda_{0,K}^{\theta'}} = \frac{1}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_{0,K}^{\theta'}}{\lambda_{0,K}^\theta}} = \frac{1}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_0^{\theta'}}{\lambda_0^\theta}} = \frac{\lambda_0^\theta}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'}}. \tag{A.47}$$

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

$$\bar\lambda^{\theta,m} = \bar\lambda_K^{\theta,m} = \frac{\lambda_{0,K}^\theta q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_{0,K}^{\theta'} q_K^{\theta'}} = \frac{q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_{0,K}^{\theta'}}{\lambda_{0,K}^\theta} q_K^{\theta'}} = \frac{q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \frac{\lambda_0^{\theta'}}{\lambda_0^\theta} q_K^{\theta'}} = \frac{\lambda_0^\theta q_K^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} q_K^{\theta'}}, \tag{A.48}$$

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.47) 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.48) 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.48), $\bar\Lambda^m = \bar\lambda^{m,\theta^m} = \bar\lambda_K^{m,\theta^m} = \frac{\gamma^{\theta^m} p^{\theta^m,m}}{\gamma^{\theta^m}(p^{\theta^m,m} + p^{\theta^m,f})} = \frac{p^{\theta^m,m}}{p^{\theta^m,m} + p^{\theta^m,f}} = \frac{\phi^N}{\phi^N + (1-\phi)^N}$, as asserted. Finally, we compare this quantity with its counterpart in Eq. (20):

$$\frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^N} = \frac{(1-\phi)^{2N} + \phi^{2N}}{\left[(1-\phi)^N + \phi^N\right]^2} <$$

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

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 $t(0), \dots, t(K)$. *Q.E.D.*

Proof of Proposition 7. 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 $\phi > \frac{1}{2}$, $p^{\theta,m} > p^{\theta^{\text{sym}},m}$. At time 0 we thus have $\lambda_0^\theta = p^{\theta,m} > p^{\theta^{\text{sym}},m} = \lambda_0^{\theta^{\text{sym}}}$. Moreover, since p_f is defined with the roles of ϕ and $1-\phi$ reversed, $p^{\theta,f} = p^{\theta^{\text{sym}},m} < p^{\theta,m} = p^{\theta^{\text{sym}},f}$.

Since $\gamma^{\theta^{\text{sym}}} = \gamma^\theta$, it follows that at time 0, if $\lambda_0^{\theta^{\text{sym}}} > \frac{C}{\gamma^{\theta^{\text{sym}}}P}$, then also $\lambda_0^\theta > \frac{C}{\gamma^\theta P}$. In addition, $p_m^\theta + p_f^\theta = p_m^{\theta^{\text{sym}}} + p_f^{\theta^{\text{sym}}}$. Thus, in the notation of Proposition 6, for $t < \min(t^\theta, t^{\theta^{\text{sym}}})$, both θ and θ^{sym} apply, and applying part 3(a) of Theorem 1 to the relevant subsequence of $(\lambda_t)_{t\geq 0}$ as in the proof of Proposition 6, $\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} = \frac{\lambda_t^{\theta^{\text{sym}}}}{\lambda_{t-1}^{\theta^{\text{sym}}}}$, 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 7, $\lambda_t^\theta = \lambda_t^{\theta^{\text{sym}}}$ 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 7, and in addition $p^{\theta > p_m^{\theta^{\text{sym}}},m}$. *Q.E.D.*