Self-image Bias and Lost Talent

Marciano Siniscalchi
Northwestern University

Pietro Veronesi
University of Chicago, NBER, and CEPR

June 18, 2026

Abstract

We develop an overlapping-generation model in which established researchers evaluate new scholars belonging to two groups, M and F. Even with group-neutral evaluations and identical productivity distributions, mild between-group heterogeneity and evaluators' self-image bias generate persistent imbalance favoring the initially dominant group. The model matches several empirical patterns: stronger and increasing imbalance in top research institutions, higher average quality among accepted F-researchers, field clustering by group, larger senior-level disparities, and lower credit for F-researchers in coauthored work. Mentorship reduces imbalance but increases talent loss among F-researchers.

Keywords: gender discrimination, self-image bias, mentoriship

JEL codes: A11, J16, J7

* Veronesi acknowledges financial support from the Fama-Miller Center for Research in Finance and by the Center for Research in Security Prices at the University of Chicago Booth School of Business. For useful comments, we thank Gadi Barlevi, Marco Bassetto, Alberto Bisin, Shereen Chaudhry, Hui Chen, Alexia Delfino, James Dow, Nicola Gennaioli, Lars P. Hansen, Christopher Hennessy, Alex Imas, Jessica Jaffers, Seema Jayachandran, Emir Kamenica, Elisabeth Kempf, Nadia Malenko, Lubos Pastor, Jane Risen, Antoinette Schoar, Oleg Urminsky, Laura Veldkamp, Adrien Verdhelan, participants at Behavioral Science workshop, the Finance workshop, and the Economics workshop at the University of Chicago, and seminar participants at the Bocconi University, Chicago Fed, Hitotsubashi University, Imperial College, LBS, MIT, Northwestern University, the 2021 NBER Economics of Culture meetings, the 2021 Adam Smith conference, the 2021 WFA meetings, the 2022 AEA meetings, the 2023 NBER meetings in asset pricing and corporate finance (joint session), and the 2023 MFA meetings. Errors are our own. The views expressed in this paper are our own and do not necessarily reflect the views of our respective employers. We declare that we have no relevant or material financial interests that relate to the research described in this paper. Contact information: Marciano Siniscalchi: marciano@northwestern.edu; Pietro Veronesi: pietro.veronesi@chicagobooth.edu.

1. Introduction

Gender imbalance in the workplace has been the subject of a vast theoretical and empirical literature for over 50 years. Prominent theories range from taste-based discrimination (Becker, 1957) to statistical discrimination (Phelps, 1972; Aigner and Cain, 1977), from conscious or unconscious biases (Bertrand and Mullainathan, 2004), to systemic discrimination (Bohren, Hull, and Imas, 2025). The empirical literature testing these theories is vast and finds evidence of bias to varying degrees (see the review by Blau and Kahn, 2017). While this work has helped shape numerous interventions around the world to reduce gender imbalance (e.g., debiasing and mentorship programs), progress has been slow (Blau and Kahn, 2017). As a case in point, we focus in this paper on the academic market for economists. The economics profession has long been male-dominated. The Committee on the Status of Women in the Economics Profession (CSWEP), a standing committee of the AEA since 1971, [1] has regularly documented the progress of women in economics (see Chevalier (2020)). Over the last 30 years, that progress has been limited. For instance, while over this period the fraction of women among undergraduate economics majors has risen to over 40%—indicating increasing interest in economics among female students—the fraction of female assistant professors—i.e., the intake for the academic career—has been nearly flat at around 24% since 1994, with a small uptick to 28% in 2024.

This very slow progress is puzzling given the numerous initiatives aimed at increasing female representation in the economics profession over the past several decades. Many of these interventions are informed by existing theories of discrimination, such as taste-based and statistical discrimination, implicit bias, and stereotyping, which we review in Section 7. From this perspective, recent empirical evidence may suggest that efforts to remove such sources of discrimination or bias have only partially succeeded. For instance, Card, DellaVigna, Funk, and Iriberri (2020) documents that acceptance rates for women-authored papers are lower conditional on quality (proxied by future citations); Sarsons (2017) and Sarsons, Gërxhani, Reuben, and Schram (2021) show that female coauthors tend to receive less credit for published papers that are joint with male coauthors; and Dupas, Modestino, Niederle, and Wolfers (2021) document a bias against female presenters in economics seminars. Large differences in women's representation exist across fields, however (e.g. Chari and Goldsmith-Pinkham, 2018 and Lundberg and Stearns, 2019), which would then suggest that gender bias is more prominent in some economic fields than others.

Besides the evidence just cited, Section 2 collects additional stylized facts about the under-representation of women in academia. Specifically, under-representation is higher

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

in research-intensive institutions—but only for tenure-track positions; the gap in under-representation between top institutions and other institutions has been increasing over the last 50 years; and under-representation is widespread in both the US and in Europe—including in Nordic European countries, which are otherwise less gender-biased.

We propose a novel theory that is consistent with this empirical evidence but that does not depend on stereotypes or gender discrimination, whether taste-based, statistical, conscious, or unconscious. In our model, gender imbalance is due to the combination of self-image bias, i.e. the tendency of individuals to place more weight on their own positive attributes when judging others, and mild population heterogeneity in equally-valuable research characteristics. Both assumptions have strong empirical and experimental support, as we discuss below. Our model, which we calibrate to the data, yields several additional predictions, that are also verified in the data. Moreover, it suggests different policy actions to increase female representation and, especially, reduce talent loss.

Specifically, our model features overlapping generations of agents that belong to one of two groups, labelled M and F. A new cohort of young M- and F-researchers appears in every period, in equal proportions. Each researcher is endowed with a type that determines his or her productivity, which we take to mean the probability of producing research that achieves its objectives. The types are randomly and symmetrically distributed in the population of young researchers, so that the distribution of the associated, type-dependent *productivities* is the same in both the M- and F-population. That is, both groups are ex-ante identical in their ability to produce quality research. However, we also assume that some types are slightly more common in the F-group and others symmetrically slightly more common in the M-group. As in the data, we let between-group heterogeneity be far smaller than within-group heterogeneity. We emphasize that we do not make any assumptions about the *origins* of these distributional differences, which can very well be socially determined (see e.g. Guiso, Monte, Sapienza, and Zingales (2008), Falk, Becker, Dohmen, Enke, Huffman, and Sunde (2018), and Andersen, Ertac, Gneezy, List, and Maximiano (2013)), 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, the referee) decides whether or not to accept the young researcher as a member of the established population—and thus as a referee of future cohorts. Importantly, each referee's perceptions of young researchers' output reflect self-image bias (Lewicki, 1983): evaluators use their own type as a yardstick to assess others' research. The referees' evaluation is group-neutral: each given referee uses the same criterion to assess young M and F researchers. If the referee's evaluation is positive, the latter becomes a recognized, permanent member of the population; otherwise, he or she leaves the model.

Our key finding is that, for a non-degenerate set of model parameters that we characterize explicitly, the combination of self-image bias and even mild between-group heterogeneity generates a persistent bias that favors young researchers who belong to the group that is initially larger, say the M-group. Moreover, there is no convergence. While researchers from the F-group are also successful, not only are they a minority: they are endogenously selected to be the ones whose types are closer to the ones of the M-researchers; this perpetuates the bias forward. Moreover, this mechanism yields talent loss, as types slightly more common in the F-population are under-represented in the limit. In addition, the same basic logic delivers all the stylized facts collected in Section 2.

We calibrate the model to the data to evaluate whether our mechanism can indeed bring about gender imbalance that is quantitatively relevant even if the distributional differences between M- and F-populations are as small as documented in the data. For this task, we need to be concrete on the notion of types. We identify types with vectors of research characteristics. Examples of such characteristics include research approach (e.g. empirical or theoretical), methodology (e.g. structural versus reduced form), field, topic, type of questions asked, depth vs. breadth, writing style, ties to reality, policy relevance, and so on. In this specification, symmetry means that for every characteristic that is slightly more prevalent in the M-group there is a characteristic that is slightly more prevalent in the F-group; furthermore, all research characteristics are equally valuable: each has the same positive effect on the likelihood of producing quality research. Our calibrated parameters match the very mild heterogeneity in characteristics between male and female populations documented in the psychology literature (see below), as well as the success rate of Economics PhD students to become assistant professors.

Under our calibrated parameters, imbalance occurs and the system converges to about 20% women in academia, which is quite close to the data. In addition, similarly to the data, institutions with higher publication frequency have significantly fewer F-researchers; the gap in F-group under-representation between top institutions and the full set of institutions increases over time; accepted F-researchers have higher objective quality; and there is clustering of M- and F-researchers across fields. Moreover, surviving F-researchers are those whose research characteristics are closer to the ones that are more prevalent among M-researchers. 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.

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

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

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

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

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

First, there is a considerable body of research studying gender differences in personality traits, preferences, and attitudes. Regarding personality traits, Hyde and Linn (2006) reviews the literature and concludes that medium-sized effects are found for aggression (Cohen's d

between 0.40 and 0.60) and activity level in the classroom ($d = 0.49$)[2]. 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 U.S. 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). Finally, Andre and Falk (2021) survey nearly 10,000 economists' opinion about the current state and preferred direction of economic research. They find that female scholars are significantly more likely to emphasize multidisciplinarity, disruptive research, and policy relevance (cf. Table 3.)

The second important assumption of our model is researchers' self-image bias. The psychological literature on self-image bias (Lewicki, 1983) suggests that, when evaluating others, individuals tend to place more weight on positive attributes that they themselves possess (or believe they possess). Hill, Smith, and Hoffman (1988) show that this is true in particular when subjects are asked to select a partner in a competitive game. Dunning, Perie, and Story (1991) argue that a similar principle is at work when judging social categories by means of prototypes (e.g., what makes a good economist?): "people may expect the 'ideal instantiation' of a desirable social category to resemble the self in its strengths and idiosyncracies" (p. 958). Story and Dunning (1998) document a "rational" source for self-image bias and self-serving prototypes: in their experiment, "those who received success feedback came to perceive a stronger relationship between 'what they had' and 'what it takes to succeed' than did those who received failure feedback" (p. 513). Translated to our environment, established researchers view their personal success in research as evidence that their own research characteristics are the right ones to produce quality research that, in addition, are valuable to society. Hence, they use the same characteristics to evaluate the research of others.

Our assumption that referees accepts young researchers who are similar to them can also be due to referees' preferences (e.g. theorists like theorists, and empiricists like empiricists) rather than self-image bias. However, this interpretation must be subject to two caveats to fit our model. First, referees' preferences do *not* take group membership into account;

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

thus, even this "homophily" interpretation of our model differs from Becker's taste-based theory of discrimination. Moreover, in this interpretation, referees do not value heterogeneity (e.g., theorists derive no benefit from interacting with empiricists, and conversely), nor the candidate's objective productivity. This strikes us as extreme.

2. Stylized Facts

We begin by discussing a collection of stylized facts about the percentage of women students and faculty in the U.S. and elsewhere.

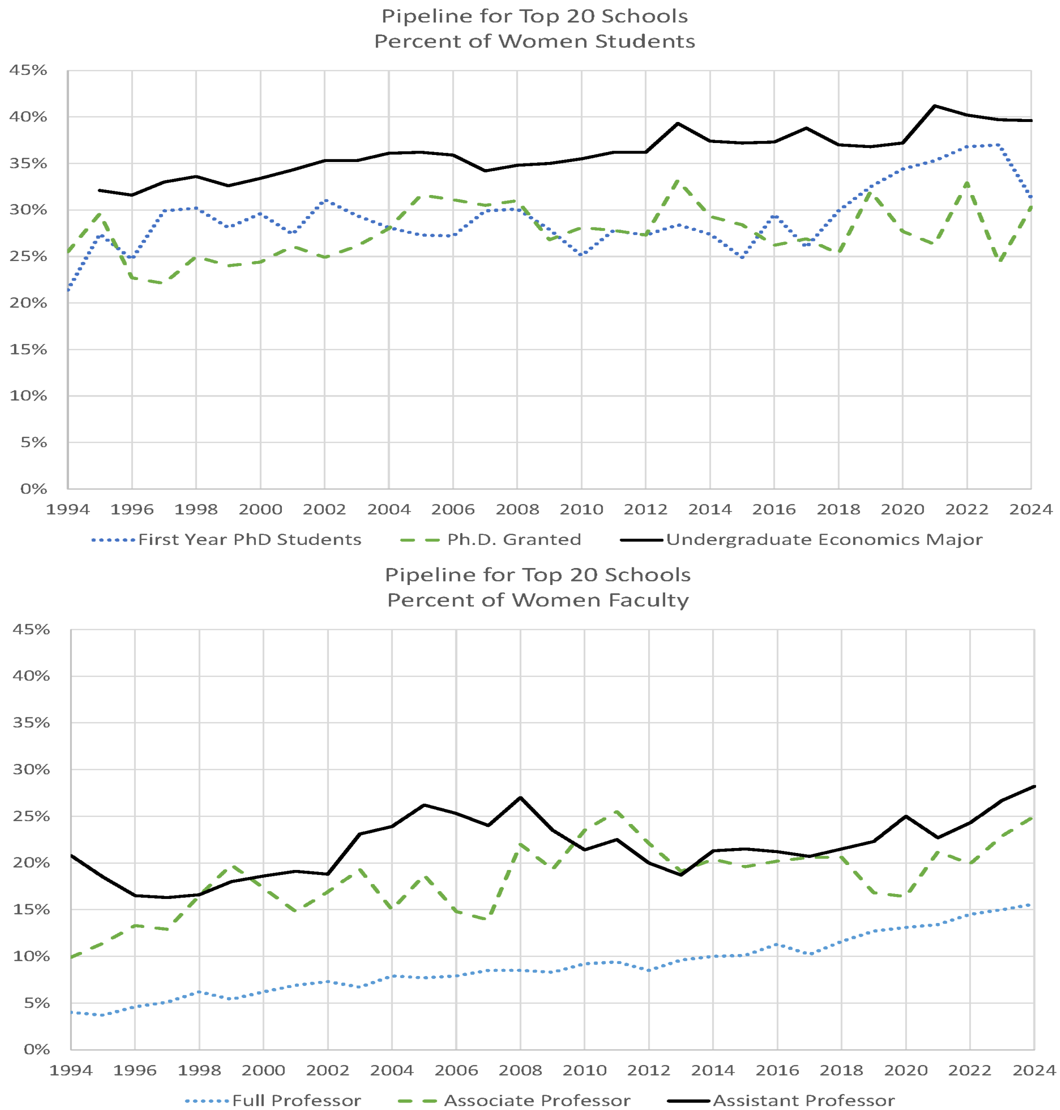
The top panel of Figure 1 shows that between 1994 and 2024, the fraction of women in undergraduate economics majors increased to over 40% in the top-20 schools. During the same period, the fraction of women PhD students has been flat at around 30%. The bottom panel is the crux of the issue: among assistant professors—i.e. the intake for the academic career—the fraction of women in top-20 schools has been flat at around 24% since 1994, albeit with a minor uptick to 28% in 2024. In other words, as discussed in the introduction, as far as representation among assistant professors is concerned, there has been very little progress in the top-20 schools over a time span of 30 years.

Even more striking, the top panel of Figure 2 shows the difference between schools with and without a PhD program, and between top institutions and all institutions. Universities without a PhD program hire over 40% of female tenure-track faculty, while universities with a PhD program hire just over 30% of women assistant professors. In addition, the top-20 schools hire only 25%, and the top-10 only 23%. In sharp contrast, the share of women among teaching faculty is quite uniform across schools, at around 34–38%. The difference between research schools and non-research schools suggests that women under-representation is associated with research intensity. Indeed, the fraction of women among teaching faculty is now close to the fraction of women in undergraduate economics majors (about 40%).

The CSWEP report refers to top 10 and top 20 universities but does not describe how this ranking is determined. The economics profession typically associates rank with research output (Chevalier (2020)). With this in mind, Table 1 shows the results of a regression of the share of women faculty at an institution on a measure of research intensity at that institution, obtained from Tilburg's ranking of U.S. institutions in terms of research output.[3] While there is considerable noise, the results suggest that an institution's research intensity is significantly negatively correlated with the percentage of women faculty in that institution.

[3]The rankings are available at https://econtop.uvt.nl/rankingsandbox.php. We run the "sandbox" in two ways. First, we use the Tilburg ranking by resetting the journals to be used, selecting USA as country, choosing to weight the results by journals impact factors, and list the top 300 universities. This is the "Tilburg ranking" in the table. Second, we selected the top 5 journals and obtained the new ranking.

Figure 1: Percentage of Women in Academia

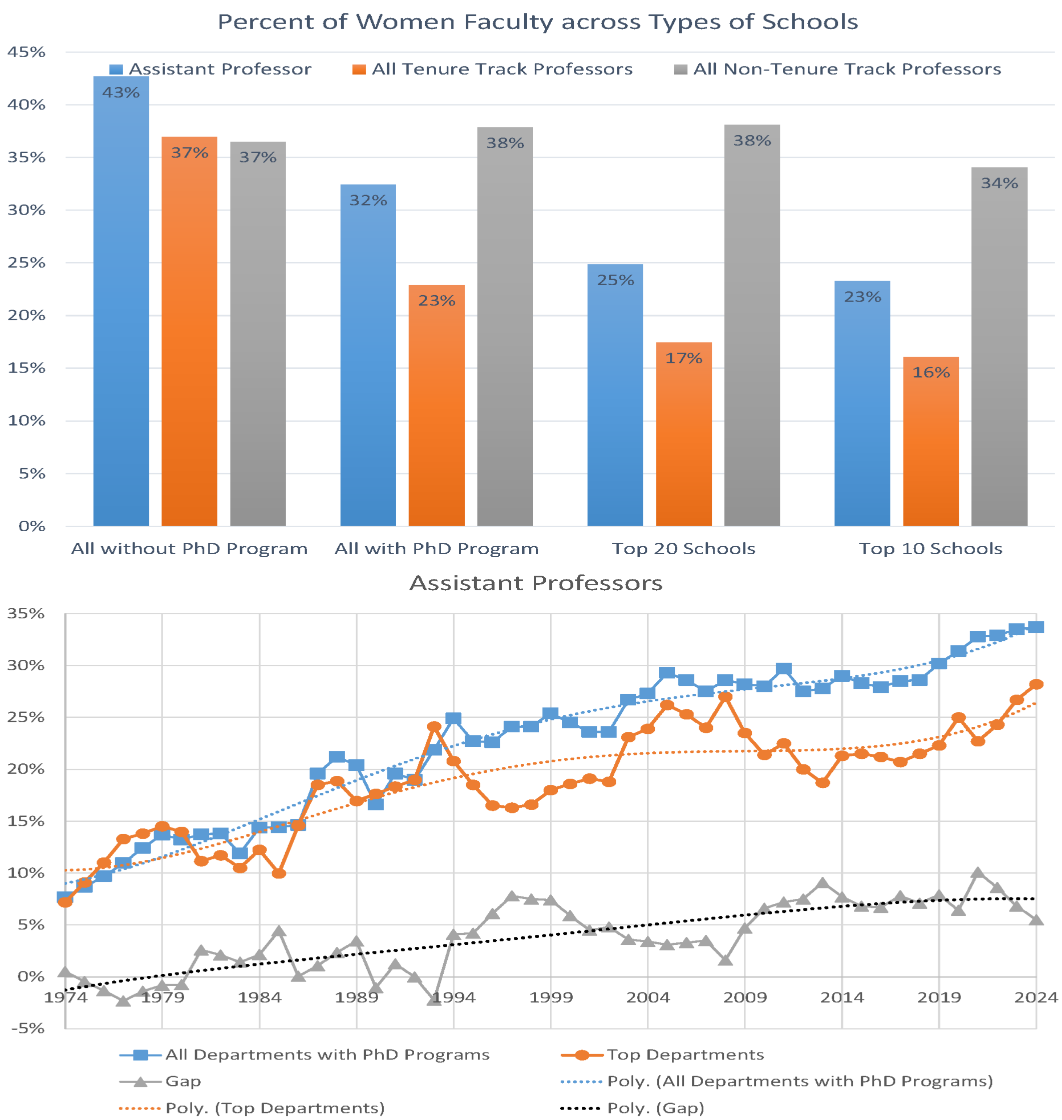


Source: CSWEP Report, 2024.

The table also shows that when we run the same regression on non-tenure track faculty, there is no relation whatsoever with the publication intensity. The negative relation is only visible for research faculty, consistently with the aggregate data displayed in Figure 2.

Focusing again on universities with PhD programs, the bottom panel of Figure 2 shows the striking dynamics of the percentage of women assistant professors between top departments and all departments since 1974. While women representation was low in the late 1970s, it was about the same percentage between top institutions and all institutions. However, over time, the percentage of women has increased, but *less* so for top institutions. The gap

Figure 2: Percentage of Women in Academia across Institutions



Notes: The top panel plots the average percentage of women in academia across types of institutions and types of academic positions. Data are from the 2024 CSWEP Report and averages are over the 2019-2024 sample. The bottom panel reports the fraction of women assistant professors in top institutions vs. all institutions with a PhD program from 1974 to 2024. Data from 1974 to 1993 were extracted from Figures 2 and 3 of the 1994 CSWEP report available at `https://www.aeaweb.org/content/file?id=682`, while data from 1994 onward are from the 2024 CSWEP report. The figure also reports 4^{th}-order polynomial trend lines as well as the difference between the two lines at the bottom.

between top and all institutions (grey line at the bottom of the graph) has been increasing in the last 50 years, which, once again, appears to indicate that research at top institutions

Table 1: Percentage of Women and Institution Research Intensity

	Ranking by Journal Impact Factor			Ranking by Top 5 Econ Journals		
	Assistant Professors	All Tenure Track Faculty	Non-Tenure Track Faculty	Assistant Professors	All Tenure Track Faculty	Non-Tenure Track Faculty
α	34.75	23.79	37.72	33.40	22.23	38.27
(S.E.)	(2.02)	(0.92)	(2.71)	(1.75)	(0.79)	(2.50)
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β	-0.0470	-0.0330	0.0003	-0.1050	-0.0670	0.0029
(S.E.)	(0.0169)	(0.0077)	(0.0223)	(0.0349)	(0.0156)	(0.0493)
(p-value)	(0.006)	(0.000)	(0.988)	(0.003)	(0.000)	(0.9526)
R^2 (%)	6.32	13.53	0.00	9.06	16.79	0.00
N	117	117	113	93	93	91

Notes: This table shows the results of the regression

$$(\text{\% women faculty})_j = \alpha + \beta \times (\text{publication score})_j + \varepsilon_j$$

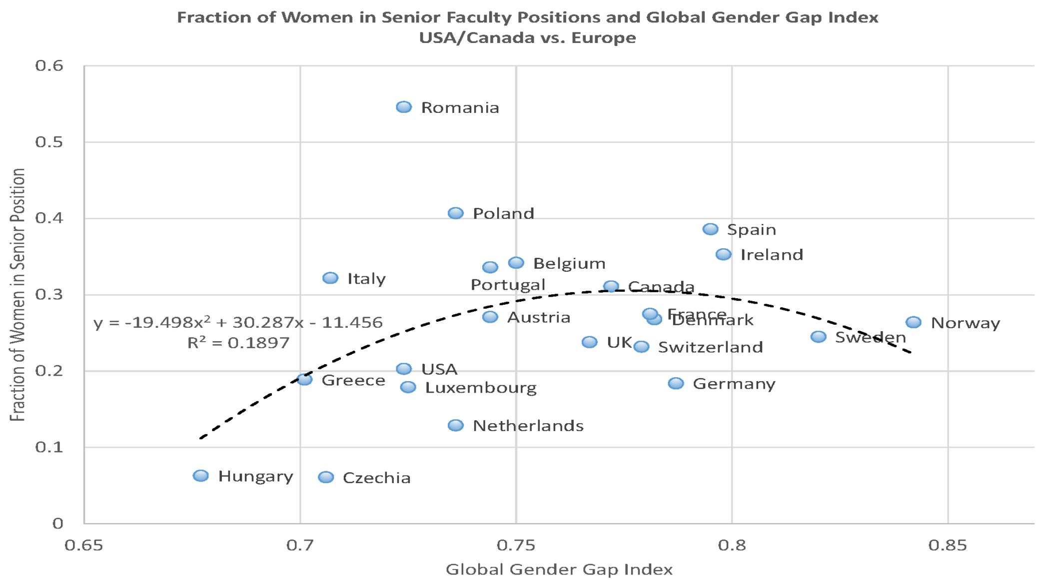
where $(\text{\% women faculty})_j$ is the average percentage of women faculty as assistant professor (columns 1 and 4), as tenure-track faculty (columns 2 and 5), as non-tenure track faculty (columns 3 and 6) in school j; and $(\text{publication score})_j$ is the research score of school j. The latter is obtained from Tilburg's ranking of U.S. institutions in terms of research output. We considered two cases: Columns 1–3 consider the base case of Tilburg rankings, using the weighting of journals by journals' impact factors. Columns 4–6 consider research output based only on publications in the top 5 economics journals (AER, Econometrica, QJE, JPE, Review of Economic Studies). Both the average percentage of female faculty and the publication score are computed over the latest sample available, namely 2016–2020. Data for the % of women faculty are from the disaggregated restricted data underlying the CSWEP report and available at https://www.icpsr.umich.edu/web/ICPSR/studies/37118/. Missing data are deleted from the sample.

is a critical component of the story. [4]

The possibility that explicit or implicit gender discrimination may not be the full story underlying the lack of women's representation in economics is also highlighted by Figure 3. This figure uses data from Auriol et al. (2021) and correlates the fraction of women in senior academic positions in economics and business in the U.S., Canada, and European countries with their respective values of the global gender gap index obtained from the World Economic Forum. A higher global gender gap index indicates *lower* differences between men and women on several dimensions (see https://www3.weforum.org/docs/WEF_GGGR_2020.pdf). As in Auriol et al. (2021), we find a positive relation between the global gender gap index and the percentage of women in academia. However, focusing now on magnitudes, even in Nordic European countries, where the global gender gap index is maximized, the percentage of women in academic positions still hovers only around 25%, which is not too different from

[4] We caveat these results, however, by noticing that from 1974 to 1993, CSWEP defines "top institution" as those above the median according to the National Research Council rankings, while in the 1994 to 2021 data, CSWEP defines "top institutions" as the top 20 schools. Still, the gap is visibly increasing also just in the latter sample.

Figure 3: Women in Academia: USA/Canada versus Europe



Source: Data of women representation are from the main dataset of Auriol, Friebel, Weinberger, and Wilhem, 2021, focusing on U.S., Canada, and European countries. The global gender gap index of each country is from the world economic forum (See Table 1 at `https://www3.weforum.org/docs/WEF_GGGR_2020.pdf`).

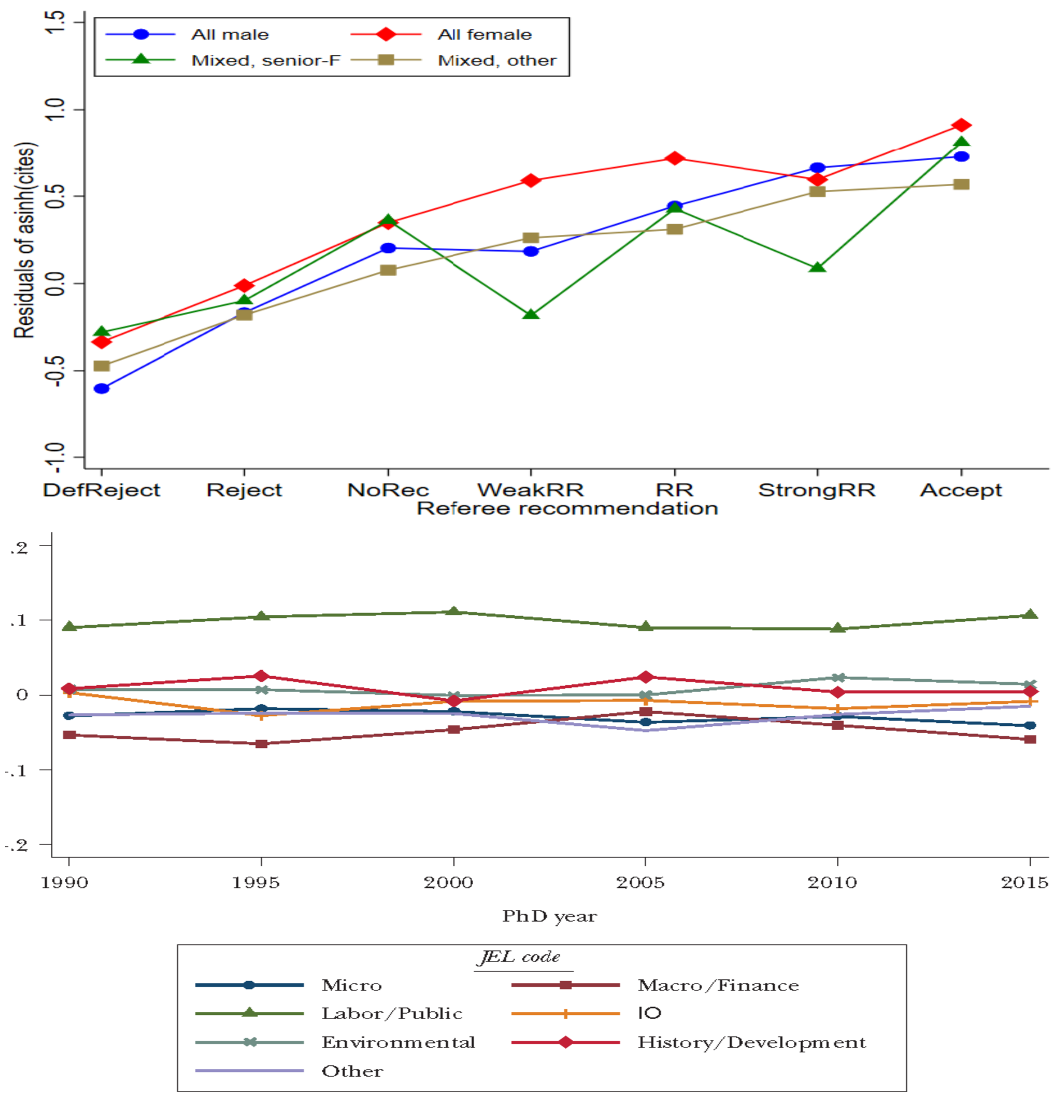
the U.S. (21%) in this dataset.[5]

Finally, Figure 4 illustrates the quality and field interests of women in academia. The top panel reports Figure 4(a) from Card et al. (2020), which shows that "[a]t nearly each referee recommendation, female-authored papers have higher citations than male-authored papers, with a 20 log point average difference. This suggests that papers by all-female authors are held to a higher bar by the referees." (Card et al. (2020), p. 296). The bottom panel reproduces Figure 5 in Lundberg and Stearns, 2019, which reports the difference between the share of women and the share of men in particular fields of economics. The data used are from the annual list of Doctoral Dissertations in Economics, from 1991–2017. As can be seen, women are relatively more frequent in Labor/Public Economics (green line) and men are relatively more frequent in Macro/Finance.[6] The striking pattern, however, is that over the 25-year sample, there is no variation at all in the relative percentages, as if the "system" has converged, consistently with the bottom panel of Figure 1.

[5]This figure is similar to Figure 4 in Auriol et al., 2021, which plots the rank-order of countries with a higher percentage of female faculty against the rank-order of countries in terms of the global gender gap index. Figure 3 makes explicit the actual percentage of women across institutions to highlight that in the majority of countries in the U.S., Canada, and Europe, this percentage is still below 30%.

[6]The figure shows "relative frequency" and not absolute frequency. Even in labor/public finance, the majority of faculty is still male.

Figure 4: Women in Academia: Quality and Fields



Source: The top panel reports Figure 4 (a) of Card, Della Vigna, Funk, and Iriberri, "Are referees and editors in economics gender neutral?" *Quarterly Journal of Economics*, 135, 2020, which plots the average (residualized) citation rate of non-desk rejected papers across types of referee recommendation. The bottom panel reports Figure 5 in Lundberg and Stearns, 2019.

3. Model

We consider an overlapping-generations model in which unit masses of two groups of young researchers, the M-group and F-group, appear at discrete times $t = 1, 2, \ldots$. Each researcher $i \in M \cup F$ is endowed with a *type*, drawn from a finite set Θ and distributed heterogeneously across M and F researchers. While systematic, these distributional differences may well be small. Assumption 1 below imposes a notion of symmetry among types that ensures that the two groups are equally productive ex-ante. 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. Let $p^{\theta,m}$ and $p^{\theta,f}$ denote the fraction of types θ in the M- and F-populations of young researchers, respectively. These fractions are strictly positive for all θ.

The researcher of type θ has probability γ^{θ} of producing quality research. "Quality" research is one that achieves its stated goals—estimating a parameter of interest, establishing a causal effect, documenting a phenomenon experimentally, or proving a theorem. We assume that whether a research paper achieves its goals is observable and can be objectively determined; this may involve, for instance, checking a formal argument regarding a theoretical claim or the application of a statistical procedure, evaluating an experimental procedure for possible biases or ambiguities, or ensuring that the formal results are clearly explained and interpreted, and that the contribution is correctly placed within its literature.

We assume that, while there are (typically small) differences in the distribution of *types* in the M- and F-populations, there are *no differences in the corresponding distribution of quality*. Specifically, we assume that type distributions in the M and F populations are symmetric in the following sense:

Assumption 1. Symmetric Distribution of Quality: for every type θ, there is a corresponding type θ' that has the same quality as θ, and that has the same mass in F (resp. M) as type θ has in M (resp. F). Formally, there exists a function $\sigma : \Theta \to \Theta$ such that

(i) for every type $\theta \in \Theta$, the corresponding type $\theta' = \sigma(\theta)$ satisfies $\gamma^{\theta} = \gamma^{\theta'}$ and $p^{\theta,m} = p^{\theta',f}$; furthermore,

(ii) for every type $\theta \in \Theta$, $\sigma(\sigma(\theta)) = \theta$.[7]

Note that, by (i) and (ii), it is also the case that, for every θ, the corresponding $\theta' = \sigma(\theta)$ satisfies $p^{\theta,f} = p^{\theta',m}$, as $p^{\theta,f} = p^{\sigma(\sigma(\theta)),f} = p^{\sigma(\theta),m}$. That is, there is full symmetry in the distributions of θ's across the M and F population.

Finally, to model heterogeneity in distributional frequencies of types across M and F population, we assume the following:

Assumption 2. Heterogeneity in M- and F-Distribution: for some type $\theta \in \Theta$, $p^{\theta,m} > p^{\theta,f}$; by (i), this implies that $p^{\theta',f} > p^{\theta',m}$ for $\theta' = \sigma(\theta)$.

That is, type θ is more frequent in M population, while the corresponding type $\theta' = \sigma(\theta)$ is more frequent in the F population. We let $p^{g} = (p^{\theta,g})_{\theta\in\Theta}$ for $g = f, m$.

[7] That is, σ is self-inverse, which implies that it is a bijection.

3.1. Objective Refereeing

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

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

$$a_t^{\theta,g} = \gamma^\theta \cdot p^{\theta,g}, \quad g \in \{f,m\}. \tag{1}$$

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

Denote by $\lambda_t^{\theta,g}$ the mass of established researchers of type θ and group g at time t. We normalize the initial mass of all established researchers to one: $\sum_\theta \sum_g \lambda_0^{\theta,g} = 1$.[8] 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. This assumption is also geared towards maximizing the impact of young researchers on the evolution of the system, and thus give the best chance for the system to converge to group balance.[9] The resulting dynamic is then described by the following equation:

$$\lambda_t^{\theta,g} = (1-a_t)\lambda_{t-1}^{\theta,g} + a_t^{\theta,g}, \quad g \in \{f,m\}. \tag{2}$$

We then obtain the following proposition:

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

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

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

We prove all Propositions containing our main results in the Appendix. For other Propositions, we provide a proof sketch in the Appendix, and a full proof in the Online Appendix. We prove all Corollaries in the Online Appendix.

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

[9] 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.

In our benchmark model with objective refereeing, initial conditions have no long-run effects. In addition, the system always converges to equal shares of M and F established researchers, and the limiting type distribution is fully characterized by the probability of producing quality research (γ^{θ}) and the relative frequency of each type in the population of young researchers ($p^{\theta,m}$ and $p^{\theta,f}$). Given the symmetry of the model, this is intuitive.

3.2. Refereeing with Self-Image Bias

Our main model differs from the benchmark in Section 3.1 in that established researchers (referees) not only evaluate young researchers on whether their research is of sufficient quality (as in the previous section), but they also use their personal research styles to guide their subjective judgment 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 on referee r's own type θ^r and the characteristics of researcher i's output, which by assumption coincides with his or her type θ^i.

Consistently with self-image bias, referee r rejects applicants whose type is far from his or her own set of characteristics. For tractability, we make the following stark assumption (we relax it in the Online Appendix):

Assumption 3. Self-Image Bias: referee r evaluates young agent i's research favorably if and only if $\theta^r = \theta^i$.

If agent i's output is favorably evaluated, i is accepted as an established researcher, and will serve as referee for future cohorts of young researchers.

Let $\lambda_t^{\theta} = \lambda_t^{\theta,f} + \lambda_t^{\theta,m}$ be the total mass of established researchers of type θ at time t; also let $\lambda_t = (\lambda_t^{\theta})_{\theta \in \Theta}$. Retaining the notation of Section 3.1, the dynamics for the mass of young researchers of type θ and group g that are accepted in round t is

$$a_t^{\theta,g} = \gamma^{\theta} \cdot \lambda_{t-1}^{\theta} \cdot p^{\theta,g}. \tag{3}$$

Importantly, whether a young researcher is accepted or not depends solely on her type θ, and not also on her group g. As in Equation (2), the total mass of established researchers of type θ and group g is given by

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

where as above $a_t = \sum_{\theta} \sum_{g} a_t^{\theta,g}$. Equations (3) and (4) 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 3.1, or if instead an inefficient outcome, characterized by group imbalance, is reached.

3.3. Type Dynamics

We begin by studying the evolution of the mass of each type in the population. Given our assumption that accepted researchers replace randomly drawn existing ones, the following assumption ensures that the mass of each type remains positive:

Assumption 4: boundedness. For every $\theta \in \Theta$, $\gamma^\theta(p^{\theta,m} + p^{\theta,f}) \leq 1$.

The following proposition—our first main result—characterizes the types that survive (i.e. have positive mass) in the limit as $t \to \infty$. All other types vanish over time.

Proposition 2 *Let*

$$\Theta^{\max} = \left\{ \theta \in \Theta \text{ such that } \lambda_0^\theta > 0 \text{ and } \theta \in \arg\max_{\theta' \in \Theta} \gamma^{\theta'}(p^{\theta',m} + p^{\theta',f}) \right\} \tag{5}$$

then:

(i) only $\theta \in \Theta^{\max}$ survive in the limit as $t \to \infty$;

(ii) $\Theta^{\max}$ preserves symmetry across types: if $\theta \in \Theta^{\max}$ and $\lambda_0^{\sigma(\theta)} > 0$, then $\sigma(\theta) \in \Theta^{\max}$.

The proposition shows that the only types θ that survive in the limit are those that had positive mass at time 0, $\lambda_0^\theta > 0$, and that maximize the product $\gamma^\theta(p^{\theta,m} + p^{\theta,f})$. Intuitively, such types θ are both objectively good (high γ^θ) and frequent in the young population (high $(p^{\theta,m} + p^{\theta,f})$). A type θ that does not have a high frequency in the young population, or it is consistently unproductive, will unlikely be part of the refereeing population. Thus, self-image bias will act against it, as, in the limit, no referee will view his/her research favorably.

Part (ii) states that the assumed symmetry of types in the M and F populations (Assumption 1) is preserved in the limit, unless specific types are not represented in the initial population λ_0. This is the case because, for any type θ, $\gamma^\theta(p^{\theta,m}+p^{\theta,f}) = \gamma^{\sigma(\theta)}(p^{\sigma(\theta),f}+p^{\sigma(\theta),m})$.

Our second main result characterizes the limiting distribution of types:

Proposition 3 *Let $\overline{\lambda}^{\theta,g}$ denote the limit of type θ's mass $\lambda_t^{\theta,g}$, for $g \in \{f, m\}$, as $t \to \infty$, and let $\overline{\lambda}^\theta = \overline{\lambda}^{\theta,m} + \overline{\lambda}^{\theta,f}$. Then, for all $\theta \in \Theta^{\max}$:*

(i) The limiting mass of θ is

$$\bar{\lambda}^{\theta} = \frac{\lambda_0^{\theta}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'}} \tag{6}$$

(ii) The limiting mass of θ in its f and m groups are:

$$\bar{\lambda}^{\theta,f} = \frac{\lambda_0^{\theta}\gamma^{\theta}p^{\theta,f}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'}\gamma^{\theta'}(p^{\theta',f}+p^{\theta',m})}; \quad \bar{\lambda}^{\theta,m} = \frac{\lambda_0^{\theta}\gamma^{\theta}p^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'}\gamma^{\theta'}(p^{\theta',m}+p^{\theta',f})} \tag{7}$$

Point (i) shows that $\theta \in \Theta^{\max}$ is more frequent in the limit than another type $\theta' \in \Theta^{\max}$ if it was relatively more frequent in the initial distribution at time 0. Intuitively, types that are more frequent in the initial distribution have more referees who "like" those same types, and thus they are more likely to self-replicate and their larger mass will persist in the limit.

In sharp contrast with Proposition 1 under objective refereeing, point (ii) of Proposition 3 shows that the final mass of types in M and F populations does depend on the initial distribution λ_0. In particular, surviving types $\theta \in \Theta^{\max}$ have a higher relative mass $\bar{\lambda}^{\theta,g}$ in group $g = m, f$ if they are more frequent initially (higher λ_0^{θ}), they have higher productivity γ^{θ}, and they are more frequent in the distribution $p^{\theta,g}$.

The ex-ante symmetry and heterogeneity of the distributions $p^{\cdot,m}$ and $p^{\cdot,f}$ in Assumptions 1 and 2 then lead to our main result for this section, namely, that for symmetric types θ and $\theta' = \sigma(\theta)$, the relative initial distribution of types determines the final relative mass of surviving types across the two groups.

The following definition is convenient.

Definition 1 *Types θ and θ' are* distinct symmetric elements of $\Theta^{\max}$ *if (i) $\theta, \theta' = \sigma(\theta) \in \Theta^{\max}$ and (ii) $p^{\theta,m} \neq p^{\theta',m}$ (or, equivalently, $p^{\theta,f} \neq p^{\theta',f}$).*

We then have the following:

Corollary 1 *Let θ, θ' be distinct symmetric elements of $\Theta^{\max}$ with $\lambda_0^{\theta} > \lambda_0^{\theta'}$ and $p^{\theta,m} = p^{\theta',f} > p^{\theta,f} = p^{\theta',m}$. Then*

$$\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta',m} > \bar{\lambda}^{\theta,f} + \bar{\lambda}^{\theta',f}. \tag{8}$$

This corollary is a key result of the paper. The initial condition $\lambda_0^{\theta} > \lambda_0^{\theta'}$ is group independent. Furthermore, from Assumption 1, type symmetry implies that $\gamma^{\theta} = \gamma^{\theta'}$ (both types are equally productive) and $p^{\theta,m} + p^{\theta',m} = p^{\theta,f} + p^{\theta',f}$ (the flow of young researchers whose type is either θ or θ' is the same in the M- and F-group). Yet, the fact that type θ, which happens to be more frequent in the M-population ($p^{\theta,m} > p^{\theta',m}$), is also more frequent in the overall population of referees at time 0 ($\lambda_0^\theta > \lambda_0^{\theta'}$) implies that, in the limit, the mass of M-researchers of type θ or θ' is greater than the mass of F-researchers having the same types. Formally, this follows from part (ii) of Proposition 3. Intuitively, M-researchers are more likely to be of type θ than type θ', and they are also more likely to be matched with a referee of type θ than of type θ', regardless of the referee's group; this "positive assortative matching" of referees and M-researchers yields a high probability of being accepted. By way of contrast, F-researchers are more likely to be of type θ' than θ, but they are still more likely to be matched with type-θ referees; thus, "negative assortative matching" leads to a lower probability of being accepted.This perpetuates the bias forward.

3.4. Model Predictions

In this section, we discuss the model's qualitative predictions under the assumption that the initial distribution of referees, λ_0, is skewed towards the types of the M-population. Specifically, we assume throughout $\lambda_0 = p^m$, a natural assumption to investigate the impact of a population of referees initially dominated by the M-group.

3.4.1. Group Imbalance in the Limit

In this subsection we expand on the group imbalance discussed in Corollary 1 above for only two symmetric types θ and θ'. In particular, let

$$\bar{\Lambda}^g = \sum_{\theta \in \Theta^{\max}} \bar{\lambda}^{\theta,g}, \quad \text{for} \quad g = m, f$$

denote the total limiting mass of researcher in group g. Then, the following result follows:

Proposition 4 *Let $\lambda_0 = p^m$.*

(a) If $\Theta^{\max}$ contains distinct symmetric elements, then

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f > 0.5. \tag{9}$$

(b) If $\Theta^{\max}$ contains no distinct symmetric elements, then $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$.

Point (a) of this proposition shows that even with ex-ante symmetric types between M and F (Assumption 1), self-image bias leads to imbalance in the limit when the initial population of referees is initially dominated by the M-population. This is consistent with the bottom panel of Figure 1, which shows that the percentage of women in top 20 economics departments has been constant at around 25% in nearly 30 years. We remark that the condition of part (a) is satisfied if a type $\theta \in \Theta^{\max}$ has $p^{\theta,m} \neq p^{\theta,f}$, and $\theta' = \sigma(\theta)$ has $\lambda_0^{\theta'} > 0$. The calibration section provides parametric conditions for this to be satisfied.

3.4.2. Higher "Bar" for F-researchers

If the initial population of referees' types is skewed towards the M-population, 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 model all researchers are ex-ante identical in terms of productivity. However, because of self-image bias, the acceptance rate of young researchers from M-population is higher than the one from the F-population.

Proposition 5 *Let $\lambda_0 = p^m$. Then for every $t > 0$, the acceptance rate of M-researchers of "quality" $\overline{\gamma}$ is higher than the one of F-researchers of the same quality:*

$$\sum_{\theta:\gamma^\theta=\overline{\gamma}} a_t^{\theta,m} \geq \sum_{\theta:\gamma^\theta=\overline{\gamma}} a_t^{\theta,f}$$

and the inequality is strict if there is a type $\theta \in \Theta$ for which $p^{\theta,m} > p^{\sigma(\theta),m}$.

This result is in line with the evidence in the top panel of Figure 4 from Card et al. (2020) that, conditional on quality (proxied by citations post-publication) women-authored papers tend to be accepted less frequently than men's.[10]

The above result suggests that, on average, the objective quality of accepted F-researchers should be higher than that of M-researchers. We have conducted extensive simulations in the parametric model of Section 4, and indeed this appears to be the case. We can actually prove this formally in the setting of the present section when Θ has only four types, namely, a "good" type θ_1 (high γ^{θ_1}), a "bad" type θ_0 (low γ^{θ_0}), and two intermediate and symmetric types θ, θ' (intermediate and identical $\gamma^\theta = \gamma^{\theta'}$, with $p^{\theta,m} = p^{\theta',f} > p^{\theta',m} = p^{\theta,f}$).

Proposition 6 *Let $\Theta = \{\theta_0, \theta, \theta', \theta_1\}$, with $\gamma^{\theta_0} < \gamma^{\theta_1}$, $\gamma^\theta = \gamma^{\theta'} \leq \frac{\gamma^{\theta_0}+\gamma^{\theta_1}}{2}$, $p^{\theta,m} = p^{\theta',f} > p^{\theta',m} = p^{\theta,f}$, and $p^{\theta_0,m} = p^{\theta_0,f} = p^{\theta_1,m} = p^{\theta_1,f}$. Finally, let $\lambda_0 = p^m$. Then:*

(i) The average quality of accepted F-researchers is higher than that of accepted M-researchers:

$$E[\gamma|f, accepted] = \sum_{\hat{\theta}} \gamma^{\hat{\theta}}\ w_t^{\hat{\theta},f} > \sum_{\hat{\theta}} \gamma^{\hat{\theta}}\ w_t^{\hat{\theta},m} = E[\gamma|m, accepted] \tag{10}$$

where

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

[10] Card et al. (2020) also show that, unconditionally, men- and women-authored papers are equally likely to be accepted. The model in this section does not generate this finding: summing over all types θ in the displayed equation of Proposition 5, one readily sees that young M researchers are more likely to be accepted on average. The model with endogenous choice in Section A1.1. yields more uniform unconditional acceptance across genders, and fewer female acceptance overall due to self-selection.

(ii) As $t \to \infty$, the average quality of both F and M converges to either $\gamma^{\theta} = \gamma^{\theta'}$ (if $\theta, \theta' \in \Theta^{\max}$) or γ^{θ_1} (otherwise).

3.4.3. Talent Loss and Clustering

One further implication of our model is that with self-image bias types that are more common in the F-group are under-represented in the limit.

Corollary 2 *Assume $\lambda_0 = p^m$. Let θ, θ' be distinct symmetric elements of $\Theta^{\max}$ with $\lambda_0^{\theta} > \lambda_0^{\theta'}$ and $p^{\theta,m} = p^{\theta',f} > p^{\theta,f} = p^{\theta',m}$. Then in the limit*

$$\frac{\bar{\lambda}^{\theta,m}}{\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta',m}} > 0.5 = \frac{\bar{\lambda}^{\theta',f}}{\bar{\lambda}^{\theta,f} + \bar{\lambda}^{\theta',f}} = \frac{\bar{\lambda}^{\theta,f}}{\bar{\lambda}^{\theta,f} + \bar{\lambda}^{\theta',f}} > \frac{\bar{\lambda}^{\theta',m}}{\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta',m}} \tag{11}$$

In the limiting distribution, the majority of established M researchers are of type θ. However, the established F-population has the *same* fraction of type θ as type θ'. This result is in stark contrast with θ' being the prevalent type in each cohort of young F-researchers (assumption $p^{\theta',f} > p^{\theta,f}$). The result is moreover independent of the actual values of $p^{\theta,g}$'s. That is, we could have a flow of young researchers with e.g. $p^{\theta',f} = 0.9 > 0.1 = p^{\theta,f}$ (and symmetrically, $p^{\theta,m} = 0.9 > 0.1 = p^{\theta',m}$), and yet both types θ and θ' would be equally represented in the F-population in the limit. The selection mechanism makes the type most prevalent among M-researchers, θ, be a frequent type in the established F-researchers (50% of the time). This implies that F-group research types are underrepresented in the limit.

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

Corollary 3 *Assume $\lambda_0 = p^m$. Let θ, θ' be distinct symmetric elements of $\Theta^{\max}$ with $\lambda_0^{\theta} > \lambda_0^{\theta'}$ and $p^{\theta,m} = p^{\theta',f} > p^{\theta,f} = p^{\theta',m}$. Then in the limit, M-researchers are relatively more frequent as type θ and F-researchers are relatively more frequent as type θ':*

$$\frac{\bar{\lambda}^{\theta,m}}{\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta,f}} = \frac{\bar{\lambda}^{\theta',f}}{\bar{\lambda}^{\theta',m} + \bar{\lambda}^{\theta',f}} > 0.5. \tag{12}$$

If, as seems plausible, some types may be better suited for some specific research fields than others, this result implies that the two groups will be differently represented across fields. This is qualitatively consistent with the evidence in the bottom panel of Figure 4 from Lundberg and Stearns (2019) (see also Chari and Goldsmith-Pinkham (2018)) documenting large gender differences across economics topics, although the result in Corollary 3 is too extreme, as women's frequency never breaks the 50% threshold in economics (although it does in other areas, such as psychology). Still, the result is consistent with this clustering across fields being true in the limit, which is consistent with the lack of variation or trends shown in the bottom panel of Figure 4.

3.4.4. Publication Success and F-Underrepresentation

Our model is also consistent with the evidence that more research-intensive universities have lower female representation, as shown in the top panel of Figure 2 and Table 1. In particular, suppose that $\Theta^{\max}$ contains only two distinct symmetric types θ and θ' with $\lambda_0^{\theta} > \lambda_0^{\theta'}$. We analyze the resulting limit economy; see Section 4 for numerical results with calibrated parameter values and a finite time horizon.

Consider an institution with an arbitrary fraction $y \in [0,1]$ of θ-researchers and a complementary fraction $(1-y)$ of θ'-researchers; since no other types survive in the limit, these are the only researcher types that an institution can employ.[11] Under self-image bias in refereeing, the probability that a researcher of type θ (resp. θ') publishes successfully is $\overline{\gamma} \cdot \bar{\lambda}^{\theta}$ (resp. $\overline{\gamma} \cdot \bar{\lambda}^{\theta'}$), where $\overline{\gamma} = \gamma^{\theta} = \gamma^{\theta'}$. Hence, the *average publication frequency* of the institution is

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

Since $\overline{\lambda}^{\theta} > \overline{\lambda}^{\theta'}$ by Proposition 3, $P(y)$ increases in y.

We assume that the type-θ and type-θ' researchers at the institution under consideration belong to the F and M groups in proportions analogous to those in the population: that is, a fraction $y\bar{\lambda}^{\theta,f}$ are of type θ and group F, a fraction $(1-y)\bar{\lambda}^{\theta',f}$ are of type θ' and group F, etc. Then, the fraction of F-researchers in an institution parameterized by y is given by:

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

Corollary 2 then implies:

Corollary 4 *Assume $\lambda_0 = p^m$ and let $\Theta^{\max}$ contain only two distinct symmetric types θ and θ', with $\lambda_0^{\theta} > \lambda_0^{\theta'}$. Then $F(y)$ is decreasing in y. That is: in the limit, institutions with higher exogenous fraction y of θ-researcher, and a complementary fraction $(1-y)$ of θ' researchers, have higher publication frequency and lower percentage of F-researchers.*

Intuitively, the result follows from the fact that the limit mass of θ-researchers in the population is higher than that of θ'-researchers: $\overline{\lambda}^{\theta} > 0.5 > \overline{\lambda}^{\theta'}$. Self-image bias implies that types θ have a higher chance of publication success, because the probability they are matched

[11] The total mass of researchers in the institution under consideration is irrelevant to the analysis, and can be considered small. In Section 4, we consider a different parameterization in which the entire population is divided into a given, fixed number of institutions.

with referees of their own type is higher. Consequently, on average, institutions with a higher fraction y of type-θ scholars are more likely on average to achieve successful publications. However, types θ are also more likely to come from the M group: this generates a negative relation between research success and F-representation. In the limit, F-researchers are least represented in "top institutions," where "top" is defined as in terms of publication record.

As a final comment, the above result pertains to researchers in the M and F populations, as referees have self-image bias and thus affect the publication process. The same scrutiny is unlikely to apply to "teaching faculty," as self-image bias in refereeing is unlikely to play any role. In this case, if types θ also relate to candidates' teaching abilities (e.g., each type θ maps into a probability t^θ of being an objectively good teacher), then the same argument as in Proposition 1 for objective refereeing yields equal representation of M and F candidates in teaching positions. Moreover, there is no correlation with the publication frequency of each institution, as shown in Figure 2 and Table 1.

3.4.5. Perceived Trade-off Between Quality and Diversity

Self-image bias also explains why the current population of referees may incorrectly perceive a trade-off between diversity and "quality" (see, e.g., First Round Review, 2022) or "merit" (see, e.g., Crosby, Iyer, Clayton, and Downing (2003)). The intuition is simple: by definition, self-image bias implies that the closer another researcher is to one's own type, the higher their subjectively perceived quality.

In particular, our preceding results imply that the population of established scholars consists mostly of types that are more prevalent in the M-group. Therefore, a randomly drawn established researcher will subjectively perceive other M-researchers to be *ex-ante* of higher average quality than F-researchers, because it is more likely that a randomly drawn young M-researcher's type will be close to their own. Hence the mistaken impression that, in order to increase F-representation, one has to "accept" a loss of quality.

This conclusion is in stark contrast with the fact that, in our model, the average *objective* quality of each cohort M- and F-researchers, which is given by the average probability of producing quality research, is exactly the same, by construction. Furthermore, as shown in Proposition 6, when the set Θ has a specific structure, the average quality of accepted F researchers is actually *higher* than that of accepted M researchers; the same is true for the calibrated model of Section 4, with multiple characteristics (see Fig. 9). Thus, in fact, increasing diversity can potentially *increase* average objective quality.

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

$$Q(M|\lambda_t) = \sum_{\theta\in\Theta} p^{\theta,m}\gamma^\theta\lambda_t^\theta \quad \text{and} \quad Q(F|\lambda_t) = \sum_{\theta\in\Theta} p^{\theta,f}\gamma^\theta\lambda_t^\theta$$

The key observation is now that $p^{\theta,g}\gamma^\theta\lambda_t^\theta = a_t^{\theta,g}$ for $g = m, f$. Therefore, summing over all possible values $\overline{\gamma}$ of the function $\theta \mapsto \gamma^\theta$ in Proposition 5 yields

Corollary 5 *Assume $\lambda_0 = p^m$ and that there is $\theta \in \Theta$ with $p^{\theta,m} > p^{\sigma(\theta),m}$. Then, the population of referees λ_t (mis)perceives the quality of a random F-researcher to be lower than the quality of a random M-researcher. That is, $Q(F|\lambda_t) < Q(M|\lambda_t)$.*

4. Calibration

The previous section provided numerous results that hold for arbitrary sets of types Θ. To calibrate the model, we need to be more specific about the definition of a type. We adopt a simple symmetric environment in which each type θ corresponds to a vector of N characteristics which can only take two values, 0 and 1: that is, $\Theta \equiv \{0,1\}^N$. For each agent of type $\theta \in \Theta$, θ_n denotes the value of the n-th characteristic. The number N of characteristics is even, characteristics are mutually independently distributed, and their distributions depend on a single parameter $\phi > 0.5$. Our main assumption, illustrated in Figure 5, is that characteristics are distributed symmetrically in the M and F population, in the sense that for $n = 1, \ldots, \frac{N}{2}$, $Pr(\theta_n = 1) = \phi$ for M-researchers and $Pr(\theta_n = 1) = (1-\phi)$ for F-researchers, and the opposite for $n = \frac{N}{2}+1, \ldots, N$. For every $\theta \in \Theta$, let $p^{\theta,f}$ (resp. $p^{\theta,m}$) denote the fraction of types in the F (resp. M) population of young researchers. Also let $p^g = (p^{\theta,g})_{\theta\in\Theta}$ for $g = f, m$. To sum up,

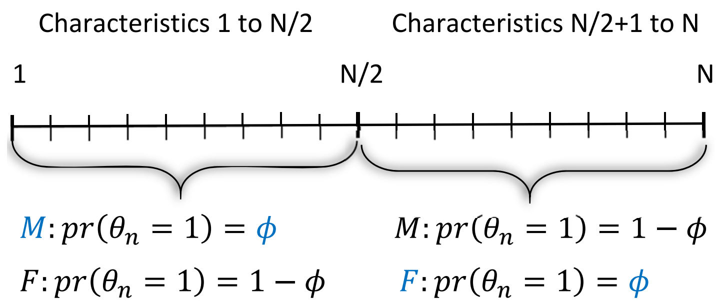
$$p^{\theta,m} = \prod_{n=1}^{N/2} \phi^{\theta_n}(1-\phi)^{1-\theta_n} \cdot \prod_{n=N/2+1}^{N} (1-\phi)^{\theta_n}\phi^{1-\theta_n}, \quad p^{\theta,f} = \prod_{n=1}^{N/2}(1-\phi)^{\theta_n}\phi^{1-\theta_n} \cdot \prod_{n=N/2+1}^{N} \phi^{\theta_n}(1-\phi)^{1-\theta_n}. \tag{13}$$

Furthermore, we assume that the probability of producing quality research is given by

$$\gamma^\theta = \gamma_0 \cdot \rho^{\frac{1}{N}\sum_{n=1}^N \theta_n},$$

where $\gamma_0 \in (0,1)$ and $\rho \in [1, \frac{1}{\gamma_0}]$. The two key features of this specification are that, first, types θ with more characteristics (i.e., more 1's) are more likely to produce quality

Figure 5: Symmetric Distribution of Research Characteristics



research; and second, every research characteristic θ_n has the same impact on our measure of objective quality. The parameter γ_0 reflects the minimum probability of successful output—the productivity of the least "capable" type $(0, \ldots, 0)$. The parameter ρ instead reflects the importance of research characteristics. For instance, $\rho = 1$ means that characteristics are irrelevant—all types are equally likely to produce quality research. A higher value of ρ implies that more 1's increase the probability of quality research.

With these assumptions, the function $\sigma : \Theta \to \Theta$ defined by $\sigma(\theta) = (\theta_{N/2+1}, \ldots, \theta_N, \theta_1, \ldots, \theta_{N/2})$ satisfies all the assumptions of Section 3. Thus, we have:

Corollary 6 *Define the threshold*

$$\bar{\rho}(\phi, N) = \frac{1}{4}\left(\left(\frac{1-\phi}{\phi}\right)^{N/2} + \left(\frac{\phi}{1-\phi}\right)^{N/2}\right)^2. \tag{14}$$

(a) If $\rho < \bar{\rho}(\phi, N)$, then only two types survive, $\Theta^{\max} = \{\theta^m, \theta^f\}$, where

$$\theta^m = (1, ..., 1, 0, ...0) \text{ and } \theta^f = (0, ...0, 1, ..., 1) \tag{15}$$

In addition, if at time 0, all referees are in the M-group with $\lambda_0 = p^m$, then the total limit mass of M and F researchers are

$$\bar{\Lambda}^m = 1 - \bar{\Lambda}^f = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}} > 0.5. \tag{16}$$

(b) If $\rho > \bar{\rho}(\phi, N)$ then, regardless of the distribution of time-0 referees, $\Theta^{\max} = \theta^* = (1, ..., 1)$. In addition, the limiting mass of M and F researcher are $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$.

(c) For any set of parameters with $\phi > 0.5$, $\gamma_0 \in (0, 1)$ and $\rho \in [1, 1/\gamma_0)$, there is N large enough such that $\rho < \bar{\rho}(\phi, N)$ and thus point (a) holds.

(d) Moreover, for any set of parameters with $\phi > 0.5, \gamma_0 \in (0,1)$ and $\rho \in [1, 1/\gamma_0)$, as $N \to \infty$, $\bar{\Lambda}^m \to 1$ and $\bar{\Lambda}^f \to 0$.

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

$$d = \frac{\mathrm{E}[\theta_n^i | i \in M] - \mathrm{E}[\theta_n^i | i \in F]}{\sigma_{\text{pooled}}(\theta_n^i)} = \frac{2\phi - 1}{\sqrt{\phi(1-\phi)}}. \tag{17}$$

For $n = \frac{N}{2} + 1, \ldots, N$, the d statistic is the negative of the above expression. Cohen (2013) suggests that values of d around 0.2 should be considered "small," values around 0.5 "medium," and values around or above 0.8 "large." However, points (c) and (d) of Corollary 6 shows that, if the number N of characteristics is sufficiently large, even small across-group differences ($\phi - 0.5$ small) still yield our main conclusions.

Given the parametric assumptions in this section, we now turn to calibrate the model. The first issue is the number of characteristics that lead to quality research and are taken into account by referees when they evaluate a candidate. We suggest that the number of characteristics is actually large. The following is but a partial list: (i) Economic motivation; (ii) "Nose" for good questions; (iii) Institutional knowledge; (iv) Ability to find new data sources; (v) Solid identification strategy; (vi) Sophisticated empirical analysis; (vii) Clever experimental design; (viii) Skilful theoretical modelling; (ix) Ability to highlight insights, strategic effects, etc. (x) Mathematical sophistication, proof techniques, etc. (xi) Ability to position within the literature; (xii) Ability to highlight policy implications; (xiii) Presentation skills; (xiv) Ability to address questions from audience; (xv) Honesty;[12] and so on. Likely, there are many others. Perhaps some of these research traits are more important than others, but as a first pass, it is indeed plausible that the positive or negative result of a review depends on a combination of research characteristics, and not just a small number. In light of these considerations, and to be conservative, we assume that $N = 10$.

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

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

This value is considered "small" and in line with the estimated group differences of the various traits discussed in the introduction. With $\phi = 0.5742$, between-group heterogeneity

[12]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).

in each characteristic is far smaller than within-group heterogeneity.[13]

Third, we need to calibrate the parameters γ_0 and ρ in the probability of producing quality research, γ^θ. We proceed as follows. First, we assume that the best researcher, $\theta^* = (1, 1, \ldots, 1)$, has 100% probability of producing quality research, i.e., $\gamma^{\theta^*} = 1$. Second, we calibrate γ_0 to match the rate at which economics PhD students succeed in getting an academic job. We compute the latter from the NSF Survey of Doctoral Recipients. We take the ratio of economics PhD recipients who are employed in 4-year educational institutions over the total of economics PhD recipients, both inside and outside the U.S.[14] That ratio is 0.462. Choosing $\gamma_0 = 0.2$ yields an objective success rate $\sum_\theta \gamma^\theta (p^{\theta,f} + p^{\theta,m})/2 = 0.462$. Interestingly, the implied $\rho = \gamma^{\theta^*}/\gamma_0 = 5$ entails that researcher θ^* is objectively five times as productive as researcher $(0, \ldots, 0)$, which is roughly in line with the evidence on research productivity reported in Conley and Önder (2014).[15]

Finally, we assume that initially F-researchers represent 10% of the total mass, which is roughly consistent with the percentage of women faculty in 1974, and is consistent with the distribution of annual inflows of young researchers, i.e., $\lambda_0 = 0.9p^m + 0.1p^f$.

4.1. Calibration Results

The calibration results are shown in Figures 6 through 8, to which we now turn.

4.1.1. F-Under-representation in the limit

Figure 6 shows that the system converges to a large imbalance between M- and F-researchers, with F-researchers representing around 20% of the population.[16] This large imbalance obtains despite the fact that the distribution of characteristics is very similar across M and F types. The hump shape visible in the figure is due to our assumption that λ_0 represents all the characteristics in proportion to the distributions $p^{\theta,m}$ and $p^{\theta,f}$. Since, in our calibration,

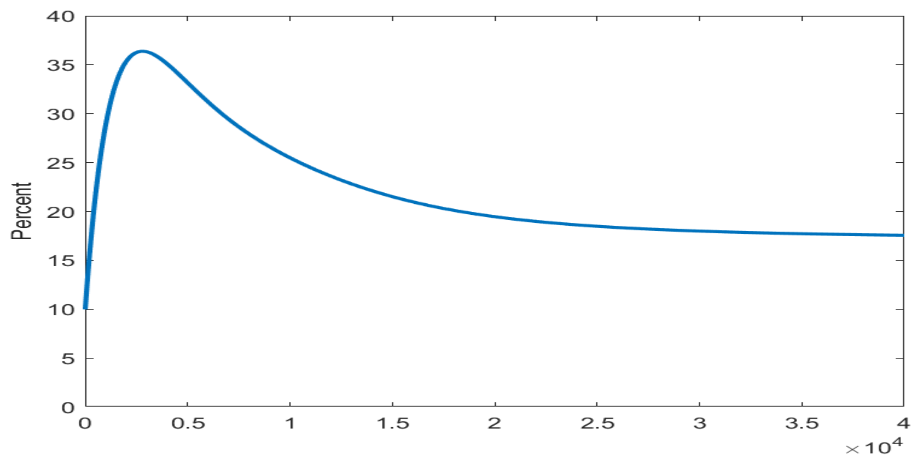
[13] We focus on effect size for a single characteristic as its magnitude has been widely documented in the psychology literature (see introduction). Unfortunately, we were unable to identify experimental studies measuring multidimensional effect sizes between genders for us to use in our calibration.

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

[15] 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 the mass of reviewers to 1, but we can choose another mass K to match the failure rate from the data. See footnote 8.

[16] By comparison, setting $\phi = 0.5742$ and $N = 10$ in Eq. (16), the limiting fraction of M researchers is $\bar{\Lambda}^m \approx 91\%$. The difference is due to the fact that Eq. (16) was derived assuming that $\lambda_0 = p^m$.

Figure 6: Percent of F-Researchers in Calibrated Model



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

the differences between p^m and p^f are small, initially, there is a sufficient mass of referees with characteristics that are common in the F-population to yield a high acceptance rate of F-researchers and thus an increase in their mass. However, as characteristics common in F-population start being weeded out, eventually the acceptance rate of F-researchers drops, and so does the overall mass of F-researchers. While such hump is not visible in the data, the limit fraction of around 20% of F-researchers is rather close to the 25% women faculty as assistant professor observed in the bottom panel of Figure 1.

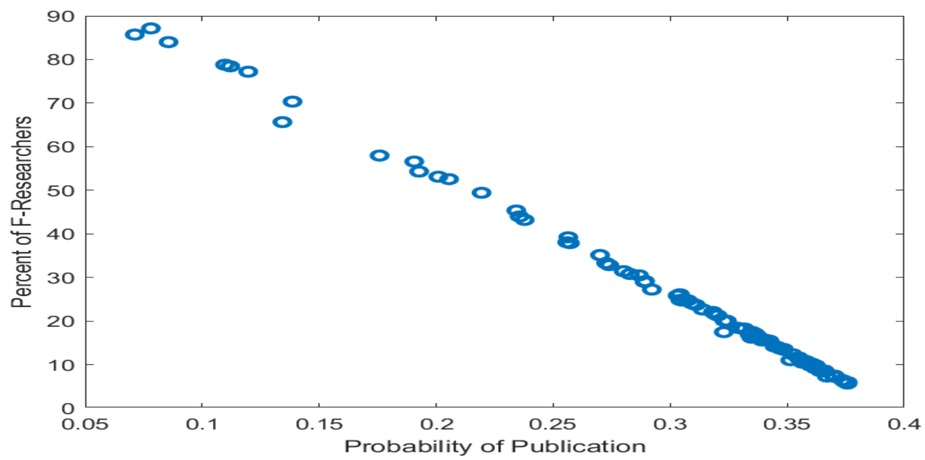
Given that our mechanism is only based on self-image bias – a bias of referees that is gender-neutral and is likely to be common across countries – our results also explain why low representation of women in economics occurs nearly everywhere, including Nordic European countries, which have a far more balanced stance towards women compared to the United States (see Figure 3).

4.1.2. Higher under-representation in top institutions

Next, we build on Section 3.4.4 and examine the publication success of researchers in institutions that differ in their type composition. Specifically, consider J institutions, and assume that each institution $j = 1, \ldots, J$ employs an exogenously specified fraction x_j^θ of all type-θ researchers, with $\sum_{j=1}^J x_j^\theta = 1$. The mass of θ researchers in institution j at time t is thus $x_j^\theta \lambda_t^\theta$. We make no assumptions about the distribution of *groups* in institutions.

Since, at each time t, institution j employs $x_j^\theta \lambda_t^\theta$ researchers of type θ, each such researcher

Figure 7: The Endogenous Negative Relation between Institutions' Publishing Intensity and the Percentage of F-researchers



This figure plots a scatterplot of 100 simulated institutions' publishing probabilities (x-axis) versus F-researcher representation in the same institution (y-axis). Initially $\lambda_0 = 0.9p^m + 0.1p^f$. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$.

publishes with probability $\gamma^\theta \lambda_t^\theta$, and the total mass of researchers employed by institution j is $\sum_{\theta\in\Theta} x_j^\theta \lambda_t^\theta$, the weighted-average probability of publication of (established) researchers in institution j is

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

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

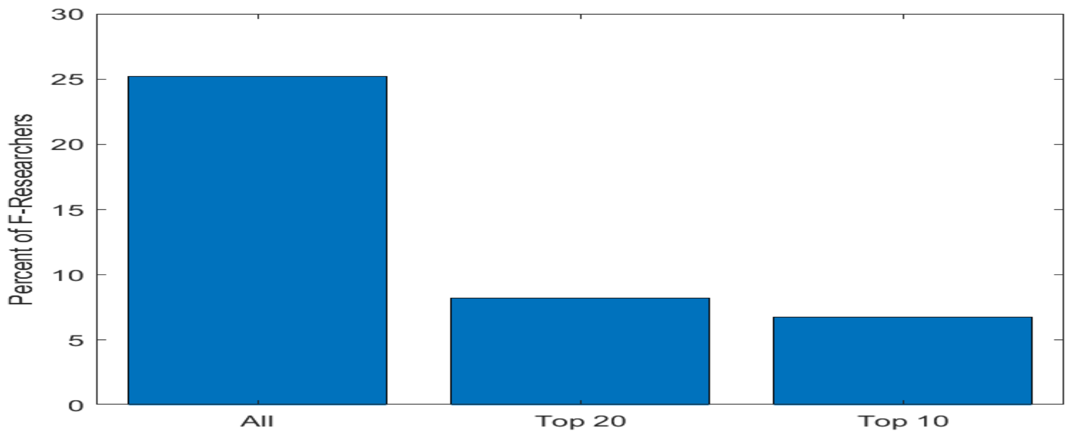
$$F_{j,t} = \frac{\sum_{\theta\in\Theta} \lambda_t^{\theta,f} x_j^\theta}{\sum_{\theta\in\Theta} \lambda_t^\theta x_j^\theta}$$

Figure 7 shows the scatterplot of $P_{j,t}$ and $F_{j,t}$ from the model simulation for a random draw of x_j's for $J = 100$ institutions, for large t (i.e., "in the limit"). As the plot shows, there is a negative relation between an institution publishing intensity (x-axis) and its F-representation ($y-$axis). The negative slope in Figure 7 is consistent with the theoretical result in Corollary 4 and with the empirical findings in Table 1, showing that indeed, the fraction of women in economics departments is negatively related to the publication intensity of the same departments.

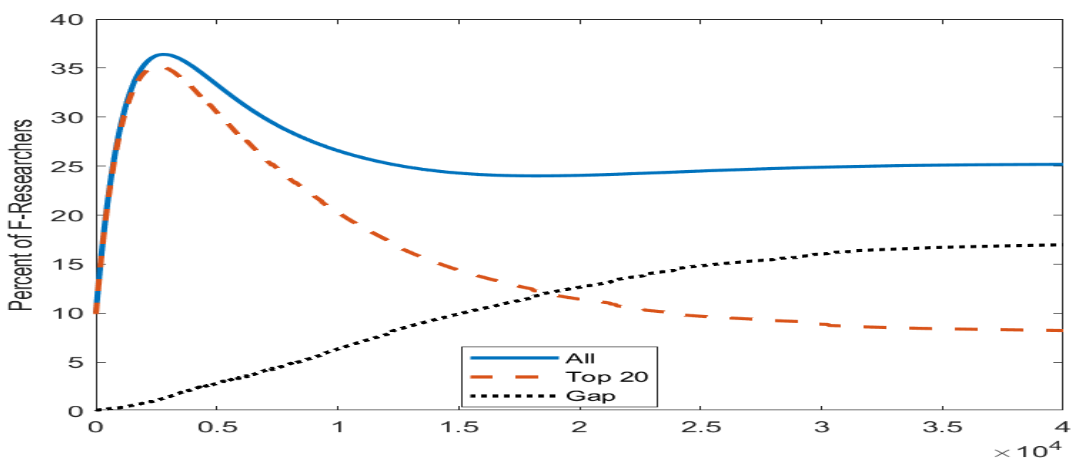
The two panels of Figure 8 replicate in the model the corresponding results collected in the two panels of Figure 2. Specifically, in Panel (a), we first rank the J institutions in terms of publishing intensity $P_{j,t}$, and then take the average fraction of F-researchers across the top 10, top 20, and, respectively, all institutions. While the absolute levels are smaller in

Figure 8: F-researchers Under-representation in Top Research Institutions

(a) Fraction of F-researchers across Institutions



(b)The Dynamics of F Researchers in Top Research Institutions:

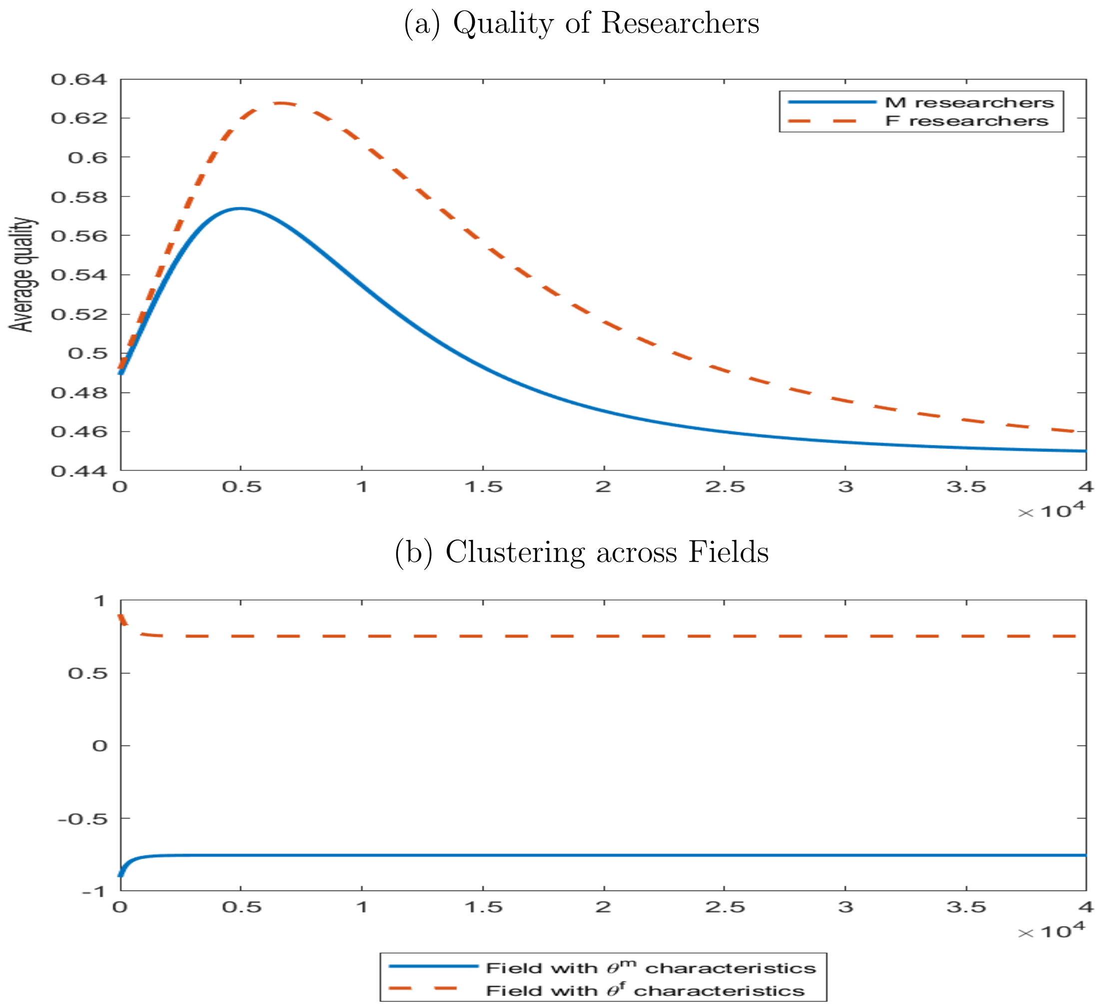


Panel (a) plots the percent of F-researchers across 100 simulated institutions, the top 20, and the top 10. Panel (b) reports the dynamics of the fraction of F-researchers in top research institutions and across all institutions as simulated from the model. Top research institutions are the top 20 out of 100 with the highest frequency of publication. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9p^m + 0.1p^f$.

our model, the close match with Panel (a) in Figure 2 is surprising, given that our model has no group bias at all.

Indeed, our model also replicates the time-series dynamics of the fraction of F-researchers in top institutions. Panel (b) of Figure 8 shows that the gap between the fraction of F-researchers in top institutions and in the full set of institutions increases over time. Intuitively, sorting institutions by their publishing intensity implicitly defines "top institutions" as those whose types are increasingly similar to the types of the majority of referees—that

Figure 9: Quality and Clustering across Fields of M- and F-researchers in Calibrated Model



Panel (a) plots the average quality of accepted M and F researchers in the model. Quality is measured as $\sum_\theta \gamma^\theta w_t^{\theta,g}$ where $\gamma^\theta = \gamma_0 \rho^{\frac{1}{N}\sum_{n=1}^{N}\theta_n}$ and $w_t^{\theta,g} = a_t^{\theta,g} / \sum_{\theta'} a_t^{\theta',g}$, $g = f, m$. Panel (b) reports clustering across fields implied by θ^m and θ^f. Parameters: $\phi = 0.5742$, $d = 0.3$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 10$. Initially $\lambda_0 = 0.9p^m + 0.1p^f$.

is, in the limit, $\overline{\lambda}^{\theta^m}$. By construction, the remaining institutions will have a larger fraction of researchers whose types are less represented in the refereeing population.

4.1.3. Higher quality of successful F-researchers and clustering

Finally, the two panels of Figure 9 replicate in the model the corresponding panels in Figure 4 in the data. Specifically, Panel (a) of Figure 9 plots the average quality of F- and M-researchers conditional on being accepted, and shows that the average quality of F-

researchers is uniformly higher than that of M-researchers. This plot is consistent with Proposition 6 and our conjecture that the result should hold for every N.[17] This result is consistent with the top panel in Figure 4.

Panel (b) of Figure 9 shows that the model produces clustering across fields, consistent with the bottom panel in Figure 4. Specifically, consider types θ^m and θ^f. Panel (b) plots the difference between the share of F- and M-researchers across the θ^f-field and the θ^m-field. As also shown in Corollary 3, F-researchers are relatively more represented in the former (top line), and M-researchers are relatively more represented in the latter (bottom line). While our model yields more extreme predictions relative to the data, with only two fields surviving in the limit (θ^f and θ^m), the lack of dynamics in the data is consistent with our model's predictions. Moreover, in the extension in Section 6, we show that, with multiple distinct types $\theta \in \Theta^{\max}$ and endogenous entry in academia, the model also predicts multiple "fields" and a higher number of M-dominated fields than F-dominated fields; this is consistent with the data.

5. The Impact of Mentoring

In this section, we discuss the impact of one particular policy action that has been proposed to address gender imbalance, namely mentoring. There is evidence that mentoring does help increase the success rate of female economists (Ginther, Currie, Blau, and Croson (2020)). We now investigate the implications of mentoring in our model.

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

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

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

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

In words, the increase in the probability of getting hired must be sufficiently high relative to the cost of undergoing mentoring. For instance, if the right-hand side is negative (type θ is already likely to succeed), no researcher of that type would pay such a cost.

[17] Again, Proposition 6 assumes that $\lambda_0 = p^m$; the results in Panel (a) of Figure 9 thus suggest that the conclusions of the proposition are robust to small changes in the initial population.

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, the higher the cost, indicating that it will take greater effort to "learn" to become a type that is likely to be hired. Note that such distance may be high as the young researcher θ may have some characteristics that are desirable from an objective standpoint, but that are not viewed as important or relevant by the majority of established researchers. The cost, in that case, is to "unlearn" what is deemed "irrelevant." The Online Appendix contains the details of the system dynamics for the model parameterization in Section 4. For brevity, we provide only the intuition here. Panel (a) of Figure 10 illustrates the dynamics under the same parameters as in Section 4 and a cost function $C(\theta, \theta') = \beta \sum_{n=1}^{N} (\theta_n - \theta'_n)^2$, with $\beta = 0.075$. We choose this cost so that not all young researchers want to pay the switching cost to become like their advisors, which seems plausible. The resulting steady state is roughly consistent with the observed female representation in economics.

Initially, the dynamics are as in the base case, as all λ_t^{θ} are small and thus no young researcher wants to pay the cost of mentoring. In this dynamics, as we know, $\lambda_t^{\theta^m}$ and $\lambda_t^{\theta^f}$ increase, with the former increasing faster, as shown in panel (c) of Figure 10. At some point, the mass of $\lambda_t^{\theta^m}$ becomes large enough to induce many young researchers, both M and F, to pay the mentoring cost, and the system (nearly) jumps. The reason is that many young researchers now expect that their advisor will likely be of type θ^m, which is also the type of established researchers who will evaluate their research. They are thus happy to pay the cost and become like their advisors.

The bottom panels of Figure 10 show, however, that the mass of young M-researchers jumps by more than the mass of F-researchers. The reason is that even though the cost function is the same for M- and F-researchers, young M-researchers are on average closer to θ^m and thus have systematically lower costs of switching than young F-researchers. For this reason, group imbalance persists forever.[18] Moreover, only type θ^m survives and therefore the research characteristics mildly more common in the F-population, but also very common in the M-population, disappear, thus yielding talent loss and loss of knowledge.

6. Extension: Endogenous Entry

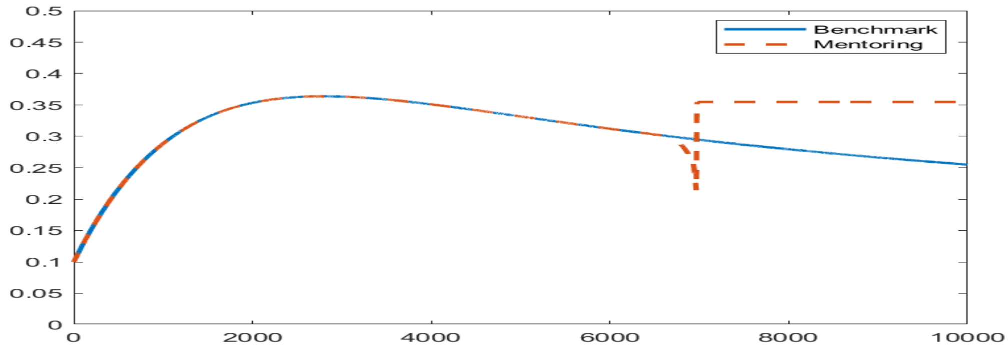
In this section we summarize the results from an extension of the model to endogenous entry.[19] We assume that candidates may choose a career in academia, in which case their

[18] If the cost function was lower, however, then *all* young researchers, M and F, would pay the cost and the system would jump to group balance.

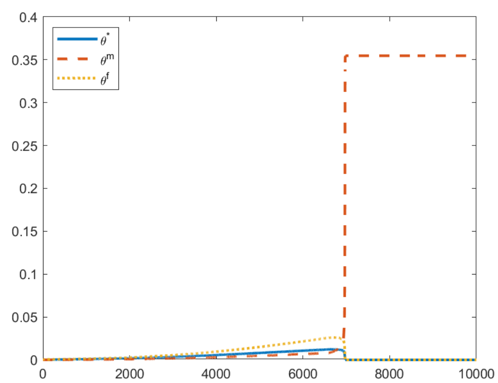
[19] The Online Appendix contains additional extensions, including a two-layer hierarchy of junior and senior established researchers; attribution of co-authored work; and more general forms of self-image bias. In

Figure 10: Costly Mentoring

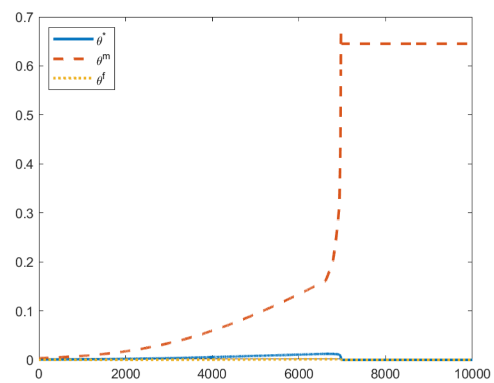
(a) Fraction of F-Researchers



(b) F researchers



(c) M researchers



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

success depends on the judgment of the established group of researchers as described in Section 3, or can opt for a different career. If they choose the academic career, candidates pay a utility cost C but receive a payoff P if successful. The outside option gives a benchmark utility of zero. We obtain several results:

1. When the relative cost C/P is small and below a cutoff, the same equilibrium and results as in Section 3.3 obtain;

2. When the relative cost C/P is intermediate, the equilibrium changes. In particular,

addition, we also show that the same equilibrium as in this section occurs when hiring institutions decide to hire candidates based on their probability of success as opposed to their objective quality.

(i) The set of surviving types $\Theta^{\max}$ shrinks: Even in environments where, for $C = 0$, $\Theta^{\max}$ in (5) contains pairs of distinct symmetric types (see Definition 1), when C/P is intermediate, some types that are more common in the F-population may not survive in the limit.

(ii) This new force increases the limit imbalance towards the M-population, and leads to further talent loss, as young researchers of certain types will not even apply for an academic career.

(iii) Moreover, assuming that there is a mapping between fields and types as in Corollary 3, the number of M-dominated fields is higher than the number of F-dominated fields, as is the case in the data (see the bottom panel of Fig. 4).

3. Finally, in the special case of the calibrated model in Section 4, we also show that the pool of applicants skews towards the M-population. That is, imbalance occurs even in the "pipeline," which may explain the low percentage of applications from women to PhD programs, for instance.

7. Literature Review

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

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

On the theoretical side, our model is related to the literature on statistical discrimination: a relative recent survey is Fang and Moro (2011). One strand within that literature, originating from Phelps (1972), posits the existence of exogenous differences between groups, either in the distribution of productivity ("Case 1"), or in the quality of signals about it ("Case 2"). In Case 2, the employer does not observe the productivity of individual applicants, but receives a signal about it. Differential average treatment of the two groups can emerge either through risk aversion of the employer (Aigner and Cain, 1977), investment in human capital (Lundberg and Startz, 1983), or if hiring occurs in a tournament (Cornell and Welch, 1996). In Conde-Ruiz, Ganuza, and Profeta (2022), the difference in signal quality leads members of the group in the minority of a hiring committee to underinvest in human capital; this perpetuates the imbalance. A recent contribution, Bardhi, Guo, and Strulovici (2019), revisits Phelp's Case 1, but assume that success or failure is observed over time and is informative about the worker's type. This can lead to large differences in ex-post treatment of the two groups, even if ex-ante productivity differences are small. Differently from this literature, in our model the ex-ante distributions of productivity are the same in the M and F group, because all characteristics are equally valuable. Furthermore, productivity is observed. In our model, standard statistical discrimination does not lead to gender imbalance.

Becker (1957)'s model of taste-based discrimination instead posits that employers may have a preference for hiring members of one specific group. This is not the case in our model: while referees only accept applicants whose research characteristics match their own, they do not take group membership into consideration at all.

Heidhues, Kőszegi, and Strack (2025) proposes a model in which an agent's ability is unobserved, both by herself and by others. Agents belong to different groups, each potentially subject to "discrimination," and are "stubbornly overconfident" about their own ability. Overconfidence leads agents to have a more favorable view of individuals in their own social group, ascribing poor performance to discrimination against them. In our model, ability is observed, and there is no exogenously imposed discrimination on either group. Incorporating (possibly biased) learning (cf. e.g. Bohren, Imas, and Rosenberg, 2019) about young researchers' characteristics is an interesting direction for future work.

8. Conclusions and Policy Implications

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

Our theoretical and numerical results are consistent with numerous empirical regularities that we collected in Section 2, which we do not repeat for brevity.

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 programs are akin to lowering the cost of doing research, but in our model, the cost is zero. Similarly, mentorship programs for female researchers may also not be effective (see Section 5), and have the unintended consequence of inducing a talent loss, as female researchers give up their research characteristics to adopt those that are prevalent in the reviewer population.

Because the problem is self-image bias, the best policy intervention must involve limiting the ability of reviewers to use their own research style as a yardstick while judging others' research. One solution is to provide strict guidelines in the refereeing process. Indeed, in light of Corollary 6 (d), editors should guide referees to limit the number of aspects of the submitted research paper they should focus on. Dunning, Meyerowitz, and Holzberg (1989) provides suggestive evidence in support of this approach.

Another solution is 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.

A. Appendix: Results for the general model

Proof of Proposition 1 Eq. (1) shows that $a_t^{\theta,g}$ is time-invariant for $g \in \{f, m\}$; hence, so is a_t^{θ}, and therefore a_t. Dropping time indices, for $g \in \{f, m\}$, a straightforward derivation shows that

$$\lambda_t^{\theta,g} = (1-a)\lambda_{t-1}^{\theta,g} + a^{\theta,g} = (1-a)^t \lambda_0^{\theta,g} + a^{\theta,g}\frac{1-(1-a)^t}{a} \to \frac{a^{\theta,g}}{a} = \frac{\gamma^\theta p^{\theta,g}}{a}, \tag{18}$$

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

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

By Assumption 1, for every θ, the type $\bar{\theta} = \sigma(\theta)$ satisfies $p^{\theta,m} = p^{\bar{\theta},f}$ and $\gamma^\theta = \gamma^{\bar{\theta}}$; hence, the above fraction equals 1. *Q.E.D.*

To prove the main results of the paper, 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 $g \in \{f,m\}$. Also define $q = q^m + q^f$. Then, for $g \in \{f,m\}$, the dynamics are given by

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}\left(1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}\right) + \lambda_{t-1}^{\theta} q^{\theta,g}; \qquad \lambda_t^\theta = \lambda_t^{\theta,m} + \lambda_t^{\theta,f}. \tag{19}$$

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

1. *if $\lambda_0^\theta = 0$, then $\lambda_t^\theta = 0$ for all $t \geq 0$;*
2. *if $\lambda_0^\theta > 0$, then $\lambda_t^\theta > 0$ for all $t \geq 0$;*
3. *for $\theta, \tilde\theta \in \Theta$ with $\lambda_0^\theta \cdot \lambda_0^{\tilde\theta} > 0$:*
   (a) *$\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} - \frac{\lambda_t^{\tilde\theta}}{\lambda_{t-1}^{\tilde\theta}} = q^\theta - q^{\tilde\theta}$ for all $t \geq 1$, and*
   (b) *$q^\theta > q^{\tilde\theta}$ implies $\frac{\lambda_t^\theta}{\lambda_t^{\tilde\theta}} \to \infty$, and $q^\theta = q^{\tilde\theta}$ implies $\frac{\lambda_t^\theta}{\lambda_t^{\tilde\theta}} = \frac{\lambda_0^\theta}{\lambda_0^{\tilde\theta}}$ for all $t \geq 0$;*
4. *define the set*
$$\Theta^{\max} = \{\theta \in \Theta \;:\; \lambda_0^\theta > 0,\; \theta \in \arg\max_{\theta'\in\Theta} q^{\theta'}\} \tag{20}$$
and let $\bar\lambda \in \Delta(\Theta)$ be such that
$$\bar\lambda^{\tilde\theta} = \begin{cases} \frac{\lambda_0^{\tilde\theta}}{\sum_{\theta\in\Theta^{\max}} \lambda_0^\theta} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} \end{cases}; \tag{21}$$
then $\lim_{t\to\infty} \lambda_t = \bar\lambda$;
5. *define*
$$\bar\lambda^{\tilde\theta,f} = \begin{cases} \frac{\lambda_0^{\tilde\theta} q^{\tilde\theta,f}}{\sum_{\theta\in\Theta^{\max}} \lambda_0^\theta q^\theta} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} \end{cases} \quad \text{and} \quad \bar\lambda^{\tilde\theta,m} = \begin{cases} \frac{\lambda_0^{\tilde\theta} q^{\tilde\theta,m}}{\sum_{\theta\in\Theta^{\max}} \lambda_0^\theta q^\theta} & \tilde\theta \in \Theta^{\max} \\ 0 & \tilde\theta \notin \Theta^{\max} \end{cases}; \tag{22}$$
then $\lim_{t\to\infty} \lambda_t^f = \bar\lambda^f$ and $\lim_{t\to\infty} \lambda_t^m = \bar\lambda^m$.

Proof: Eq. (19) implies that

$$\lambda_t^\theta = \left(1 - \sum_{\theta'\in\Theta} \lambda_{t-1}^{\theta'} q^{\theta'}\right) \lambda_{t-1}^\theta + \lambda_{t-1}^\theta q^\theta. \tag{23}$$

By assumption $\lambda_0 \in \Delta(\Theta)$. Inductively, suppose $\lambda_{t-1} \in \Delta(\Theta)$ and $\lambda_{t-1}^m, \lambda_{t-1}^f \in \mathbb{R}_+^\Theta$. Summing over Θ on both sides of Eq. (23) yields $\sum_\theta \lambda_t^\theta = (1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'})(\sum_\theta \lambda_{t-1}^\theta) + \sum_\theta \lambda_{t-1}^\theta q^\theta = (1 - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}) + \sum_\theta \lambda_{t-1}^\theta q^\theta = 1$. Furthermore, since $\lambda_{t-1} \in \Delta(\Theta)$, $\sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} \in [\min_{\theta'} q^{\theta'}, \max_{\theta'} q^{\theta'}] \subseteq [0,1]$; moreover, $q^\theta \geq 0$ and $\lambda_{t-1}^\theta \geq 0$, so Eq. (23) implies that $\lambda_t^\theta \geq 0$

as well. By the same argument, $q^\theta \geq 0$ and $\lambda_{t-1}^{\theta,g} \geq 0$ for $g \in \{f,m\}$ imply $\lambda_t^{\theta,g} \geq 0$ for $g \in \{f,m\}$ as well by Eq. (19). Thus, $\lambda_t \in \Delta(\Theta)$, and $\lambda_t^g \in \mathbb{R}_+^\Theta$ for each g.

Claim 1 is immediate. For Claim 2, again we argue by induction. For $t=0$, the claim is trivially true. Inductively, assume $\lambda_{t-1}^\theta > 0$. By Eq. (23), 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. (23) 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. (23) for type θ by λ_{t-1}^θ, which is assumed to be positive; this yields

$$\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} = 1 + q^\theta - \sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'}. \tag{24}$$

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. (24), 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 second equality 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 $0 \le \lambda_t^{\theta,g} \le \lambda_t^\theta$ for all $t \ge 0$, 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. (19) and the fact that $\sum_{\theta'} \lambda_{t-1}^{\theta'} q^{\theta'} \in [0,1]$ and $0 \le \lambda_{t-1}^{\theta,g} \le \lambda_{t-1}^\theta \le 1$ 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 $\sum_{\theta'} \bar{\lambda}^{\theta'} q^{\theta'} \ge \bar{\lambda}^{\theta^+} q^{\theta^+} > 0$.) 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. (19) can be written as

$$\lambda_t^{\theta,g} = (1 - q_{t-1}) \lambda_{t-1}^{\theta,g} + \lambda_{t-1}^\theta q^{\theta,g}. \tag{25}$$

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

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

$$\lambda_t^{\theta,g} = \lambda_T^{\theta,g} \prod_{s=T}^{t-1} (1 - q_s) + q^{\theta,g} \sum_{s=T}^{t-1} \lambda_s^\theta \prod_{r=s+1}^{t-1} (1 - q_r). \tag{26}$$

For $t = T+1$, this follows from Eq. (25). Inductively, assume it holds for $t-1 > T$. Then, by Eq. (25) and the inductive hypothesis,

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

as claimed.

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

Since $\lambda_t^\theta \to \bar{\lambda}^\theta$ and $q_t \to \bar{q}$, there is $T \ge 0$ 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. (26) 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.*

Proof of Propositions 2 and 3: by assumption, $\gamma^{\theta}(p^{\theta,m} + p^{\theta,f}) \leq 1$ for every θ. Hence, part (i) of Proposition 2 and Proposition 3 follow from Theorem 1 parts 4 and 5, by setting $q^{\theta,g} = \gamma^{\theta} p^{\theta,g}$ for $g = m, f$ and $\theta \in \Theta$. For part (ii) of Proposition 2, note that, if $\theta \in \arg\max_{\theta' \in \Theta} \gamma^{\theta'}(p^{\theta',m} + p^{\theta',f})$, then $\gamma^{\sigma(\theta)} = \gamma^{\theta}$ and $p^{\sigma(\theta),m} + p^{\sigma(\theta),f} = p^{\theta,f} + p^{\theta,m}$ imply that also $\sigma(\theta) \in \arg\max_{\theta' \in \Theta} \gamma^{\theta'}(p^{\theta',m} + p^{\theta',f})$. *Q.E.D.*

Proof sketch of Proposition 4 (see Online Appendix A2. for a detailed proof): write

$$\bar{\Lambda}^g = \sum_{\theta \in \Theta^{\max}:\sigma(\theta)=\theta} \bar{\lambda}^{\theta,g} + \sum_{\theta \in \Theta^{\max}:\sigma(\theta)\neq\theta} \bar{\lambda}^{\theta,g} = \sum_{\theta \in \Theta^{\max}:\sigma(\theta)=\theta} \bar{\lambda}^{\theta,g} + \frac{1}{2} \sum_{\theta \in \Theta^{\max}:\sigma(\theta)\neq\theta} (\bar{\lambda}^{\theta,g} + \bar{\lambda}^{\sigma(\theta),g}).$$

By the properties of symmetric types, if $\Theta^{\max}$ is homogeneous then the above expression is the same for $g = m, f$, and so necessarily $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$. Otherwise, for at least one θ, we have $\theta, \sigma(\theta) \in \Theta^{\max}$ and $p^{\theta,m} \neq p^{\sigma(\theta),m}$, and it is wlog to assume that $p^{\theta,m} > p^{\sigma(\theta),m}$. Corollary 1 shows that, for such $\theta, \sigma(\theta)$, $\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\sigma(\theta),m} > \bar{\lambda}^{\theta,f} + \bar{\lambda}^{\sigma(\theta),f}$. Therefore, $\bar{\Lambda}^m > \bar{\Lambda}^f$, which implies that $\bar{\Lambda}^m > \frac{1}{2}$.

Proof sketch of Proposition 5 (see Online Appendix A2. for a detailed proof): we first show that, for all θ and $t > 0$, $p^{\theta,m} \geq p^{\sigma(\theta),m}$ iff $\lambda^{\theta}_{t-1} \geq \lambda^{\sigma(\theta)}_{t-1}$. For $t = 1$, this is by assumption. For $t > 1$, this follows by inductively invoking part 3(a) of Theorem 1. By direct calculation, this in turn implies that for every $\theta \in \Theta$ and $t \geq 1$, $a^{\theta,m}_t + a^{\sigma(\theta),m}_t \geq a^{\theta,f}_t + a^{\sigma(\theta),f}_t$. The argument is completed by showing that, for $g = m, f$, we can write

$$\sum_{\theta:\gamma^\theta=\bar{\gamma}} a^{\theta,g}_t = \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta=\sigma(\theta)} a^{\theta,g}_t + \frac{1}{2} \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta\neq\sigma(\theta)} [a^{\theta,g}_t + a^{\sigma(\theta),g}_t].$$

Proof sketch of Proposition 6 (see Online Appendix A2. for a detailed proof): as in the proof of Proposition 5, we have $a^{\theta,m}_t + a^{\theta',m}_t > a^{\theta,f}_t + a^{\theta',f}_t$ for the intermediate types θ, θ'. On the other hand, $a^{\theta_0,m}_t = a^{\theta_0,f}_t$ and $a^{\theta_1,m}_t = a^{\theta_1,f}_t$ for the highest and lowest types θ_0, θ_1. This implies that the weight on the intermediate quality $\gamma^\theta = \gamma^{\theta'}$ is higher for accepted M researchers. We then show inductively that $a^{\theta_1,g}_t > a^{\theta_0,g}_t$ for both $g = m, f$. A direct calculation and comparison of the expressions for the expected qualities $E[\gamma|M]$ and $E[\gamma|F]$ yields the result.

References

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

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

Peter Andre and Martin Falk. What's worth knowing? economists' opinions about economics. ECONtribute Discussion Paper 102, University of Bonn and University of Cologne, Reinhard Selten Institute (RSI), Bonn and Cologne, 2021. URL http://hdl.handle.net/10419/237347.

Emmanuelle Auriol, Guido Friebel, Alisa Weinberger, and Sascha Wilhem. Women in economics: Europe and the world. CEPR Discussion Paper 16686, CEPR, November 2021. URL https://cepr.org/publications/dp16686.

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

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

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

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.

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

J. Aislinn Bohren, Peter Hull, and Alex Imas. Systemic discrimination: Theory and measurement. *The Quarterly Journal of Economics*, 140(3):1271–1341, 2025. doi: 10.1093/qje/qjaf010.

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

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

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

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

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

J. Ignacio Conde-Ruiz, Juan José Ganuza, and Paola Profeta. Statistical discrimination and committees. *European Economic Review*, 141:103994, 2022. doi: 10.1016/j.euroecorev.2021.103994.

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

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

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

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

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

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

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

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

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

Pascaline Dupas, Alicia Sasser Modestino, Muriel Niederle, and Justin Wolfers. Gender and the dynamics of economics seminars. NBER Working Paper 28494, National Bureau of Economic Research, February 2021. URL https://www.nber.org/papers/w28494.

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

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

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

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

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

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

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

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

Paul Heidhues, Botond Kőszegi, and Philipp Strack. Overconfidence and prejudice. *The Review of Economic Studies*, 92(2):968–1010, 2025. doi: 10.1093/restud/rdaf048.

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

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

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

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

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

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

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

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

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

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

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

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

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

Self-Image Bias and Talent Loss
On-Line Appendix

Marciano Siniscalchi
Northwestern University
Pietro Veronesi
University of Chicago

A1.. Extensions

A1.1.. Endogenous Choice of Young Researchers

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

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

$$\gamma^\theta \lambda_t^\theta (P - C) + (1 - \gamma^\theta \lambda_t^\theta)(-C) > 0. \tag{A.27}$$

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

$$a_t^{\theta,g} = \begin{cases} \gamma^\theta \cdot \lambda_{t-1}^\theta \cdot p^{\theta,g} & \text{if } \gamma^\theta \lambda_{t-1}^\theta \geq \frac{C}{P} \\ 0 & \text{otherwise} \end{cases} \tag{A.28}$$

$$\lambda_t^{\theta,g} = \lambda_{t-1}^{\theta,g}\,(1 - a_t) + a_t^{\theta,g} \tag{A.29}$$

Expression (A.28) shows that if the mass of type-θ reviewers drops below $\frac{C}{\gamma^\theta P}$ at time $t - 1$, both M and F young type-θ researchers will not apply at date t. From Eq. (A.29), this implies that the total mass of such types will decrease, at least weakly, because some type-θ established researchers will have to retire in order to make room for researchers of other types who are accepted. In fact, the mass of such types will decrease strictly, except in case no young researcher wants to apply.

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

Proposition A.1 *Let*

$$\Theta^{\max} = \left\{ \theta \in \Theta \text{ such that } \lambda_0^\theta \geq \frac{C}{\gamma^\theta P} \text{ and } \theta \in \arg\max_{\theta' \in \Theta} \gamma^{\theta'} (p^{\theta',m} + p^{\theta',f}) \right\} \tag{A.30}$$

then:

(i) only $\theta \in \Theta^{\max}$ survive in the limit as $t \to \infty$;

(ii) $\Theta^{\max}$ need not preserve symmetry across types: if $\theta \in \Theta^{\max}$, then $\sigma(\theta) \in \Theta^{\max}$ if and only if $\lambda_0^{\sigma(\theta)} \geq \frac{C}{\gamma^\theta P}$.

(iii) for every $\theta \in \Theta^{\max}$,

$$\bar{\lambda}^\theta = \frac{\lambda_0^\theta}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'}}. \tag{A.31}$$

(iv) for every $\theta \in \Theta^{\max}$,

$$\bar{\lambda}^{\theta,m} = \frac{\lambda_0^\theta \gamma^\theta p^{\theta,m}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} \gamma^\theta (p^{\theta,m} + p^{\theta,f})} \quad \text{and} \quad \bar{\lambda}^{\theta,f} = \frac{\lambda_0^\theta \gamma^\theta p^{\theta,f}}{\sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} \gamma^\theta (p^{\theta,m} + p^{\theta,f})}. \tag{A.32}$$

While the general structure of the limit distribution of types is similar to that in Propositions 2 and 3, the key difference is part (ii), which has implications for Eqs. (A.31) and (A.32). Two types θ and $\theta' = \sigma(\theta)$ may be symmetric, and yet differences in their initial frequencies $\lambda_0^\theta, \lambda_0^{\theta'}$ may imply that one survives and the other doesn't. This may exacerbate group imbalance.

Take for instance Corollary 1, a key result in our model. Consider $\theta, \theta' = \sigma(\theta)$ with $p^{\theta,m} > p^{\theta',m}$, and assume that both θ and θ' are in $\Theta^{\max}$. Then the conclusion of that Corollary holds verbatim, because its proof only relies on Eq. (7), which is structurally identical to Eq. (A.32). (The denominators will be different because of differences in the sets $\Theta^{\max}$, but this is immaterial to the argument.) However, if in addition type θ' does not survive because $\lambda_0^{\theta'} < \frac{C}{\gamma^\theta P}$, there is an additional force contributing to group imbalance, as shown next.

Corollary A.1 *Assume $\lambda_0 = p^m$. Let $\theta \in \Theta^{\max}$ be such that $p^{\theta,m} > p^{\theta,f}$. Then, there is a relative cost C/P such that $\theta' = \sigma(\theta) \notin \Theta^{\max}$ and $\bar{\lambda}^{\theta,m} > \bar{\lambda}^{\theta,f}$. Moreover, in the special case in which θ is the only surviving type, M-imbalance occurs:*

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

Finally, in this case, M-imbalance is larger than in the case with zero cost $C = 0$. That is:

$$\bar{\lambda}^{\theta,m} - \bar{\lambda}^{\theta,f} > \left(\bar{\lambda}^{\theta,m}_{C=0} + \bar{\lambda}^{\theta',m}_{C=0} \right) - \left(\bar{\lambda}^{\theta,f}_{C=0} + \bar{\lambda}^{\theta',f}_{C=0} \right)$$

where the subscript $C = 0$ denotes the case with zero cost, in which case $\theta' = \sigma(\theta) \in \Theta^{\max}$ [see Proposition 2 part (ii)].

Proposition A.1 and Corollary A.1 formalize statements (i)–(iii) in Section 6. Statement (i) corresponds to part (ii) of Proposition A.1. Statement (ii) follows from Corollary A.1 (greater imbalance than for $C = 0$) and the definition of $\Theta^{\max}$ in Proposition A.1 (types more common in the F group may choose not to apply). Finally, Statement (iii) follows by noting that, in Corollary 3, if $\theta, \theta' = \sigma(\theta)$ are distinct elements of $\Theta^{\max}$ when $C = 0$, with $p^{\theta,m} > p^{\theta,f}$, then "field" θ is M-dominated and, symmetrically, "field" θ' is F-dominated; this remains true for $C > 0$ if $\theta, \theta' \in \Theta^{\max}$, but if $\theta' \notin \Theta^{\max}$, then "field" θ is M-dominated by Corollary A.1 and there is no corresponding F-dominated field.

For the special case of the N-characteristics model in Section 4, we have:

Corollary A.2 *Assume that at time 0, all referees are from the M-group with $\lambda_0 = p^m$.*

(a.1) If $\rho < \bar{\rho}(\phi, N)$ and $\frac{C}{P} \leq (1-\phi)^N \gamma_0 \sqrt{\rho}$, then the steady state is as in Corollary 6 (a).

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

$$\bar{\Lambda}^m = \lim_{t \to \infty} \sum_{\theta} \lambda_t^{m,\theta} = \frac{\phi^N}{\phi^N + (1-\phi)^N} > \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + \left(\frac{\phi}{1-\phi}\right)^{2N} + 2\left(\frac{\phi}{1-\phi}\right)^{N}}. \tag{A.33}$$

(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 Corollary 6 (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 4 for the same two conditions about ρ, respectively. This is intuitive. The only difference is that all types other than surviving ones drop out in finite time, rather than only in the limit.

The interesting new part is (a.2). In this case, the only type that survives in the long-run is θ^m, the most prevalent type in the M-population. In particular, θ^f now disappears. Thus, the characteristics that are mildly more frequent in the F-population, but also common in the M-population, eventually disappear. In this case, endogenous entry greatly exacerbates the loss of talent compared to the base case. Indeed, the total mass of M researchers, $\bar{\Lambda}^m$, is now even larger than in its counterpart without endogenous entry, whose expression is in Eq. (16) in Corollary 6. 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 an 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.

A1.1.1.. 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, Corollary A.2 shows that the steady state either remains the same, if the cost C is sufficiently low, as in case (a.1), or it becomes even more skewed towards the M group, as in case (a.2). In the latter case, the limiting fraction of M-researchers is $\bar{\Lambda}^m = \phi^N/(\phi^N + (1-\phi)^N) = 95\%$.

We now illustrate how endogenous choice may prevent convergence to group balance even when group balance would in fact attain in the basic model. We use the same parameterization as in Section 4, except that the number of characteristics is $N = 8$ instead of $N = 10$. With these parameter values, Corollary 6 part (b) implies that the system will converge to an equal mass of M and F researchers, because $\rho = 5 > 3.61 = \bar{\rho}(\phi, N)$. The solid and dashed lines in Figure A.1 confirm this.

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

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

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

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

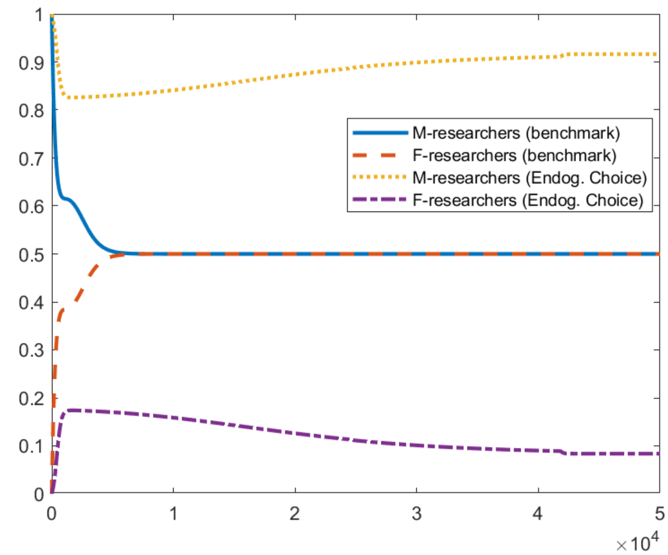
On the other hand, type θ^m (M or F) does:

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

Therefore, while other types are also willing to apply, type θ^m will prevail, which will lead to a severe imbalance between M and F researchers in the limit, as shown in Figure A.1. 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 A.2 shows that both F and M researchers are of type θ^m in the long run.

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

Figure A.1: Fraction of M and F Researchers with Endogenous Entry

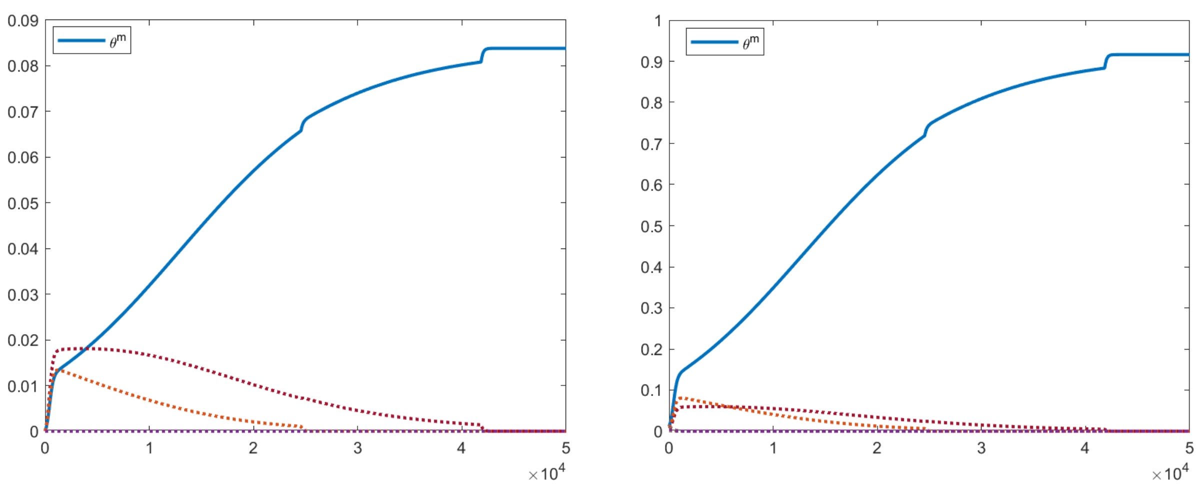


Fraction of M and F researchers when $\lambda_0 = p^m$. Parameters: $\phi = 0.5742$ $(d = 0.3)$, $\gamma_0 = 0.2$, $\rho = 5$, $N = 8$, $P = 1000$, and $C = 4$.

Figure A.2: Types of Established F and M Researchers with Endogenous Entry

(a) F researchers

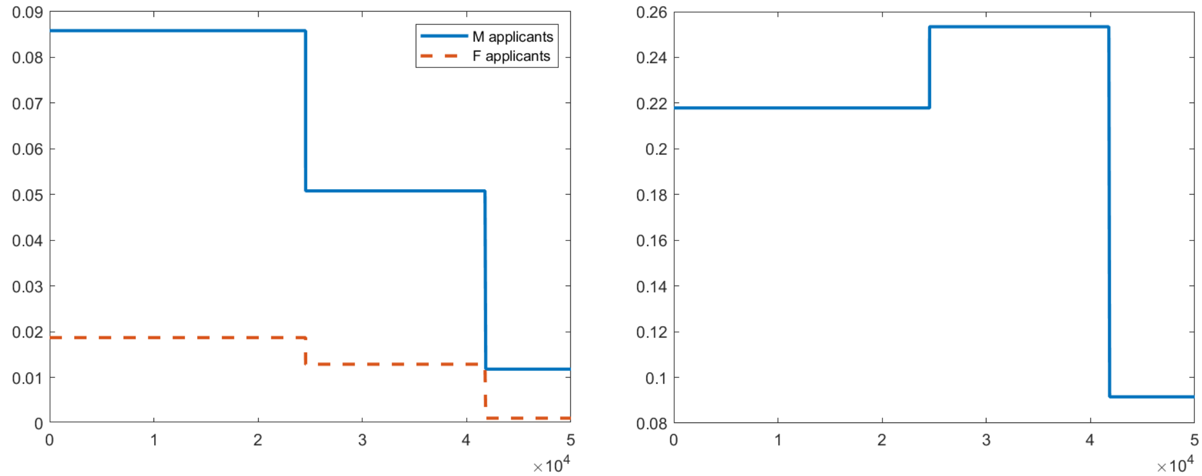
(b) M researchers



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

Figure A.3: Endogenous entry: applicants

(a) M and F Applicants (b) Fraction of F Applicants



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

A1.1.2.. Characterization of the Applicant Pool

Due to variation in the distribution of characteristics, Corollary A.2 also has implications for the mass of young M and F researchers who decide to apply for an academic job:

Proposition A.2 *For every t, let*

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

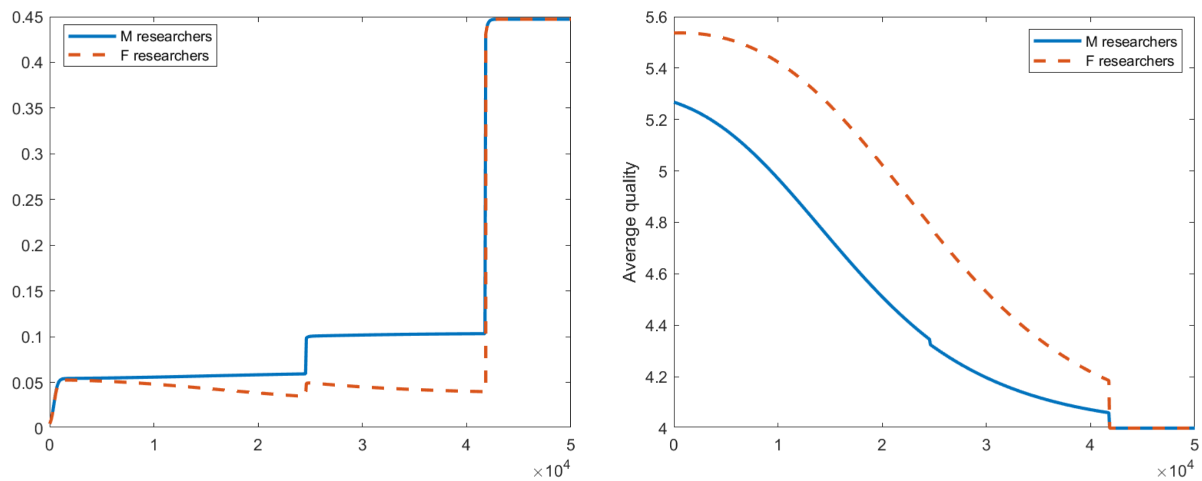
Then $A_t^m \geq A_t^f$. Moreover, if $\lambda_0^{\theta^m} > \frac{C}{\gamma_0 \sqrt{\rho} P} > \lambda_0^{\theta^f}$, then $A_t^f \to 1 - \bar{\Lambda}^m$, where $\bar{\Lambda}^m$ is as in part (a.2) of Corollary A.2.

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

Figures A.3a and A.3b 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 A.1. Consistently with Corollary A.2, the mass of M applicants is always greater than that of F applicants; furthermore, the latter declines over time. The discrete jumps in these masses occur whenever, for some type θ, the population fraction λ_t^θ falls below the cutoff $C/(\gamma^\theta P)$. In the limit, the fraction of F applicants equals the fraction

Figure A.4: Endogenous entry: Acceptance Rates

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



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

of F researchers of the only surviving type θ^m over the total:

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

Finally, the left panel of Figure A.4 shows the total acceptance rates of M and F applicants. In the initial period, the acceptance rates of M and F applicants are similar. They though diverge in the intermediate period, in which M applicants are accepted more often than the (fewer) F applicants, and then they finally converge, when only type θ^m survives. Interestingly, the right panel shows that the average quality of F researchers is uniformly higher until the time of convergence. This implies that in the initial period our model predicts similar acceptance rates of M and F researchers, even if the latter have higher objective quality. This result is reminiscent of Card et al. (2020), who show that unconditionally, acceptance rates of men- and women-authored papers are similar, but that the average quality of accepted women-authored papers, proxied by their future citations, is higher.

A1.2.. Endogenous Selection by Hiring Institutions

The previous section demonstrates that endogenizing the choice of entry into academia may shrink the supply of talent. We now show that the a similar mechanism operates on the demand side: when hiring decisions are based on the expectation of academic success, the anticipation of self-image bias in the refereeing process (Section 3.2) induces institutions to hire only those types θ that can produce research that is more likely to be "accepted" by the established refereeing population.

Consider the following alternative interpretation of our model. When a hiring institution evaluates a candidate, it takes into account whether or not the candidate will produce quality work that the profession recognizes, or—in the language of Section 3.2—"accepts." A candidate who is accepted by the profession yields a payoff P to the institution; this reflects e.g. visibility, grant money, or increased ability to attract top students. Hiring a candidate involves a cost C, which may be monetary but may also reflect mentoring resources and/or opportunity cost. This cost is borne by the institution whether or not the candidate is eventually accepted, and it is the same for M and F researchers. If the candidate is eventually not accepted or if the institution does not hire any candidate, the institution's payoff is zero. As above, a candidate of type θ produces quality work with probability γ^θ. To analyze demand effects, we reinterpret the key assumption of Section 3.2 as follows: the hiring institution anticipates that referees are subject to self-image bias, so that a type-θ researcher will be accepted by the profession with probability $\gamma^\theta \, \lambda_t^\theta$ at the end of time t.

Under these conditions, the institution hires a young researcher of type θ if and only if

$$\gamma^\theta \lambda_t^\theta \, (P - C) + \left(1 - \lambda_t^\theta \gamma^\theta\right)(-C) > 0 \tag{A.34}$$

This is the same condition as in Equation (A.27) in the previous section. Thus, the mass of established researchers λ_t^θ follows the system dynamics described by Equations (A.28) - (A.29). Proposition A.1 then applies and group imbalance and loss of talent obtains.

Moreover, in the special case of the model of Section 4, under the conditions of case (a.2) Corollary A.2, the system converges, in finite time, to a steady state in which only type θ^m survives. That is, if institutions *only* take acceptance by the profession into account at the hiring stage, type θ^f eventually disappears, even when such type would survive without endogenous selection. Again, this implies talent loss: research characteristics that are (mildly) more common in the F-population disappear.

We can also re-interpret the example in subsection A1.1.1. as a consequence of the hiring practices of hiring institutions. In the absence of endogenous selection, the parametric choices in that example lead to group balance, with both types θ^m and θ^f being represented in the limit. However, if institutions wish to hire only young researchers who are sufficiently likely to be accepted by the *current* population of referees, then group imbalance emerges, as in Figure A.1. Again, in this example type θ^f then disappears completely, as in Figure A.2.

This mechanism with endogenous choice further explains the patterns documented in Section 2

A1.3.. Seniors and Juniors

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

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

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

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

$$a_{1,t}^{\theta,g} = \gamma^\theta \cdot p^{\theta,g} \cdot (\lambda_{1,t-1}^\theta + \lambda_{2,t-1}^\theta) \tag{A.35}$$

$$a_{2,t}^{\theta,g} = \gamma^\theta \cdot \lambda_{1,t-1}^{\theta,m} \cdot \lambda_{2,t-1}^\theta. \tag{A.36}$$

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

$$\lambda_{1,t}^{\theta,g} = \lambda_{1,t-1}^{\theta,m} \left(1 - \frac{1}{1-\sigma}(a_{1,t} - a_{2,t})\right) + a_{1,t}^{\theta,g} - a_{2,t}^{\theta,g} \tag{A.37}$$

$$\lambda_{2,t}^{\theta,g} = \lambda_{2,t-1}^{\theta,g} \left(1 - \frac{1}{\sigma} a_{2,t}\right) + a_{2,t}^{\theta,g} \tag{A.38}$$

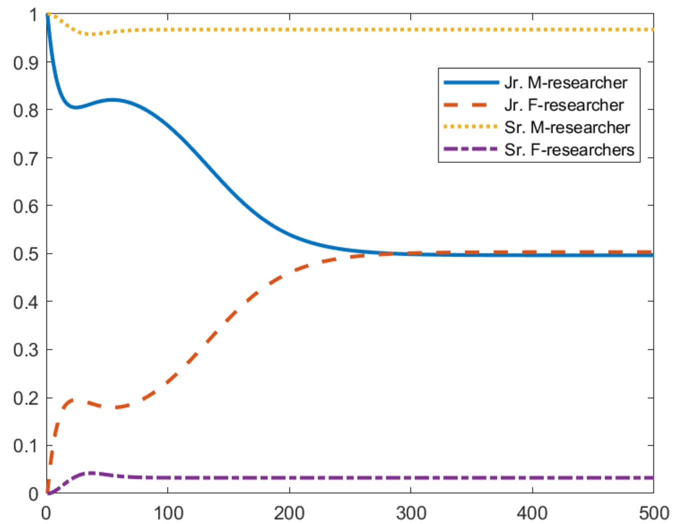
for $g \in \{f, m\}$, where $a_{j,t} = \sum_\theta (a_{j,t}^{\theta,f} + a_{j,t}^{\theta,m})$ for $j = 1, 2$.

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

A1.3.1.. Leaky Pipeline

Here we focus on the most interesting case, namely, the fact that this extension can also account for the "leaky pipeline" pattern highlighted in the CSWEP report (Chevalier, 2020).

Figure A.5: 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 $\sigma = 0.5$.

Figure A.5 provides a stark illustration: under the given parametric assumptions, group balance attains among juniors, but not among seniors. A rough intuition is that the self-image bias may not be strong enough to result in a prevalence of θ^m types among juniors, given the constant influx of new researchers with a more balanced distribution of types. However, it may be strong enough if the candidates' types are themselves more biased towards the M researchers' distribution—as is the case for junior up for promotion to the senior rank.

A1.4.. Co-authorship

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

Proposition A.3 *Let the economy be at its steady state with only types θ^f and θ^m surviving. For each researcher of type θ, define $L(\theta) = \sum_{n=1}^{N} \theta_n$ its objective quality. Let a research that is coauthored by type θ^a and θ^b be of type $\theta = \theta^a \vee \theta^b$, where $\vee$ denotes the component-wise maximum. Let researcher $a \in M$ and $b \in F$. Then, conditional on* *acceptance of the joint work, i.e.* $\theta^a \vee \theta^b \in \{\theta^m, \theta^f\}$*, we have*

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

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

A1.5.. Generalized self-image bias

In this section we discuss two extensions of the model to investigate the case in which referees do not only accept researchers who have characteristics identical to their own.

First, in the setting of Section 3, we assume that the set Θ of types is partitioned into subsets Θ_j; a referee of type θ^r drawn from subset Θ_j only accepts applicants of types $\theta^a \in \Theta_j$. That is, we replace Assumption 3 in the main text with

Assumption 3′: there exist disjoint sets $\Theta_1, \ldots, \Theta_J$ such that $\Theta = \Theta_1 \cup \ldots \Theta_J$ and, for every $j = 1, \ldots, J$, referees of type $\theta^r \in \Theta_j$ accept applicants of type θ^a if and only if $\theta^a \in \Theta_j$.

One interpretation is that types in each partition element Θ_j are in some sense "close" or "similar." Another is that each Θ_j represents a "field." In this interpretation, a field may involve different research attributes, but each attribute is only useful in one field.

The dynamics in Eqs. (3)–(4) must then be modified as follows: for every $j = 1, \ldots, J$ and $\theta \in \Theta_j$,

$$a_t^{\theta,g} = \gamma^\theta \lambda_{t-1}^j p^{\theta,g}, \qquad \lambda_t^{\theta,g} = (1 - a_t)\lambda_{t-1}^{\theta,g} + \gamma^\theta \lambda_{t-1}^j p^{\theta,g}, \tag{A.39}$$

where $\lambda_{t-1}^j = \sum_{\theta' \in \Theta_j} \lambda_{t-1}^{\theta'}$.

While this generalization of our model is not covered by our results, we can still invoke them indirectly to analyze it. First, summing over all $\theta \in \Theta_j$ yields

$$a_t^{j,g} = \gamma^j \lambda_{t-1}^j p_\gamma^{j,g}, \qquad \lambda_t^{j,g} = (1 - a_t)\lambda_{t-1}^{j,g} + \gamma^j \lambda_{t-1}^j p_\gamma^{j,g},$$

where $a_t^{j,g} = \sum_{\theta' \in \Theta_j} a_t^{\theta,g}$, $\lambda_\tau^{j,g} = \sum_{\theta' \in \Theta_j} \lambda_\tau^{\theta',g}$, $\gamma^j = \sum_{\theta' \in \Theta_j} \gamma^{\theta'}$, and $p_\gamma^{j,g} = \sum_{\theta' \in \Theta_j} \gamma^{\theta'} p^{\theta',g}/\gamma^j$. These equations are exactly like Eqs. (3)–(4), except that types $\theta \in \Theta$ are replaced with subsets or "fields" Θ_j, $j = 1 \ldots, J$.

Second, replace the symmetry assumption with the following

Assumption 1′: there is a function $\sigma : \Theta \to \Theta$ such that (i) for all $j, k \in \{1, \ldots, J\}$ and $\theta, \theta' \in \Theta_j$, $\sigma(\theta) \in \Theta_k$ implies $\sigma(\theta') \in \Theta_k$; (ii) for all $j = 1, \ldots, J$, if $\sigma(\theta) \in \Theta_k$ for all $\theta \in \Theta_j$, then $\gamma^j = \gamma^k$ and $p_\gamma^{j,m} = p_\gamma^{k,f}$; and (iii) $\sigma(\sigma(\theta)) = \theta$.

Finally, replace the heterogeneity and boundedness assumptions with

Assumption 2′: for some $j \in \{1, \ldots, J\}$, if k is such that $\sigma(\theta) \in \Theta_k$ for all $\theta \in \Theta_j$, then $p^{j,m} > p^{k,m}$.

Assumption 4′: for all $j \in \{1, \ldots, J\}$, $\gamma^j \cdot (p^{j,m}_{\gamma} + p^{j,f}_{\gamma}) = \sum_{\theta \in \Theta_j} \gamma^{\theta}(p^{m,\theta} + p^{f,\theta}) \leq 1$.

With these assumptions, our results go through unmodified, but apply to subsets, or fields $\Theta_1, \ldots, \Theta_J$, rather than to individual types θ. The dynamics of individual types can still be retrieved from Eq. (A.39).

A1.5.1.. Similarity in research characteristics

An alternative approach to modeling generalized self-image bias is to consider the parametric model of Section 4, introduce a notion of distance between types, and assume that referees accept young researchers whose characteristics are close enough to their own. Specifically, we assume that referee r of type θ^r accepts the research of young researcher θ if

$$D(\theta^r, \theta) = \sum_n (\theta^r_n - \theta_n)^2 \leq \eta \tag{A.40}$$

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 model calibrated in Section 4 corresponds to $\eta = 0$. If instead $\eta > 0$, the dynamics for λ^{θ}_t are still as in Eq. (4), but the mass $a^{\theta,g}_t$ of accepted researchers of type θ in group $g \in \{f, m\}$ is given by

$$a^{\theta,g}_t = \gamma^{\theta} \sum_{\theta^r : D(\theta^r, \theta) \leq \eta} \lambda^{\theta^r}_{t-1} p^{\theta,g} \tag{A.41}$$

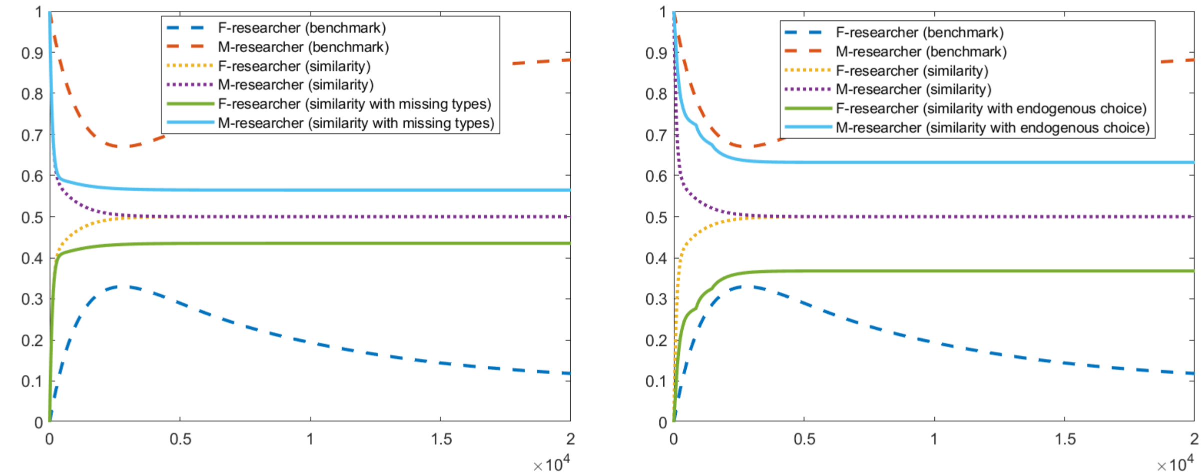
Unfortunately, obtaining general analytical results in this case seems difficult. Therefore, we consider illustrative special cases.

Connected Set of Types The set Θ of types we have considered in the calibration Section 4 enjoys a special structure that is relevant to the relaxed definition of "acceptance" in Eq. (A.40). For every $\eta \geq 1$, and every pair $\theta, \theta' \in \Theta$, there is a finite ordered list $\theta_1, \ldots, \theta_K \in \Theta$ such that $\theta_1 = \theta$, $\theta_K = \theta'$, and $D(\theta_k, \theta_{k+1}) \leq \eta$ for all $k = 1, \ldots, K-1$. In this sense, we say that $\Theta = \{0, 1\}^N$ is η-*connected* for every $\eta \geq 1$. Of course, being 1-connected implies being η-connected for $\eta > 1$; we shall see in the next subsection that a subset of $\{0, 1\}^N$ may be η-connected for some $\eta > 1$, but for any smaller integer η' (including $\eta' = 1$).

With $\Theta = \{0, 1\}^N$, and for the parameter values used in the examples of Section 4, the relaxed acceptance criterion in Eq. (A.40) leads to convergence. For instance, the left panel of Figure A.6 illustrates the parameterization used in Section 4. The dashed lines represent the

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

(a) Missing Types (b) Endogenous Entry



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

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.

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 the left panel of Figure A.6 represent the same parameterization as in the previous subsection, with $\eta = 1$, but applied to a state space Θ obtained by randomly removing 20% of the elements of $\{0, 1\}^N$ and suitably renormalizing probabilities. As expected, the system does not attain group balance in the limit.

A model with a disconnected set of types is related to the partitional model considered at the beginning of this section—it is, in a sense, a generalization of that model, in the context of the specific setting of Section 4. In the partitional model, a referee θ^r accepts type θ if and only if they belong to the same partition element Θ_j. In the similarity model, the distance between θ^r and θ may be greater than η, but there may be some intermediate type θ^ρ that θ^r accepts, and that in turn accepts θ. The similarity-based model essentially allows the sets Θ_j to overlap. The numerical result in this subsection indicates that even with these overlap, convergence to group balance may not be achieved.

A1.5.2.. 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 A1.1.. In this case, even if the connected set of types would lead to convergence (see subsection A1.5.1.), the endogenous entry prevents such convergence, as shown in Section A1.1.1.. 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, respectively, the research similarity case ($\eta = 1$). The solid lines now show the the fraction of M- and F-researchers under research similarity ($\eta = 1$) but with endogenous entry. The intuition is the same as the one given in Section A1.1..

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

A2.. Omitted or sketched proofs

Proof of Corollary 1: Let $D = \sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} \gamma^{\theta'} (p^{\theta',m} + p^{\theta',f})$; then Eqs. (7) can be written as $\bar{\lambda}^{\tilde{\theta},g} = D^{-1} \lambda_0 \gamma^{\tilde{\theta}} p^{\tilde{\theta},g}$ for $\tilde{\theta} = \theta, \theta'$ and $g = f, m$. Since $\theta' = \sigma(\theta)$, $\gamma^{\theta'} = \gamma^{\theta}$. Letting $K = \gamma^\theta D^{-1}$, we have $\bar{\lambda}^{\tilde{\theta},g} = K p^{\tilde{\theta},g}$ for $\tilde{\theta} = \theta, \theta'$ and $g = f, m$. Then

$$\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta',m} = K\left(\lambda_0^{\theta} p^{\theta,m} + \lambda_0^{\theta'} p^{\theta',m}\right) = K\left(\lambda_0^{\theta} p^{\theta',f} + \lambda_0^{\theta'} p^{\theta,f}\right) > K\left(\lambda_0^{\theta'} p^{\theta',f} + \lambda_0^{\theta} p^{\theta,f}\right) = \bar{\lambda}^{\theta,f} + \bar{\lambda}^{\theta',f},$$

where the first equality follows from $p^{\theta,m} = p^{\theta',f}$ and $p^{\theta',m} = p^{\theta,f}$ because $\theta' = \sigma(\theta)$, and the inequality follows because, by assumption, $p^{\theta',f} > p^{\theta,f}$ and $\lambda_0^{\theta} > \lambda_0^{\theta'}$. *Q.E.D.*

Proof of Proposition 4: Write

$$\bar{\Lambda}^g = \sum_{\theta \in \theta^{\max} : \sigma(\theta) = \theta} \bar{\lambda}^{\theta,g} + \sum_{\theta \in \Theta^{\max} : \sigma(\theta) \neq \theta} \bar{\lambda}^{\theta,g} = \sum_{\theta \in \theta^{\max} : \sigma(\theta) = \theta} \bar{\lambda}^{\theta,g} + \frac{1}{2} \sum_{\theta \in \Theta^{\max} : \sigma(\theta) \neq \theta} (\bar{\lambda}^{\theta,g} + \bar{\lambda}^{\sigma(\theta),g}).$$

Letting $D = \sum_{\theta' \in \Theta^{\max}} \lambda_0^{\theta'} \gamma^{\theta'} (p^{\theta',m} + p^{\theta',f})$, $\bar{\lambda}^{\theta,g} = D^{-1} \lambda_0^\theta \gamma^\theta p^{\theta,g}$.

By definition, $\gamma^\theta = \gamma^{\sigma(\theta)}$, $p^{\theta,m} = p^{\sigma(\theta),f}$; moreover, since $\sigma(\sigma(\theta)) = \theta$, also $p^{\sigma(\theta),m} = p^{\theta,f}$.

Thus, consider first $\theta \in \Theta^{\max}$ such that $\sigma(\theta) = \theta$: then $\bar{\lambda}^{\theta,m} = D^{-1}\lambda_0^\theta \gamma^\theta p^{\theta,m} = D^{-1}\lambda_0^\theta \gamma^\theta p^{\sigma(\theta),f} = D^{-1}\lambda_0^\theta \gamma^\theta p^{\theta,f} = \bar{\lambda}^{\theta,f}$. Thus, $\sum_{\theta \in \Theta^{\max}:\sigma(\theta)=\theta} \bar{\lambda}^{\theta,m} = \sum_{\theta \in \Theta^{\max}:\sigma(\theta)=\theta} \bar{\lambda}^{\theta,f}$.

Now consider $\theta \in \Theta^{\max}$ with $\sigma(\theta) \neq \theta$. By the definition of σ, $p^{\theta,m} = p^{\sigma(\theta),f}$. Furthermore, if $p^{\theta,m} = p^{\sigma(\theta),m}$, then also $p^{\theta,m} = p^{\sigma(\theta),m} = p^{\theta,f}$, so $p^{\theta,m} = p^{\sigma(\theta),m} = p^{\theta,f} = p^{\sigma(\theta),f}$ and so $D^{-1}\lambda_0^\theta \gamma^\theta p^{\theta,m} + D^{-1}\lambda_0^{\sigma(\theta)} \gamma^{\sigma(\theta)} p^{\sigma(\theta),m} = D^{-1}\lambda_0^\theta \gamma^\theta p^{\theta,f} + D^{-1}\lambda_0^{\sigma(\theta)} \gamma^{\sigma(\theta)} p^{\sigma(\theta),f}$. Therefore, for such θ, $\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\sigma(\theta),m} = \bar{\lambda}^{\theta,f} + \bar{\lambda}^{\sigma(\theta),f}$.

If $\Theta^{\max}$ is homogeneous, then for all $\theta \in \Theta^{\max}$, either $\sigma(\theta) = \theta$, or $\sigma(\theta) \neq \theta$ but $p^{\theta,m} = p^{\sigma(\theta),m}$. In this case, the above decomposition of $\bar{\Lambda}^g$ implies that $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$ (because the total mass of established researchers is 1).

Otherwise, there is at least one $\theta \in \Theta^{\max}$ for which $\sigma(\theta) \in \Theta^{\max}$ and $p^{\theta,m} \neq p^{\sigma(\theta),m}$, and it is wlog to assume that $p^{\theta,m} > p^{\sigma(\theta),m}$. Corollary 1 shows that, for such types $\theta, \sigma(\theta)$, $\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\sigma(\theta),m} > \bar{\lambda}^{\theta,f} + \bar{\lambda}^{\sigma(\theta),f}$. Therefore, $\bar{\Lambda}^m > \bar{\Lambda}^f$, which implies that $\bar{\Lambda}^m > \frac{1}{2}$. *Q.E.D.*

Proof of Proposition 5: We first claim that, for every $\theta \in \Theta$,

$$a_t^{\theta,m} + a_t^{\sigma(\theta),m} \geq a_t^{\theta,f} + a_t^{\sigma(\theta),f}. \tag{A.42}$$

Notice that, if $\sigma(\theta) = \theta$, the above inequality just says that $a_t^{\theta,m} \geq a_t^{\theta,f}$. Recall that, by the definition and properties of σ, $\gamma^\theta = \gamma^{\sigma(\theta)}$, $p^{\theta,m} = p^{\sigma(\theta),f}$, and $p^{\theta,f} = p^{\sigma(\theta),m}$, so also $p^{\theta,m} + p^{\theta,f} = p^{\sigma(\theta),m} + p^{\sigma(\theta),f}$.

Suppose that $p^{\theta,m} \geq p^{\sigma(\theta),m}$. Then, at time 0, $\lambda_0^\theta = p^{\theta,m} \geq p^{\sigma(\theta),m} = \lambda_0^{\sigma(\theta)} > 0$. Since, in the notation of Theorem 1, $q^\theta = \gamma^\theta(p^{\theta,m} + p^{\theta,f}) + \gamma^{\sigma(\theta)}(p^{\sigma(\theta),m} + p^{\sigma(\theta),f}) = q^{\sigma(\theta)}$, by part 3(a) of that Theorem, for every $t > 0$, $\frac{\lambda_t^\theta}{\lambda_{t-1}^\theta} = \frac{\lambda_t^{\sigma(\theta)}}{\lambda_{t-1}^{\sigma(\theta)}}$, and hence $\frac{\lambda_t^\theta}{\lambda_t^{\sigma(\theta)}} = \frac{\lambda_{t-1}^\theta}{\lambda_{t-1}^{\sigma(\theta)}} = \frac{\lambda_0^\theta}{\lambda_0^{\sigma(\theta)}} \geq 1$. Thus, $\lambda_t^\theta \geq \lambda_t^{\sigma(\theta)}$ for all $t > 0$ as well.

The argument just given applies verbatim if one replaces "$\geq$" in "$p^{\theta,m} \geq p^{\sigma(\theta,m)}$" and "$\lambda_t^\theta \geq \lambda_t^{\sigma(\theta)}$" with "$>$", "$<$" or "$=$." Thus, for all $t \geq 0$, $\lambda_t^\theta > \lambda_t^{\sigma(\theta)}$ if $p^{\theta,m} > p^{\sigma(\theta),m}$; $\lambda_t^\theta < \lambda_t^{\sigma(\theta)}$ if $p^{\theta,m} < p^{\sigma(\theta),m}$; and $\lambda_t^\theta = \lambda_t^{\sigma(\theta)}$ if $p^{\theta,m} = p^{\sigma(\theta),m}$. Therefore, letting $\bar{\gamma} = \gamma^\theta = \gamma^{\sigma(\theta)}$,

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

where the last step follows from $p^{\theta,m} = p^{\sigma(\theta),f}$ and $p^{\theta,f} = p^{\sigma(\theta),m}$.

If $p^{\theta,m} = p^{\sigma(\theta),m}$, then both terms in square brackets equal zero, so equality obtains; in particular, this is true if $\theta = \sigma(\theta)$. If $p^{\theta,m} > p^{\sigma(\theta),m}$, then both terms are positive, and if $p^{\theta,m} < p^{\sigma(\theta),m}$, then both terms are negative. Thus, in all cases, the last inequality, and hence Eq. (A.42), holds, and is strict unless $p^{\theta,m} = p^{\sigma(\theta),m}$; furthermore, if $\theta = \sigma(\theta)$, then $a_t^{\theta,m} = a_t^{\theta,f}$.

Now consider an arbitrary $\bar{\gamma} \in \{\gamma^\theta : \theta \in \Theta\}$. Then

$$\begin{aligned}
\sum_{\theta:\gamma^\theta=\bar{\gamma}} a_t^{\theta,m} &= \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta=\sigma(\theta)} a_t^{\theta,m} + \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta\neq\sigma(\theta)} a_t^{\theta,m} = \\
&= \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta=\sigma(\theta)} a_t^{\theta,m} + \frac{1}{2} \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta\neq\sigma(\theta)} [a_t^{\theta,m} + a_t^{\sigma(\theta),m}] \geq \\
&\geq \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta=\sigma(\theta)} a_t^{\theta,f} + \frac{1}{2} \sum_{\theta:\gamma^\theta=\bar{\gamma},\theta\neq\sigma(\theta)} [a_t^{\theta,f} + a_t^{\sigma(\theta),f}] = \sum_{\theta:\gamma^\theta=\bar{\gamma}} a_t^{\theta,f}.
\end{aligned}$$

The second equality follows from the fact that $\gamma^{\sigma(\theta)} = \gamma^\theta$, so that adding $a_t^{\theta,m} + a_t^{\sigma(\theta),m}$ over all θ with $\theta \neq \sigma(\theta)$ with productivity $\bar{\gamma}$ counts each such type twice. The inequality follows from Eq. (A.42) and the fact that $a_t^{\theta,m} = a_t^{\theta,f}$ if $\theta = \sigma(\theta)$. This inequality is strict if there is some type θ for which $p^{\theta,m} > p^{\sigma(\theta),m}$, which implies $\theta \neq \sigma(theta)$, because in that case we showed above that $a_t^{\theta,m} + a_t^{\sigma(\theta),m} > a_t^{\theta,f} + a_t^{\sigma(\theta),f}$. Finally, the last equality follows by repeating the first two steps backwards, for F-group researchers. *Q.E.D*

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

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

Proof: From Eq. (4), $\lambda_t^\theta = \lambda_t^{\theta,m} + \lambda_t^{\theta,f} = (\lambda_{t-1}^{\theta,m} + \lambda_{t-1}^{\theta,f})(1 - a_t) + \gamma^\theta(p^{\theta,m} + p^{\theta,f})$, which yields the first equation because $\lambda_\tau^\theta > 0$ for all θ and τ. From Eq. (3), for $t \geq 2$,

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

Q.E.D.

Proof of Proposition 6: Let $a_t^g = \sum_{\hat{\theta}} a_t^{\hat{\theta},g}$ and, to simplify the notation, $a^{s,g} = a^{\theta,g} + a^{\theta',g}$ ("s" stands for "symmetric types"). As in the proof of Proposition 5, since by assumption $p^{\theta,m} > p^{\theta',m}$, for all t, $a_t^{s,m} > a_t^{s,f}$. On the other hand, since $\sigma(\theta_1) = \theta_1$ and $\sigma(\theta_0) = \theta_0$, $a_t^{\theta_1,m} = a_t^{\theta_1,f}$, and $a_t^{\theta_0,m} = a_t^{\theta_0,f}$. Therefore, $a_t^m > a_t^f$, which implies that the weight on $\gamma^\theta = \gamma^{\theta'} \equiv \gamma^s$ for accepted M researchers is

$$\frac{a_t^{s,m}}{a_t^m} = 1 - \frac{a_t^{\theta_1,m} + a_t^{\theta_0,m}}{a_t^m} = 1 - \frac{a_t^{\theta_1,f} + a_t^{\theta_0,f}}{a_t^m} > 1 - \frac{a_t^{\theta_1,f} + a_t^{\theta_0,f}}{a_t^f} = \frac{a_t^{s,f}}{a_t^f}.$$

Similarly, $a_t^m > a_t^f$ and $a_t^{\theta_0,m} = a_t^{\theta_0,f}$, $a_t^{\theta_1,m} = a_t^{\theta_1,f}$ imply

$$\frac{a_t^{\theta_0,m}}{a_t^m} < \frac{a_t^{\theta_0,f}}{a_t^f}, \qquad \frac{a_t^{\theta_1,m}}{a_t^m} < \frac{a_t^{\theta_1,f}}{a_t^f}.$$

Moreover, we claim that, $a_t^{\theta_1,g} > a_t^{\theta_0,g}$. For $t = 0$, $a_0^{\theta_1,g} = p^{\theta_1,m}\gamma^{\theta_1}p^{\theta_1,g} > p^{\theta_0,m}\gamma^{\theta_0}p^{\theta_0,g} = a_0^{\theta_0,g}$, because $p^{\theta_0,g} = p^{\theta_1,g}$ but $\gamma^{\theta_1} > \gamma^{\theta_0}$. Inductively, from Lemma A.1 (proof: Appendix A2.),

$$a_t^{\theta_1,g} = a_{t-1}^{\theta_1,g} \cdot \frac{a_t^{\theta_1,g}}{a_{t-1}^{\theta_1,g}} = a_{t-1}^{\theta_1,g}\left(1 - a_{t-1} + \gamma^{\theta_1}(p^{\theta_1,m} + p^{\theta_1,f}\right) > a_{t-1}^{\theta_1,g}\left(1 - a_{t-1} + \gamma^{\theta_0}(p^{\theta_0,m} + p^{\theta_0,f}\right) >$$

$$>a_{t-1}^{\theta_0,g}\left(1 - a_{t-1} + \gamma^{\theta_0}(p^{\theta_0,m} + p^{\theta_0,f}\right) = a_{t-1}^{\theta_0,g}\frac{a_t^{\theta_0,g}}{a_{t-1}^{\theta_0,g}} = a_t^{\theta_0,g}.$$

It then follows that

$$\begin{aligned} 0 <&\frac{a_t^{\theta_0,f}}{a_t^f} - \frac{a_t^{\theta_0,m}}{a_t^m} = \frac{a_t^{\theta_0,f}}{a_t^f} - \frac{a_t^{\theta_0,f}}{a_t^m} < \left(\frac{a_t^{\theta_1,f}}{a_t^{\theta_0,f}}\right)\cdot\left(\frac{a_t^{\theta_0,f}}{a_t^f} - \frac{a_t^{\theta_0,f}}{a_t^m}\right) = \\ =&\frac{a_t^{\theta_1,f}}{a_t^f} - \frac{a_t^{\theta_1,f}}{a_t^m} = \frac{a_t^{\theta_1,f}}{a_t^f} - \frac{a_t^{\theta_1,m}}{a_t^m}; \end{aligned}$$

the first inequality follows from $a_t^{s,f} < a_t^{s,m}$ and $a_t^{\theta_0,f} = a_t^{\theta_0,m}$ and $a_t^{\theta_1,f} = a_t^{\theta_1,m}$, the next equality from $a_t^{\theta_0,m} = a_t^{\theta_0,f}$, the second inequality from $a_t^{\theta_1,f} > a_t^{\theta_0,f} > 0$ and the fact that the difference of fractions is positive, and the last equality from $a_t^{\theta_1,m} = a_t^{\theta_1,f}$. Now

$$E[\gamma|F] = \frac{\gamma^{\theta_0}a_t^{\theta_0,f} + \gamma^s a_t^{s,f} + \gamma^{\theta_1}a_t^{\theta_1,f}}{a_t^f}; \quad E[\gamma|M] = \frac{\gamma^{\theta_0}a_t^{\theta_0,m} + \gamma^s \times a_t^{s,m} + \gamma^{\theta_1}a_t^{\theta_1,m}}{a_t^m}$$

which, since $a_t^{s,g} = 1 - a_t^{\theta_0,g} - a_t^{\theta_1,g}$, implies

$$E[\gamma|F] = -(\gamma^s - \gamma^{\theta_0})\frac{a_t^{\theta_0,f}}{a_t^f} + \gamma^s + (\gamma^{\theta_1} - \gamma^s)\frac{a_t^{\theta_1,f}}{a_t^f}; \quad E[\gamma|M] = -(\gamma^s - \gamma^{\theta_0})\frac{a_t^{\theta_0,m}}{a_t^m} + \gamma^s + (\gamma^{\theta_1} - \gamma^s)\frac{a_t^{\theta_1,m}}{a_t^m}$$

and therefore, since $\frac{a_t^{\theta_1,f}}{a_t^f} - \frac{a_t^{\theta_1,m}}{a_t^m} > \frac{a_t^{\theta_0,f}}{a_t^f} - \frac{a_t^{\theta_0,m}}{a_t^m}$ and $\gamma^s - \gamma^{\theta_0} \leq \gamma^{\theta_1} - \gamma^s$,

$$E[\gamma|F] - E[\gamma|M] = -(\gamma^s - \gamma^{\theta_0})\left(\frac{a_t^{\theta_0,f}}{a_t^f} - \frac{a_t^{\theta_0,m}}{a_t^m}\right) + (\gamma^{\theta_1} - \gamma^s)\left(\frac{a_t^{\theta_1,f}}{a_t^f} - \frac{a_t^{\theta_1,m}}{a_t^m}\right) > 0.$$

This yields (i). For (ii), if $\Theta^{\max}$ is heterogeneous, then it must contain θ and θ' only; and if not, it can only contain θ_1 because $p^{\theta_0,m} + p^{\theta_0,f} = p^{\theta_1,m} + p^{\theta_1,f}$ but $\gamma^{\theta_1} > \gamma^{\theta_0}$. *Q.E.D.*

Proof of Corollary 2: From Eq. 7 in Proposition 3, since $\gamma^{\theta'} = \gamma^{\sigma(\theta)} = \gamma^{\theta}$,

$$\frac{\bar{\lambda}^{\theta,m}}{\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta',m}} = \frac{\lambda_0^{\theta}p^{\theta,m}}{\lambda_0^{\theta}p^{\theta,m} + \lambda_0^{\theta'}p^{\theta',m}} = \frac{[\lambda_0^{\theta}]^2}{[\lambda_0^{\theta}]^2 + [\lambda_0^{\theta'}]^2} > 0.5:$$

the second equality follows from $\lambda_0 = p^m$, and the inequality follows from $\lambda_0^{\theta} > \lambda_0^{\theta'}$. On the other hand,

$$\frac{\bar{\lambda}^{\theta,f}}{\bar{\lambda}^{\theta,f} + \bar{\lambda}^{\theta',f}} = \frac{\lambda_0^{\theta}p^{\theta,f}}{\lambda_0^{\theta}p^{\theta,f} + \lambda_0^{\theta'}p^{\theta',f}} = \frac{p^{\theta,m}p^{\theta,f}}{p^{\theta,m}p^{\theta,f} + p^{\theta',m}p^{\theta',f}} = \frac{p^{\theta,m}p^{\theta,f}}{p^{\theta,m}p^{\theta,f} + p^{\theta,f}p^{\theta,m}} = 0.5:$$

the second equality follows from $\lambda_0 = p^m$, and the third from $p^{\theta',m} = p^{\sigma(\theta'),f} = p^{\theta,f}$ and similarly $p^{\theta',f} = p^{\theta,m}$. The remaining equality and inequality are immediate. *Q.E.D.*

Proof of Corollary 3: From Eq. (7) in Proposition 3,

$$\frac{\bar{\lambda}^{\theta,m}}{\bar{\lambda}^{\theta,m} + \bar{\lambda}^{\theta,f}} = \frac{\lambda_0^\theta p^{\theta,m}}{\lambda_0^\theta p^{\theta,m} + \lambda_0^\theta p^{\theta,f}} = \frac{p^{\theta,m}}{p^{\theta,m} + p^{\theta,f}} = \frac{p^{\theta,m}}{p^{\theta,m} + p^{\theta',m}} = \frac{\lambda_0^\theta}{\lambda_0^\theta + \lambda_0^{\theta'}} > 0.5 :$$

the third equality follwos from $p^{\theta,f} = p^{\sigma(\theta),m} = p^{\theta',m}$ and the fourth from the assumption that $\lambda_0^\theta > \lambda_0^{\theta'}$.

The other equality in the Corollary follows because, if $\lambda_0 = p^m$, then Eq. (7) implies that $\bar{\lambda}^{\theta',f} = \bar{\lambda}^{\theta,m}$ and $\bar{\lambda}^{\theta',m} = \bar{\lambda}^{\theta,f}$. *Q.E.D.*

Proof of Corollary 4: rewrite $F(y)$ as

$$\begin{aligned}
F(y) &= \frac{(y\overline{\lambda}^{\theta,f} + (1-y)\overline{\lambda}^{\theta',f})}{(y\overline{\lambda}^{\theta} + (1-y)\overline{\lambda}^{\theta'})} = \frac{(y \cdot 0.5(\overline{\lambda}^{\theta,f} + \overline{\lambda}^{\theta',f}) + (1-y)0.5(\overline{\lambda}^{\theta,f} + \overline{\lambda}^{\theta',f}))}{(y\overline{\lambda}^{\theta} + (1-y)\overline{\lambda}^{\theta'})} = \\
&= \frac{0.5(\overline{\lambda}^{\theta,f} + \overline{\lambda}^{\theta' f})}{(y\overline{\lambda}^{\theta} + (1-y)\overline{\lambda}^{\theta'})} = \frac{0.5\overline{\gamma}(\overline{\lambda}^{\theta,f} + \overline{\lambda}^{\theta' f})}{P(y)}
\end{aligned}$$

where the second equality follows from Corollary 2, and the fourth from the definition of $P(y)$. This shows that $F(y)P(y)$ is a constant, independent of y. Since $P(y)$ increases in y, $F(y)$ must decrease in y. *Q.E.D.*

Proof of Corollary 5: the argument was given in the main text.

Lemma A.2 *Assume that, for every $\theta \in \Theta$, γ^θ, $p^{\theta,m}$ and $p^{\theta,f}$ are as defined in Section 4. 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 $\rho > \bar{\rho}(\phi, N)$. (Recall that $\bar{\rho}(\cdot)$ is defined in Eq. (14).)*
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.*

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 θ^f, θ^*:

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

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

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

which reduces to the condition in 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. (14) 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.*

Proof of Corollary 6: For part (a), Eq. (15) follows from part 1 of Lemma A.2 and part (i) of Proposition 2. Finally, for Eq. (16), part (ii) of Proposition 3 and the fact that $\lambda_0 = p^m$ and $\gamma^{\theta^m} = \gamma^{\theta^f}$ imply that

$$\begin{aligned}\bar{\lambda}^{\theta^m,m} &= \frac{\lambda_0^{\theta^m}\gamma^{\theta^m}p^{\theta^m,m}}{\lambda_0^{\theta^m}\gamma^{\theta^m}(p^{\theta^m,m} + p^{\theta^m,f}) + \lambda_0^{\theta^f}\gamma^{\theta^f}(p^{\theta^f,m} + p^{\theta^f,f})} = \\ &= \frac{\phi^{2N}}{\phi^N(\phi^N + (1-\phi)^N) + (1-\phi)^N((1-\phi)^N + \phi)^N} = \frac{\phi^{2N}}{(\phi^N + (1-\phi)^N)^2}\end{aligned}$$

and analogously

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

Therefore,

$$\bar{\Lambda}^m = \bar{\lambda}^{\theta^m,m} + \bar{\lambda}^{\theta^f,m} = \frac{\phi^{2N} + (1-\phi)^{2N}}{(\phi^N + (1-\phi)^N)^2} = \frac{1 + \left(\frac{\phi}{1-\phi}\right)^{2N}}{1 + 2\left(\frac{\phi}{1-\phi}\right)^N + \left(\frac{\phi}{1-\phi}\right)^{2N}}$$

and Proposition 4 ensures that $\bar{\Lambda}^m > 0.5$ regardless of the parameterization. This completes the proof of part (a).

For part (b), part 1 of Lemma A.2 and part (i) of Proposition 2 imply that $\Theta^{\max} = \{\theta^*\}$. Then, since $p^{\theta^*,m} = p^{\theta^*,f} = \phi^{N/2}(1-\phi)^{N/2}$, part (ii) in Proposition 3 implies that $\bar{\lambda}^{\theta^*,m} = \bar{\lambda}^{\theta^*,f} = \frac{1}{2}$. Consequently, $\bar{\Lambda}^m = \bar{\Lambda}^f = \frac{1}{2}$ as well, and these conclusions are independent of λ_0.

Part (c) now follows from part 3 of Lemma A.2 and part (a) of this Corollary. Finally, part (d) follows from Eq. (16) by taking limits as $N \to \infty$. *Q.E.D.*

A3.. Proof of the results in Online Appendix A1.

Proof of Proposition A.1: 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 $\lambda^m_{\tau,j}, \lambda^f_{\tau,j}, \lambda_{\tau,j} \in \mathbb{R}^\Theta_+$ such that

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

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

(iii) $\lambda^\theta_{\tau,j} < \frac{C}{\gamma^\theta P}$ for $0 \leq \tau \leq t(j) - t(j-1)$ and all $\theta \in \Theta \setminus \Theta_j$, and $\lambda^{\theta_0}_{t(j)-t(j-1),j} < \frac{C}{\gamma^{\theta_0} P}$ 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. (19)–(19) 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 $\lambda^{\theta_0}_{\bar\tau,k} < \frac{C}{\gamma^{\theta_0} P}$ and $\lambda^\theta_{\tau,k} \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_k$ and $0 \leq \tau < \bar\tau$. Let $t(k) = t(k-1) + \bar\tau$. Then, for each group $g = f, m$, the dynamics in Eqs. (19)–(19) induced by the vectors q^f_k, q^m_k for the subsequence $(\lambda^g_{\tau,k})_{\tau=0,\dots,\bar\tau}$ coincide with those in Eq. (A.29) for the subsequences $(\lambda^g_t)_{t=t(k-1),\dots,t(k)}$; 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 $\theta \in \Theta \setminus \Theta_k$, $q^{\theta,m}_k + q^{\theta,f}_k \leq q^{\theta'}_{m,k} + q^{\theta'}_{f,k}$, which by part 3(a) in Theorem 1 implies that $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k} \leq \lambda^{\theta'}_{\tau+1,k}/\lambda^{\theta'}_{\tau,k}$. Therefore, it must be the case that $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k} \leq 1$ for all $\theta \in \Theta \setminus \Theta_k$: 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 $\tau = 0, \dots, \bar\tau$ and for any such θ. Thus, in this case Conditions $C(k)$ hold.

If instead $\lambda^\theta_{\tau,k} \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_k$ and $\tau \geq 0$, then for each group $g = f, m$, the dynamics in Eqs. (19)–(19) 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. (A.29) 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, $\lambda^\theta_{\tau,K} \geq \frac{C}{\gamma^\theta P}$ for all $\theta \in \Theta_K$ and $\tau \geq 0$. For $k = 0, \dots, K$, let $\Theta^{\max}_k = \arg\max\{q^{\theta,m}_k + q^{\theta,f}_k : \theta \in \Theta\}$. Since by definition $q^{\theta,g}_k = \gamma^\theta p^{\theta,g} > 0$ for $\theta \in \Theta_k$ and $q^{\theta,g}_k = 0$ for $\theta \in \Theta \setminus \Theta_k$, $\Theta^{\max}_k =$

$\arg\max\{\gamma^\theta(p^{\theta,m}+p^{\theta,f}) : \theta\in\Theta_k\}$. In particular, by the definition of Θ_0, $\Theta_0^{\max}=\Theta^{\max}$, where $\Theta^{\max}$ is as defined in Eq. (A.30).

For every $k=0,\ldots,K-1$, and every $\theta\in\Theta_k^{\max}$, $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k}\geq 1$ for $0\leq\tau<t(k)-t(k)$; otherwise, by part 3(a) in Theorem 1 and the definition of $\Theta_k^{\max}$, $\lambda^\theta_{\tau+1,k}/\lambda^\theta_{\tau,k}<1$ for all such τ and all θ, so $\sum_{\theta\in\Theta}\lambda^\theta_{\tau+1,k}<\sum_{\theta\in\Theta}\lambda^\theta_{\tau,k}=1$, which contradicts the fact that $\lambda_{\tau+1}\in\Delta(\Theta)$ per Theorem 1.

It follows that, for every $\theta\in\Theta_0^{\max}$,

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

so $\theta\in\Theta_k$ for all $k=0,\ldots,K$. Therefore, since $\Theta_0\supsetneq\Theta_1\supsetneq\ldots\supsetneq\Theta_K$ and $\Theta_k^{\max}=\arg\max\{\gamma^\theta(p^{\theta,m}+p^{\theta,f}):\theta\in\Theta_k\}$, $\Theta_0^{\max}=\Theta_k^{\max}$ for all $0\leq k\leq K$.[1] Hence $\Theta^{\max}=\Theta_K^{\max}$.

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

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

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

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

which is Eq. (A.31). Similarly, for $\theta\in\Theta^{\max}$, letting $q^\theta_K=q^{\theta,m}_K+q^{\theta,g}_K$, part (5) in the same Theorem implies that

$$\bar\lambda^{\theta,m}=\bar\lambda^{\theta,m}_K=\frac{\lambda^\theta_{0,K}q^{\theta,m}_K}{\sum_{\theta'\in\Theta^{\max}}\lambda^{\theta'}_{0,K}q^{\theta'}_K}=\frac{q^{\theta,m}_K}{\sum_{\theta'\in\Theta^{\max}}\frac{\lambda^{\theta'}_{0,K}}{\lambda^\theta_{0,K}}q^{\theta'}_K}=\frac{q^{\theta,m}_K}{\sum_{\theta'\in\Theta^{\max}}\frac{\lambda^{\theta'}_0}{\lambda^\theta_0}q^{\theta'}_K}=\frac{\lambda^\theta_0 q^{\theta,m}_K}{\sum_{\theta'\in\Theta^{\max}}\lambda^{\theta'}_0 q^{\theta'}_K},$$

and analogously for $\bar\lambda^{\theta,f}$. Since, for $\theta\in\Theta_k^{\max}$, $q^{\theta,g}_K=\gamma^\theta p^{\theta,g}$, this yields Eq. (A.32). *Q.E.D.*

Proof of Corollary A.1: since $\lambda_0=p^m$, $\lambda^\theta_0=p^{\theta,m}>p^{\theta,f}=p^{\sigma(\theta),m}=\lambda_0^{\sigma(\theta),m}$; thus, for C sufficiently small, $\sigma(\theta)\notin\Theta^{\max}$. In this case, from Eq. (A.32), $\bar\lambda^{\theta,m}/\bar\lambda^{\theta,f}=p^{\theta,m}/p^{\theta,f}>1$, so $\bar\lambda^{\theta,m}>\bar\lambda^{\theta,f}$. Furthermore, if $\Theta^{\max}=\{\theta\}$, then Eq. (A.32) reduces to $\bar\lambda^{\theta,g}=\frac{p^{\theta,g}}{p^{\theta,m}+p^{\theta,g}}$, which implies that the first displayed equation in the Corollary holds.

[1] Pick $\theta_k\in\Theta_k^{\max}$ arbitrarily. For every k, we just showed that $\theta_0\in\Theta_k$; furthermore, $\Theta_k\subsetneq\Theta_0$. Since $\theta_0,\theta_k\in\Theta_0$ and $\theta_0\in\Theta_0^{\max}$, $\gamma^{\theta_0}(p^{\theta_0,m}+p^{\theta_0,f})\geq\gamma^{\theta_k}(p^{\theta_k,m}+p^{\theta_k,f})$. Since $\theta_0,\theta_k\in\Theta_k$ and $\theta_k\in\Theta_k^{\max}$, $\gamma^{\theta_k}(p^{\theta_k,m}+p^{\theta_k,f})\geq\gamma^{\theta_0}(p^{\theta_0,m}+p^{\theta_0,f})$. Therefore $\theta_0\in\Theta_k$ and $\theta_k\in\Theta_0$, which implies the claim.

To prove the last inequality, let $D = \lambda_0^{\theta}\gamma p^{\theta,m} + \lambda_0^{\theta}\gamma p^{\theta,f} + \lambda_0^{\theta'}\gamma p^{\theta',m} + \lambda_0^{\theta'}\gamma p^{\theta',f}$ where $\gamma = \gamma^{\theta} = \gamma^{\theta'}$. Then, from Proposition 3 part (ii),

$$\begin{aligned}
\left(\bar{\lambda}_{C=0}^{\theta,m} + \bar{\lambda}_{C=0}^{\theta',m}\right) - \left(\bar{\lambda}_{C=0}^{\theta,f} + \bar{\lambda}_{C=0}^{\theta',f}\right) &= \left(\frac{\lambda_0^{\theta}\gamma p^{\theta,m}}{D} + \frac{\lambda_0^{\theta'}\gamma p^{\theta',m}}{D}\right) - \left(\frac{\lambda_0^{\theta}\gamma p^{\theta,f}}{D} + \frac{\lambda_0^{\theta'}\gamma p^{\theta',f}}{D}\right) \\
&= \left(\frac{\lambda_0^{\theta}\gamma\left(p^{\theta,m} - p^{\theta,f}\right)}{D}\right) - \left(\frac{\lambda_0^{\theta'}\gamma\left(p^{\theta',f} - p^{\theta',m}\right)}{D}\right) \\
&= \left(p^{\theta,m} - p^{\theta,f}\right)\left[\frac{\left(\lambda_0^{\theta} - \lambda_0^{\theta'}\right)\gamma}{\lambda_0^{\theta}\gamma p^{\theta,m} + \lambda_0^{\theta}\gamma p^{\theta,f} + \lambda_0^{\theta'}\gamma p^{\theta',m} + \lambda_0^{\theta'}\gamma p^{\theta',f}}\right] \\
&= \frac{\left(p^{\theta,m} - p^{\theta,f}\right)}{\left(p^{\theta,m} + p^{\theta,f}\right)}\left[\frac{\lambda_0^{\theta} - \lambda_0^{\theta'}}{\lambda_0^{\theta} + \lambda_0^{\theta'}}\right]:
\end{aligned}$$

the third equality follows by noting that, in the second term in parentheses, $p^{\theta',f} - p^{\theta',m} = p^{\sigma(\theta),f} - p^{\sigma(\theta),m} = p^{\theta,m} - p^{\theta,f}$; and the fourth equality follows similarly, by noting that the two terms of the denominator D multiplied by $\lambda_0^{\theta'}$ are $p^{\theta',m} = p^{\theta,f}$ and $p^{\theta',f} = p^{\theta,m}$. Finally, the first term in the last line equals $\bar{\lambda}^{\theta,m} - \bar{\lambda}^{\theta,f}$ when $\Theta^{\max} = \{\theta\}$, and the second term is between 0 and 1 because we assume that $\lambda_0 = p^m$ and $p^{\theta,m} > p^{\theta,f} = p^{\theta',m}$. This yields the required conclusion. *Q.E.D.*

Proof of Corollary A.2: Assume that $\lambda_0 = p^m$. In (a.1), the assumption implies that $\Theta^{\max} = \{\theta^m, \theta^f\} \subseteq \Theta_0$. Substituting $\lambda_0^{\theta^m} = \phi^N$ and $\lambda_0^{\theta^f} = (1-\phi)^N$ in Eq. (A.31) yields $\bar{\lambda}^{\theta^m} = \frac{\phi^N}{\phi^N + (1-\phi)^N}$. Similarly, substituting for γ^{θ}, p^m and p^f in Eq. (A.32) yields the same expression for $\bar{\lambda}^{\theta^m,g}$ and $\bar{\lambda}^{\theta^f,g}$, $g = m, f$, as in Corollary 6. This yields the required expression for $\bar{\Lambda}^m$.

For (a.2), the assumption implies that $\Theta^{\max} = \{\theta^m\}$. This immediately implies that $\bar{\lambda}^{\theta^m} = 1$. Furthermore, from Eq. (A.32), $\bar{\Lambda}^m = \bar{\lambda}^{\theta^m,m} = \frac{\gamma^{\theta^m} p^{\theta^m,m}}{\gamma^{\theta^m}(p^{\theta^m,m} + p^{\theta^m,f})} = \frac{p^{\theta^m,m}}{p^{\theta^m,m} + p^{\theta^m,f}} = \frac{\phi^N}{\phi^N + (1-\phi)^N}$, as asserted. Finally, we compare this quantity with its counterpart in Eq. (16):

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

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

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

The statements about t^{θ} for $\theta \notin \Theta^{\max}$ follow from the construction of $t(0), \ldots, t(K)$ in the proof of Proposition A.1. *Q.E.D.*

Proof of Proposition A.2. For part 1, the key step is analogous to the proof of Proposition 5, 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 Corollary A.2, 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 A.1, $\frac{\lambda_t^{\theta}}{\lambda_{t-1}^{\theta}} = \frac{\lambda_t^{\theta^{\text{sym}}}}{\lambda_{t-1}^{\theta^{\text{sym}}}}$, 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 A.2, $\lambda_t^{\theta} = \lambda_t^{\theta^{\text{sym}}}$ 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 A.2, and in addition $p^{\theta > p_m^{\theta^{\text{sym}}},m}$. *Q.E.D.*

Proof of Proposition A.3 Let θ^a and θ^b be the types of the two young researchers. We assume that the type of the joint project is the elementwise maximum of θ^a and θ^b: that is, the project displays characteristics i if and only if at least one of the researchers displays it.

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

$$p^{\theta,m} = p^{\bar{\theta},f} \qquad \text{and} \quad p^{\theta',f} = p^{\bar{\theta}',m}. \tag{A.43}$$

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

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

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

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

First,

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

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

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

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

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

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

as asserted.

Q.E.D