

# UC Berkeley

## UC Berkeley Electronic Theses and Dissertations

### Title

Gatekeepers and Trailblazers: Essays in Microeconomics

### Permalink

<https://escholarship.org/uc/item/8tj1x6q3>

### Author

Muller-Itten, Michele C.

### Publication Date

2016

Peer reviewed|Thesis/dissertation

**Gatekeepers and Trailblazers: Essays in Microeconomics**

by

Michèle C Müller-Itten

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Haluk Ergin, Co-chair  
Professor William Fuchs, Co-chair  
Professor David Ahn  
Professor Brett Green

Fall 2016

**Gatekeepers and Trailblazers: Essays in Microeconomics**

Copyright 2016  
by  
Michèle C Müller-Itten

## Abstract

Gatekeepers and Trailblazers: Essays in Microeconomics

by

Michèle C Müller-Itten

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Haluk Ergin, Co-chair

Professor William Fuchs, Co-chair

This thesis studies dynamic markets with endogenous entry, where access hurdles are either an explicit feature of the market or an implicit result of productive externalities.

The first chapter concerns markets of influence, where formal gatekeepers are in charge of selecting and promoting the most promising ideas. I model this as a two-sided matching market between a continuum of experts and a finite number of gatekeepers under sequential directed search. Real-world examples include academic publishing, venture capitalism or political agenda setting. Uniqueness of the resulting equilibrium allows for clear-cut predictions: First, sorting may fail in equilibrium. Second, gatekeepers may have an incentive to add artificial delay. Such red tape occurs in equilibrium – and only at the top – when the impact of two gatekeepers is very different. Third, artificial delay may improve equilibrium sorting and thereby enhance welfare. Finally, the bottom gatekeeper may endogenously specialize on a quality-irrelevant attribute.

In other applications, there are no formal gatekeepers who regulate market access, yet endogenous access hurdles play a similarly selective role. In this vein of research, the second chapter studies the evolution of labor market composition under mentoring externalities. This chapter is joint work with Aniko Oery. We provide a continuous time, overlapping generations framework to analyze the costs and benefits of affirmative action policies if mentoring complementarities are present. Senior workers reduce the young population's educational cost through mentoring, and thus act as 'trailblazers' for future generations. In such a framework, the main trade off is between using the most able workers and reducing mentorship misallocation. We identify conditions under which persistent market intervention is warranted to improve long-term surplus. We also contrast different channels through which the planner can affect market outcomes, and highlight the benefits of educational fellowships over hiring restrictions.

To my parents Claudia and Alex Itten.

# Contents

<b>1</b>	<b>Gatekeeping under Asymmetric Information</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	The Model . . . . .	5
1.3	Equilibrium Analysis . . . . .	8
1.4	Sorting Failures . . . . .	15
1.5	Gatekeeper Competition . . . . .	17
1.6	Applications and Potential Extensions . . . . .	27
1.7	Conclusion . . . . .	31
<b>2</b>	<b>Mentoring and the Dynamics of Affirmative Action</b>	<b>33</b>
2.1	Model . . . . .	36
2.2	Affirmative Action Policy Instruments . . . . .	43
2.3	Conclusion . . . . .	47
	<b>Bibliography</b>	<b>48</b>
<b>A</b>	<b>Supplementary Material to Chapter 1</b>	<b>52</b>
A.1	Additional Proofs . . . . .	52
A.2	Comparative Statics . . . . .	69
<b>B</b>	<b>Supplementary Material to Chapter 2</b>	<b>72</b>

## Acknowledgments


I would not have been able to write this thesis without the tremendous support of my advisors, colleagues, my family, and friends.

I thank my thesis committee Haluk Ergin, William Fuchs, David Ahn and Brett Green for their invaluable support not only in shaping this research agenda, but also in discussing the less fruitful ideas that ultimately didn't make it to this stage. It has been an honor to learn from you and observe your razor sharp minds at work even when I presented you with half-fleshed ideas. I specifically thank Haluk Ergin for spontaneous and open-ended research discussions whenever he was on campus. I thank William Fuchs for stepping in during the other times, and especially for his encouragement and strategic support during the job market. My thanks also go to David Ahn for his honest advice and practical support throughout my studies, GSI appointments and job search. I deeply appreciate his initiative to create the theory lunch, which was critical for my transition from student to researcher. Brett Green joined my committee shortly before the job market, and his contagious enthusiasm was a wonderful motivation boost during a stressful time.

I am indebted to Shachar Kariv, Adam Szeidl and Botond Köszegi for guiding my first steps in active research by including me in their own projects during my first summer at Berkeley. I thank Chris Shannon and Ulrike Malmendier for being trail-blazers for female economists in academia, and for their personal advice and support whenever needed. I also want to acknowledge Ben Hermalin, Bob Anderson, Alper Atamturk, Ted Miguel, Fred Finan, Tymofiy Mylovanov and Yuichiro Kamada for their research input at various stages of my PhD. My research was generously supported by a Fulbright Fellowship from the US Department of State, a grant from the National Science Foundation (SES 1227707) and my own personal role-model, trail-blazer and (actual) godmother Corinne Itten.

My time at Berkeley was marked by great colleagues and friendships. You have sharpened my research, preserved my sanity, and will be responsible for my lifelong nostalgia for this place. Aniko Oery, Tadeja Gracner, Satoshi Fukuda, Joseph Root, Alisa Tazhitdinova, Ivan Balbuzanov, Takeshi Murooka, Liang Bai, Aluma Dembo, Matthew Leister, Omar Nayeem, Sandra Polanía Reyes, David Echeverry, Patrick and Cathy Lapid, Brad and Jasmin Miller, TJ Tsai, Erice Liu, Christina Vo, Caitlin Appert, Dave Nguyen, Amy Banas, Maciej Kotoski, Antoine Peiffer and Anna Baranskaya – thank you for the wonderful times we shared. I am undoubtedly forgetting to name important people here, which is really a testament to the abundance of blessings you all have been for me. To avoid future embarrassment, let me thus draw a blank to fill in as needed, because really, you are the best, \_\_\_\_\_.

All of the above people were crucial to the completion of this dissertation, but there is one person without whom I wouldn't have written any dissertation at all: Marc, given the incredible support, love and encouragement you have been to me throughout this adventure, it is only fitting that this thesis carries your name too. Thank you for delighting in my abilities, for inspiring me through our conversations, for pulling your full weight at home, for cherishing and investing in our kids. I adore you.

To my two sons, who put every other accomplishment to shame: Thank you from the bottom of my heart. Micah, you changed the rhythm of my life and the face of my research. Now there are little trains all over my proof sketches. Look, here's one just for you!  – And Jonah, my little darling, thank you for accompanying me throughout the craziness of the job market. Your constant presence made the distant travels less daunting, and your gentle kicks reminded me that any place will be a happy place with you three by my side. Thank you, kiddos, for your patience when mommy had to work, and for filling my free time with laughter, stories and never-ending laundry – you have stolen time away from the worries that plagued many of my colleagues.

Lastly, I want to thank my own mom and dad for instilling in me the curiosity, self confidence and optimism I needed to pursue this degree. I know Berkeley is too far away for your comfort, yet you've never failed to tell me how proud you are of me. I feel close to you despite the distance, and I cherish every moment we have together. You are an amazing team, and this thesis is dedicated to you.



# Chapter 1

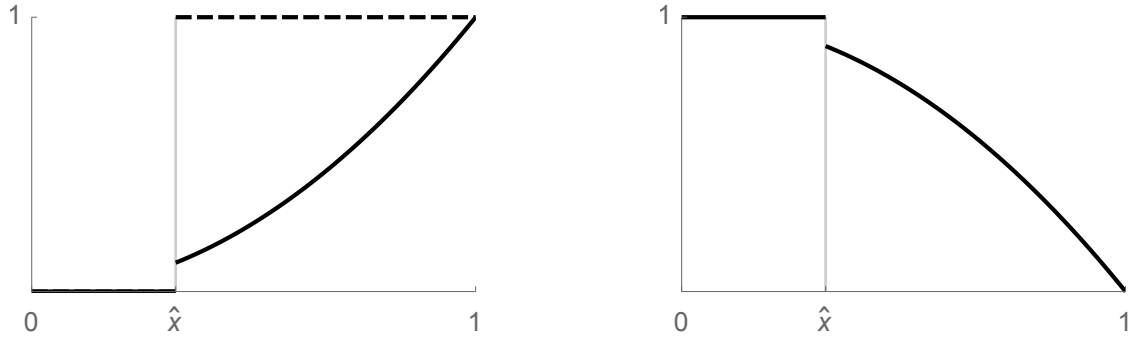
## Gatekeeping under Asymmetric Information

### 1.1 Introduction

Access to influence is often controlled by a small number of gatekeepers: Political motions must be endorsed by a member of parliament, academics vie for publication in a handful of top scientific journals, and hosts of startup companies compete for financial support from a limited number of venture capitalists. In all of these applications, market participants are heterogeneous with respect to their desirability and information is imperfect and often asymmetric. The impact of this informational friction is felt on both sides of the market: Uninformed gatekeepers have to base all decisions on noisy signals, and experts on the other side can no longer fully anticipate match outcomes. This risk of rejection forces experts to strategically target their search, which affects the equilibrium both in terms of match outcomes and strategic incentives.

Understanding the allocation process is crucial since realized matches ultimately shape public opinion, lawmaking, the direction of future research and technological progress. Yet, several relevant questions remain without satisfactory answers: For instance, under what conditions is matching assortative? How can gatekeepers attract better matches? Do these strategic considerations differ for the best and the worst gatekeeper? What are the welfare implications of this nonstandard competitive environment?

To address these questions, this paper proposes a tractable dynamic matching model where a continuum of experts seek to promote ideas through one of finitely many channels, each of which is controlled by a separate gatekeeper. An expert can propose to at most one gatekeeper at any given time, and then has to stand by idly while the gatekeeper makes her final match decision. Being unable to fully anticipate the success of a given proposal and eager to find a match as quickly as possible, the impatient expert thus has to balance high-impact options against less competitive ‘safe bets’. Three characteristics determine his optimal proposal order over gatekeepers: the impact of each channel, response times, and his



- (a) The top gatekeeper receives proposal of exactly those experts with high enough quality  $x \geq \hat{x}$  (dashed line). Among these proposals, better experts generate better signals and thus match more frequently (solid line).
- (b) The bottom gatekeeper receives proposals from experts with quality below  $\hat{x}$  as well as from those that were previously rejected by the top gatekeeper. Since she accepts all proposals, her proposal and match distribution coincide.

**Figure 1.1:** Proposal and match distribution for the stylized example discussed in the introduction.

equilibrium acceptance odds. Since time is particularly precious to an expert with several promising avenues, the cost of each proposal is endogenously determined.

A gatekeeper's objective is straightforward: She seeks to match with the highest-quality experts subject to a finite flow capacity. Due to informational frictions, she however has to base her decision on an imperfect signal of expert quality – taking into account that her acceptance criteria not only affects which proposals she *accepts*, but also who *proposes* to her in the first place. Contrary to the impatient experts, the long-lived gatekeepers do not discount the future.

**Example 1.** To illustrate how experts' directed search affects match outcomes, consider a stylized case with two gatekeepers. The impact of the bottom gatekeeper is normalized to one and she *immediately* accepts *all* proposals. The top gatekeeper has an impact of two, but takes six months before deciding on any proposal and only accepts those with a high enough signal. As a result, only the highest-quality experts propose to the top gatekeeper first: Their signal draws are usually high, so they expect their proposal to be accepted. Lower-quality experts instead prefer to match immediately with the bottom gatekeeper in order to avoid the delay of a rejected proposal.

Consequently, the top gatekeeper here only receives proposals from the upper tail of the quality spectrum. She however rejects even good experts with positive probability. These experts go on to propose to the bottom gatekeeper, who therefore receives proposals of all qualities, as illustrated in [Figure 1.1](#).  $\diamond$

The discussion of expert incentives in this simple setup may suggest that sorting naturally ensues in equilibrium, in the sense that high impact gatekeepers are more selective than their

low-impact counterparts and that only the best experts first propose to the top gatekeeper. When endogenizing gatekeeper’s acceptance strategies however, both sides of the market may display nonmonotone equilibrium behavior. Specifically, I identify limiting conditions that either guarantee or prevent each of the sorting failures and illustrate via examples how these situations may occur under realistic parameter values. This serves as a cautionary tale when an uninformed outsider makes inferences from observed equilibrium behavior. In doing so naively, he may well end up drawing false conclusions on agents’ actual desirability.

The strategic targeting by experts shapes gatekeepers’ proposal distribution, and it seems natural that they may seek to exploit it in their favor. In the simple example of [Figure 1.1](#), experts’ self-selection essentially filters out the lowest-quality proposals for the top gatekeeper. The filtering here is perfect due to the unconditional acceptance by the bottom gatekeeper. In an actual equilibrium, this effect is less pronounced but the basic intuition maintains: A gatekeeper may prefer to *decrease* her own desirability in order to reduce her inflow of proposals, but then accept the remaining proposals *more easily*. By guiding experts’ proposal strategy through incentives, she partially circumvents her informational disadvantage.

To incorporate such manipulation, I allow gatekeepers to artificially delay acceptance decisions beyond the (exogenously given) time required for signal generation, which I refer to as ‘red tape’. In a setup with two gatekeepers, I show that such red tape occurs in equilibrium and only at the top, exactly when the impact differential is large. In reality, multiple forms of red tape are imaginable, ranging from unnecessary delay over specific customization requirements to gratuitous access hurdles that make it more costly to contact to a gatekeeper.<sup>1</sup> The intuition presented here applies to all of them, even though the exact impact on the match distribution may vary slightly.

It is worth highlighting that these results obtain even when signal generation is completely costless. If reviewing proposals is instead taxing for gatekeepers, this yields an additional reason to employ disincentives: Fewer proposals means less resources are wasted to evaluate experts that are ultimately rejected. Any such consideration would only strengthen the result.

The existence of red tape has important efficiency implications and may come at a significant social cost. Nevertheless, I identify conditions under which red tape improves match assortativity and thereby actually *raises* welfare. As such, even when technological innovations allow drastic reductions in response times, it may not be socially optimal to fully exploit these possibilities.

When experts differ on an observable, quality-irrelevant attribute, gatekeepers have yet another way of exploiting expert self-selection: By conditioning proposal acceptance on the

---

<sup>1</sup>Monetary proposal costs also act as a (more efficient) form of red tape. Indeed, such “money burning” actually completely discourages the worst experts from proposing, in addition to pushing some intermediate ones towards other gatekeepers first. However, gatekeepers may be opposed to such policies for fairness considerations if experts face heterogeneous budget constraints ([Cotton, 2013](#)) or if proposal costs are borne by someone other than the expert, for example because research departments directly reimburse journal submission costs.

attribute, the gatekeeper becomes more easily accessible to experts within a target group, and in turn receives more proposals from that group. Such horizontal differentiation occurs for instance within academic publishing, where lower-ranked journals often *specialize* in a specific sub-discipline while the top journals span the entire field. This asymmetry can be motivated from the nonstandard competitive framework that they are participating in.

Indeed, there is an immediate negative effect from specialization: Nontarget proposals are now rejected to make room for target proposals with a *lower* signal. However, as experts react to these new acceptance rules, the gatekeeper also secures additional proposals in the target group and loses some nontarget proposals. As such, specialization partially replaces a gatekeeper's own proposals with those that previously went to her competitor first. Now, since experts from the upper tail of the quality distribution first propose to the top gatekeeper, the sign of this indirect effect depends on the strategic position of the gatekeeper: It is *positive* for the bottom gatekeeper, but *negative* for the top (who would rather conserve her own proposals). As such, the top gatekeeper doesn't have any incentives for specialization, but the bottom gatekeeper endogenously limits his scope when the indirect effect outweighs.

Unfortunately, equilibrium strategies are less straightforward since the top gatekeeper may still *react* to a differential acceptance policy of her competitor, even if she never unilaterally specializes. As such, the *asymmetry* between gatekeepers' taste for specialization is an off-equilibrium result – the endogenous *occurrence* of specialization however maintains even in equilibrium.

**Related Literature.** Equilibrium sorting is a central concern in the decentralized matching literature. Limiting my discussion to the case of non-transferable utility where match surplus is divided exogenously, [Gale and Shapley \(1962\)](#) establish that aligned preferences alone are sufficient to ensure assortative matching in the absence of search frictions. When search is time-consuming but match opportunities arise randomly, [Smith \(2006\)](#) shows that assortative matching ensues under the preference structure considered here: Equilibrium matches are characterized by an interval-partition over the type space, in the sense that agents within the same interval match up while all others keep searching. Stochastic assortative matching remains an equilibrium even under informational frictions, where agents merely observe a noisy signal of their potential partner's type ([Chade, 2006](#)). The introduction of directed search however breaks assortativity, as [Chade et al. \(2014\)](#) show in closely related model with costly but *simultaneous*, one-shot search. These nonmonotonic equilibrium strategies maintain under the *sequential* search considered here. However, they no longer occur independently for the two sides. More generally, the main advantage of the present framework lies in its tractability: Contrary to the former model, equilibrium is (essentially) unique and well defined for any finite number of gatekeepers. It is thanks to these clear equilibrium predictions that a meaningful discussion of gatekeepers' competitive incentives becomes possible. Moreover, the multiplicity in their setting is generated by the fact that experts can make proposals at a *single point in time* – they cannot later pursue new options if their initial proposals do not work out. In applications where no such hard

deadline exists, the current framework may thus offer a better fit.

Previous papers have explored competition among academic journals, albeit in a partial equilibrium setup.<sup>2</sup> The driving factor in these models is the assumption of a *costly* refereeing process, resulting in direct savings from *any* reduction in submission rates. In contrast, gatekeepers here are motivated purely by improvements in their match outcomes and as such the present analysis supplements, rather than restates, these previous findings. Moreover, by embedding these tradeoffs into a general equilibrium framework, I am able to identify an asymmetry which is consistent with the empirical observation that editorial delays are substantially longer at top journals.

The paper is organized as follows: [Section 1.2](#) establishes the formal model and [Section 1.3](#) identifies the unique equilibrium, which may or may not entail sorting failures depending on conditions specified in [Section 1.4](#). Extending the model, [Section 1.5](#) then analyzes gatekeeper behavior in this non-standard competitive environment. [Section 1.6](#) highlights several applications of the model and discusses potential extensions motivated by their idiosyncrasies.

## 1.2 The Model

This paper studies a decentralized, two-sided matching market with heterogeneity on both sides and a common preference. What sets it apart from the classic setup is the strategic meeting technology and the asymmetric information structure.

Formally, I consider a continuous time game where short-lived, impatient *experts* seek to broadcast their ideas. For expositional simplicity, I refer to an expert by the quality  $x \in X = [0, 1]$  of his idea. Experts are born continuously at unit flow rate: During each time interval  $dt$ , a mass  $dt$  of experts arrive according to distribution  $F$ . I assume that  $F$  is continuous with full support over the quality space  $X$  and admits a differentiable density  $f$ .

On the other side, let  $J$  denote the finite set of long-lived *gatekeepers* who control separate channels of influence. The size of gatekeeper  $j$ 's sphere of influence, or her *impact*, is publicly known and denoted by the exogenous parameter  $\gamma_j > 0$ .<sup>3</sup> Within her limited per-period capacity  $\kappa_j \in [0, 1]$ , a gatekeeper seeks to accept (and promote) the most worthwhile ideas.<sup>4</sup>

---

<sup>2</sup>Weitzman (1979), Oster (1980), Heintzelman and Nocetti (2009) and Salinas and Munch (2015) characterize the optimal submission strategies for academic authors. The decision problem they solve is identical to the one faced by individual experts in this paper. Baghestanian and Popov (2015) further endogenize authors' effort level and explore the impact of exogenous changes in the publishing process. Leslie (2005), Azar (2005, 2007) and Cotton (2013) discuss the possible benefits of long editorial decision times and monetary submission costs at a single journal.

<sup>3</sup>In [Section 1.6](#), I discuss an extension where impact is endogenously determined.

<sup>4</sup>The qualitative results carry over if gatekeepers' cardinal utility ranking is different – as long as they all share the same *ordinal* preference which is strictly increasing in  $x$ , even the worst expert is weekly better than a vacancy and exceeding capacity is costly enough to discourage additional matches even with the most attractive expert  $x = 1$ . For simplicity, I henceforth assume that the utility from accepting expert  $x$  is exactly equal to  $x$ .

Total capacity is insufficient for all proposals,  $\sum_{j \in J} \kappa_j < 1$ , which rules out trivial situations where the least attractive gatekeeper summarily accepts all proposals.

I abstract away from any goodness-of-fit issues by assuming that guaranteed and instantaneous acceptance by gatekeeper  $j$  carries the same utility  $\gamma_j$  for all experts. Although experts share this common preference, they do not behave in the same way because acceptance is *neither* guaranteed *nor* instantaneous. Indeed, while experts are perfectly aware of their own type, gatekeepers only observe a noisy signal  $\sigma$  of a proposer's quality, independently drawn from  $G(\sigma|x)$ . For tractability, I assume  $G(\cdot|x)$  has full support over a closed interval  $\Sigma \subseteq \mathbb{R}$  and admits a density  $g(\sigma|x)$  that is differentiable in both arguments.<sup>5</sup> I assume that signals obey the strict monotone likelihood ratio property (MLRP), meaning that  $\frac{g(\tilde{\sigma}|x)}{g(\sigma|x)}$  is increasing in the expert's type  $x$  for all signals  $\tilde{\sigma} > \sigma$ .<sup>6</sup> Generating this signal is a time-intensive but costless process, resulting in a nontrivial *response* time  $T_j > 0$ , during which the expert cannot explore any other options. However, once reached, a gatekeeper's acceptance decision is *final* in the sense that the expert is prohibited from approaching the same gatekeeper again. Experts expire upon a successful match or after unsuccessfully exploring all options.

Since promotion decisions cannot be fully anticipated by experts and are costly in terms of wasted time, experts face the dilemma of wanting to be accepted by the most prestigious gatekeeper while trying to avoid lengthy rejection cycles. Formally, an expert arriving at time  $t$  and being accepted by gatekeeper  $j$  at time  $t + \Delta t$  exits the game with utility  $e^{-r\Delta t}\gamma_j$ , where  $r > 0$  is the common discount rate. For notational simplicity, when there is no ambiguity I sometimes refer to  $e^{-rT_j}$  as  $\delta_j$ . An expert chooses when to propose to an individual gatekeeper  $j$ , but then has to wait until he hears back  $T_j$  units of time later. For some applications, it is more reasonable to interpret  $T_j$  as an active time investment for establishing contact, attending meetings or customizing the product. In this case, a higher discount rate reflects the increased opportunity cost.

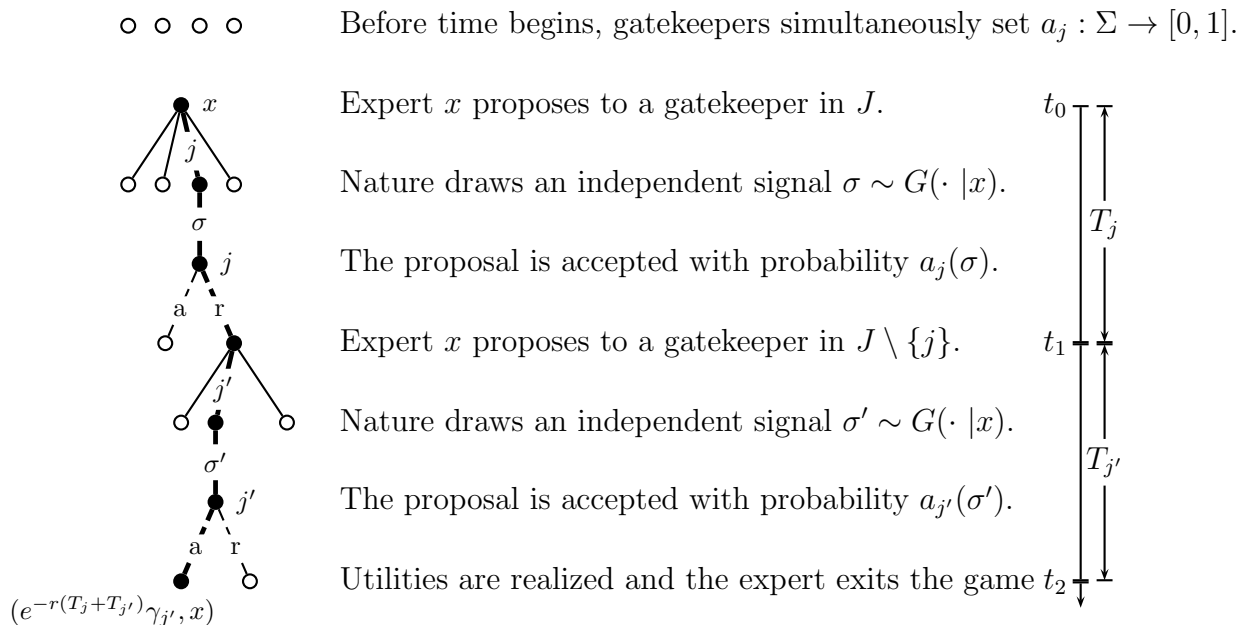
For tractability, I want to focus on the situation where expert history is completely private, and thus also need to shut down any inferences thereon that could be made from calendar time.<sup>7</sup> For that reason, I assume that gatekeepers can condition their **acceptance strategy** *only* on the observed signal  $a_j : \Sigma \rightarrow [0, 1]$ , where  $a_j(\sigma)$  denotes the acceptance

<sup>5</sup>It is without loss of generality to restrict signals to the unit interval  $\Sigma = [0, 1]$ , but allowing for standard exponential signals simplifies exposition later on.

<sup>6</sup>The model can be adapted to include gatekeeper-specific signal distributions  $G_j(\cdot|x)$ . Most qualitative results carry over, with the exception of the asymmetry in [Theorems 3](#) and [5](#) (*both* gatekeepers may now use red tape in equilibrium or deviate to a specialization strategy) and the conditions that guarantee monotone equilibria in [Theorem 2](#) (these are no longer sufficient since the double secant property fails).

<sup>7</sup>Apart from realism, this assumption greatly simplifies the inference problem for gatekeepers as the signal is now the sole base of inference on a specific expert's type. In contrast, if it were known that an expert has already been rejected five times, the previously generated signals must have been rather unattractive. Thus, even if the most recent signal is very high, a gatekeeper has cause for suspicion. As gatekeepers incorporate this source of information, acceptances become less and less likely for later proposals, thereby impacting experts' optimization problem.

Such considerations partially survive even if expert history and arrival times are private, as any gatekeeper approached at  $t = 0$  knows that she is an expert's first choice.



**Figure 1.2:** Sample history for an expert arriving at time  $t_0$  and selecting proposal order  $\omega = (j, j', \dots)$ .

probability of a proposal with signal  $\sigma$ . Formally, I consider an extensive form game where gatekeepers simultaneously commit to time-invariant acceptance strategies before time begins,<sup>8</sup> and experts thereafter choose the timing of their individual proposals.

Time-invariance of the acceptance decision leads to two immediate simplifications: No impatient expert ever delays proposing to a new gatekeeper, and experts of quality  $x$  face the same tradeoffs irrespective of their arrival time. As a result, optimal expert behavior is characterized by a (possibly random) **proposal order**  $\psi_x \in \Delta\Pi(J)$ , where a specific permutation  $\omega = (\omega_1, \dots, \omega_{|J|}) \in \Pi(J)$  represents the strategy of proposing first to gatekeeper  $\omega_1$  and moving on to  $\omega_{k+1}$  as soon as the expert is rejected by  $\omega_k$ . To clarify, Figure 1.2 represents a sample history for a specific expert.

Consequently, let  $\alpha_j(x) = \int_{\Sigma} a_j(\sigma)g(\sigma|x)d\sigma$  denote the (expected) **acceptance rate** among proposals of a specific quality. Under any pure proposal order  $\omega$ , the fraction of experts  $x$  eventually proposing to gatekeeper  $j$  is given by the cumulative rejection probability of all previous proposals,

$$\pi_j(x|\omega) = \prod_{k=1}^{\omega^{-1}(j)-1} (1 - \alpha_{\omega_k}(x)),$$

<sup>8</sup> In the absence of commitment, the suggested time-invariant equilibria remain subgame perfect as long as off-equilibrium beliefs have everybody expect equilibrium play from that point forward. (The only exception are the spontaneous specialization results in Section 1.5.2, which require gatekeeper commitment.) However, additional time-variant equilibria may arise in the no-commitment case.

where  $\omega^{-1}(j)$  denotes  $j$ 's position in permutation  $\omega$ . For general proposal orders  $\psi$ , the **proposal rate** is given by the expectation  $\pi_j(x) = \sum_{\omega \in \Pi(J)} \psi_x(\omega) \pi_j(x|\omega)$ . In turn, the realized **match rate** between gatekeeper  $j$  and experts  $x$  is equal to the product  $\mu_j(x) = \alpha_j(x) \pi_j(x)$ . Whenever I want to stress the dependence on a specific parameter  $\xi$ , I include it in the notation as  $\mu_j(x|\xi)$ .

As a result, I can write an expert's payoff under any pure proposal order  $\omega \in \Pi(J)$  as

$$U(\omega|x) = \sum_{k=1}^{|\mathcal{J}|} e^{-r \sum_{m=1}^k T^m} \mu_{\omega_k}(x|\omega) \gamma_{\omega_k} \quad (1.1)$$

which is the sum over his match rate  $\mu_j(x)$  with each gatekeeper  $j$ , multiplied by the instantaneous payoff  $\gamma_j$  and discounted by the total delay of this specific match. An expert who exits the game unmatched yields a payoff of zero. On the other hand, gatekeeper  $j$ 's per-unit **total matches** are given by  $M_j(a_j) = \int_X \mu_j(x|(a_j, a_{-j})) f(x) dx$ , yielding her a total match **value** of

$$V_j(a_j) = \int_X x \mu_j(x|(a_j, a_{-j})) f(x) dx.$$

In addition, gatekeeper  $j$  faces a unit penalty of  $K > 1$  on any flow rate in excess of  $\kappa_j$ , bringing her final payoff to  $V_j - K \cdot \max\{0, M_j - \kappa_j\}$ .<sup>9</sup>

Consequently, a **(subgame perfect) equilibrium** of the game

$$((\gamma_j, T_j, \kappa_j)_{j \in \mathcal{J}}, \{G(\cdot|x)\}_{x \in X}, F)$$

specifies an acceptance strategy  $a_j : \Sigma \rightarrow [0, 1]$  for each gatekeeper  $j$  and a contingent proposal order  $\psi_x : ([0, 1]^\Sigma)^J \rightarrow \Delta \Pi(J)$  for each expert  $x$ , such that

- (i) For each vector of acceptance strategies  $a = (a_j)_{j \in \mathcal{J}}$ , the proposal order  $\psi_x(a)$  maximizes the payoff of expert  $x$ , meaning  $U(\omega|x, a) \geq U(\omega'|x, a)$  for any  $\omega, \omega' \in \Pi(J)$  with  $\psi_x(a)(\omega) > 0$ .
- (ii) The acceptance strategy  $a_j$  maximizes the payoff for gatekeeper  $j$ , meaning

$$V_j(\tilde{a}_j|a_{-j}, \{\psi_x(\tilde{\alpha}_j, \alpha_{-j})\}_{x \in X}) - K \cdot \max\{0, M_j(\tilde{a}_j|a_{-j}, \{\psi_x(\tilde{\alpha}_j, \alpha_{-j})\}_{x \in X}) - \kappa_j\}$$

is maximized at  $\tilde{a}_j = a_j$ .

## 1.3 Equilibrium Analysis

### 1.3.1 Expert behavior

Since there is a continuum of agents, experts are only indirectly competing with each other through their impact on gatekeeper capacity, but face independent decision problems *conditional* on gatekeepers' acceptance strategies. This simplifies equilibrium analysis, as it is

<sup>9</sup>This flexible way to introduce capacity constraints has been used previously by, e.g. [Che and Koh \(2015\)](#).



sufficient to analyze an isolated expert with given acceptance chances  $\alpha_j(x)$  at each gatekeeper  $j$ . Indeed, a close look at [Equation \(1.1\)](#) reveals that those are the only ingredients that determine his payoff  $U(\omega|x)$ .

Decomposing the payoff into the individual proposals, one observes that the expected utility of submitting to gatekeeper  $j$  is history dependent: While high-prestige options remain, it is a bad move to approach a low-impact gatekeeper with almost certain acceptance since that may cost the possibility of reaching for the most attractive channel. Similarly, a lengthy response time is particularly costly when multiple other venues have to be delayed. Fortunately, the underlying preference ordering is more straightforward, and it is possible to reformulate experts' payoffs to get rid of that history dependence while preserving their strategic behavior. This observation has previously been made by [Heintzelman and Nocetti \(2009\)](#) (drawing largely on [Weitzman \(1979\)](#)'s optimal search theory), who study the expert's decision problem in a partial equilibrium framework. I offer an independent proof in the appendix.

**Lemma 1.3.1.** *Let*

$$z_j(\alpha_j) = \frac{e^{-rT_j}\alpha_j\gamma_j}{1 - e^{-rT_j}(1 - \alpha_j)}.$$

*Then expert  $x$ 's optimal proposal order is according to decreasing  $z_j(\alpha_j(x))$  score.*

*Proof.* See [Appendix A.1.1](#). □

Intuitively, the  $z_j$  score captures both the expected discounted benefit  $e^{-rT_j}\alpha_j\gamma_j$  from an isolated proposal to  $j$  as well as the externality imposed from delaying any further proposals by  $T_j$  in the case of rejection, which happens with probability  $1 - \alpha_j$ . Consequently, high impact makes a gatekeeper more attractive but doesn't affect the externality, while fast response times or high acceptance rates both increase the value of a proposal and decrease its externality. The simplification pins down which proposal rates can occur in equilibrium and how these are affected by changes in external parameters.

**Corollary 1.3.2.** *Proposal rates  $\pi_j(x)$  are nondecreasing in  $\alpha_j(x)$  and  $\gamma_j$  and nonincreasing in  $T_j$  in the sense that any feasible  $\pi_j(x)$  is weakly lower than all feasible  $\tilde{\pi}_j(x)$  under  $\alpha_j(x) \leq \tilde{\alpha}_j(x)$ ,  $\gamma_j \leq \tilde{\gamma}_j$  or  $T_j \geq \tilde{T}_j$ . Changes in opponent parameters have the opposite effect:  $\pi_j(x)$  is nonincreasing in  $\alpha_k(x)$  and  $\gamma_k$  and nondecreasing in  $T_k$  for all  $k \neq j$ .*

*Proof.* See [Appendix A.1.1](#). □

The dependence of  $\pi_j(x)$  on external parameters is entirely driven by their impact on the relative  $z_j$  scores. For that reason, one game parameter is conspicuously absent from [Corollary 1.3.2](#): the overall discount factor  $r$ . Since changes in expert impatience affect all  $z_j$ -scores *simultaneously*, changes in this parameter can attract *more* proposals from experts at one end of the quality spectrum and *reduce* those from the other end.

Finally, each expert eventually exhausts all options, and exits the game unmatched if and only if all gatekeepers reject his proposals. Since both the individual signal draws  $\sigma$  and

the acceptance probabilities  $a_j(\sigma)$  are history-independent, changes in an expert's proposal order have no bearing on his overall match probability.<sup>10</sup>

**Observation 1.** The overall match probability for a given expert is independent of proposal order and equal to  $1 - \prod_{j \in J} (1 - \alpha_j(x))$ .

It is thanks to this property that equilibrium uniqueness can be established in [Theorem 1](#).

### 1.3.2 Gatekeeper behavior

Having thus characterized experts' best responses for any vector of gatekeepers acceptance strategies, it follows that gatekeepers optimally employ cutoff strategies in the first stage.<sup>11</sup>

**Lemma 1.3.3.** *In any equilibrium under private expert history, gatekeepers employ cutoff rules. More formally, the optimal strategy  $a_j : \Sigma \rightarrow [0, 1]$  of gatekeeper  $j$  is equivalent to a cutoff strategy in the sense that there exists a cutoff  $\underline{\sigma}_j \in \Sigma$  such that*

$$a_j(\sigma) = \mathbf{1}_{\underline{\sigma}_j}(\sigma) = \begin{cases} 1 & \text{if } \sigma \geq \underline{\sigma}_j \\ 0 & \text{otherwise} \end{cases}$$

at almost all  $\sigma_j \in \Sigma$ .

*Proof.* See [Appendix A.1.2](#). □

This result is nontrivial because gatekeepers move first and as such have to internalize the *indirect effect* of their acceptance decision on the experts' behavior.<sup>12</sup> However, since proposal rates are nondecreasing in  $\alpha_j(x)$ , they actually act as an amplifier – and any acceptance strategy that favors high-quality ideas (which the cutoff rule does best) has a doubly positive impact on the match distribution.<sup>13</sup>

<sup>10</sup>This would be true even if one adds a small monetary cost  $c_j > 0$  to each option, as an expert would only ever pursue gatekeeper  $j$  if and only if  $\delta_j \gamma_j \alpha_j \geq c_j$ . As such, the set of proposals is still independent of the order.

<sup>11</sup>Incidentally, it is due to this result that only the *ordinal* preference of a gatekeeper matters. The proof of [Lemma 1.3.3](#) does not depend on any specific utility function and hence cutoff strategies are generally optimal. If in addition even the worst idea is preferred over a vacancy, the capacity constraint uniquely pins down a gatekeeper's best response.

<sup>12</sup>Under simultaneous proposals ([Chade et al., 2014](#)), this observation follows more easily since experts always match with the best acceptance offer they receive. Moreover, since proposals are made *prior* to the gatekeeper's decision, she doesn't incorporate the impact of her acceptance strategy onto the *distribution* of proposals she receives. As a result, [Chade et al. \(2014\)](#) consider expected expert quality conditional on proposing to  $j$ , generating signal  $\sigma$  and ultimately matching with  $j$ , and show that this quality is increasing in the signal  $\sigma$ . In the sequential case, experts propose *after* gatekeepers fix their strategy, and as such, gatekeeper  $j$ 's own acceptance strategy impacts the distribution of experts, precluding a similar approach.

<sup>13</sup>As a corollary from this observation, the basic model maintains whether or not experts observe gatekeeper decisions. Indeed, if gatekeeper strategies are private, a deviation is not detected by experts and the amplifying effect of  $\pi_j(x)$  is absent. Still, cutoffs favor high-quality proposals, and hence the equilibrium analysis remains valid. However, as soon as I turn to more interesting strategic considerations such as red tape in [Section 1.5.1](#), the observability of gatekeeper strategies becomes crucial.

An important assumption of the model is that no proposals are of negative quality. In equilibrium, each gatekeeper thus either exhausts her capacity,  $M_j = \int_X \mu_j(x)f(x)dx = \kappa_j$ , or accepts all proposals,  $\underline{\sigma}_j = \min \Sigma$ .<sup>14</sup> However, since joint capacity is insufficient and overall match probability is independent of proposal order, the latter does not occur in equilibrium.<sup>15</sup>

**Lemma 1.3.4.** *In any equilibrium,  $\underline{\sigma}_j \in \text{int}(\Sigma)$  for each  $j \in J$ , and hence all gatekeepers fill capacity in equilibrium.*

*Proof.* Certain acceptance by even a single gatekeeper implies that no expert remains unmatched (see [Observation 1](#)), creating a contradiction with the model's assumption that  $\sum_{j \in J} \kappa_j < 1$ .<sup>16</sup> Categorical rejection  $\underline{\sigma}_j = \max \Sigma$  on the other hand violates optimality since *all* matches are beneficial to the gatekeeper, and a high enough interior cutoff creates strictly positive value while respecting capacity.  $\square$

### 1.3.3 Equilibrium Characterization

Based on the previous characterization of agent's best response, it follows that the game admits an essentially unique equilibrium. More specifically, gatekeeper cutoffs are uniquely determined, while indifferent experts may split in a number of ways, as long as total matches are maintained for each gatekeeper. This indeterminacy is inconsequential when only an  $F$ -measure zero of experts is indifferent across submission orders, as is true for exponential signals under arbitrary cutoffs (see [Lemma A.1.2](#) in the appendix). In such a case, the game can be analyzed in reduced form as a simultaneous game between gatekeepers, admitting the succinct proof given here. I complete this proof in the appendix for general signal functions.

**Theorem 1.** *An equilibrium exists. Moreover, gatekeeper cutoffs are uniquely determined and vary continuously in all external parameters.*

*Partial proof.* To distill the main ideas, I here assume that the optimal submission order is unique for almost all experts under any cutoffs  $\underline{\sigma} \in \Sigma$ , which in turn unambiguously determines proposal rates  $\pi_j(x|\underline{\sigma})$  and total matches  $M_j(\underline{\sigma})$ . An extension to general signal functions is relegated to [Appendix A.1.3](#), as is the discussion of continuity.

Part I: Best responses are monotone.

Total matches for gatekeeper  $j$  are unambiguously given by the function

$$M_j(\underline{\sigma}) = \int_X \mu_j(x|\underline{\sigma})f(x)dx.$$

<sup>14</sup>Capacity manipulation is not beneficial since the match rate is increasing in  $\underline{\sigma}_j$  at *every* quality level  $x$ . This rules out situations as in [Sönmez \(1997\)](#) where match participants can *increase* matches with higher types by underreporting capacity.

<sup>15</sup>The assumption of nonnegative quality can be relaxed in a situation where total capacity is far smaller than the inflow of worthwhile proposals, as long as all gatekeepers still fill capacity in equilibrium.

<sup>16</sup>In a model with small enough monetary cost  $c_j < \delta_j \gamma_j$ , the same argument goes through.

By continuity of the  $z$  scores and the signal distribution  $G$ , this function is continuous. Moreover, both acceptance and proposal rates are decreasing in  $\underline{\sigma}_j$ , while the latter is weakly increasing in  $\underline{\sigma}_{-j}$  by [Corollary 1.3.2](#). In other words, total matches are strictly decreasing in  $\underline{\sigma}_j$  and weakly increasing in  $\underline{\sigma}_{-j}$ . As a consequence, the best response function  $\Phi : \Sigma^J \rightarrow \Sigma^J$  with

$$\Phi_j(\underline{\sigma}) = \min \{ \underline{\sigma}_j \in \Sigma \mid M_j(\underline{\sigma}_j, \underline{\sigma}_{-j}) \leq \kappa_j \}$$

is monotone over the complete lattice  $(\Sigma^J, \leq)$ .<sup>17</sup>

Part II:  $\Phi$  has a unique fixed point.

By Tarski's fixed point theorem, the fixed points of  $\Phi$  thus also form a nonempty and complete lattice along with the partial order  $\leq$ . Any such fixed point represents an equilibrium of the game when accompanied by any proposal orders that follow the  $z_j(x)$  scores.

Finally, by [Observation 1](#), the overall match probability for expert  $x$  is equal to  $1 - \prod_{j \in J} G(\underline{\sigma}_j | x)$ . As such, any two distinct ordered fixed points  $\underline{\sigma} < \underline{\sigma}'$  differ in the total flow of matches

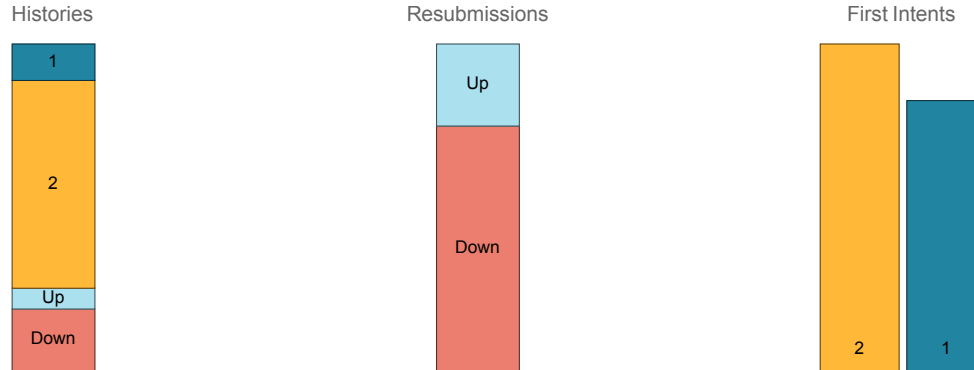
$$\sum_j M_j(\underline{\sigma}) = \int_X \left( 1 - \prod_{j \in J} G(\underline{\sigma}_j | x) \right) f(x) dx > \int_X \left( 1 - \prod_{j \in J} G(\underline{\sigma}'_j | x) \right) f(x) dx = \sum_j M_j(\underline{\sigma}').$$

However, since that sum equals total capacity  $\sum_j \kappa_j$  by [Lemma 1.3.4](#), equilibrium is unique.  $\square$

The previous theorem relies on the strategic complementarity among gatekeepers. The intuition for this is twofold: As an expert's acceptance rate at a competing gatekeeper  $k$  decreases, so does his  $z_k$  score and he might now propose to gatekeeper  $j$  *earlier*. In addition, even *conditional* on proposal order, the expert proposes to  $j$  more often since his previous proposals are now more likely rejected. To counter her surge in proposals, gatekeeper  $j$  therefore also accepts fewer signals. Equilibrium existence follows from Tarski's Fixed Point Theorem. Uniqueness on the other hand follows from the fact higher signals lower total matches irrespective of experts' proposal orders.

The proof is more involved when a nontrivial mass of experts may be indifferent across proposal orders. The indeterminacy of  $\pi_j(x)$  in this case doesn't allow a reduced form discussion of the form above. In [Appendix A.1.3](#), I instead approximate the  $z_j$  scores by appropriately chosen step functions  $z_j^n$  that preclude expert indifference. Letting  $\underline{\sigma}$  denote the limit of the respective equilibrium cutoffs, I then construct expert strategies that implement  $\underline{\sigma}$  as equilibrium cutoffs. The delicacy of this part lies in the fact that expert strategies are defined over an uncountable set  $X$  and may not converge pointwise. To circumvent this issue, I let all experts with identical limiting preference adopt the same mixed proposal strategy in a way that maintains gatekeepers' total matches. While the individual expert strategies are no longer best responses to the *perturbed* scores  $z_j^n$ , they admit a convergent subsequence by

<sup>17</sup>In the case of an unbounded signal space, a complete lattice is obtained by replacing  $\Sigma$  with  $\tilde{\Sigma} = [\inf \Sigma, \sup \Sigma] \subseteq \mathbb{R} \cup \{\pm\infty\}$ . Continuity of  $M_j(\underline{\sigma})$  and [Lemma 1.3.4](#) ensure that all fixed points are interior.



- (a) Among all published papers, most were first submitted to the bottom journal.
- (b) Successful resubmission flows are oriented downstream in 75% of all cases.
- (c) The bottom journal publishes more first intents (77%) than the top journal (64%).

**Figure 1.3:** Equilibrium outcomes that reproduce the findings of [Calcagno et al. \(2012\)](#). Here, a small top journal  $(\gamma_1, \kappa_1) = (1.3, 0.027)$  competes against a larger bottom journal  $(\gamma_2, \kappa_2) = (1, 0.13)$ . Response times are nine months each and the yearly discount rate is  $r = 10\%$ . Ex-ante quality and signals are distributed exponentially,  $x \sim \exp(1)$  and  $\sigma|x \sim \exp(1/x)$ . Equilibrium cutoffs are  $\underline{\sigma}_1 = 5$  and  $\underline{\sigma}_2 = 2$ . As in the empirical findings, 75% of published papers are first intents.

Bolzano-Weierstrass<sup>18</sup> and the limit represents a best response under score  $z$  and cutoffs  $\underline{\sigma}$ . Intuitively, any strict preference is robust to small enough perturbations, and any proposal order is optimal under exact indifference.

Uniqueness follows from a generalization of [Observation 1](#): Any two distinct gatekeeper cutoffs  $\underline{\sigma} \neq \tilde{\underline{\sigma}}$  (not necessarily ordered) differ in total matches for a subset of gatekeepers ([Lemma A.1.3](#)).

Lastly, I consider any convergent parameter sequence  $\xi^n \rightarrow \xi$  and let  $\underline{\sigma}^n$  and  $\underline{\sigma}$  denote the corresponding equilibrium cutoffs. The limit of any convergent subsequence  $\underline{s} = \lim_{k \rightarrow \infty} \underline{\sigma}^{n_k}$  can be implemented as an equilibrium under  $\xi$  by the argument above. By uniqueness, this implies  $\underline{s} = \underline{\sigma}$ , and convergence follows since a bounded sequence with a unique subsequence limit is itself convergent. This completes the proof of continuity.

To illustrate how this formal framework may sharpen our understanding of such matching markets, let me reconsider an empirical study originally published in *Science* and give an alternative explanation for one of their main empirical puzzles.

**Example 2.** [Calcagno et al. \(2012\)](#) study flows of research manuscripts among journals by surveying successful authors where their published paper had previously been submitted. They observe two empirical regularities: First, among papers that were first rejected by one

<sup>18</sup>Indeed, since all experts with the same limit preference behave in the same way, this reduces the number of strategies to at most  $|J|^2$  (the number of weak orders over  $J$ ).

journal and then published at another journal (here called ‘resubmissions’), most – but not all – ended up at a lower-impact journal than where they were first submitted to. Second, authors generally target journals efficiently, in the sense that 75% of all publications were first submitted to the journal that ultimately published them (‘first-intents’). Together, these two observations lead them to expect that high-impact journals primarily publish first-intents, while lower journals receive and publish more resubmissions. Unexpectedly, they find the opposite pattern: the proportion of first-intent publications *decreases* with impact factor, except at the very top. To make sense of these findings, the authors suggest that top journals have a denser competitive neighborhood due to their general focus, and as such receive resubmissions from more sources.

However, instead of citing network effects, those findings can also be reconciled parsimoniously within the current model. [Figure 1.3](#) lists the relevant parameter values and illustrates key equilibrium predictions. Intuitively, consider that the high-ranked journal has a small capacity, and the bulk of papers are of relatively low quality. Given their meager acceptance chances, most experts first submit to the bottom journal, are rejected and thereafter submit to the top. As a result, the top journal receives many more resubmissions than first-intents. So many in fact, that despite her high equilibrium cutoff, she ultimately publishes quite a few recycled papers. Conversely, only the very best papers are first submitted to the top, and in turn there is a small absolute number of papers that are first rejected at the top and then accepted at the bottom journal.

Put differently, the third finding may owe to averaging effects: In absolute numbers, the top journal publishes fewer resubmissions than the bottom,<sup>19</sup> but since its capacity is small in comparison, those make up a larger *percentage* of total publications in [Figure 1.3c](#).  $\diamond$

In the next two sections, I discuss equilibrium sorting and analyze the nonstandard competitive environment among gatekeepers. To do so, I restrict attention to the case of just two gatekeepers for tractability. In particular, this restriction ensures that the effect of an exogenous parameter change on equilibrium cutoffs is unambiguous and intuitive.

**Lemma 1.3.5.** *With only two gatekeepers, a raise in  $\gamma_j$  or a drop in  $T_j$  causes a raise in gatekeeper  $j$ ’s cutoff and a drop in her opponent’s cutoff. The changes are strict if there exists at least one indifferent interior expert  $\hat{x} \in (0, 1)$  with  $z_j(\hat{x}) = z_k(\hat{x})$ .*

*Proof.* See [Appendix A.2](#).  $\square$

In the interest of brevity, the remaining comparative statics results are relegated to [Appendix A.2](#).

---

<sup>19</sup>This is visible from [Figure 1.3a](#). Papers with history ‘Up’ correspond to resubmissions published at the top journal and those with a ‘Down’ history are published at the bottom journal. Histories 1 and 2 correspond to first-intents at journal 1 and 2 respectively.

## 1.4 Sorting Failures

A naive interpretation of the common preference assumption may suggest assortative matching, at least under some weakened notion that accounts for partial information. After all, all gatekeepers rank experts by proposal quality  $x$ , and any expert with certain acceptance everywhere ranks gatekeepers according to the present discounted value of future impact  $z_j(1) = e^{-rT_j}\gamma_j$ , which I refer to as a gatekeeper's **appeal**.

In particular, two natural properties for equilibrium sorting come to mind:

- (i) More appealing gatekeepers should be more selective than their less appealing counterparts.
- (ii) Only the best experts should propose first to the top gatekeeper.

As the following result shows, both monotonicity properties can fail in equilibrium: A gatekeeper might set a higher cutoff than a *more appealing* competitor. Conversely, experts with particularly poor acceptance rates everywhere have a low opportunity cost of waiting and therefore are less responsive to differences in response times. As a result, they might directly propose to the top gatekeeper even when more qualified experts first go after a faster, low-impact gatekeeper.

**Theorem 2.** *Consider a setup with two gatekeepers, such that  $e^{-rT_1}\gamma_1 > e^{-rT_2}\gamma_2$ .*

- (i) *If the bottom gatekeeper is small (large) enough relative to the top, she is also more selective (accepting) in the sense that  $\underline{\sigma}_2 > \underline{\sigma}_1$  ( $\underline{\sigma}_2 < \underline{\sigma}_1$ ).*

*Moreover, for any capacity  $k_2 \in (0, 1)$ , there exists an upper bound  $\bar{k}_1 \in (0, 1 - k_2)$  such that equilibrium cutoffs are monotone whenever the top gatekeeper is small enough  $\kappa_1 < \bar{k}_1$ , and these bounds are independent of the parameters  $(T_j, \gamma_j)$ .*

- (ii) *If the bottom gatekeeper responds much faster than at the top, such that  $\delta_2 > \frac{\gamma_1}{\delta_1\gamma_1 + (1-\delta_1)\gamma_2}\delta_1$ , there exists a MLRP signal function and capacities  $(\kappa_1, \kappa_2)$  such that experts behave non-monotonically in equilibrium, i.e.  $z_1(x) > z_2(x)$  and  $z_1(x') < z_2(x')$  for some  $x < x'$ .*

*This nonmonotonic behavior is ruled out when the low-impact gatekeeper  $\gamma_2 < \gamma_1$  is not too much faster or sets a lower equilibrium cutoff.*

*Proof.* See [Appendix A.1.4](#). □

It is intuitive that a gatekeeper of disappearing capacity will set a high cutoff independent of her competitive position, and that this eventually upsets the first type of equilibrium sorting. However, the following example illustrates that gatekeepers may behave nonmonotonically even when they are of *equal* capacity.

**Example 3.** Consider a setup with two gatekeepers of equal capacity,  $(\gamma_1, \tau_1, \kappa_1) = (1.18, 2, 0.2)$ ,  $(\gamma_2, \tau_2, \kappa_2) = (1, 1, 0.2)$  and discount rate  $r = 10\%$ . Signals are distributed exponentially,  $G(\sigma|x) = 1 - x^\sigma$ . Expert quality is distributed with density  $f(x, \lambda) = (\lambda + 1)(1 - x)^\lambda$ , where  $\lambda > 0$  measures the concentration of expert quality. The distribution converges to a point mass at 0 as  $\lambda \rightarrow \infty$  and to the standard uniform distribution as  $\lambda \rightarrow 0$ . For this example, fix  $\lambda = 3$ .

In equilibrium, gatekeepers behave nonmonotonically since  $\underline{\sigma}_1 = 0.82 < 0.87 = \underline{\sigma}_2$ . Intuitively, experts apply first to the bottom gatekeeper if and only if  $x \leq \hat{x} = 0.44$ , but due to the steep decrease of quality, this represents 90.3% of all experts. In order to counteract this high proposal rate, the bottom gatekeeper needs to be *more* selective than the top.  $\diamond$

Experts as well may violate the natural sorting when facing a fast and selective bottom gatekeeper against a slow but more accepting top gatekeeper. In this comparison, only the fast turnaround time speaks for the bottom gatekeeper – she accepts fewer proposals and yields a lower payoff upon acceptance. Her quick decision increases the expected proposal payoff for all agents by the same factor, but the externality reduction on the delayed top proposal is most pronounced when rejection at the bottom and acceptance at the top is a likely outcome.<sup>20</sup> As the following example illustrates, ‘average’ experts may then first choose this low-yield, low-externality option, while the best and worst experts first pursue the top gatekeeper – the best being motivated by high success rates and the worst by a low opportunity cost of time.<sup>21</sup>

**Example 4.** Consider a situation with two gatekeepers

$$(\gamma_1, \tau_1, \kappa_1) = (1.12, 2, 0.30), \quad (\gamma_2, \tau_2, \kappa_2) = (1, 1, 0.27)$$

and discount rate  $r = 10\%$ . Signals and quality are distributed as before, with  $\lambda = 0.27$ .

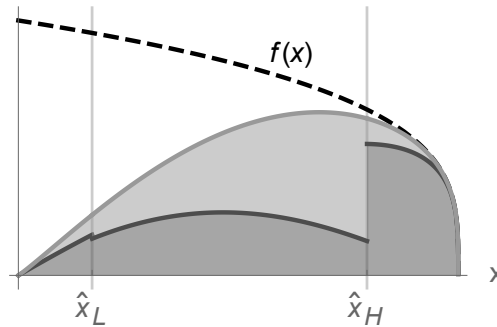
In equilibrium, the bottom gatekeeper is more selective than the top ( $\underline{\sigma}_1 = 1 < 1.3 = \underline{\sigma}_2$ ) which causes experts to behave nonmonotonically. Specifically, experts first propose to the bottom gatekeeper if and only if  $x \in [0.17, 0.79]$ , while both lower- and higher-quality experts first propose to the top. [Figure 1.4](#) depicts the corresponding equilibrium.  $\diamond$

These sorting failures show that monotone equilibrium behavior is not the ‘default outcome’. As such, crude inferences from a gatekeeper’s rejection rate to her innate appeal may be faulty, as may those from an expert’s proposal order to his quality. Neither are they a mere artifact of the exogenous impact parameter. To illustrate this, the examples in this section are chosen such that the endogenous mean quality of matches at either gatekeeper

<sup>20</sup>Put simply, when acceptance at the bottom journal is almost certain, experts care little about additional delay *conditional* on rejection. Similarly, when acceptance by the top gatekeeper is extremely unlikely, it doesn’t matter if the proposal has to be further delayed.

<sup>21</sup>The actual proof is more mechanical: I consider the secant slopes of the indifference curve and identify necessary and sufficient conditions under which they move in the same direction. When they do, I derive a signal structure that ensures multiple crossings between indifference and acceptance curve. When they do not, the double secant property ([Chade et al., 2014](#), Theorem 1) guarantees a single crossing, and thereby monotone proposal strategies.





**Figure 1.4:** Example equilibrium outcomes for [Example 4](#), where both the best  $x \geq \hat{x}_H$  and the worst  $x \leq \hat{x}_L$  experts first propose to the top gatekeeper. Depicted are the equilibrium flow rate of matches  $\mu_j(x)f(x)$  at the top (dark gray) and bottom (light gray) gatekeeper for each quality level  $x$ .

is proportional to the exogenous impact ratio  $\gamma_1/\gamma_2$ . In other words, if gatekeepers act as signaling devices for expert quality, then the relative value for an expert to be matched with gatekeeper  $j$  is exactly equal to the assumed impact  $\gamma_j$ .

Furthermore, while both of these sorting failures also occur under simultaneous proposals ([Chade et al., 2014](#)), the necessary and sufficient conditions are different for both cases. In a simultaneous setup, sorting failures occur independently for experts and gatekeepers, while they here occur among experts *only if* gatekeepers also set nonmonotone cutoffs. The latter however is more frequent now: [Example 3](#) shows nonmonotone gatekeeper behavior under *equal* capacities, which for simultaneous proposals is restricted to cases where  $\kappa_2$  is significantly smaller than  $\kappa_1$ .<sup>22</sup> In [Appendix A.1.4](#), I further show a third type of sorting failure that is unique to sequential proposals: Even under monotone equilibrium behavior, some low-quality experts may match *more frequently* with the top gatekeeper than experts of higher quality.

## 1.5 Gatekeeper Competition

Experts' directed search generates a non-standard competitive environment for gatekeepers. After simplification however, optimal gatekeeper behavior reduces to finding the capacity-clearing cutoff and fails to incorporate other realistic and subtle strategic tradeoffs. For that

<sup>22</sup>The proof mechanics are also more subtle under sequential proposals since the top gatekeeper is less shielded from cutoff decisions at the bottom: In the simultaneous setup, experts first make any number of proposals, based on which gatekeepers make *match offers* that are finally accepted or rejected by experts. Although each expert can only accept a single offer, all experts prefer being matched with the top gatekeeper. As such, and *conditional* on expert proposals, changes in the bottom's offer strategy have no bearing on the proposals received by the top nor on the acceptances of offers made. In a sequential world, this isolation breaks down; for even *conditional* on experts' proposal order, acceptance changes at the bottom directly affect how many experts will make further proposals.

reason, I now restrict attention to monotone equilibria and expand the model to a richer strategy space.

### 1.5.1 Red Tape

Among the real-world choices of gatekeepers are various considerations that determine their attractiveness in the eyes of experts: Ease of access, regulatory hurdles, customization requirements or response times. I study one such consideration by allowing gatekeepers some freedom in choosing their turnaround times  $T_j$ . In particular, I assume that there exists an exogenously given *minimal response time*  $\tau_j$ , but a gatekeeper can costlessly set any longer response time  $T_j \geq \tau_j$ .<sup>23</sup> I am interested in knowing whether gatekeepers ever choose to set unnecessarily long  $T_j > \tau_j$ . When they do, I refer to such spurious delay as **red tape**, a term that generically describes ‘seemingly unnecessary tasks or actions that are required in order to get something.’<sup>24</sup> Indeed, while adding red tape may seem purely wasteful at first sight, it may actually benefit the infinitely patient gatekeeper  $j$  by nudging experts towards strategies that improve her proposal distribution.

Endogenizing response times in this way does not upset equilibrium uniqueness and as such allows clear predictions regarding the presence and amount of red tape under monotone strategies. Most importantly, the competitive environment between gatekeepers creates an asymmetry where the bottom gatekeeper always minimizes turnaround times in equilibrium, while the top chooses a strictly positive amount of red tape when the impact differential  $\gamma_1 - \gamma_2$  is sufficiently large.

**Theorem 3.** *Consider a setup with two gatekeepers who can add arbitrary red tape in the form of delayed turnaround,  $T_j \geq \tau_j$ . Label gatekeepers such that the ‘top’ gatekeeper 1 is more attractive in the absence of red tape,  $e^{-r\tau_1}\gamma_1 \geq e^{-r\tau_2}\gamma_2$ .*

*If  $\kappa_1$  is sufficiently small, a unique equilibrium exists. There is no red tape at the bottom,  $T_2 = \tau_2$  and the optimal response time  $T_1^*$  at the top is increasing in  $\gamma_1$  and unbounded. The equilibrium thus exhibits red tape at the top  $T_1^* > \tau_1$  under sufficiently high impact  $\gamma_1$ .*

*Proof.* See [Appendix A.1.5](#). □

To illustrate the central proof ideas, consider a marginal increase in red tape  $dT_j > 0$  accompanied by a capacity-preserving cutoff reduction  $d\sigma_j < 0$ . Any strict expert preference

<sup>23</sup>Certainly, assuming an exogenous lower bound on response times is still a rather coarse assumption. Reality may be better described by a decreasing cost function that captures the organizational challenges to speeding up proposal evaluations. The lower-bound assumption corresponds to a situation where costs for  $T_j < \tau_j$  exceed the maximal gatekeeper payoff of one, and drop to zero thereafter. Any additional cost reductions beyond  $\tau_j$  only make longer response times even more attractive in the eyes of the gatekeeper. One advantage of the lower-bound assumption is that it allows me to clearly identify excessive delay as  $T_j > \tau_j$ .

<sup>24</sup>According to “Definition of Red Tape”, Merriam-Webster, accessed October 02, 2015, [www.merriam-webster.com/dictionary/red%20tape](http://www.merriam-webster.com/dictionary/red%20tape).

$z_1(\alpha(x|\underline{\sigma}_1)) \geq z_2(\alpha(x|\underline{\sigma}_2))$  is maintained after this marginal change, implying that the proposal rate  $\pi_j(x)$  remains unaltered for all but the marginal expert  $\hat{x}$ . Pointwise higher acceptance rates through the cutoff decrease thus imply a higher match rate  $d\mu_j(x) = g(\underline{\sigma}_j|x)\pi_j(x)$  for all  $x \neq \hat{x}$ .

To maintain  $j$ 's total matches  $M_j$ , the marginal expert  $\hat{x}$  thus has to propose strictly *less* often, so that the net effect for  $j$  is a substitution away from matches with  $\hat{x}$  towards matches with other experts  $x \neq \hat{x}$ . The net impact on her payoff is proportional to

$$\int_X (x - \hat{x})g(\underline{\sigma}_j|x)\pi_j(x)f(x)dx \quad (1.2)$$

where  $x - \hat{x}$  is the payoff difference between matches with  $x$  and  $\hat{x}$ , which are substituted at rate  $g(\underline{\sigma}_j|x)\pi_j(x)f(x)$ . The deviation is thus profitable if and only if  $\hat{x}$  is low enough.

Up until this point, there is no difference across the two gatekeepers, and the analysis thus shows that even the bottom gatekeeper's best response involves red tape in off-equilibrium situations. An asymmetry arises however from experts' proposal behavior and the monotone equilibrium assumption: Since  $\pi_1(x) \geq \pi_2(x)$  if and only if  $x \geq \hat{x}$ , the top gatekeeper receives more (less) proposals with a positive (negative) quality difference  $x - \hat{x}$ . Moreover, the monotone equilibrium cutoffs<sup>25</sup>  $\underline{\sigma}_2 < \underline{\sigma}_1$  together with MLRP imply that  $g(\underline{\sigma}_2|x)$  is 'dominated' by  $g(\underline{\sigma}_1|x)$ , loosely speaking. In other words, even if the bottom gatekeeper faced the same proposal rates  $\pi_1(x)$  as the top, additional red tape would be less profitable for her. Together, the two arguments imply that unless Equation (1.2) is *negative* for the bottom gatekeeper (she wants to *reduce* delay),  $\hat{x}$  is low enough that the top gatekeeper actually wants to further increase red tape. This of course rules out red tape at the bottom in equilibrium.<sup>26</sup>

The proof also highlights the top gatekeeper's rationale for adding red tape. Indeed, as in any filtering problem, noisy quality perception causes her to balance two types of mistakes: Accepting low-quality proposals and rejecting high-quality ones. In a situation where almost everyone first proposes to the top gatekeeper (which occurs as  $\gamma_1$  grows), a combination of red tape and cutoff adjustment can reduce both error types; red tape reduces low-quality proposals while the cutoff drop increases acceptance everywhere. Since matches are jointly determined by proposal and acceptance rate, red tape allows the gatekeeper to reduce matches with low-quality experts in favor of those with higher quality.

However, while discouraging early proposals through red tape reduces the inflow of proposals, their distribution is now *adversely selected*. Indeed, the top gatekeeper now hears only from previously *rejected* experts, implying that proposal rates have gone down over some range  $x \in [\tilde{x}, \hat{x}]$ , but the change  $\Delta\pi_j(x) < 0$  is *increasing* in quality  $x$ . Naturally, this limits the attractiveness of delay as red tape. The problem of adverse selection can be partially

<sup>25</sup>Indeed, the capacity bound  $\bar{k}_1$  identified in Theorem 2 ensures monotone equilibrium cutoffs *independent* of response times.

<sup>26</sup>More generally, if gatekeepers face finite marginal costs to reducing response times below  $\bar{\tau}_j$ , the bottom gatekeeper competes on time in order to attract better proposals.

avoided for other forms of red tape, such as a purely wasteful *monetary cost*  $c_j$  that is due upon proposal and destroyed thereafter. In contrast to delay, the lowest-quality experts with acceptance rates below  $e^{rT_j}c_j/\gamma_j$  now refrain from proposing altogether. Similar to delay however, such red tape also affects the optimal proposal order of higher-quality experts,<sup>27</sup> and as such reduces proposals both from the very bottom *and* from the marginal expert  $\hat{x}$ . Adverse selection only affects the latter reduction and consequently, gatekeepers generically prefer monetary proposal costs to delay,<sup>28</sup> as long as they can ensure that proposal costs are borne *privately* by the individual expert. At least within academic publishing however, long time lags are more frequent than submission fees, presumably due to the fact that experts are often reimbursed for any monetary costs. In other applications, such as political lobbying, ethical reasons preclude (or at least discourage) monetary fees: Politicians have to expect serious scrutiny if it becomes known that they demand payments before even considering their constituents' concerns. Both of these reasons motivate my primary focus on delays rather than fees.

The main takeaway of [Theorem 3](#) is that unnecessarily long delays may occur precisely *because* gatekeepers are competing with each other for experts' proposals. Extending the basic intuition beyond the formal two-gatekeeper case, the result seems to suggest that delays are particularly long among high-ranked gatekeepers. Empirical data from economic journals reported in [Table 1.1](#) indeed exhibits high correlation between impact factor and first response time. A similar effect is reported by [Ellison \(2002\)](#) who analyses determinants for the slow-down in economic publishing and attributes most weight to the negative link between review speed and journal rank.

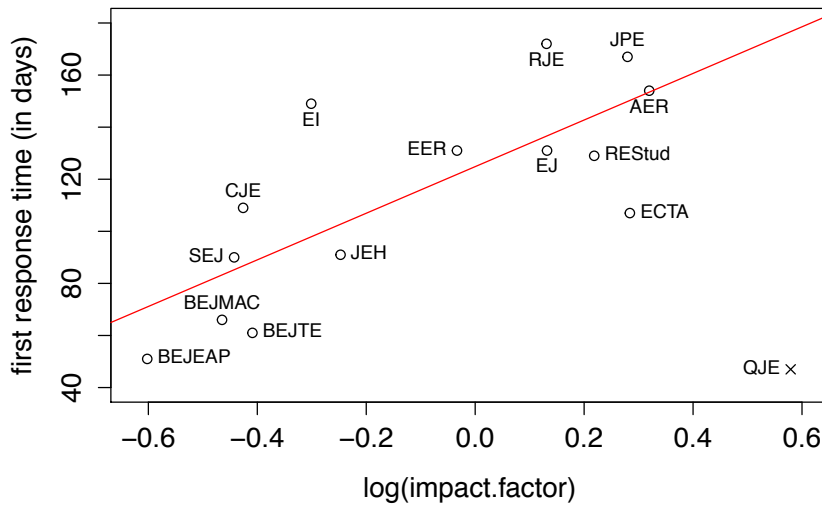
Although red tape may be beneficial to the top gatekeeper, I have yet to consider its externality on the other agents. Impatient experts are now facing longer match delays, and the competitive gain by the top gatekeeper may lower the bottom gatekeeper's payoff. As it turns out however, equilibrium effects are more subtle: A red tape equilibrium is accompanied by cutoff adjustments that may actually *increase* payoff for those experts who value easier

<sup>27</sup>[Heintzelman and Nocetti \(2009\)](#) include such a monetary cost in their analysis and show that the scores assigned to gatekeepers are now given by  $z_j(\alpha) = (\delta_j\alpha\gamma_j - c_j)/(1 - \delta_j(1 - \alpha))$ .

<sup>28</sup>Intuitively, there always exists a profitable, capacity-preserving deviation  $dc_j > 0$ ,  $dT_j < 0$  and  $d\hat{x}_j < 0$  that maintains  $\hat{x}$  as the indifferent expert and replaces some of the worst proposals with higher-quality matches. This generalizes [Heintzelman and Nocetti \(2009\)](#)'s observation by incorporating gatekeepers' capacity concern.

<sup>29</sup>I use response times reported in [Azar \(2007\)](#), restricting my sample to economic journals and ignore the four journals in accounting and finance, mainly because finance journals distinguish themselves through high submission fees. For some journals, the author lists both mean and median first response times over several subsamples of all publications (e.g. rejections only, revisions only). To avoid oversampling individual journals based on data availability, I restrict myself to the mean first response time for first submissions only, or the closest report thereof.

I retrieved the 2001 impact factors from the Journal Citation Reports, ISI Web of Knowledge, <http://admin-apps.webofknowledge.com/JCR/JCR> on September 19th, 2015. Missing data for Berkeley Economics Journals was replaced by 2009 ISI impact factor data compiled by the Tepper School of Business, accessed September 20th, 2015, <https://server1.tepper.cmu.edu/barnett/rankings.html>.



	excluding QJE	including QJE
$\log_{10}(\text{impact.factor})$	89.49*** (22.11)	45.27 (28.87)
Intercept	124.84*** (7.52)	113.30*** (10.42)

**Table 1.1:** First response times at economic journals as a function of impact factor.<sup>29</sup> There is a highly significant positive relationship between a journal’s impact factor and editorial delays (p-value 0.16%), once the obvious outlier representing the Quarterly Journal of Economics  $\times$  is removed. The coefficient remains positive but loses its significance when the outlier is included.

access to the top above a fast response. Similarly, the bottom gatekeeper now matches more frequently with experts close to  $\hat{x}$  with an ambiguous impact on his payoff.

Since red tape equilibria are not Pareto ordered, I consider the alternative welfare measure of **social impact**

$$W = \sum_{j \in J} \gamma_j \int_X x \mu_j(x) f(x) dx. \tag{1.3}$$

True to the modeling framework, this notion assumes that ideas promoted by gatekeeper  $j$  attain an impact of  $\gamma_j$ , fixing the social value of an  $x$ - $j$  match to  $x\gamma_j$ . In other words, it is socially more efficient that the best ideas be promoted through the most influential channels.

As [Theorem 4](#) hereafter shows, red tape can actually *help* to achieve this goal by incentivizing experts to sort more efficiently. Intuitively, if the more influential gatekeeper responds too quickly, she becomes flooded with proposals and consequently has to set an extremely high signal cutoff. Red tape diverts some low-quality proposals to the bottom gatekeeper, allowing both of them to employ more similar cutoffs. For signal distributions

satisfying the **strong monotone likelihood ratio** property<sup>30</sup>

$$\frac{g(\underline{\sigma}_1|x)G(\underline{\sigma}_2|x)}{g(\underline{\sigma}_2|x)G(\underline{\sigma}_1|x)} < \frac{g(\underline{\sigma}_1|x')G(\underline{\sigma}_2|x')}{g(\underline{\sigma}_2|x')G(\underline{\sigma}_1|x')} \quad \forall \underline{\sigma}_1 > \underline{\sigma}_2, \forall x < x' \quad (\text{sMLRP})$$

this improves the overall quality distribution of matched experts  $\mu_1(x) + \mu_2(x)$  in the sense of first-order stochastic dominance. [Lemma A.1.8](#) in the appendix establishes this property for any exponential signal distributions. In this case, red tape simultaneously improves the mean quality of experts matched with the top gatekeeper *as well as* the total quality of matched experts, and as such necessarily improves social impact.

**Theorem 4.** *Consider a setup with two gatekeepers such that  $\gamma_1 > \gamma_2$ , with  $\kappa_1$  sufficiently small compared to  $\kappa_2$  so as to ensure monotone equilibrium strategies. If the signal distribution satisfies (sMLRP) and  $\tau_1$  is low enough, then some red tape is welfare-enhancing.*

*Proof.* See [Appendix A.1.5](#). □

The welfare result in [Theorem 4](#) implies that there exists a lower bound on the optimal response time  $T_1$ . Whenever the minimal response time  $\tau_1$  lies below this bound, red tape is welfare enhancing. Admittedly, this bound partly depends on the second response time  $\tau_2$ , and as such the statement doesn't speak to the welfare impact of red tape under *joint* reductions in minimal response times. However, a closer look at the proof identifies red tape as beneficial exactly in situations where the marginal expert  $\hat{x}$  is particularly low. In other words, red tape improves welfare if and only if the vast majority of experts first propose to the top gatekeeper, which occurs if the impact differential  $\gamma_1 - \gamma_2$  is large, if experts are particularly patient or if response times are generally short.<sup>31</sup> As such, a stronger conclusion is warranted: Even if technological innovations drastically lower *all* minimal response times  $\tau_j$ , it is socially preferable that the top gatekeeper *do not* reduce delay as much as technically feasible. By keeping some red tape  $T_1 > \tau_1$ , she nudges the lowest quality experts towards the bottom gatekeeper first, thereby improving total match surplus.

Simultaneously, this analysis is also relevant to a market designer who controls gatekeepers' impact  $\gamma_j$ . By way of example, committee membership in the senate is an important determinant of a politician's influence ([McCubbins et al., 1994](#), p.18). Ideally, the designer would like to appoint the gatekeeper with the highest-quality matches to a key position. However, shifting impact from gatekeeper 2 towards gatekeeper 1 upsets equilibrium sorting due to a surge in low-quality proposals. Fortunately, red tape allows the designer to maintain equilibrium matches by ensuring that the two gatekeepers retain a similar score *in the*

<sup>30</sup>Property (sMLRP) is equivalent to log-supermodularity of the reverse hazard rate.

<sup>31</sup>Indeed,  $z_1(x) - z_2(x) \rightarrow \infty$  as  $\gamma_1 - \gamma_2 \rightarrow \infty$ , and  $z_j(x) \rightarrow \gamma_j$  as either  $r \rightarrow \infty$  or  $T_j \rightarrow 0$ . Under any of these limits, all experts first propose to the top gatekeeper.

experts' eyes even as the actual impact differential grows.<sup>32</sup> This observation runs contrary to popular wisdom, where red tape and aggregation of power is often seen as a hallmark of corruption.<sup>33</sup> In the present setup instead, red tape is used as an allocation mechanism in the absence of price discrimination, and it may be valuable if the brunt of the delay is privately borne by experts and of minimal concern for welfare.

Indeed, by considering only the final match distribution, social impact (1.3) does not directly associate any cost to the slow-down of the match process incurred from red tape. As such, it does not incorporate the notion that experts' ideas may become obsolete before a match occurs. This is true of the model more generally, as I assume that expert quality  $x$  does not decline over time and exits are voluntary. Nevertheless, the result generalizes for a patient social planner with sufficiently low discount factor  $\rho > 0$ , since

$$W' = \sum_{j \in J} \gamma_j \int_X e^{-\rho \bar{T}_j(x)} x \mu_j(x) f(x) dx \longrightarrow W \quad \text{as } \rho \rightarrow 0^+,$$

where  $(\bar{T}_1, \bar{T}_2)(x) = (T_1, T_1 + T_2)$  for  $x > \hat{x}$  and  $(\bar{T}_1, \bar{T}_2)(x) = (T_1 + T_2, T_2)$  for  $x < \hat{x}$  measures the total delay incurred before match  $x - j$  is realized.

The role of editorial delays has been studied before in a partial equilibrium framework (Leslie, 2005; Azar, 2007; Heintzelman and Nocetti, 2009), mostly with a focus on contrasting them to monetary submission costs.<sup>34</sup> Associating an explicit cost to refereeing (i.e. signal generation), these papers articulate why journals want to raise submission hurdles. However, the arguments presented here do not rely on reductions in refereeing load, and as such are *in addition* to the previously identified channels. Finally, since none of these previous models incorporates adjustments made by other journals, they do not allow welfare statements of the type considered here.

---

<sup>32</sup>The appropriate increase in red tape maintains indifference for the marginal expert  $\hat{x}$ , that is  $z_1(\hat{x}|\gamma_1, T_1) = z_2(\hat{x}|\gamma_2, T_2)$ . In the present setup, the planners problem is only well-defined under either an upper bound on  $T_1$  or a lower bound on  $\gamma_2$ . This is because as  $\gamma_2 \rightarrow 0$ , the marginal expert remains indifferent if and only if  $T_1 \rightarrow \infty$ .

<sup>33</sup>Seminal papers in development economics identify two main channels: In Shleifer and Vishny (1993), a central bureaucrat with monopoly power raises the official costs for services through red tape, and then charges bribes in exchange for lower access hurdles. In Banerjee (1997), a corrupt official is tasked with the allocation of goods to cash-constrained buyers. The official has to allocate goods efficiently for fear of detection by the government. However, rather than identifying buyers' valuation through price discrimination, the official charges maximal fees and implements red tape as a sorting mechanism.

<sup>34</sup>If such proposal fees are feasible, they represent a more efficient deterrent than red tape since they simultaneously delay proposals from  $\hat{x}$  and fully discourage proposals from the bottom end of the quality distribution. I here concentrate on situations where monetary payments are ruled out for ethical or fairness concerns, or do not serve as a deterrent for instance because submission costs are paid for directly by the department.

## 1.5.2 Second-tier Specialization

Another way in which actual gatekeepers set themselves apart is through horizontal differentiation, meaning that a gatekeeper caters primarily to a specific subset of the expert population. By way of example, academic journals often focus on a specific subdiscipline, even when the subject matter is not related to an article's quality *per se*.

To capture such tradeoffs, I here assume that incoming experts have an exogenous, quality-irrelevant type  $\theta \in \{A, B\}$  and that gatekeepers can condition their acceptance decision on  $\theta$ . Specifically, I assume that in each time interval  $dt$ , an equal mass of experts of either type arrives, each representing the same quality distribution  $F$ . I say that gatekeeper  $j$  **specializes** in  $\theta$  if she more easily accepts proposals of that type, i.e.  $\underline{\sigma}_j^\theta < \underline{\sigma}_j^{\theta'}$  for  $\theta \neq \theta'$ .

The following example illustrates the possible advantage from such specialization.

**Example 5.** Consider a situation with two gatekeepers

$$(\gamma_1, \tau_1, \kappa_1) = (1.1, 0.5, 0.22) \quad \text{and} \quad (\gamma_2, \tau_2, \kappa_2) = (1, 1, 0.41)$$

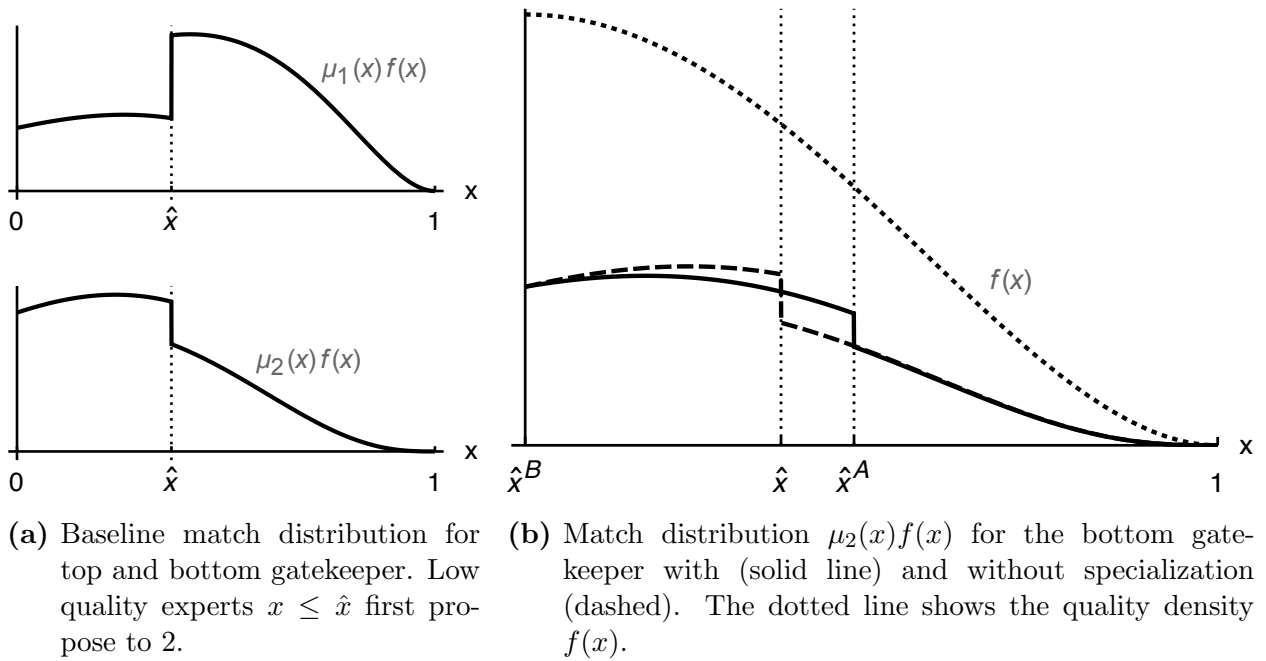
and common discount rate  $r = 10\%$ . Signals are distributed exponentially,  $G(\sigma|x) = 1 - e^{-(1-x)\sigma}$ , and the density function of expert quality is  $f(x) = \frac{15}{8}(1-x)^2(1+x)^2$ .

In the baseline scenario where experts only differ with respect to quality, the unique equilibrium with red tape is given by  $((T_1, \underline{\sigma}_1), (T_2, \underline{\sigma}_2)) = ((10.33, 2), (1, 1))$ . [Figure 1.5a](#) depicts the resulting match distribution of each gatekeeper.

When gatekeepers can condition their acceptance on expert type  $\theta \in \{A, B\}$ , the bottom gatekeeper has a strict incentive to deviate from her previous strategy and instead specialize on an arbitrary target type. [Figure 1.5b](#) depicts the change in her match distribution if she instead sets her cutoffs at  $(\underline{\sigma}_2^A, \underline{\sigma}_2^B) = (0.58, 1.58)$ . By contrast, any specialization worsens the total match quality of the top gatekeeper.  $\diamond$

The asymmetry among gatekeepers in the previous example has an intuitive explanation. Indeed, specialization affects a gatekeeper's match distribution through both a *direct* and an *indirect* effect. The former is considering the impact of specialization *conditional* on experts proposal order and is always negative: When a gatekeeper specializes on type  $A$ , she starts accepting such proposals with signals *below* the prior cutoff  $\underline{\sigma}_j$  and instead rejects  $B$ -proposals with signals above  $\underline{\sigma}_j$ . The *indirect* effect however comes from changes in experts' proposal order: Her change in cutoffs raises the acceptance rate and hence the  $z_j$  score for experts of type  $A$ , while lowering that of  $B$  types. As a result, the expert receives more first proposals in her target group and loses some in her nontarget group. By specializing, the bottom expert thus gains  $A$  proposals *above*  $\hat{x}$  and loses  $B$  proposals *below*, where  $\hat{x}$  denotes the indifferent expert in a monotone equilibrium. In contrast, when the top gatekeeper contemplates specialization, she expects to recruit any additional  $A$  proposals *below*  $\hat{x}$  while losing  $B$  proposals *above*. Intuitively speaking, the bottom gatekeeper thus specializes whenever the positive indirect effect outweighs the negative direct effect. The top gatekeeper on the other side is affected negatively by both effects, and hence generally prefers to maintain a global signal cutoff.





**Figure 1.5:** An example where the bottom gatekeeper can improve her match distribution by specializing.

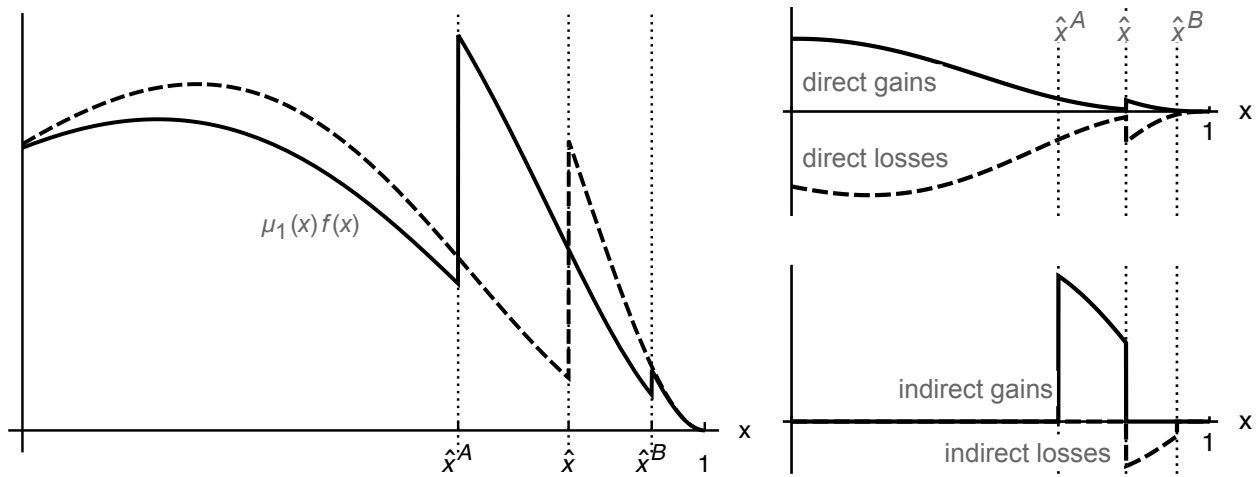
Unfortunately, the intuition misses one technical aspect that complicates the formal analysis: While the total matches are maintained by the capacity constraint, some direct gains might offset indirect losses or vice versa. For extreme cases, this may offset the negative conclusion reached above for gatekeeper 1, as illustrated by the next example.

**Example 6.** Maintaining the baseline example from above, suppose the lower bound  $\tau_1 = 1.5$  on gatekeeper 1's response time is so high that she does not employ red tape. In this case, the top gatekeeper can improve her payoff by specializing, as illustrated in Figure 1.6.

Intuitively, in this situation the value of Equation (1.2) is negative for the top gatekeeper, meaning that the marginal expert  $\hat{x} = 0.8$  is much higher than the mean quality  $\tilde{x}$  lost by a raise in threshold. Prohibited from reducing her response times, the top gatekeeper instead turns to specialization in order to replace some direct losses (around  $\tilde{x}$ ) with indirect gains (around  $\hat{x}$ ).  $\diamond$

In such a situation, the equilibrium presence of red tape comes in handy since it ensures that any marginal direct change is of equal mean quality as the marginal indirect change. This enables me to formally rule out deviations through specialization for the top gatekeeper.

**Theorem 5.A.** *Assume that signals are distributed exponentially. Consider a monotone equilibrium  $\underline{\sigma}_1 \geq \underline{\sigma}_2$  where expert types  $\theta \in \{A, B\}$  are unobservable and the top gatekeeper*



(a) Match distribution for the top gatekeeper with (solid line) and without specialization (dashed). (b) Direct net losses are here compensated with indirect net gains.

**Figure 1.6:** An example where the top gatekeeper can improve her match distribution by specializing.

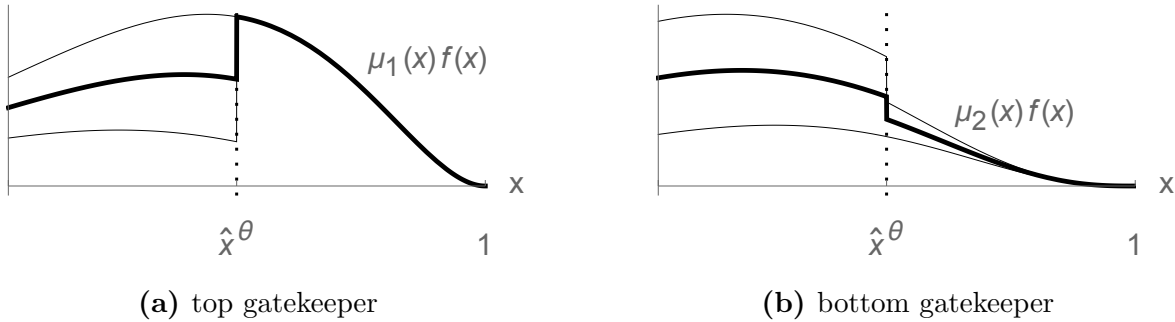
employs red tape,  $T_1 > \tau_1$ . Then as expert types become observable, the top gatekeeper cannot improve his payoff through specialization.

*Proof.* See [Appendix A.1.6](#). □

Since the bottom gatekeeper never employs red tape in equilibrium by [Theorem 3](#), her marginal direct change may be of strictly lower quality than the marginal indirect change. Thus, to give a sufficient condition for a profitable specialization, one needs to ensure both that the direct effect is strong enough and that indirect net losses do not compensate direct net gains. Both are satisfied when there are many experts of indifferent quality ( $f(\hat{x})$  is high) and the slope of the density at that point is not too negative in comparison. Both conditions can be achieved without altering the tails of the quality distribution, and indeed are satisfied often for moderate values as highlights [Example 5](#) above.

**Theorem 5.B.** Fix any monotone equilibrium  $(\underline{\sigma}_1, \underline{\sigma}_2)$  with the possibility of red tape. Denote the equilibrium response times at the top by  $T_1$  and the indifferent expert by  $\hat{x}$ . When  $f(\hat{x})$  is large enough and either  $f'(\hat{x}) \geq 0$  or  $f'(\hat{x}) = o(f(\hat{x})^2)$ , then the bottom gatekeeper has a strict incentive for specialization.

More formally, consider any sequence of ex-ante quality distributions  $F^n$  and capacities  $(\kappa_1^n, \kappa_2^n)$  such that the same gatekeeper cutoffs  $(\underline{\sigma}_1, \underline{\sigma}_2)$  along with the response times  $(T_1, \tau_2)$  remain an equilibrium for each  $n$ . Suppose further that for any  $c > 0$ , there exists  $N \in \mathbb{N}$  such that  $f^n(\hat{x}) > c$  and  $-(f^n)'(\hat{x}) < c \cdot f^n(\hat{x})^2$  for all  $n \geq N$ . In this case, the bottom gatekeeper can strictly improve her payoff through specialization for  $n$  large enough.



**Figure 1.7:** An equilibrium with specialization at the bottom. Here, the bottom gatekeeper specializes on either type  $\theta \in \{A, B\}$  with equal probability. The resulting match distributions in both the target and nontarget group are drawn in thin lines, with the mean in bold.

*Proof.* See [Appendix A.1.6](#). □

The second part of [Theorem 5](#) has important implications for equilibrium behavior as well, for it gives sufficient conditions under which at least one gatekeeper specializes. However, although the top gatekeeper cannot benefit by *unilaterally* specializing, this does not mean she will not *respond* to a differential strategy by the bottom gatekeeper. For that reason, equilibria will generally involve mixing as illustrated below.

**Example 5 (continued).** In the first example of this section, there exists an equilibrium with observable expert types  $\theta \in \{A, B\}$  such that the bottom gatekeeper specializes with equal probability on either type  $(\underline{\sigma}_2^\theta, \underline{\sigma}_2^{\tilde{\theta}}) = (0.58, 1.60)$ , while the top gatekeeper maintains an unique cutoff  $\underline{\sigma}_1 = 2.01$ . [Figure 1.7](#) illustrates the equilibrium match distributions for each gatekeeper. ◇

Ruling out equilibrium specialization for the top gatekeeper is unfortunately not generally possible.<sup>35</sup> Thus, while [Theorem 5](#) identifies sufficient conditions for the endogenous *emergence* of specialization, it does not rule out specialization by *both* gatekeepers in equilibrium. What it does say, however, is that the first impetus for specialization always comes from the bottom gatekeeper, and never from the top.

## 1.6 Applications and Potential Extensions

By referring to the two sides of the market as *experts* and *gatekeepers*, I purposely adopt generic language to encompass a wide range of real-world instances.

<sup>35</sup>The main reason for this is that any equilibrium now includes at least two marginal experts  $\hat{x}^\theta$  and  $\hat{x}^{\tilde{\theta}}$ . As a result, the profitability of non-marginal deviations becomes dependent on the value of the quality density  $f(\cdot)$  *everywhere*, among all other parameters.

A first instance of the framework is familiar to all academics and pertains to **scientific publishing** or **grant applications**. Authors (experts) are usually restricted from submitting their papers to multiple journals or grant agencies (gatekeepers) simultaneously<sup>36</sup> and wait several months until a publication decision is reached (Ellison, 2002). I focus on noise in the refereeing process as the main informational friction, a phenomenon that has been documented by several empirical studies.<sup>37</sup> As such, I abstract away from bias or noise in the author’s personal assessment of a manuscript’s quality and take journal impact as exogenous,<sup>38</sup> but I briefly explore how to incorporate either two-sided imperfect information or endogenous impact at the end of this section.

A second instance of the model concerns **technology adoption**, whereby inventors approach managers of big corporations in the hope of selling a product or technology, or **venture capitalism** where startups seek funding from investors. In both instances, the inventor is more familiar with the venture, and approaching potential customers or investors is taxing, both in terms of active time investment (search costs, customization, meetings) and passive delays while waiting to hear back from a potential partner. As search here is not necessarily exclusive, these models carry features of both simultaneous and sequential proposals. However, I argue hereafter that the pertinent features of such a setting are better captured by the present model than by a purely simultaneous setting as in Chade et al. (2014).

Third, gatekeepers are also present in **political debate**: In an indirect democracy such as the United States or the United Kingdom, only elected officials can suggest new laws. Thus, any interest group (expert) wishing to draft a new bill needs to find a lawmaker (gatekeeper) willing to sponsor it. Simplifying reality, I assume the only disagreement about the social value of a specific proposal stems from imperfect knowledge, which is more pronounced on the side of the lawmaker. Lawmakers vary in their degree of political influence as determined by committee membership or their persuasion power.<sup>39</sup>

Finally, the model also offers insights with respect to high-end **labor markets**, where job applicants spend significant time preparing for each individual application, customizing their portfolio and sometimes even earning additional, industry-specific credentials.

As with any model, some of the assumptions made in this paper are rather stylized. The actual applications motivate several important extensions, some more tractable than others.

---

<sup>36</sup>For grant applications, the most common restrictions concern simultaneous submissions within the same agency, see for example “Frequently Asked Questions”, National Science Foundation, accessed August 14th, 2015, <http://www.nsf.gov/pubs/gpg/faqs.pdf>.

<sup>37</sup>See for example Bohannon (2013); Langfeldt (2001); Rothwell and Martyn (2000); Welch (2014).

<sup>38</sup>Two other notable features I abstract away from are reputation effects on the part of the scientist (each individual submission is evaluated on its own merit) and added value from the revision process. However, as long as the latter is positively correlated with a journal’s ranking, it can partially be incorporated in the impact value.

<sup>39</sup>This study of the political process differs from the existing gatekeeping literature in political economy (see Crombez et al. (2006)). While I focus on the interaction between constituents and politicians, previous papers analyze the *interior* organization of legislative bodies and disagree on whether individual committees can block proposals before they enter the general debate.

Let me briefly consider some of them here:

First, I assume that a gatekeeper’s impact is exogenously determined and not affected by the quality of her realized matches. In instances where  $\gamma_j$  measures grant amounts or political persuasion power, that assumption is more easily accepted than in academic publishing for example, where journal impact is determined by the quality of prior publications.<sup>40</sup> To capture this feedback effect, the mean quality of matches seems to be an appropriate measure of influence, i.e.  $\gamma_j = V_j/M_j$ . Incorporating such a notion would give insight into new strategic considerations such as misrepresenting capacity and the extent of vertical differentiation. On the technical side however, the main downside to this approach is an indeterminacy brought about by multiplicity of subgame equilibria. Specifically, there can be several ways in which experts may respond to gatekeeper strategies which are not payoff equivalent. This creates issues when studying a gatekeepers strategic tradeoffs since he cannot anticipate whether a deviation would increase or decrease her payoff. To avoid this issue, in future work I plan to consider an ‘overlapping-generation’ framework where gatekeepers can choose different strategies for each generation of experts, but impacts are determined by the realized matches of the *previous* generation. Since experts now again face *exogenous* impact, their search behavior is uniquely determined (up to indifference in the  $z_j$ -scores). On the other hand, this framework is rich enough to consider gatekeepers’ long-term strategic positioning.

A second modeling assumption concerns the informational environment. By postulating one-sided imperfect information, I assume that experts are perfectly informed about their own desirability. For grant applications, political ventures or the technological potential of startup firms, it seems reasonable to assume that experts are more adept judges of quality. In academic publishing however, opinions diverge on whether authors or reviewers see more clearly. [Chade et al. \(2014\)](#) make the important observation that a world with two-sided imperfect information is equivalent to one where gatekeepers draw *affiliated* signals. Intuitively, referring to the expert’s signal ‘ $x$ +noise’ as his *type*, gatekeeper evaluation now has a common component ‘–noise’. Since experts now simultaneously learn about their true quality, the equivalence to a history-independent  $z_j$ -score breaks down, and it seems necessary to reduce the analysis to just two gatekeepers for tractability. For that case however, it can be shown that conditional on marginal acceptance probabilities, more experts first approach the high-impact gatekeeper when signals are more affiliated (more uncertainty is on the expert side) – leading one to expect that such situations exhibit more crowding at the top. Intuitively, acceptance at the bottom is now bitter-sweet, since it simultaneously raises the expert’s own quality estimate, which causes him to regret his low aim. Unfortunately, a meaningful comparison about equilibrium behavior is complicated by the fact that changes in relative noise have a heterogeneous effect on experts’ marginal acceptance probabilities, and the extent depends largely on the specific cutoffs employed in equilibrium. All-in-all, while solving a specific two-gatekeeper game with two-sided asymmetric information seems tractable, it appears difficult to draw insightful equilibrium comparisons.

---

<sup>40</sup>Even so, [Card and DellaVigna \(2013\)](#) mentions that journal reputations are sticky and remain broadly stable over the years, so reality is probably a mixture between exogenous and endogenous factors.

In this paper, I assume that match utilities are non-transferable throughout and equal to the quality or impact of one’s match partner. While this is not an uncommon assumption, it fits some applications better than others. For venture capital, ownership negotiations are an integral part of the match process. As such, it may seem more suitable to assume that match surplus is determined jointly by a supermodular<sup>41</sup> function  $\phi(x, \gamma_j)$  which can be split arbitrarily across the two match partners. Importantly however, if surplus is split according to exogenously determined shares  $(\beta, 1 - \beta)$  between experts and gatekeepers,<sup>42</sup> then the qualitative findings from this paper carry over, albeit under modified  $z_j$  scores that take into account the new payoff structure,

$$z_j(\alpha_j, x) = \frac{e^{-rT_j} \alpha_j \beta \phi(x, \gamma_j)}{1 - e^{-rT_j} (1 - \alpha_j)}.$$

Since experts’ own quality now directly affects their payoff, they no longer share the same score function – higher-quality experts are now even more eager to match quickly – but the qualitative reaction to parameter changes are maintained. Similarly, gatekeeper’s cardinal preferences now differ, but they still seek to match with the highest-quality experts and as such the qualitative tradeoffs remain: red tape might still occur (possibly both at the top and the bottom) and by improving assortativity, it can still be welfare-improving.

In situations where experts can approach multiple gatekeepers simultaneously but proposals are costly, experts face a portfolio choice problem (Chade and Smith, 2006). However, this doesn’t imply that these applications are better described by the one-shot model in Chade et al. (2014), for although experts are no longer restricted to a single proposal *per moment of time*, they are also not restricted to a single *opportunity* for proposals. Indeed, a defining feature of their college application model is that all proposals *must* be made simultaneously, and if an expert forgoes proposing to a specific gatekeeper  $j$ , he can no longer seek out  $j$  in case his other proposals don’t work out. It is this possibility of regret that generates equilibrium multiplicity and thus prevents an unambiguous reduced-form discussion of gatekeeper competition.<sup>43</sup> Such a strict deadline is generally absent in grant applications, venture capitalism, technology adoption or political debate, where experts can pursue new options if the initial proposals don’t work out. Therefore, while both models simplify reality

---

<sup>41</sup>Supermodularity maintains assortative matching as the socially efficient outcome.

<sup>42</sup>Such a split is obtained under Nash bargaining with threat points zero, which fits a situation where match decisions are made *before* surplus negotiations take place and any match dissolution implies foregone capacity for the gatekeeper and an immediate exit for the expert.

It also seems like an interesting and nontrivial extension to implement more general solutions of Nash bargaining, where the threat points are determined endogenously from continuation payoffs, which would need to go hand-in-hand with public history to ensure that everybody can correctly anticipate their match payoffs.

<sup>43</sup>Intuitively, consider a mediocre expert who proposes to both gatekeepers. As his acceptance rate at the top increases, his marginal benefit from a ‘safety proposal’ to the bottom gatekeeper decreases and he may limit himself to only proposing to the top gatekeeper. As such, although his acceptance rate weakly increases at both gatekeepers, the expert’s equilibrium *match* probability goes down, contrary to the monotonicity resulting from [Observation 1](#).

for the sake of tractability, I argue that assuming a limit on the number of *simultaneous* rather than on the number of *sequential* proposals both better captures the essence of the applications considered here and offers greater tractability benefits.

Throughout this analysis, I have assumed that gatekeepers have no control over the *precision* of their screening technology by exogenously restricting the conditional signal distributions to  $G(\cdot|x)$ . Interesting questions can be raised if gatekeepers can instead improve their signal precision through either monetary investments or a more time-intensive review process.<sup>44</sup> In general, such an increase in signal precision will have a doubly positive effect on matches: A *direct* improvement of the gatekeeper’s own proposal filtering and an *indirect* raise in the quality of proposals.<sup>45</sup> The relative strength of these incentives for the top or bottom gatekeeper will however crucially depend on distributional assumptions regarding ex-ante quality and signal structures. Moreover, since the indirect part of a gatekeeper’s match improvement stems from a reordering of expert proposals, the private incentives outweigh social welfare considerations, and there will generally be inefficiently high investment in such screening technology.

Lastly, one may care to relax the assumption of private expert history, at least to the point of making arrival times public. Indeed, especially since the advent of internet technology, it is often possible to detect how long an idea has already circulated. To avoid inference issues, the model promises to actually be *more* tractable if the full proposal history is known rather than just the arrival time.<sup>46</sup> As acceptance histories become conditional on proposal history, experts’ proposal order gets computationally more heavy, but a setup with just two gatekeepers seems tractable and insightful. It seems of particular policy interest to analyze how this impacts match delays. On one side, there is likely going to be an acceptance penalty for second submissions (due to previous low signal draw), forcing experts to target their first proposal more strategically. At the same time, this also raises the potential returns from proposing boldly, as the top gatekeeper becomes particularly optimistic about his first proposals. It seems likely that either effect can outweigh for an individual expert depending on the exact distributional assumptions.

## 1.7 Conclusion

This paper develops a model for sequential directed search in a two-sided matching market. The framework captures important tradeoffs that guide decisions in many areas including

---

<sup>44</sup>I here concentrate on situations where gatekeepers commit to a signal precision before the game begins. An alternative setup would be one where gatekeepers solve an optimal stopping problem to determine how deeply to investigate each proposal. Unfortunately, the latter problem requires explicit distributional assumptions even for a single gatekeeper.

<sup>45</sup>Indeed, any capacity-clearing cutoff  $\sigma'_j$  will raise acceptance rates above some quality  $x_0$  and lower those below. As a result, experts both propose and match more often with  $j$  if and only if  $x \geq x_0$ .

<sup>46</sup>The latter may generate incentives for experts to wait idly in order to pool with those of a different proposal history. However, for the most tractable case of two gatekeepers, the two are equivalent since first proposals always happen right upon arrival.

academic publishing, political decision-making, venture capitalism and grant allocations. Heterogeneous experts all seek to match with the most influential gatekeepers, but inherent quality differences endogenously create an endogenous distinction in their sensitivity to delay. In turn, the long-lived gatekeepers may exploit this aversion by using unnecessary delay or horizontal differentiation to influence experts' search decisions. Apart from these new findings, the main contribution of the paper is to provide a tractable equilibrium framework that speaks to a broad range of applications and offers exciting opportunities for future research.



## Chapter 2

# Mentoring and the Dynamics of Affirmative Action

For decades, affirmative action has remained a topic of heated debate and wildly opposing opinions. While some view it as a form of discrimination against historically favored social classes, others see in it the only way of confronting the reality that race and gender still matter for educational achievement and employment outcomes (see e.g. [Sotomayor \(2014\)](#)). Part of the political discourse focuses on arguments of justice and righting historical wrongs, topics which cannot be addressed through economic research. However, the economic angle of *productivity tradeoffs* is least as relevant for policy decisions.

There is empirical evidence that race and gender affect an individual's career prospects irrespective of his or her innate ability ([Milkman et al., 2014](#); [Ellison and Swanson, 2009](#)). Such differences in hiring rates may arise through *preference bias*, where decision makers innately favor members of a certain groups, or it may come through *statistical discrimination*, where rational utility-maximizers infer imperfectly observable information on productivity from correlated, but utility-irrelevant characteristics such as race or gender (for a comprehensive survey, see [Fang and Moro \(2011\)](#)). In particular, certain groups may be stuck in an equilibrium with little skill investment and poor employment prospects, while more fortunate types owe their high returns to education to favorable equilibrium beliefs.

We here focus on a third channel that comes from complementarities among productive members of the same type. Specifically, we consider mentorship relations between senior professionals and young students. It has been empirically shown that mentoring relationships are stronger and more common between members of the same demographic group ([Ibarra, 1992](#); [Dreher and Cox Jr., 1996](#)) and that the availability of similar role models affects the academic performance of individuals ([Carrell et al., 2010](#)). In such a world, an uneven composition of the senior workforce may result in tangible productivity differences among junior workers that affect schooling and employment decisions.

This paper considers a model that incorporates the trade-off between the strong mentoring complementarities in a homogeneous work force and the optimal use of innate ability

that is common to workers of all types. Our results can be applied to minorities of any type, such as gender, race, disability status or other demographic characteristics. Specifically, we consider a continuous-time overlapping generations model where established mentors lower the cost of education to the incoming cohort. This mentorship boost differs across types and is increasing in the availability of mentors of one's own type.

Due to the lack of suitable mentors, minority students face more hurdles in their education and are less likely to obtain a degree than their peers of equal ability.<sup>1</sup> This affects their employment prospects in the unsaturated labor market during the latter part of their professional life. We assume that firms are either too small or myopic in order to internalize the long-run mentorship dynamics, and thus care only about short-term productivity without regard to the fact that today's hires affect tomorrow's candidate pool.

The properties of the mentorship boost function determine the dynamics of this system, including steady state properties and instantaneous surplus. We show that both a completely homogeneous and a balanced work force can be stable steady states of the economy. A homogeneous work force makes maximal use of the mentoring complementarities, while a balanced base of mentors optimally promotes the innate ability of all workers and often increases total labor force participation by harnessing the talent of both types.

The goal of the paper is to develop conditions under which regulations can be used to increase total productivity of the economy and to compare the effectiveness of different policies. In particular, we are interested in the following questions: Under which conditions does regulation improve long-term labor productivity? How radical and long-lived should the optimal policy be? Who are the losers and winners of different policies?

Our main result shows that a sufficiently patient planner may intervene persistently in favor of the minority. Indeed, the long-term benefits of additional minority mentors may outweigh the subsidy required to attract them. This is in contrast to much of the public discourse that sees affirmative action as a temporary policy tool to correct a historical imbalance.<sup>2</sup> An efficient strategy achieves this goal quickly, the reasoning goes, and thereby renders itself obsolete in a relatively short amount of time. Our model is able to capture the history-dependence that underlies this argument. It can formalize how temporary programs may divert the long-time convergence from one steady state to a more efficient one. However, it also proposes another rationale for affirmative action, beyond the mere correction of a suboptimal starting point: Myopic individuals fail to fully internalize mentoring tradeoffs. This prevents them from implementing the surplus-maximizing level of diversity on their own, and justifies ongoing market intervention.

---

<sup>1</sup>Along these lines, the academic senate at Berkeley writes: "Information about grants, the politics of a department, or opportunities for participation in professional activities are often shared with younger colleagues upon meeting them in social settings to which some groups may have less access. Lacking power and access to the occasions in which power is shared, some groups are offered different opportunities for participation in the academic world." (SWEM (2012), p.18)

<sup>2</sup>Consider for example the debate following the Supreme Court decision to uphold a ban on affirmative action in Michigan's public university admissions (Kahlenberg et al., 2014). Most arguments center around the question of whether or not historical racial discrimination has been overcome.

We also compare different policies instruments, such as educational subsidies and workplace hiring quotas under varying assumptions on wage determination. In our framework, educational subsidies are the most versatile tool to influence labor force evolution. When it comes to *maintaining* the optimal workforce composition however, hiring quotas are just as effective as long as the competitive environment allows for type-specific signing bonuses. When wage setting is restricted for exogenous reasons (such as cultural norms or firm-intern politics), hiring quotas however cause a significant crowding out of majority workers. The dire employment prospects dampen their investment, yet some of them still pursue an ex-post worthless education.

**Related Literature.** Our analysis provides a rationale for effective affirmative action, which is in contrast to some previous theoretical research. Our opposing predictions stem from differing assumptions on the source of the hiring imbalance. Indeed, when taste-based discrimination is at work, affirmative action is essentially a zero-sum game where the benefit to the minority is offset by a direct utility loss of the majority. Under statistical discrimination, larger minority participation need not translate into updated beliefs. Quite to the contrary: Under certain parameter values, employment quotas may actually reinforce negative stereotypes against certain groups [Coate and Loury \(1993\)](#). The intuition is simple: When minority employment is mandated by law, firms may have to hire minority members even if they are unskilled. This in turn may actually *reduce* the minority's returns to education and thereby further lower equilibrium skill investment. A similar conclusion is reached when agents infer their personal success probability from their own type's employment history as in [Chung \(2000\)](#). In an unregulated market, observing successful people with similar characteristics sheds a positive light on one's own prospects in the labor market and hence encourages investment. However, these positive inferences disappear under temporary hiring restrictions, and agents will not be any more optimistic once the employment constraints are lifted. It is important to emphasize that these arguments rely purely on informational inferences and assume no direct productivity benefit from relatable role models.

Together, these models seem to suggest that affirmative action is futile at best, or downright harmful at worst. If individuals however receive a tangible productivity boost from mentoring, we show that a more positive view is warranted.

The group complementarities at the heart of our model have been observed in the empirical literature. [Dreher and Cox Jr. \(1996\)](#) find that not only were female MBA graduates and students of color less likely to form mentoring partnerships with white men, but these missing relationships also had a tangible impact on later compensation. Indeed, students mentored by white men earned on average \$16,840 more annually than those with mentors of other demographic profiles. In a similar vein, [Ibarra \(1992\)](#) analyzes the professional network within an advertising firm and finds that differential patterns of network connectivity helped men reap greater network returns than a woman in the same position. [Bettinger and Long \(2005\)](#) find that an increased share of female faculty positively influences course selection and major choice for female undergraduates in some (though not all) disciplines where

women are historically underrepresented. Observations by [Carrell et al. \(2010\)](#) also seem to support the hypothesis of a measurable productivity boost from relatable role models. They show that assigning a female professor to mandatory introductory science and math classes significantly increases course grades for female students without hurting male performance, to the point of eradicating the gender gap in grades and STEM majors.

Contrary to [Krishnamurthy and Edlin \(2014\)](#), we assume no ex-ante differences across the two groups, which we argue is the more interesting case for long-term policy design.<sup>3</sup> The structure of our model is similar to that of [Athey et al. \(2000\)](#), as they study the intertemporal promotion decision in long-lived firms. As in our model, they assume that senior workers offer an additive mentorship boost to junior employees, and that the size of this boost is increasing function in the availability of same-type mentors. We believe that our two papers are complementary: They offer additional theoretical insight into the family of boost functions that admit certain steady states, while we restrict our attention to a specific mentoring function and contrast different ways in which a social planner can guide myopic firms to a more productive equilibrium. They highlight the possibility of a “glass ceiling” (a stable steady state that is less diverse than the overall population), and we establish a rationale for ongoing market intervention. Lastly, by studying the case of an unsaturated labor market (which seems particularly relevant for high-skill sectors), we show how these mentoring externalities affect both the intensive and the extensive margin of social surplus.

## 2.1 Model

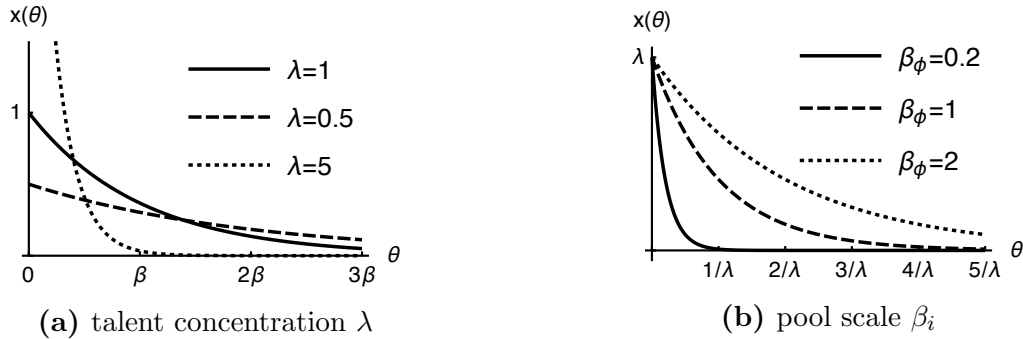
We study the continuous-time evolution of a heterogeneous, high-skill labor force, with an infinite pool of new potential recruits at each moment in time. Upon birth, each junior has a one-time opportunity to invest into costly education, and thereafter seeks employment in a competitive and unsaturated labor market. For the sake of simplicity, we assume that education is instantaneous and the only determinant of later productivity.<sup>4</sup> Formally, each unit mass of educated worker contributes one unit to a firm’s profit flow, while uneducated workers do not improve firm output. Assuming free entry of firms, this ensures that wages for educated workers equal one per unit of time.

A worker invests into education if his expected lifetime earnings outweigh the cost of education. Life expectancy follows a standard exponential distribution with parameter 1, simplifying expected lifetime earnings to  $w = 1$ . Workers are heterogeneous with respect

---

<sup>3</sup>By assuming that minorities have innately lower skill levels, [Krishnamurthy and Edlin \(2014\)](#) suggest that colleges should set *higher* admission standards for minority students in order to ensure that the same inferences are made on the ability of its graduates. The analysis fails to explain how those skill differences are created in the first place, and we believe that any policy recommendations won from such an analysis could thus be optimal only over a very short horizon.

<sup>4</sup>Real-world examples that fit this (admittedly stylized) description include any sectors where a diploma is the main hiring criterion. One may think of specialized exams such as the Bar license for lawyers or the PE license for engineers, which in and of themselves don’t increase a candidate’s effectiveness but are required to perform certain functions.



**Figure 2.1:** Ability distribution as a function of parameters  $\lambda$  and  $\beta_i$ .

to educational cost. Formally, the cost of education  $c$  is decreased both by the worker's individual ability  $x$  and the type-specific strength of mentoring  $\mu_i$ . For simplicity, we assume that there are no complementarities between ability and mentoring efficiency. As a result, investment is favored whenever

$$c - x - \alpha\mu_i \leq w, \quad (2.1)$$

where  $\alpha > 0$  measures the relative importance of mentoring versus innate skill.

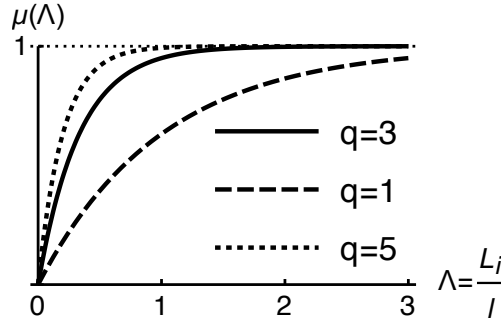
While the two groups may be of different size, there are no differences with respect to innate ability: Among the infinite pool of juniors born at time  $t$ , ability is distributed according to a decreasing *talent function*  $x : [0, \infty) \rightarrow [0, \infty)$  irrespective of type. By choosing the functional form

$$x(\theta) = \lambda e^{-\lambda\theta},$$

we are able to isolate the role of talent *concentration*  $\lambda > 0$ . To account for differences in pool size, we further introduce *scale* parameters  $\beta_1 \geq \beta_2 > 0$  and refer to type 1 as the *dominant* type. Formally, a mass  $\beta_i\theta$  of each type has ability larger than  $x(\theta)$ , for all  $\theta > 0$ . The two parameters measure complimentary features of the talent pool. Indeed, the latter corresponds to total talent in the workforce,  $\int_0^\infty x(\theta/\beta_i)d\theta = \beta_i$ , while the former dictates how much of it is concentrated among the top candidates, as illustrated in [Figure 2.1](#).

Active workers form mentoring relationships with incoming students, which also reduces the cost of education. However, mentoring is more effective between workers of the same type, and mentorship assignment is imperfect.<sup>5</sup> Moreover, there are decreasing returns to scale from mentorship for an individual student. To capture these effects parsimoniously, we consider first a discrete matching market with  $nL_i$  seniors of type  $i$  and  $nl$  students (of either type). Mentor matching is random, and the likelihood of an individual link depends on the ratio of seniors to students. Specifically, we assume that each mentor-student link exists with probability  $\frac{q}{nl}$ , where  $q$  denotes the average number of students per mentor. A student of type  $i$  enjoys a mentorship boost of 1 if and only he or she is being mentored by at

<sup>5</sup>Reasons for this matching friction could pertain to the availability of mentors within a certain sub-discipline, geographical area or individual firm.



**Figure 2.2:** Mentorship boost as a function of mentor availability  $\Lambda = L_i/l$ , at various mentor capacity levels  $q$ .

least one same-type mentors.<sup>6</sup> By the Law of Rare Events, the number of same-type mentors per students can be approximated by the Poisson distribution as  $n$  grows. The probability of reaching the critical mentoring mass, and hence the expected mentorship boost, is therefore equal to

$$\mu(\Lambda) = 1 - e^{-q\Lambda},$$

where  $\Lambda = L_i/l$  is the ratio of mentors to mentees. Figure 2.2 illustrates how mentor capacity  $q$  affect this mentorship boost function.

Close inspection of the mentorship boost function reveals interesting dynamics: Since mentors are increasingly busy during periods of labor growth, students compete more fiercely for mentorship opportunities. Moreover, all mentorship relations are (weakly) beneficial in the sense that adding senior workers of either type does no harm to either type, *ceteris paribus*. Indirectly however, seniors of the opposite type hurt juniors indirectly. These new mentors don't reduce the junior's *own* cost of education, but attract additional students from the opposite type, and thereby dilute the market for mentors.

Two restrictions on the cost parameter are necessary for realism and tractability, so as to ensure that labor supply never completely dries out or explodes. For the remainder of the paper, we therefore assume

$$0 < c - 1 - \alpha\mu(1) < \lambda. \quad (\text{H1})$$

On page 40, we provide intuition for the dynamic implications of these assumptions.

### 2.1.1 Unregulated Market Dynamics

Since workers' lifespan follows a standard exponential distribution, the total labor force at time  $t$  can be written as

$$L_i(t) = \int_{-\infty}^t e^{-(t-\tau)} l_i(\tau) d\tau,$$

<sup>6</sup>For  $n$  large enough, the boundary conditions are satisfied in so far as the link probabilities are non-degenerate,  $0 < \frac{q}{nl} < 1$ .

where  $l_i(\tau)$  is the mass of type- $i$  workers born at time  $\tau$  who choose to invest into education. The advantage of exponential life expectancy is that the mass of workers at a moment  $t$  is a sufficient statistic for future labor force development. Indeed, by Leibniz' rule, the law of motion is given by

$$\dot{L}_i(t) = l_i(t) - L_i(t) \quad \forall i \in \{1, 2\}. \quad (2.2)$$

At each moment in time, individual incentives pin down the mass of new recruits  $(l_1, l_2)$  as a function of mentor availability  $(L_1, L_2)$  as

$$\begin{cases} c - x \left( \frac{l_1}{\beta_1} \right) - \alpha \mu \left( \frac{L_1}{l_1 + l_2} \right) = w & \text{or } l_1 = 0 \text{ and } c - \lambda - \alpha \mu \left( \frac{L_1}{l_2} \right) > w, \\ c - x \left( \frac{l_2}{\beta_2} \right) - \alpha \mu \left( \frac{L_2}{l_1 + l_2} \right) = w & \text{or } l_2 = 0 \text{ and } c - \lambda - \alpha \mu \left( \frac{L_2}{l_1} \right) > w, \end{cases} \quad (2.3)$$

where wage  $w$  equals 1 for the case without market intervention.

By a single crossing argument, these equations uniquely pin down the labor supply.

**Lemma 2.1.1.** *Let  $\mathbb{R}_+^2 = \{(L_1, L_2) \in \mathbb{R}^2 \mid L_1 \geq 0, L_2 \geq 0 \text{ and } (L_1, L_2) \neq (0, 0)\}$ . Over this set, Equation (2.3) admits a unique solution  $(l_1, l_2) \in \mathbb{R}_+^2$  for any senior labor force  $(L_1, L_2) \in \mathbb{R}_+^2$ .*

*Proof.* Consider a total student body of size  $l$ , and note that  $\mu(L_i/l)$  is strictly decreasing in  $l$ . Thus, larger values of  $l$  attracts fewer students  $l_i$  of either type. In other words, as  $l$  grows, school enrollment  $l_1 + l_2$  (weakly) drops, implying a single crossing  $l = l_1 + l_2$ .  $\square$

We say that the economy is in a *steady state* when  $\dot{L}_i(t) = 0$  for both types  $i \in \{1, 2\}$ . We call such a steady state *stable* whenever a small perturbation does not affect the long-term convergence, i.e. when there exists  $\varepsilon > 0$  such that  $\lim_{t \rightarrow \infty} L(t) = \hat{L}$  for all  $L(0) \in \mathbb{R}_+^2$  with  $\|L(0) - \hat{L}\| < \varepsilon$ .

Whenever any one type is completely absent in the labor force, we refer to it as a *homogeneous* labor force  $L \in \mathbb{R}_{>0} \times \{0\} \cup \{0\} \times \mathbb{R}_{>0}$ . As it can easily be seen, the labor supply of a type may dry out in the long run if and only if even the most able individuals require a mentorship boost,

$$c - \lambda \geq 1. \quad (\text{H2})$$

As the following result shows, two homogeneous steady states exist if and only if H2 holds, and they are stable whenever the inequality is strict. Moreover, there exists at least one mixed steady state  $\hat{L} \in \mathbb{R}_{>0}^2$  if and only if

$$c - \lambda - \alpha \mu(0.5) < 1, \quad (\text{H3})$$

and one of them is stable whenever  $q$  is big enough.

**Theorem 6.** *The economy admits two homogeneous fixed points*

$$(\beta_1 x^{-1}(c - 1 - \alpha \mu(1)), 0) \quad \text{and} \quad (0, \beta_2 x^{-1}(c - 1 - \alpha \mu(1)))$$

whenever [Hypothesis H2](#) holds. They are stable if and only if the inequality [H2](#) is strict.

The economy also admits a mixed steady state  $(\hat{\phi}\hat{L}, (1 - \hat{\phi})\hat{L})$  with  $\hat{\phi} \in (0, 1)$  and

$$c - x \left( \frac{\hat{\phi}\hat{L}}{\beta_1} \right) - \alpha\mu(\hat{\phi}) = 1 \quad \text{and} \quad c - x \left( \frac{(1 - \hat{\phi})\hat{L}}{\beta_2} \right) - \alpha\mu(1 - \hat{\phi}) = 1 \quad (2.4)$$

if and only if [Hypothesis H3](#) holds. As long as

$$c - 1 - \alpha > 0, \quad (\text{H1}^+)$$

one of them is stable for high enough mentoring capacity  $q$ .

*Proof.* See [Appendix B](#). □

Proving the existence of stable homogeneous steady states is straightforward and follows immediately from [Hypothesis H2](#). Indeed, whenever one type's mentor availability dips below a certain positive threshold, all educational investment ceases. This further lowers mentor availability, until that type eventually abstains completely. On the other hand, a mixed steady state requires a continuous investment by the minority, which occurs exactly when [Hypothesis H3](#) holds. The assumption [H1](#)<sup>+</sup> ensures that [Hypothesis H1](#) maintains a strict inequality in the limit  $\lim_{q \rightarrow \infty} \mu(q) = 1$ . It rules out the knife-edge case where educational investment grows without bounds as mentor availability improves. If [H1](#)<sup>+</sup> holds, the mass of students never exceeds  $\beta_i x^{-1}(c - 1 - \alpha) < \infty$ . Whatever the ratio of mentors to students, it comes arbitrarily close to that level as the mentoring capacity grows, since  $\mu(\Lambda) \xrightarrow{q \rightarrow \infty} 1$  pointwise. The proof formally shows that this implies the existence of a stable steady state near this upper bound.

At this point, it is illuminating to discuss the role of the cost assumptions in [Hypothesis H1](#). Indeed, a homogeneous labor force minimizes mentoring frictions, yielding a steady-state mentorship boost of  $\mu(1)$ . No positive steady state can possibly exist unless at least the most able junior obtains an education under this maximal sustainable mentorship boost,

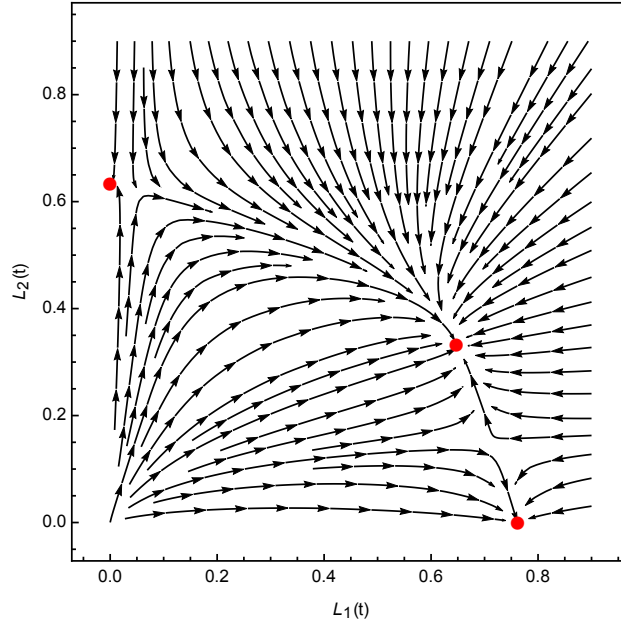
$$c - \lambda - \alpha\mu(1) < 1. \quad (\text{H5})$$

Conversely, we want to rule out infinite growth of the labor force. We achieve this by ensuring that the cost of education is prohibitive for zero-ability individuals under  $\mu(1)$ ,

$$c - \alpha\mu(1) > 1. \quad (\text{H6})$$

[Figure 2.3](#) highlights the unregulated dynamics of a sample economy with three stable steady states.





**Figure 2.3:** Labor force evolution  $\dot{\mathbf{L}}(t)$  for any initial labor force composition  $\mathbf{L}(t)$ . The three stable steady states  $\hat{\mathbf{L}} \in \{(0.76, 0), (0.65, 0.33), (0, 0.64)\}$  are highlighted in red. Parameter values are  $c = 2.1$ ,  $\alpha = 0.6$ ,  $\lambda = 1$ ,  $\beta_1 = 1.2$ ,  $\beta_2 = 1$ ,  $k = 1$ ,  $q = 3$ .

### 2.1.2 Welfare

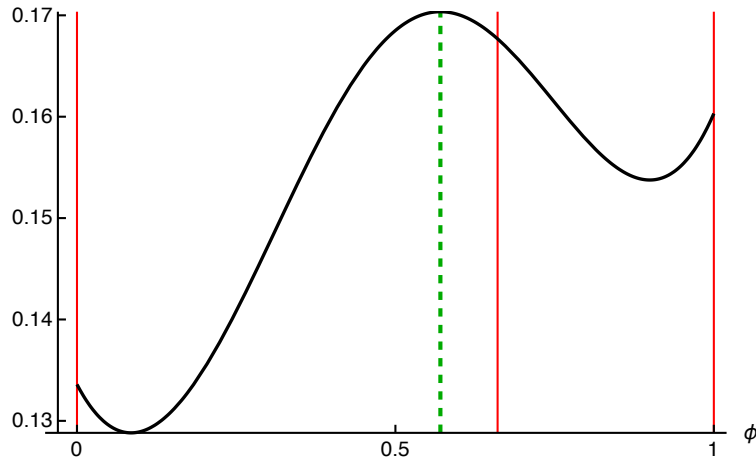
Due to perfect competition in the hiring market, the surplus is entirely captured by high-ability students whose expected lifetime earnings outweigh their cost of education. As such, instantaneous surplus is given by

$$\begin{aligned} \pi(\mathbf{L}, \mathbf{l}) &= \sum_{i=1}^2 \int_0^{l_i} 1 - c + x(\theta/\beta_i) + \alpha\mu \left( \frac{L_i}{l_1 + l_2} \right) d\theta \\ &= \sum_{i=1}^2 \beta_i (1 - e^{-\lambda \frac{l_i}{\beta_i}}) + l_i \left( 1 - c + \alpha\mu \left( \frac{L_i}{l_1 + l_2} \right) \right), \end{aligned}$$

where  $\mathbf{L} = \mathbf{L}(t) \in \mathbb{R}_+^2$  denotes the instantaneous labor force and  $\mathbf{l} = \mathbf{l}(t)$  the new recruits. Without intervention,  $\mathbf{l}(t)$  is defined implicitly by Equation (2.3) and wage  $w = 1$ .

Labor force composition affects the unregulated surplus through both the extensive and the intensive margin. Specifically, as mentoring improves, *more* students invest into education and the cost of education decreases at each ability level. The optimal market intervention trades off these benefits for the minority against the corresponding reduction in majority mentoring. Formally, a benevolent dictator with full control over individual investments and discount rate  $r > 0$  solves the Hamilton-Jacobi-Bellman equation

$$rV(\mathbf{L}; r) = \max_{\mathbf{l} \in \mathbb{R}_+^2} \pi(\mathbf{L}, \mathbf{l}) + V_{\mathbf{L}}(\mathbf{L}; r) \cdot (\mathbf{l} - \mathbf{L}). \quad (2.5)$$



**Figure 2.4:** Steady-state welfare  $\max_L \pi((\phi L, (1 - \phi)L), (\phi L, (1 - \phi)L))$  under the optimal labor size at any given mixture  $\phi$ . Steady-state mixtures are indicated by solid red lines. As highlighted by the dashed green line, the optimal labor force composition is more diverse ( $\phi \approx 0.57$ ) than in the mixed steady state ( $\phi \approx 0.66$ ). Parameters are as in Figure 2.3.

Let the solution to this problem be  $\mathbf{l}^*(L; r)$ . The optimal labor force evolution  $\mathbf{L}^*(t)$  is then given by the differential equation

$$\begin{cases} \dot{\mathbf{L}}(t) = \mathbf{l}^*(\mathbf{L}(t); r) - \mathbf{L}(t), \\ \mathbf{L}(0) = \mathbf{L}_0. \end{cases}$$

Solving for the optimal continuation for arbitrary initial conditions  $\mathbf{L}_0$  is computationally infeasible. However, it is possible to characterize the optimum among a more restricted class of policies, namely those where labor force is artificially constrained to a constant level  $\mathbf{L} \in \mathbb{R}_+^2$ .

**Theorem 7.** *Steady-state welfare  $\pi(\mathbf{L}, \mathbf{L})$  is maximized at some  $\mathbf{L}^* \in \mathbb{R}_{>0}^2$ . An interior solution corresponds to a steady state if and only if  $\beta_1 = \beta_2$ .*

*Proof.* See Appendix B. □

This result has important practical consequences. In particular, it shows that a sufficiently patient planner would intervene persistently in favor of the minority, as long as he is faced with two groups of unequal size, and skill recruitment is important enough so as to warrant positive labor supply from both types. Indeed, the result shows one particular intervention that – while not necessarily optimal – generates positive social surplus, and hence dominates a ‘laissez-faire’ regime. Figure 2.4 highlights the optimal steady-state welfare along with the ‘native’ stable steady states for the previously discussed parameters.

### 2.1.3 Alternative Interpretation: Job Appeal

The additive productivity boost through mentoring is just one interpretation of our model. Although supported by empirical evidence, the salience of this feature may vary across sectors. To underline the relevance of our model, we here discuss an entirely different setup without mentoring tradeoffs that nonetheless yields the same dynamics.

In particular, suppose that the *appeal* of a particular profession depends on the race and gender representation of its current workforce. For instance, recent political events have drawn criticism about the underrepresentation of African Americans in the police force. It is often argued that the racial imbalance in law enforcement is both cause and result of poor trust ratings among minorities.<sup>7</sup> In a similar way, young girls may steer away from predominantly male fields not because they doubt their own chances of success (as in Chung (2000)), but because they inherently enjoy having professors and colleagues of the same gender.<sup>8</sup> By interpreting  $\mu$  as type-specific extra utility enjoyed from working in the high-skilled industry, such a setup is equivalent to the previous model.

## 2.2 Affirmative Action Policy Instruments

Having thus shown conditions under which the policy maker can improve welfare through market intervention, we here turn our focus to the practical implementation of such a policy. Since educational decisions are taken individually, a social planner can merely nudge the independent actors into the right direction. In other words, we have shown that long-term welfare is higher under different investment decisions, but it remains to show that they *can* be modified in such a way.

In this section we contrast two options that allow the policy maker to implement desired investment levels: Direct modification in the cost of education (through a tuition raise and fellowships), as well as labor market regulations.

Before we define the specific policies, we add slightly more notation in order to simplify the discussion. First, we call a hiring quota  $\hat{\phi}$  **binding** at a state  $\mathbf{L} \in \mathbb{R}_+^2$  if it requires a more diverse workforce than what would myopically be implemented, i.e.  $\hat{\phi} < \frac{l_1}{l_1+l_2}$ . Moreover, while an unsaturated labor market guarantees employment security for all students, this may no longer be the case in a regulated market. To clarify this effect, we thus use  $\bar{\mathbf{l}}(t) \geq \mathbf{l}(t) \in \mathbb{R}_+^2$

---

<sup>7</sup>In a recent article published in the *Wall Street Journal*, Kesling and McWhirter (2015) mention the struggle to recruit and retain African-Americans, which they argue is at least partly explained by “black distrust of police departments”. Diversity at individual departments may vary specifically *because* high-level recruiters seek to counteract this imbalance. Indeed, citing the recruitment officer for St.Louis County Police Department, they write that good quality minority applicants “can write their own ticket” by filling out applications for multiple departments and then picking the most appealing one. This preferential treatment of *qualified* minority applicants under Affirmative Action efforts gives further credence to the type-dependent wage assumption in Section 2.2.2.

<sup>8</sup>The effects measured in Bettinger and Long (2005) may in fact measure either expected mentoring benefits or pure taste preference.

to denote the mass of **educated** and **employed** individuals of each type born at time  $t$ , respectively.

### 2.2.1 Educational incentives

The most direct market intervention directly modifies the cost-benefit analysis of prospective students. Because the labor market remains unrestricted, all educated workers find employment,  $\bar{l} = l$ , and expected returns to education remain equal to  $w = 1$ . As such, a combination of fellowship and/or tuition hikes  $\Delta \in \mathbb{R}^2$  with

$$c + \Delta_i - x \left( \frac{l_i}{\beta_i} \right) - \alpha \mu \left( \frac{L_i}{l_1 + l_2} \right) = 1$$

allows the policy maker to implement any desired labor market inflow  $l \in \mathbb{R}_+^2$ .<sup>9</sup>

One critical observation from our model is that whenever we discuss fellowships  $\Delta_i > 0$ , those are assumed to be available to *all* interested minority students. It is straightforward to see that ability-based fellowships *only* affect the extensive margin if the available pool exceeds the unregulated student supply obtained from Equation (2.3). This may explain why studies such as Prenovitz et al. (2015) fail to observe additional minority recruitment for competitive scholarship programs with a very limited budget.

### 2.2.2 Labor Force Quotas

Alternatively, the policy maker can restrict the recruitment decisions of firms by setting caps on the type composition of new hires. We here frame quotas as upper bounds on hiring from the dominant type, which of course is equivalent to minimum participation requirements from the minority type. Specifically, we assume that a quota  $\hat{\phi} \in [0, 1]$  specifies the a *maximal* fraction of majority workers among all new qualified hires.<sup>10</sup> As firm competition ultimately decides on market wage and the size of the labor force, the outcomes of such a quota depend on whether the market allows for wage differentials based on minority membership.

**Type-specific wage.** When wages are determined *only* through market forces, any imbalance in the supply of educated workers,  $(1 - \hat{\phi})\bar{l}_1 > \hat{\phi}\bar{l}_2$ , will affect the market wage for majority and minority workers. In such a situation, educated minority workers are in short supply and firms pay a premium in order to attract them, while the oversupply of majority workers drives down wages. In equilibrium, educational investment guarantees employment.<sup>11</sup>

<sup>9</sup>If budget balance is a concern, note that the policy maker can always recoup any imbalance through a general tax levied on all agents irrespective of type or education status.

<sup>10</sup>Only quotas with restrictions on qualification can be effective. Otherwise, firms could always costlessly meet any quota by hiring unqualified minority workers at a wage of zero.

<sup>11</sup>By contradiction, any oversupply of educated workers of type  $i$  forces  $\Delta_i = -1$  by firm competition. However, facing zero expected lifetime earnings, education is wholly unattractive to all type- $i$  individuals.

In our model, both individuals and firms are risk-neutral and have access to the same information with regards to future policy changes. Moreover, we restrict the job search of each individual to a single moment in time, thereby ruling out later employment changes or anticipatory investment on the eve of expected policy changes. As such, it is here without loss of generality to assume that any wage differentials take the form of (positive or negative) *hiring bonuses*  $\Delta \in \mathbb{R}$  that are paid upon initial appointment.

Adopting a general equilibrium approach, we write down the market-clearing equations for educated labor  $(\hat{\phi}l, (1 - \hat{\phi})l)$  under a binding quota  $\hat{\phi}$  as

$$\begin{cases} 0 = \hat{\phi}\Delta_1 + (1 - \hat{\phi})\Delta_2, \\ 1 + \Delta_1 = c - x\left(\frac{\hat{\phi}l}{\beta_1}\right) - \alpha\mu\left(\frac{L_1}{l}\right), \\ 1 + \Delta_2 = c - x\left(\frac{(1-\hat{\phi})l}{\beta_2}\right) - \alpha\mu\left(\frac{L_2}{l}\right). \end{cases} \quad (2.6)$$

The first equation is the zero-profit condition for competitive firms, stating that the marginal benefit of hiring a high-productivity worker (at the required diversity level) is equal to zero. The latter equations stem from workers' individual rationality constraints: Facing lifetime earnings of  $1 + \Delta_i$ , the marginal worker is indifferent about investing in education. Taken together, these market clearing conditions uniquely determine the instantaneous cohort  $\bar{l}(t) = \mathbf{l}(t) = (\hat{\phi}, 1 - \hat{\phi})l(t) \in \mathbb{R}_{>0}^2$  as well as type-specific hiring bonuses  $\Delta_1 < 0 < \Delta_2$ .<sup>12</sup>

Imposing such a quota is less flexible than educational incentives as it delegates the decision over labor force size to myopic firms. Nevertheless, for the specific intervention highlighted in [Theorem 7](#) (a persistent implementation of a non-steady state workforce  $\mathbf{L}^*$ ), this is without efficiency loss as firms and planner agree on the optimal labor force size  $|\mathbf{L}^*|$ .

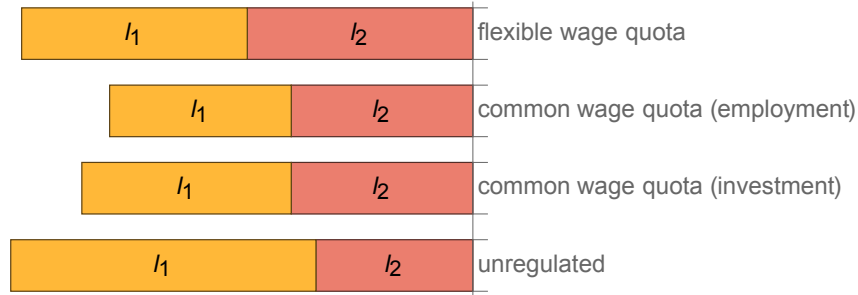
**Theorem 8.** *Consider a case where the optimal steady-state labor force  $\mathbf{L}^* \in \mathbb{R}_+^2$  is more diverse than what the market would implement, i.e.  $\frac{l_1}{l_1+l_2} > \frac{L_1^*}{L_1^*+L_2^*}$ , where  $\mathbf{l}$  is obtained from  $\mathbf{L}^*$  by solving [Equation \(2.3\)](#).*

*The policy maker can maintain this labor force through a hiring quota  $\phi^* = \frac{L_1^*}{L_1^*+L_2^*}$  as long as type-specific hiring bonuses are feasible.*

*Proof.* See [Appendix B](#). □

**Common wage.** In some industries, social or legal pressure prohibits paying unequal wage to employees in the same position. In this case, the zero-profit condition rules out hiring bonuses and forces the market wage to  $w = 1$ . Instead, any differential job market prospects stem from employment insecurity.

<sup>12</sup>The bounds on individual wage are obtained since the quota is *binding*, which rules out  $\mathbf{w} = (1, 1)$ . By the zero-profit condition, one wage is thus bigger and one smaller than one. And since  $w_1 > 1 > w_2$  would attract *even more* majority employees, expected equilibrium wages have to satisfy the bounds outlined above.



**Figure 2.5:** Labor force participation under different types of market intervention when starting with the steady-state labor force  $\mathbf{L} \approx (0.65, 0.33)$  and implementing a binding quota  $\hat{\phi} = 0.5$ . Common wages compress the total labor force more than flexible wages ( $l \approx 0.39$  versus  $l \approx 0.96$ ), and lead to an excess supply  $\bar{l}_1 - l_1 \approx 0.06$  of educated majority workers. For comparison, the last line plots the investment decisions in an unregulated market  $\mathbf{l} = \mathbf{L}$ . Parameter values are as in [Figure 2.3](#).

Indeed, a binding quota caps the demand for type-1 workers at  $\frac{\hat{\phi}}{1-\hat{\phi}}\bar{l}_2$ , while all  $\bar{l}_2$  educated minority workers are hired. Workers factor this into their cost-benefit analysis as they contemplate education. As a result, majority investment is dampened and minority investment ticks up slightly due to reduced competition for mentors. [Figure 2.5](#) illustrates how overall employment responds to labor market quotas, and how the outcome depends on the availability of signing bonuses.

**Theorem 9.** *When type-specific wages are infeasible, binding labor force quotas reduce total investment. More specifically, minority investment and recruitment increases, but less so than the reduction in majority investment and recruitment.*

*Furthermore, there is now a positive mass of majority workers who invest into education and subsequently fail to secure employment.*

*Proof.* See [Appendix B](#). □

This result highlights the negative externalities that arise in the presence of common wages: Workers of the dominant type invest into an ex-post worthless education, wasting their own resources in the process and diluting mentoring efficiency for everybody else. This greatly reduces the appeal of workplace quotas in situations where wage is sticky or subject to social scrutiny. Now, real-world firms can implement favorable work conditions for a minority without relying on wage premiums. Instead, there may be workplace or schedule accommodations that particularly appeal to the target minority and hence serve a similar purpose. The reality thus likely lies in between the two situations outlined here, making workforce quotas an effective, but less versatile and politically more challenging policy tool when compared to scholarships.

## 2.3 Conclusion

By analyzing the far-reaching impact of workplace complementarities on long-term welfare, we hope to contribute to the discussion on affirmative action by reconciling some arguments from both sides. For instance, opponents of positive discrimination often criticize the *persistency* of such policies. The very fact that lower admission standards for minority are *still* viewed as necessary after decades, the argument goes, demonstrates their ineffectiveness in bringing about lasting change.<sup>13</sup> Our analysis highlights the shortcomings of such an inference: Indeed, there is room for welfare improvement through policy design in many situations, and some of them involve *ongoing* market intervention to correct systematic undervaluation of mentoring externalities.

That being said, minority support programs *can* be completely ineffective at minority recruitment if they are chosen ‘too small’. As we highlight in [Section 2.2](#), merit-based minority scholarships must be available in big enough quantity in order to affect the extensive margin of minority education. Similarly, market interventions must alter labor force composition enough so as to affect its long-term convergence.

However, our results also warrant some caution when it comes to implementing specific policies. When mentoring is far more important than innate skill, a homogeneous workforce may actually prove more efficient. Moreover, when workplace quotas are implemented, a thorough understanding of wage determination is crucial. Wage stickiness significantly reduces potential surplus gains, and the resulting employment uncertainty harms the majority. Crowding out is far more pronounced under these circumstances, and over-investment now arises even in an unsaturated labor market.

---

<sup>13</sup>See for instance the argument made in Stanford Magazine ([Sacks and Thiel, 1996](#)).

# Bibliography

- Susan Athey, Christopher Avery, and Peter Zemsky. Mentoring and diversity. *American Economic Review*, 90(4):765–786, 2000. 36
- Ofer H. Azar. The review process in economics: Is it too fast? *Southern Economic Journal*, 72(2):pp. 482–491, 2005. 5
- Ofer H. Azar. The Slowdown in first-response times of economic journals: Can it be beneficial? *Economic Inquiry*, 45(1):179–187, 2007. 5, 20, 23
- Sascha Baghestanian and Sergey V Popov. On Publication, Refereeing, and Working Hard. Available at <http://dx.doi.org/10.2139/ssrn.2424868>, September 2015. 5
- Abhijit V Banerjee. A theory of misgovernance. *The Quarterly Journal of Economics*, 112(4):1289–1332, 1997. 23
- Eric P Bettinger and Bridget Terry Long. Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, pages 152–157, 2005. 35, 43
- John Bohannon. Who’s afraid of peer review? *Science*, 342(6154):60–65, 2013. 28
- V. Calcagno, E. Demoinet, K. Gollner, L. Guidi, D. Ruths, and C. de Mazancourt. Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, 338(6110):1065–1069, 2012. 13
- David Card and Stefano DellaVigna. Nine Facts about Top Journals in Economics. *Journal of Economic Literature*, 51(1):144–61, 2013. 29
- Scott E Carrell, Marianne E Page, and James E West. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144, 2010. 33, 36
- Hector Chade. Matching with noise and the acceptance curse. *Journal of Economic Theory*, 129(1):81–113, 2006. 4
- Hector Chade and Lones Smith. Simultaneous search. *Econometrica*, 74(5):1293–1307, 2006. 30



- Hector Chade, Gregory Lewis, and Lones Smith. Student portfolios and the college admissions problem. *The Review of Economic Studies*, 81(3):971–1002, 2014. 4, 10, 16, 17, 28, 29, 30, 58
- Yeon-Koo Che and Youngwoo Koh. Decentralized college admissions. *Journal of Political Economy*, 2015. forthcoming. 8
- Kim-Sau Chung. Role models and arguments for affirmative action. *American Economic Review*, pages 640–648, 2000. 35, 43
- Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993. 35
- Committee on the Status of Women and Ethnic Minorities (SWEM), Academic Senate, Berkeley Division. Guide to faculty advancement and promotion at u.c. berkeley. Available at [http://academic-senate.berkeley.edu/sites/default/files/committees/swem/swem\\_tenure\\_guide\\_final\\_2-2012.pdf](http://academic-senate.berkeley.edu/sites/default/files/committees/swem/swem_tenure_guide_final_2-2012.pdf), retrieved on 11/30/2016., 2012. 34
- Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009. 55
- Christopher Cotton. Submission fees and response times in academic publishing. *American Economic Review*, 103(1):501–09, 2013. 3, 5
- Christophe Crombez, Tim Groseclose, and Keith Krehbiel. Gatekeeping. *Journal of Politics*, 68(2):322–334, 2006. 28
- George F. Dreher and Taylor H. Cox Jr. Race, gender, and opportunity: A study of compensation attainment and the establishment of mentoring relationships. *Journal of Applied Psychology*, 81(3):297 – 308, 1996. 33, 35
- Glenn Ellison. The slowdown of the economics publishing process. *Journal of Political Economy*, 110(5):pp. 947–993, 2002. 20, 28
- Glenn Ellison and Ashley Swanson. The gender gap in secondary school mathematics at high achievement levels: Evidence from the american mathematics competitions. Technical report, National Bureau of Economic Research, 2009. 33
- Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. In Jess Benhabib, Matthew O. Jackson, and Alberto Bisin, editors, *Handbook of Social Economics, Vol. 1A*, pages 133–200. North-Holland, The Netherlands, 2011. 33
- D. Gale and L.S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. 4

- Martin Heintzelman and Diego Nocetti. Where should we submit our manuscript? an analysis of journal submission strategies. *The BE Journal of Economic Analysis & Policy*, 9(1), 2009. 5, 9, 20, 23
- Herminia Ibarra. Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly*, 37(3):pp. 422–447, 1992. 33, 35
- GJO Jameson. Counting zeros of generalised polynomials: Descartes’ rule of signs and laguerre’s extensions. *The Mathematical Gazette*, pages 223–234, 2006. 54
- Richard D. Kahlenberg, Halley Potter, Tanya Katerí Hernández, Haibo Huang, F. Michael Higginbotham, Jennifer Lee, and Angel L. Harris. Should affirmative action be based on income? *The New York Times*, April 27 2014. 34
- Ben Kesling and Cameron McWhirter. Percentage of African-Americans in U.S. Police Departments remains flat since 2007. *The Wall Street Journal*, May 14 2015. 43
- Prasad Krishnamurthy and Aaron Edlin. Affirmative action and stereotypes in higher education admissions. Working Paper 20629, National Bureau of Economic Research, October 2014. 36
- Liv Langfeldt. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, 31(6):820–841, 2001. 28
- Derek Leslie. Are delays in academic publishing necessary? *The American Economic Review*, 95(1):pp. 407–413, 2005. 5, 23
- Mathew D. McCubbins, Roger G. Noll, and Barry R. Weingast. Legislative intent: The use of positive political theory in statutory interpretation. *Law and Contemporary Problems*, 57(1):pp. 3–37, 1994. 22
- Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations. unpublished, April 2014. 33
- Sharon Oster. The optimal order for submitting manuscripts. *The American Economic Review*, 70(3):pp. 444–448, 1980. 5
- Sarah J. Prenovitz, Gary R. Cohen, Ronald G. Ehrenberg, and George H. Jakubson. An evaluation of the mellon mays undergraduate fellowship’s effect on phd production at non-unconf institutions. Working Paper 21451, National Bureau of Economic Research, August 2015. 44
- Peter M Rothwell and Christopher N Martyn. Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone? *Brain*, 123(9):1964–1969, 2000. 28

- David Sacks and Peter Thiel. The case against affirmative action. *Stanford Magazine*, September/October 1996. 47
- Santiago Salinas and Stephan B. Munch. Where should i send it? optimizing the submission decision process. *PLoS ONE*, 10(1):e0115451, 01 2015. 5
- A. Shleifer and R.W. Vishny. Corruption. *The Quarterly Journal of Economics*, 108(3): 599–617, 1993. 23
- Lones Smith. The marriage model with search frictions. *Journal of Political Economy*, 114 (6):pp. 1124–1144, 2006. 4
- Tayfun Sönmez. Manipulation via capacities in two-sided matching markets. *Journal of Economic Theory*, 77(1):197 – 204, 1997. 11
- Sonia Sotomayor. *Schuette v. Coalition to Defend Affirmative Action*. 572 U.S., 2014. 33
- Martin L Weitzman. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654, 1979. 5, 9
- Ivo Welch. Referee recommendations. *Review of Financial Studies*, 2014. 28

# Appendix A

## Supplementary Material to Chapter 1

### A.1 Additional Proofs

#### A.1.1 Proofs concerning Section 1.3.1

**Proof of Lemma 1.3.1.** For any  $j, k \in J$  and permutation  $\omega$  over  $J' \subseteq J \setminus \{j, k\}$ , simple algebra shows that  $U((j, k, \omega)) \geq U((k, j, \omega))$  can be rewritten as

$$\begin{aligned} 0 &\leq \delta_j \alpha_j \gamma_j + \delta_j \delta_k (1 - \alpha_j) \alpha_k \gamma_k + \delta_j \delta_k (1 - \alpha_j) (1 - \alpha_k) U(\omega) \\ &\quad - (\delta_k \alpha_k \gamma_k + \delta_j \delta_k (1 - \alpha_k) \alpha_j \gamma_j + \delta_j \delta_k (1 - \alpha_j) (1 - \alpha_k) U(\omega)) \\ \iff 0 &\leq \frac{\delta_j \alpha_j \gamma_j}{1 - \delta_j (1 - \alpha_j)} - \frac{\delta_k \alpha_k \gamma_k}{1 - \delta_k (1 - \alpha_k)} = z_j(\alpha_j) - z_k(\alpha_k). \end{aligned}$$

Let  $\omega^*$  be an optimal permutation over  $J$  and assume by contradiction that there exists some  $j, k \in J$  with  $z_j(\alpha_j) < z_k(\alpha_k)$  and  $\omega^* = (\omega_{(1)}, j, k, \omega_{(2)})$ . By the previous argument,

$$U((j, k, \omega_{(2)})) < U((k, j, \omega_{(2)})),$$

and by the definition of  $U(\cdot)$ , this contradicts optimality of  $\omega^*$ .  $\square$

**Lemma A.1.1.** *Whenever experts are best responding, proposal rates  $\pi_j(x)$  satisfy*

$$\prod_{k \in \mathcal{P}_j^{\geq}(x)} (1 - \alpha_k(x)) \leq \pi_j(x) \leq \prod_{k \in \mathcal{P}_j^{>}(x)} (1 - \alpha_k(x)), \quad (\text{A.1})$$

where  $\mathcal{P}_j^{\triangleright}(x) = \{k \neq j \mid z_k(\alpha_k(x)) \triangleright z_j(\alpha_j(x))\}$  for  $\triangleright \in \{\geq, >\}$  denotes upper contour sets according to the  $z_j$ -score.<sup>1</sup>

*Proof.* The bounds obtain for the degenerate lotteries that place weight only on optimal permutations  $\omega \in \Pi(J)$  where  $\omega^{-1}(j)$  is maximal or minimal respectively.  $\square$

<sup>1</sup>In line with convention, any product over the empty set is assumed to be one, i.e.  $\prod_{k \in \emptyset} * = 1$  for any factor  $*$ .

**Proof of Corollary 1.3.2.** Rewriting

$$z_j(\alpha_j) = \gamma_j \left( 1 - \left( 1 + \frac{1}{e^{rT_j} - 1} \alpha_j \right)^{-1} \right)$$

makes it apparent that  $\frac{\partial z_j}{\partial \alpha_j} > 0$ ,  $\frac{\partial z_j}{\partial \gamma_j} \geq 0$  and  $\frac{\partial z_j}{\partial T_j} \leq 0$  with strict inequality for all  $\alpha_j > 0$ . (And since  $z_k(\alpha_k) = 0$  if and only if  $\alpha_k = 0$ , the value  $\pi_j(x)$  is uniquely determined and independent of  $\gamma_j$  and  $T_j$  even when  $\alpha_j(x) = 0$ .) Monotonicity follows from [Lemma A.1.1](#) as a raise in  $z_j < \tilde{z}_j$  increases the upper contour sets in the sense that  $\mathcal{P}_j^> \supseteq \tilde{\mathcal{P}}_j^>$  and hence  $\pi_j \leq \prod_{k \in \mathcal{P}_j^>} (1 - \alpha_k) \leq \prod_{k \in \tilde{\mathcal{P}}_j^>} (1 - \alpha_k) \leq \tilde{\pi}_j$ . Similarly, when another gatekeeper becomes more attractive through  $\tilde{\alpha}_k \geq \alpha_k$ ,  $\tilde{\gamma}_k \geq \gamma_k$  or  $\tilde{T}_k \leq T_k$ , the strict increase in  $z_k < \tilde{z}_k$  implies  $\tilde{\mathcal{P}}_j^> \supseteq \mathcal{P}_j^>$  and hence  $\pi_j \geq \tilde{\pi}_j$ .  $\square$

### A.1.2 Proofs concerning [Section 1.3.2](#)

**Proof of Lemma 1.3.3.** For any gatekeeper strategy  $a_j : \Sigma \rightarrow [0, 1]$ , let  $\underline{\sigma}_j \in \Sigma$  denote the unique cutoff that leads to the same total match rate  $M_j$ .<sup>2</sup>

Using the simple structure of the cutoff rule, the change in acceptance rate at  $x$  is given by

$$\begin{aligned} (1 - G(\underline{\sigma}_j|x)) - \alpha_j(x) &= - \int_{s_j < \underline{\sigma}_j} a_j(s_j) g(s_j|x) ds_j + \int_{s_j \geq \underline{\sigma}_j} (1 - a_j(s_j)) g(s_j|x) ds_j \\ &= g(\underline{\sigma}_j|x) \left[ - \int_{s_j < \underline{\sigma}_j} a_j(s_j) \frac{g(s_j|x)}{g(\underline{\sigma}_j|x)} ds_j + \int_{s_j \geq \underline{\sigma}_j} (1 - a_j(s_j)) \frac{g(s_j|x)}{g(\underline{\sigma}_j|x)} ds_j \right]. \end{aligned}$$

For any  $x' > x$ , the monotone likelihood ratio property implies  $\frac{g(s_j|x)}{g(s'_j|x)} > \frac{g(s_j|x')}{g(s'_j|x')}$  for any  $s_j \leq s'_j$ . Unless  $a_j(\cdot) \equiv \mathbf{1}_{\underline{\sigma}}(\cdot)$  almost everywhere, this therefore means that

$$\begin{aligned} (1 - G(\underline{\sigma}_j|x')) - \alpha_j(x') &> g(\underline{\sigma}_j|x') \left[ - \int_{s_j < \underline{\sigma}_j} a_j(s_j) \frac{g(s_j|x)}{g(\underline{\sigma}_j|x)} ds_j + \int_{s_j \geq \underline{\sigma}_j} (1 - a_j(s_j)) \frac{g(s_j|x)}{g(\underline{\sigma}_j|x)} ds_j \right] \\ &= \frac{g(\underline{\sigma}_j|x')}{g(\underline{\sigma}_j|x)} (1 - G(\underline{\sigma}_j|x) - \alpha_j(x)). \end{aligned}$$

This implies the single crossing property  $1 - G(\underline{\sigma}_j|x) \geq \alpha_j(x) \Rightarrow 1 - G(\underline{\sigma}_j|x') > \alpha_j(x') \quad \forall x' \geq x$ . By monotonicity of  $\pi_j(\cdot)$  in  $\alpha_j(x)$ , the same holds for match rates  $\mu_j(x) = \alpha_j(x)\pi_j(x)$ .

<sup>2</sup>The match rate  $\mu_j(x)$  is strictly decreasing in  $\underline{\sigma}_j$  by virtue of the strong monotonicity of  $G(\cdot|x)$  and the weak monotonicity  $\pi_j(x)$  in  $\alpha_j$ . While this establishes uniqueness, existence is not as straightforward: Indeed, the function  $\pi_j(x|\underline{\sigma}_j)$  contains discontinuities with respect to  $\underline{\sigma}_j$ . However, monotonicity still implies the existence of a decreasing and convergent sequence of cutoffs  $\underline{\sigma}_j^n \rightarrow \underline{\sigma}_j$  such that matches are below  $M_j(\underline{\sigma}_j^n) \leq M_j(a_j) \leq M_j(\underline{\sigma}_j)$ . The subsequent argument highlights that a deviation to a cutoff rule  $\underline{\sigma}_j^n$  will eventually be profitable for large enough  $n$ . This rules out any other strategies as equilibrium candidates.

Since total matches are constant across the two, either the two policies are equal for almost all  $x$ ,<sup>3</sup> or the value to gatekeeper  $j$  strictly improves under the cutoff strategy, as low-quality matches are replaced by more worthwhile ones.  $\square$

### A.1.3 Proofs concerning Section 1.3.3

**Lemma A.1.2.** *Exponential signals  $G(\sigma|x) = 1 - e^{-\phi(x)\sigma}$  satisfy the MLRP for any decreasing parameter function  $\phi : X \rightarrow \mathbb{R}_+$ . Moreover, suppose no two gatekeepers have exactly the same impact,  $\gamma_j \neq \gamma_k$  for any  $j \neq k \in J$ . In that case, the set*

$$\bigcup_{j \neq k \in J} \{x \mid z_j(1 - G(\underline{\sigma}_j|x)) = z_k(1 - G(\underline{\sigma}_k|x))\}$$

has  $F$ -measure zero for any gatekeeper cutoffs  $\underline{\sigma} \in \Sigma^J$ .

*Proof.* MRLP follows immediately from the definition, since  $g(\underline{\sigma}|x)/g(\underline{\sigma}'|x) = e^{-\phi(x)(\sigma - \sigma')}$  is increasing in  $x$  for any  $\sigma > \sigma'$ . The exponential distribution also implies that acceptance rates relate to each other as  $\alpha_j = \alpha_k^{\sigma_j/\sigma_k}$  for all experts  $x$ . By expanding the  $z$  score, one can rewrite the indifference condition

$$z_j(\alpha_k^{\sigma_j/\sigma_k}) = z_k(\alpha_k) \quad \text{if and only if} \quad \Gamma_1 \alpha_k^{\sigma_j/\sigma_k} + \Gamma_2 \alpha_k^{1+\sigma_j/\sigma_k} + \Gamma_0 \alpha_k = 0$$

for  $\Gamma_0 = \delta_k(1 - \delta_j)\gamma_k$ ,  $\Gamma_1 = (1 - \delta_k)\gamma_j$  and  $\Gamma_2 = \delta_j\delta_k(\gamma_k - \gamma_j)$ . If  $\gamma_j \neq \gamma_k$ , all three coefficients are positive, implying that the Dirichlet polynomial has at most three real zeros (Jameson, 2006).  $\square$

**Lemma A.1.3.** *Consider two interior cutoff vectors  $\underline{\sigma}, \tilde{\underline{\sigma}} \in \text{int}(\Sigma^J)$  and let  $J^< = \{j \in J \mid \tilde{\underline{\sigma}}_j < \underline{\sigma}_j\}$ .*

*Joint total matches over any nonempty  $J^<$  strictly decrease for any feasible proposal rates  $\pi, \tilde{\pi} : X \rightarrow [0, 1]^J$ ,  $\sum_{j \in J^<} (\tilde{M}_j - M_j) < 0$ .*

*Proof.* Fix any expert type  $x \in X$  and let  $\alpha, \tilde{\alpha}$  denote the vectors of marginal acceptance probabilities for expert  $x$  under cutoffs  $\underline{\sigma}$  and  $\tilde{\underline{\sigma}}$  respectively, and let  $z, \tilde{z}$  denote the corresponding scores. Finally, consider any arbitrary pure optimal proposal order  $\omega, \tilde{\omega} \in \Pi(J)$ .

The proof proceeds in three steps. First, I maintain the proposal order  $\omega$  and update acceptance rates one-by-one, showing that the value of

$$\mu^< = \sum_{j \in J^<} \alpha_j \prod_{n=1}^{\omega^{-1}(j)-1} (1 - \alpha_{\omega(n)})$$

weakly decreases in each round, and strictly for dimensions  $j \in J^<$ . Indeed,  $\mu^<$  is weakly decreasing in the marginal acceptance rate for any  $k \notin J^<$ , and by definition of  $J^<$ ,  $\tilde{\alpha}_k \geq \alpha_k$ .

<sup>3</sup>This happens only if  $a_j(s_j) = \mathbf{1}_{\underline{\sigma}_j}(s_j)$  for almost all signals  $s_j$ .

As for any  $k \in J^<$  with  $m_k = \omega^{-1}(k)$  split  $\mu^<$  at summand  $k$  as in

$$\mu^< = \sum_{\substack{n < m_k \\ \omega(n) \in J^<}} \alpha_{\omega(n)} \prod_{m=1}^{n-1} (1 - \alpha_{\omega(m)}) + \prod_{m=1}^{m_k-1} (1 - \alpha_{\omega(m)}) \cdot \left[ \alpha_k + (1 - \alpha_k) \sum_{\substack{n > m_k \\ \omega(n) \in J^<}} \alpha_{\omega(n)} \prod_{m=m_k+1}^{n-1} (1 - \alpha_{\omega(m)}) \right].$$

The first term is independent of  $\alpha_k$  and the square brackets contain a weighted sum over two terms. In addition, note that

$$\sum_{\substack{n > m_k \\ \omega(n) \in J^<}} \alpha_{\omega(n)} \prod_{m=m_k+1}^{n-1} (1 - \alpha_{\omega(m)}) \leq \sum_{n > m_k} \alpha_{\omega(n)} \prod_{m=m_k+1}^{n-1} (1 - \alpha_{\omega(m)}) = 1 - \prod_{m=m_k+1}^{|J|} (1 - \alpha_{\omega(m)}),$$

where the equality follows by the overall success rate of independent Bernoulli trials. For any interior cutoffs, the second term of the weighted sum is thus strictly below one, and hence as  $\alpha_j$  decreases to  $\alpha_j$ , weight is shifted towards the smaller term, *strictly* decreasing the value of  $\mu^<$ .

Second, having thus updated all acceptance rates (informally expressed as  $\mu_\omega^< > \tilde{\mu}_\omega^<$ ), I now swap neighboring elements of the proposal order until the permutation agrees with  $\tilde{\omega}$ . By the Bubblesort algorithm (Cormen, 2009), such swaps need to be performed only over neighboring elements  $(j, k)$  that switch their relative order, i.e.  $\omega^{-1}(j) < \omega^{-1}(k)$  and  $\tilde{\omega}^{-1}(j) > \tilde{\omega}^{-1}(k)$ . Moreover, any such swap only affects the match rate at gatekeepers  $j$  and  $k$ , leaving all others unchanged. It follows immediately that swaps over  $j, k \notin J^<$  have no bearing on the value of  $\mu^<$ . Neither do swaps  $j, k \in J^<$  since  $\tilde{\alpha}_j + (1 - \tilde{\alpha}_j)\tilde{\alpha}_k = \tilde{\alpha}_k + (1 - \tilde{\alpha}_k)\tilde{\alpha}_j$ . Finally, since  $z_j > \tilde{z}_j$  and  $z_k \leq \tilde{z}_k$ , any swap between the two elements must be such that  $j$  moves *down* in the proposal order, i.e.  $\omega^{-1}(j) < \omega^{-1}(k)$  and  $\tilde{\omega}^{-1}(j) > \tilde{\omega}^{-1}(k)$ . The impact on  $\mu^<$  is therefore given by an additional factor  $(1 - \tilde{\alpha}_k) < 1$  in the  $j$ th term, which again weakly lowers its value.

Finally, since  $\mu_\omega^< > \tilde{\mu}_\omega^< \geq \tilde{\mu}_{\tilde{\omega}}^<$  for *any* feasible proposal orders, the inequality also holds over lotteries – and therefore  $\sum_{j \in J^<} \mu_j(x) > \sum_{j \in J^<} \tilde{\mu}_j(x)$ . By integrating over  $X$ , the desired result follows.  $\square$

**Complement to the proof of Theorem 1.** In the main text, I assume that the measure of indifferent experts is zero for all signal cutoffs  $\underline{\sigma}_j \in \Sigma$ . This simplification uniquely pins down proposal rates and as such allows for the most crisp exposition of the proof details.

To establish general existence, consider step function scores  $z_j^n : [0, 1] \rightarrow \mathbb{R}_+$  as  $z_j^n = \frac{1}{n}(\lfloor n z_j \rfloor + \frac{j}{|J|+1})$ , where  $\lfloor z \rfloor = \max \{n \in \mathcal{N} \mid n \leq z\}$ . By adding the second term, I ensure that no expert is ever indifferent across submission orders,<sup>4</sup> yet the approximation is such that  $z_j^n \rightarrow z_j$  uniformly for all  $j$ . The proof outlined in the main text implies existence of an unique fixed point  $\underline{\sigma}^n \in \Sigma^J$ . By Bolzano-Weierstrass, there exists a convergent subsequence  $\underline{\sigma}^{n_k} \rightarrow \underline{\sigma}$ . I will now construct expert proposal rates that implement cutoffs  $\underline{\sigma}$  as an equilibrium.

<sup>4</sup>Indeed,  $n z_j^n$  only takes on values within  $\mathbb{N} + \left\{ \frac{j}{|J|+1} \right\}$ , and these sets are disjoint across gatekeepers.

To do so, consider any weak order  $\preceq \subseteq J^2$  and let

$$X^{\preceq} = \{x \mid z_j(\alpha(x|\underline{\sigma}_j)) \geq z_k(\alpha(x|\underline{\sigma}_k)) \Leftrightarrow j \preceq k\} \subseteq X$$

denote the set of experts whose  $z$ -score agrees with  $\preceq$ .

Now for any  $x \in X^{\preceq}$ , let  $\omega^n(x)$  denote the unique optimal proposal order under cutoffs  $\underline{\sigma}^n$  and score  $z^n$ , and call the set of all such orders  $\Omega^n = \{\omega^n(x) \mid x \in X^{\preceq}\}$ . Total matches  $M(X^{\preceq}) \subseteq [0, 1]^J$  over  $X^{\preceq}$  are given by  $\int_{X^{\preceq}} \alpha(x|\underline{\sigma}_j) \pi_j(x|\omega) f(x) dx$ . This implies that the set of feasible total proposal rates is spanned by the convex hull of those resulting from  $\omega_n \in \Omega^n$ , and as such there exists a lottery over proposal orders  $\ell^n \in \Delta\Omega^n$  such that total matches  $M(X^{\preceq})$  are maintained even if *all* experts employ  $\ell^n$ . In doing this, I essentially ensure that all experts with the same (weak) limiting preference over gatekeepers employ the same strategy (which may however now involve mixing). While this no longer describes a best response under  $(z^n, \underline{\sigma}^n)$ , it is useful for the following two reasons:

- Since there are finitely many weak orders  $\preceq \subseteq J^2$ , these orders partition the experts into at most  $|J|^2$  sets, all of which employ the same strategy  $\ell_{\preceq}^n$  (equal to a lottery over permutations of  $J$ ). By Bolzano-Weierstrass, there exists therefore a subsequence  $n_k$  such that these expert strategies converge to some limiting  $\ell_{\preceq}$ . I will now establish that this limit describes a best response for all experts under score  $z$  and cutoffs  $\underline{\sigma}$ . Since capacities are maintained by definition of  $\ell_{\preceq}^n$ , they therefore establish equilibrium existence.
- To show that  $\ell_{\preceq}$  describes a best response under preference  $\preceq$ , I need to show that the probability of proposing to  $k$  prior to  $j$  is zero whenever  $k \prec j$ .

To do so, let  $m \in \mathcal{N}$  and consider the set of  $x \in X^{\preceq}$  such that  $g(\sigma|x) \leq m$  and

$$\Delta z_{jk}^m = z_j(\alpha(x|\underline{\sigma}_j)) - z_k(\alpha(x|\underline{\sigma}_k)) > \frac{1}{m}.$$

I refer to the set of all qualifying  $x$  as  $X_m^{\preceq} \subseteq X^{\preceq}$ . Since the derivative of the  $z_j$  score is bounded above by  $\gamma_j(e^{rT_j} - 1)^{-1}$  and that of  $\alpha(x|\cdot)$  is bounded below by  $-m$ , there exists  $n_0^m \in \mathcal{N}$  large enough (and independent of  $x$ ) such that

$$z_j(\alpha(x|\underline{\sigma}_j^n)) - z_k(\alpha(x|\underline{\sigma}_k^n)) \quad \forall x \in X_m^{\preceq} > \frac{1}{2m} \forall n \geq n_0^m.$$

Moreover, by uniform convergence of  $z^n$ , there also exists  $n_1^m > n_0^m$  large enough such that

$$z_j^n(\alpha(x|\underline{\sigma}_j^n)) > z_k^n(\alpha(x|\underline{\sigma}_k^n)) \quad \forall x \in X_m^{\preceq} > \frac{1}{2m} \forall n \geq n_1^m.$$

In other words, for  $n$  large enough, all experts in  $X_m^{\preceq}$  propose first to gatekeeper  $j$  as a best response to  $(z^n, \underline{\sigma}^n)$ . This limits the weight placed on any proposal order with  $\omega^{-1}(j) > \omega^{-1}(k)$  through  $F(X^{\preceq} \setminus X_m^{\preceq})$ . In particular, as  $m \rightarrow \infty$ , this bound goes to zero and establishes the desired result.



As for uniqueness, consider by contradiction that  $\underline{\sigma} \neq \tilde{\sigma}$  both describe equilibrium cutoffs. Up to relabeling, the set  $J^< = \{j \in J \mid \tilde{\sigma}_j < \underline{\sigma}_j\}$  is nonempty, and hence [Lemma A.1.3](#) implies that total matches for  $j \in J^<$  differ over the two cutoffs. This contradicts the equilibrium condition  $\sum_{j \in J^<} M_j = \sum_{j \in J^<} \kappa_j = \sum_{j \in J^<} \tilde{M}_j$ .

As for continuity of total matches, consider any convergent parameter sequence

$$(T^n, \gamma^n, \kappa^n) \rightarrow (T, \gamma, \kappa).$$

Let  $\underline{\sigma}^n, \underline{\sigma} \in \Sigma^J$  be the corresponding equilibrium cutoffs. By the same argument as above, there exists a convergent subsequence  $\underline{\sigma}^{n_k} \rightarrow \tilde{\sigma}$  along with proposal orders that implement  $\tilde{\sigma}$  as an equilibrium under  $(T, \gamma, \kappa)$ . By [Lemma A.1.3](#), the limit is unique  $\tilde{\sigma} = \underline{\sigma}$ . Assume by contradiction that  $\underline{\sigma}^n$  doesn't converge. By definition, this implies that there exists  $\varepsilon > 0$  and a subsequence  $n_m$  such that  $\|\underline{\sigma}^{n_m} - \underline{\sigma}\| > \varepsilon$  for all  $m$ . However, this sequence also admits a convergent subsequence  $\underline{\sigma}^{n_{m_t}} \rightarrow \tilde{\sigma}$  by Bolzano-Weierstrass, and the construction above implements  $\tilde{\sigma}$  as equilibrium cutoffs under  $(T, \gamma, \kappa)$ . Since  $\|\tilde{\sigma} - \underline{\sigma}\| \geq \varepsilon$  however, this creates a contradiction with [Lemma A.1.3](#). As a result, equilibrium cutoffs converge.  $\square$

## A.1.4 Proofs concerning [Section 1.4](#)

### Proof of [Theorem 2](#).

- (i) Fix any  $\kappa_1 \in (0, 1)$  and consider what happens as  $\kappa_2 \rightarrow 0$ . One of two things has to happen in equilibrium: The bottom gatekeeper either categorically rejects all proposals by raising his cutoff  $\underline{\sigma}_2 \rightarrow \sup \Sigma$ , or the top gatekeeper accepts all proposals  $\underline{\sigma}_1 \rightarrow \inf \Sigma$ . The limited capacity at the top rules out the latter possibility and implies instead that the top gatekeeper cutoff converges to an interior  $\tilde{\sigma}_1 < \sup \Sigma = \lim_{\kappa_2 \rightarrow 0} \underline{\sigma}_2$ .<sup>5</sup> By continuity of the equilibrium cutoffs ([Theorem 1](#)), there exists  $\bar{k}_2 > 0$  such that  $\kappa_2 < \bar{k}_2$  implies  $\underline{\sigma}_2 > \underline{\sigma}_1$ .

Conversely, as  $\kappa_2 \rightarrow 1 - \kappa_1$ , almost all proposals get accepted somewhere. In other words, one of the gatekeepers has to lower her standards completely by setting  $\underline{\sigma}_j \rightarrow \inf \Sigma$ . By the same argument as above,  $\underline{\sigma}_1$  is bounded below and hence  $(\underline{\sigma}_1, \underline{\sigma}_2) \rightarrow (\tilde{\sigma}'_1, \min \Sigma)$  with  $\tilde{\sigma}'_1 > \inf \Sigma = \lim_{\kappa_2 \rightarrow 1 - \kappa_1} \underline{\sigma}_2$ . By continuity, there exists  $\underline{k}_2 < 1 - \kappa_1$  such that  $\kappa_2 \geq \underline{k}_2$  implies  $\underline{\sigma}_2 \leq \underline{\sigma}_1$ .

While these bounds depend on all parameters of the game, there exist more robust bounds for  $\kappa_1$  to ensure monotonicity. To determine their value, let  $\tilde{M}_j$  denote total matches for  $j$  when all experts approach the *bottom* gatekeeper first. In this situation, matches for the bottom gatekeeper are independent of the top cutoff  $\sigma_1$ , allowing me to write

$$\tilde{M}_1(\sigma_1, \sigma_2) = \int_X \alpha(x|\sigma_1)(1 - \alpha(x|\sigma_2))f(x)dx \quad \text{and} \quad \tilde{M}_2(\sigma_2) \equiv \int_X \alpha(x|\sigma_2)f(x)dx.$$

<sup>5</sup>Since  $z_1(\alpha_1)$  is increasing in  $\alpha_1$  with limit  $z_1(1) > z_2(1)$ , there exists  $a < 1$  large enough such that any expert with  $\alpha_1(x) \geq a$  proposes first to 1. This implies that  $\underline{\sigma}_1$  is bounded from below.

By the full support assumption,  $\widetilde{M}_2$  is strictly decreasing in  $\sigma_2$ , and since the signal density varies continuously with  $x$ , the inverse  $s = M_2^{-1}(\kappa_2)$  exists. Let  $\bar{k}_1 = \widetilde{M}_1(s, s) \in (0, 1 - \kappa_2)$  denote the bottom enrollment under identical cutoffs and exogenously fixed expert strategies. Whenever  $\kappa_1 \leq \bar{k}_1$ , it follows that

$$\kappa_1 \leq \bar{k}_1 = \widetilde{M}_1(s, s) \leq M_1(s, s), \quad \kappa_2 = \widetilde{M}_2(s) \geq M_2(s, s)$$

and

$$\kappa_1 + \kappa_2 = M_1(s, s) + M_2(s, s),$$

since total matches are independent of the proposal order. By the monotonicity of total matches, equilibrium cutoffs are monotone  $\underline{\sigma}_1 \geq s \geq \underline{\sigma}_2$ .<sup>6</sup>

- (ii) Consider first the case  $\delta_2 > \frac{\gamma_1}{\delta_1\gamma_1 + (1-\delta_1)\gamma_2}\delta_1$  or equivalently  $\delta_1\gamma_1(1-\delta_2) < \delta_2\gamma_2(1-\delta_1)$ . Together with  $\delta_1\gamma_1 > \delta_2\gamma_2$ , this condition implies  $\delta_1 < \delta_2$  and  $\gamma_1 > \gamma_2$ . Moreover, this implies that  $\lim_{\alpha \rightarrow 0^+} \frac{z_1(\alpha)}{z_2(\alpha)} = \frac{\delta_1\gamma_1(1-\delta_2)}{\delta_2\gamma_2(1-\delta_1)} < 1$ . In other words, there exists  $\alpha_0 > 0$  small enough such that  $z_1(\alpha_0) < z_2(\alpha_0)$ , meaning that the indifference curve  $\phi$  starts out *below* the diagonal. As for the *acceptance* curve, note that any exponential signal  $G(\sigma | x) = 1 - e^{-\psi(x)\sigma}$  with decreasing  $\psi : X \rightarrow [0, \infty)$  satisfies the monotone likelihood ratio. Moreover, the acceptance function parametrized by  $(\alpha(x|\sigma_1), \alpha(x|\sigma_2))$  simplifies to  $\alpha_1^{\sigma_2/\sigma_1}$ , which starts out horizontally whenever  $\sigma_2 > \sigma_1$ . By choosing two cutoffs close enough together such that  $\underline{\sigma}_1 < \underline{\sigma}_2 < \frac{\log \phi(\alpha_0)}{\log \alpha_0} \underline{\sigma}_1$  ensures that  $\psi(\alpha_0) > \phi(\alpha_0)$ , implying that the acceptance function crosses the indifference function in the *upwards* direction somewhere below  $\alpha_0$ . In other words, there exist  $0 < x < x' = \alpha^{-1}(\alpha_0 | \underline{\sigma}_1)$  with  $z_1(x) > z_2(x)$  and  $z_1(x') < z_2(x')$ . By setting  $\kappa_j = M_j(\underline{\sigma}_1, \underline{\sigma}_2)$ , this describes an equilibrium.

Conversely, assume that  $\gamma_2 < \gamma_1$ . The Implicit Function Theorem bounds the slope of the indifference curve as  $\phi'(\alpha) = \frac{\phi(\alpha)\gamma_2(\gamma_1 - z_1(\alpha))}{\alpha\gamma_1(\gamma_2 - z_1(\alpha))} > \frac{\phi(\alpha)}{\alpha}$ .<sup>7</sup> Whenever  $\phi(\alpha) \geq \alpha$ , this implies for the derivatives of the secants that

$$\left[ \frac{\phi(\alpha)}{\alpha} \right]' = \frac{\alpha\phi'(\alpha) - \phi(\alpha)}{\alpha^2} \geq 0$$

and

$$\left[ \frac{1 - \phi(\alpha)}{1 - \alpha} \right]' = \frac{1 - \phi(\alpha) - (1 - \alpha)\phi'(\alpha)}{(1 - \alpha)^2} < \frac{\alpha - \phi(\alpha)}{\alpha(1 - \alpha)^2} \leq 0.$$

In other words, the secants move in opposite direction whenever the indifference curve lies weakly above the diagonal. Since [Chade et al. \(2014, Theorem 1\)](#) establishes that the secants of the *acceptance* curve are both monotone, the two can cross at most

<sup>6</sup>Indeed,  $|M_j(\underline{\sigma}_1, \underline{\sigma}_2) - \kappa_j| \geq |M_j(s, s) - \kappa_j|$  for all  $\underline{\sigma}_1 \leq s \leq \underline{\sigma}_2$  with equality only if  $\underline{\sigma}_j = s$ . Moreover, total matches are strictly increasing in  $(\underline{\sigma}_1, \underline{\sigma}_2)$ , ruling out all other cutoff combinations.

<sup>7</sup>Simple algebra shows that  $z_j'(\alpha) = \frac{z_j(\alpha)}{\alpha} (1 - \frac{z_j(\alpha)}{\gamma_j})$  and hence  $\phi'(\alpha) = \frac{z_1'(\alpha)}{z_2'(\alpha)\phi(\alpha)}$  can be rewritten as above.

once over that region. No other crossings are possible whenever  $\delta_2 < \frac{\gamma_1}{\delta_1\gamma_1 + (1-\delta_1)\gamma_2}\delta_1$  or  $\underline{\sigma}_2 < \underline{\sigma}_1$ , since the two conditions make sure that either the indifference or the acceptance curve lie entirely above the diagonal, respectively. And since  $z_1(1) > z_2(1)$ , those experts with *higher* quality propose to the top first.  $\square$

Moreover, the final match allocation may exhibit nonmonotonicities *even under monotone strategies*. Intuitively, experts are *adversely selected* after the first round of proposals, and this effect may dominate the informativeness of an individual signal when gatekeepers have similar impact.

**Lemma A.1.4.** *Consider a setup with two gatekeepers, such that  $e^{-rT_1}\gamma_1 > e^{-rT_2}\gamma_2$ . If  $z_1(1) - z_2(1)$  is small enough, there exists a MLRP signal function and monotone cutoffs  $\underline{\sigma}_2 < \underline{\sigma}_1$  under which equilibrium match rates at the top,  $\mu_1(\cdot)$ , decrease over an interval.*

*Proof.* Consider again exponential signals with acceptance curve  $\psi(\alpha) = \alpha^\beta$  for  $\beta = \underline{\sigma}_2/\underline{\sigma}_1 \in (0, 1)$ . Whenever experts first approach the bottom gatekeeper, match rates at the top are given by  $\mu_1(x) = (1 - \alpha_2(x))\alpha_1(x)$ , which is decreasing whenever  $\alpha_1 > \frac{1}{2}$ .<sup>8</sup> Fixing some  $\tilde{\alpha}_1 \in (\frac{1}{2}, 1)$ , note that  $\frac{z_1(\tilde{\alpha}_1)}{z_2(1)} = \frac{\delta_1\gamma_1}{\delta_2\gamma_2} \frac{\tilde{\alpha}_1}{(1-\delta_1)+\delta_1\tilde{\alpha}_1}$  which eventually drops below one as  $\delta_2\gamma_2 \rightarrow \delta_1\gamma_1$ . As a result, one can find  $\tilde{\alpha}_2 \in (\tilde{\alpha}_1, 1)$  and  $\beta = \frac{\ln \tilde{\alpha}_2}{\ln \tilde{\alpha}_1}$  such that the acceptance curve lies above the indifference curve over an interval within  $(1/2, \tilde{\alpha}_0)$ . To implement this acceptance curve, consider for instance the exponential signal distribution  $G(\sigma | x) = 1 - F(x)^\sigma$ , choose any  $x_0 \in \text{int}(X)$  and let  $\underline{\sigma}_j = \frac{\ln(1-\tilde{\alpha}_j)}{\ln F(x_0)}$ . In this monotone-strategy equilibrium, match rate  $\mu_j(x)$  decreases around  $x_0$ .  $\square$

### A.1.5 Proofs concerning Section 1.5.1

From now on, I label gatekeepers by decreasing (maximal) appeal, i.e. such that  $e^{-rT_1}\gamma_1 \geq e^{-rT_2}\gamma_2$ .

**Lemma A.1.5.** *Let  $X \subseteq \mathbb{R}$  describe a compact space. Suppose  $g_i : X \rightarrow \mathbb{R}_{>0}$  for  $i \in \{1, 2\}$  satisfies the monotone likelihood ratio property, meaning that  $\frac{g_1(x)}{g_2(x)}$  is weakly increasing. Let  $\phi : X \rightarrow \mathbb{R}$  be a continuous function with finite integral over  $X$  that satisfies the single crossing property*

$$\phi(x) \geq 0 \implies \phi(x') > 0 \quad \forall x' \geq x.$$

*In this case,*

$$0 \succeq \int_X g_1(x)\phi(x)dx \implies 0 \succeq \int_X g_2(x)\phi(x)dx$$

*for  $\succeq \in \{\geq, >\}$ . Moreover, if  $\frac{g_1(x)}{g_2(x)}$  is strictly increasing over a set of nonzero measure, then the latter inequality is strict.*

---

<sup>8</sup>Indeed,  $(1 - \alpha_1^\beta)\alpha_1$  is decreasing whenever  $1 < \alpha_1^\beta(\beta + 1)$ . Since  $\alpha_1 > \frac{1}{2}$ , the claim is equivalent to  $2^\beta < \beta + 1$ , which binds at  $\beta \in \{0, 1\}$  and holds for intermediate  $\beta$  by convexity of  $2^\beta$ .

*Proof.* By the Intermediate Value Theorem, the image of  $\phi$  is an interval. I consider first the case where  $0 \in \text{Im}(\phi)$ , and fix any  $x_0 \in \phi^{-1}(0)$ . By the full support assumption of the signal function, the inequality  $0 \succeq \int_X g_1(x)\phi(x)dx$  is preserved when both sides are multiplied by  $g_2(x_0)/g_1(x_0) > 0$ . The result now follows from the monotone likelihood ratio property,

$$\begin{aligned} 0 \succeq \frac{g_2(x_0)}{g_1(x_0)} \int_X g_1(x)\phi(x)dx &= g_2(x_0) \left( \int_{\leq x_0} \underbrace{\frac{g_1(x)}{g_1(x_0)}}_{\leq \frac{g_2(x)}{g_2(x_0)}} \underbrace{\phi(x)}_{\leq 0} dx + \int_{\geq x_0} \underbrace{\frac{g_1(x)}{g_1(x_0)}}_{\geq \frac{g_2(x)}{g_2(x_0)}} \underbrace{\phi(x)}_{\geq 0} dx \right) \\ &\geq \int_X g_2(x)\phi(x)dx. \end{aligned}$$

Moreover, when the likelihood is strictly monotone over a nonzero set, the second inequality is strict.

Finally, if  $\phi(\cdot)$  is always positive or negative, the same equation holds in the limit as  $x_0 \rightarrow 0$  or  $x_0 \rightarrow 1$  respectively.  $\square$

**Lemma A.1.6.** *Consider an equilibrium strategy profile with monotone expert behavior, and let*

$$\hat{x} = \inf \{x \mid z_1 \circ \alpha(x \mid \underline{\sigma}_1) > z_2 \circ \alpha(x \mid \underline{\sigma}_2)\} \in X \cup \{\infty\}.$$

*The threshold  $\hat{x}$  is defined so that exactly those experts above  $\hat{x}$  first propose to the top gatekeeper.*

*If  $\hat{x} > \inf X$ , a marginal increase in turnaround time  $T_j$  affects gatekeeper  $j$ 's matches. By accompanying such additional delay with a capacity-preserving change in the signal cutoff  $\underline{\sigma}_j$ , gatekeeper  $j$  improves her payoff if and only if*

$$\int_X (x - \hat{x})g(\underline{\sigma}_j \mid x)\pi_j(x)f(x)dx > 0. \quad (\text{A.2})$$

*If the sign is in the opposite direction,  $j$ 's payoff is negatively affected by such a compensated marginal change.*

*Finally, if  $\hat{x} = \inf X$ , the top gatekeeper can always improve her payoff by adding delay.*

*Proof.* A marginal raise in  $T_j$  affects gatekeeper  $j$ 's matches only indirectly through the change in expert behavior. Experts react to a change in  $T_j$  if and only if  $\partial\hat{x}/\partial T_j \neq 0$ , which is the case for interior  $\hat{x} > \inf X$  since  $\partial z_j(\alpha_j)/\partial T_j < 0$  whenever  $\alpha_j > 0$ .<sup>9</sup> As a result, her total matches and payoff decrease by

$$\frac{\partial B_j}{\partial T_j} = -\hat{\alpha}_1 \hat{\alpha}_2 \hat{f} \left| \frac{\partial \hat{x}}{\partial T_j} \right| < 0 \quad \text{and} \quad \frac{\partial z_j}{\partial T_j} = -\hat{x} \hat{\alpha}_1 \hat{\alpha}_2 \hat{f} \left| \frac{\partial \hat{x}}{\partial T_j} \right| < 0,$$

<sup>9</sup>A change in  $T_j$  affects the indifference condition for all but the worst expert,  $z_1 \circ \alpha_1(\hat{x}) = z_2 \circ \alpha_2(\hat{x})$  and hence  $\partial\hat{x}/\partial T_1 > 0$  and  $\partial\hat{x}/\partial T_2 < 0$ .

where  $\hat{\alpha}_* = \alpha_*(\hat{x})$ . The absolute values capture the fact that  $\hat{x}$  moves in opposite direction when the top or the bottom gatekeeper add red tape, but in both instances it lowers the deviator's own matches. A marginal raise in the cutoff  $\underline{\sigma}_j$  on the other hand lowers acceptance rates everywhere, and hence

$$\left( \frac{\partial B_j / \partial \underline{\sigma}_j}{\partial z_j / \partial \underline{\sigma}_j} \right) = - \int_X \binom{1}{x} g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx - \binom{1}{\hat{x}} \hat{\alpha}_1 \hat{\alpha}_2 \hat{f} \left| \frac{\partial \hat{x}}{\partial \underline{\sigma}_j} \right|.$$

By mass-conservation, the two effects cancel out with respect to total matches,

$$dT_j \frac{\partial B_j}{\partial T_j} + d\underline{\sigma}_j \frac{\partial B_j}{\partial \underline{\sigma}_j} = - \left( dT_j \left| \frac{\partial \hat{x}}{\partial T_j} \right| + d\underline{\sigma}_j \left| \frac{\partial \hat{x}}{\partial \underline{\sigma}_j} \right| \right) \hat{\alpha}_1 \hat{\alpha}_2 \hat{f} - d\underline{\sigma}_j \int_X g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx = 0. \quad (\text{A.3})$$

The combined impact on gatekeeper  $j$ 's payoff can thus be written as

$$\begin{aligned} dT_j \frac{\partial z_j}{\partial T_j} + d\underline{\sigma}_j \frac{\partial z_j}{\partial \underline{\sigma}_j} &= - \left( dT_j \left| \frac{\partial \hat{x}}{\partial T_j} \right| + d\underline{\sigma}_j \left| \frac{\partial \hat{x}}{\partial \underline{\sigma}_j} \right| \right) \hat{x} \hat{\alpha}_1 \hat{\alpha}_2 \hat{f} - d\underline{\sigma}_j \int_X x g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx \\ &\stackrel{(\text{A.3})}{=} -d\underline{\sigma}_j \int_X g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx \left( \frac{\int_X x g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx}{\int_X g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx} - \hat{x} \right). \end{aligned} \quad (\text{A.4})$$

For artificial delay that is accompanied by a nontrivial cutoff adjustment  $dT_j > 0 > d\underline{\sigma}_j$ , the above has the same sign as  $\int_X (x - \hat{x}) g(\underline{\sigma}_j | x) \pi_j(x) f(x) dx$ .

Finally, consider the case where  $\hat{x} = \inf X$ , in which case [Equation \(A.2\)](#) trivially holds. As the top adds more and more delay  $T_1 \rightarrow \infty$  and compensates this with capacity-preserving signal cutoffs,  $\hat{x}$  eventually becomes interior and [\(A.2\)](#) continues to hold by continuity, at least initially. This however implies that adding (sufficient) red tape is a profitable deviation for the top gatekeeper.  $\square$

**Lemma A.1.7.** *Consider a capacity-filling, monotone strategy profile  $\underline{\sigma}_2 \leq \underline{\sigma}_1$  such that the threshold type  $\hat{x}$  defined above is interior to  $X$  and almost no experts are indifferent. If the bottom gatekeeper has a (weak) incentive to add marginal red tape, the top gatekeeper strictly prefers to add additional delay.*

*Proof.* By the proof of [Theorem 2\(ii\)](#), the monotone behavior of gatekeepers inspires experts to do the same and rules out expert indifference almost everywhere. When the threshold type  $\hat{x}$  is interior, it follows that  $z_1 \circ \alpha(x | \underline{\sigma}_1) < z_2 \circ \alpha(x | \underline{\sigma}_2)$  for  $x \in (\inf X, \hat{x})$ . In other words,  $\frac{\partial}{\partial x} z_1 \circ \alpha(x | \underline{\sigma}_1)|_{x=0} < \frac{\partial}{\partial x} z_2 \circ \alpha(x | \underline{\sigma}_2)|_{x=0}$  and additional delay affects matches. Suppose the bottom gatekeeper has a weak incentive to add red tape, implying by [Lemma A.1.6](#) that

$$0 \leq \int_X (x - \hat{x}) g(\underline{\sigma}_2 | x) \pi_2(x) f(x) dx,$$

Since  $\pi_2(x) < 1$  if and only if  $x > \hat{x}$ , removing  $\pi_2(x)$  from the right side increases the value of the integrand exactly when it is positive, hence  $0 \leq \int_X (x - \hat{x}) g(\underline{\sigma}_2 | x) f(x) dx$ . As detailed

in [Lemma A.1.5](#), the monotone likelihood ratio property further implies that this sign is preserved under a cutoff increase,  $0 \leq \int_X (x - \hat{x})g(\underline{\sigma}_1 | x)f(x)dx$ . Lastly,  $\pi_1(x) < 1$  if and only if  $x < \hat{x}$ , so introducing  $\pi_1(x)$  reduces *negative* values of the integrand. This strictly<sup>10</sup> raises the value of the integral, implying

$$0 < \int_X (x - \hat{x})g(\underline{\sigma}_1 | x)h(x)f(x)dx.$$

By [Lemma A.1.6](#), the top gatekeeper has incentives to add spurious delay.  $\square$

**Proof of Theorem 3.** Assume that  $\kappa_1 \leq \underline{k}_1(\kappa_2)$  according to the bound established in [Theorem 2](#), ensuring monotone equilibrium cutoffs no matter the turnaround times  $(T_1, T_2)$ . The proof proceeds in three steps: First, I reduce the candidate equilibria by showing that it cannot entail artificial delay by the bottom gatekeeper. Among the remaining strategy profiles, I then show that there exists a unique equilibrium. Finally, I show that red tape is increasing in the impact differential.

Part I: No red tape at the bottom. By [Lemma A.1.6](#), any monotone equilibrium candidate has an interior threshold type  $\hat{x}$ . Moreover, whenever  $T_2 > \tau_2$ , the absence of profitable deviations for the bottom competitor implies that  $0 = \int_X (x - \hat{x})g(\underline{\sigma}_2 | x)\pi_2(x)f(x)dx$ . By [Lemma A.1.7](#), this however implies that the top gatekeeper has strict incentives to delay turnaround even further, contradicting the equilibrium assumption.

Part II: An unique equilibrium exists. Having thus simplified the number of equilibrium parameters to  $(T_1, \underline{\sigma}_1, \underline{\sigma}_2)$ , consider the continuous function  $\Phi : [\tau_1, \infty) \rightarrow \mathbb{R}$  with

$$T_1 \mapsto \Phi(T_1) = \int_X (x - \hat{x})g(\underline{\sigma}_1 | x)\pi_1(x)f(x)dx,$$

where  $\hat{x}$  and the cutoffs  $(\underline{\sigma}_1, \underline{\sigma}_2)$  are given by the unique equilibrium in [Theorem 1](#), vary continuously in  $T_1$  and exhibit monotone strategies by [Theorem 2](#). The comparative statics in [Lemma 1.3.5](#) imply that as the turnaround time increases to  $\tilde{T}_1 > T_1$  and assuming that  $\hat{x}$  is interior, the top gatekeeper's cutoff goes down  $\tilde{\sigma}_1 < \underline{\sigma}_1$ . Since total promotions remain constant, it follows immediately that  $\underline{\sigma}_2$  must go up simultaneously  $\tilde{\sigma}_2 > \underline{\sigma}_2$ . Similarly, as the top gatekeeper becomes more accepting and the bottom more selective, there is only one way the top may still respect capacity: Equilibrium must be such that fewer experts propose to her first, that is  $\tilde{x} > \hat{x}$ .

Taken together, I now show that these observations imply the single crossing property

$$\Phi(T_1) \leq 0 \quad \implies \quad \Phi(T'_1) < 0 \quad \forall T'_1 > T_1.$$

<sup>10</sup>Indeed, let  $\inf X < x_0 < x_1 < \hat{x}$  and let  $g_{\min} = \min_{x \in [x_0, x_1]} g(\underline{\sigma}_1 | x)$  and  $\varepsilon = (x_1 - \hat{x})g_{\min}\alpha_2(x_0) < 0$ . The value of the integrand is lowered by at least  $\varepsilon$  over the set of positive measure  $(x_0, x_1)$ , hence its total value strictly decreases.

Indeed, whenever  $\hat{x} = \inf X$ , the integral is always positive and hence  $\Phi(T_1) > 0$ . As such, the premise first implies an interior threshold type  $\hat{x}$ . Moreover, replacing  $\pi_1(x) = 1 - \mathbf{1}\{x \leq \hat{x}\} \alpha(x|\underline{\sigma}_2)$  with  $\tilde{\pi}_1(x) = 1 - \mathbf{1}\{x \leq \tilde{x}\} \alpha(x|\underline{\sigma}_2)$  adds more weight to negative integrands  $x < \hat{x}$  and removes weight from positive values  $x \in (\hat{x}, \tilde{x})$ , thereby weakly reducing the value of the integral. Replacing  $(x - \hat{x})$  by the smaller  $(x - \tilde{x})$  strictly reduces values for all  $x$ , hence

$$\int_X (x - \tilde{x}) g(\underline{\sigma}_1|x) \tilde{\pi}_1(x) f(x) dx < 0.$$

This strict inequality is maintained under the lower cutoff  $\tilde{\sigma}_1 < \underline{\sigma}_1$  by [Lemma A.1.5](#).

Lastly, since  $z_1(1) \rightarrow 0$  as  $T_1 \rightarrow \infty$ , every expert eventually proposes to the bottom gatekeeper first.<sup>11</sup> However, as  $\hat{x} \rightarrow \sup X$ , all integrands in  $\Phi$  become negative and hence  $\lim_{T_1 \rightarrow \infty} \Phi(T_1) < 0$ .

Taken together, it follows that either  $\Phi(T_1) < 0$  for all  $T_1 \geq \tau_1$ , in which case red tape is ruled out and equilibrium turnaround times are  $T_1 = \tau_1$ , or there exists a unique  $T_1 \geq \tau_1$  at which [Equation \(A.2\)](#) binds and no further deviations are profitable. The unique equilibrium is given by this turnaround time, along with  $T_2 = \tau_2$  and the cutoffs  $(\underline{\sigma}_1, \underline{\sigma}_2)$  from [Theorem 1](#).<sup>12</sup>

Part III: Monotonicity in  $\gamma_1$ . Suppose optimal turnaround at impact level  $\gamma_1$  is equal to  $T_1 > \tau_1$ , and let  $\hat{\alpha}_1 = \alpha(\hat{x} | \underline{\sigma}_1)$  describe the equilibrium acceptance rate of the indifferent expert. In this case, the optimal amount of red tape at any  $\tilde{\gamma}_1 > \gamma_1$  can be given in closed form as

$$\tilde{T}_1 = T_1 + \frac{1}{r} \ln \left( \frac{\tilde{\gamma}_1}{\gamma_1} + (1 - \hat{\alpha}_1) \left(1 - \frac{\tilde{\gamma}_1}{\gamma_1}\right) e^{-rT_1} \right), \quad (\text{A.5})$$

which is increasing in  $\tilde{\gamma}_1$ . This condition is equivalent to

$$z_1(\hat{\alpha}_1 | T_1, \gamma_1) = \frac{e^{-r\tilde{T}_1} \tilde{\gamma}_1 \hat{\alpha}_1}{1 - e^{-r\tilde{T}_1} (1 - \hat{\alpha}_1)} = \frac{e^{-rT_1} \gamma_1 \hat{\alpha}_1}{1 - e^{-rT_1} (1 - \hat{\alpha}_1)} = z_1(\hat{\alpha}_1 | \tilde{T}_1, \tilde{\gamma}_1),$$

implying that  $\hat{x}$  is preserved as the indifferent expert. Moreover, [Equation \(A.5\)](#) is strictly increasing in  $\hat{\alpha}_1$ , implying that *more* red tape is necessary to maintain indifference for better types. In other words,

$$z_1(\alpha_1 | \tilde{T}_1, \tilde{\gamma}_1) \geq z_1(\alpha_1 | T_1, \gamma_1) \quad \iff \quad \alpha_1 \geq \hat{\alpha}_1.$$

However then,  $((\tilde{T}_1, \underline{\sigma}_1), (\tau_2, \underline{\sigma}_2))$  describes an equilibrium under  $\tilde{\gamma}_1$  since proposal and acceptance rates are unchanged at all quality levels  $\hat{x}$ ,<sup>13</sup> and the delay is optimal since [Equation \(A.2\)](#) is unchanged and thus still binding.  $\square$

<sup>11</sup>Since strategies are monotone,  $z_2(1) = e^{-rT_2} \gamma_2 > e^{-rT_1} \gamma_1$  is enough to ensure that *all* experts first propose to the bottom gatekeeper.

<sup>12</sup>Capacity clearing follows from [Theorem 1](#) and the contrapositive of [Lemma A.1.7](#) rules out profitable deviations for the bottom gatekeeper.

<sup>13</sup>Indeed, the same proposal order remains optimal for experts since  $z_1(\alpha_1(x) | \tilde{T}_1, \tilde{\gamma}_1) \geq z_1(\alpha_1(x) | T_1, \gamma_1) \geq z_2(\alpha_2(x))$  for all  $x \geq \hat{x}$  and  $z_1(\alpha_1(x) | \tilde{T}_1, \tilde{\gamma}_1) \leq z_1(\alpha_1(x) | T_1, \gamma_1) \leq z_2(\alpha_2(x))$  for all  $x \leq \hat{x}$ .

**Lemma A.1.8.** *Property (sMLRP) is satisfied for exponential signals  $G(\sigma|x) = 1 - e^{-\phi(x)\sigma}$  defined by any decreasing function  $\phi : [0, 1] \rightarrow \mathbb{R}_+$ .*

*Proof.* For  $\underline{\sigma}_2 < \underline{\sigma}_1$ , one can write

$$\frac{g(\underline{\sigma}_1|x)G(\underline{\sigma}_2|x)}{g(\underline{\sigma}_2|x)G(\underline{\sigma}_1|x)} = \frac{\phi(x)e^{-\phi(x)\underline{\sigma}_1}(1 - e^{-\phi(x)\underline{\sigma}_2})}{\phi(x)e^{-\phi(x)\underline{\sigma}_2}(1 - e^{-\phi(x)\underline{\sigma}_1})} = \frac{e^{\phi(x)\underline{\sigma}_2} - 1}{e^{\phi(x)\underline{\sigma}_1} - 1} = 1 + \frac{1 - e^{\phi(x)(\underline{\sigma}_1 - \underline{\sigma}_2)}}{e^{\phi(x)(\underline{\sigma}_1 - \underline{\sigma}_2)} - e^{-\phi(x)\underline{\sigma}_2}},$$

which is strictly increasing in  $x$ .  $\square$

**Proof of Theorem 4.** As before, assume that  $\kappa_1 \leq k_1(\kappa_2)$  according to the bound established in Theorem 2. The proof establishes three claims: The top gatekeeper's *equilibrium* cutoff is strictly increasing in  $T_1$  for  $\tau_1$  small enough, the total quality of matches is monotonically increasing in red tape, and both of these properties together imply that red tape is welfare enhancing.

Part I: Red tape improves the top gatekeeper's equilibrium payoff. Since overall acceptance is independent of submission order, there exists a continuous and strictly decreasing function  $s_2(\underline{\sigma}_1) : \Sigma \rightarrow \Sigma$  such that all equilibrium cutoffs satisfy  $\underline{\sigma}_2 = s_2(\underline{\sigma}_1)$ .<sup>14</sup> By virtue of its monotonicity, this function is almost everywhere differentiable.

Taking into account the equilibrium adjustment of the signal cutoff  $\underline{\sigma}_2$ , additional red tape  $T_1 > \tau_1$  enhances the top gatekeeper's *equilibrium* payoff if and only if

$$\int_X (x - \hat{x}) [g(x|\underline{\sigma}_1)\pi_1(x) - s'_2(\underline{\sigma}_1)\alpha(x|\underline{\sigma}_1)g(x|\underline{\sigma}_2)\mathbf{1}\{x \leq \hat{x}\}] f(x) dx > 0. \quad (\text{A.6})$$

The proof is virtually identical to that of Lemma A.1.6, with the only difference being that  $\underline{\sigma}_2$  is no longer independent of  $\underline{\sigma}_1$ . Since the additional term in the integral is nonpositive everywhere, it follows that Equation (A.6) is a stronger condition than Equation (A.2). Nevertheless, both of them are satisfied for  $\hat{x} > \inf X$  small enough, which occurs in equilibrium when the turnaround  $T_1 = \tau_1$  is short enough.<sup>15</sup>

Part II: Total match quality is increasing in red tape. As  $T_1$  increases, the comparative statics in Lemma 1.3.5 imply that  $\underline{\sigma}_1$  strictly decreases ( $\partial \underline{\sigma}_1 < 0$ ) and  $\underline{\sigma}_2$  strictly increases ( $\partial \underline{\sigma}_2 > 0$ ), and by the capacity assumption and Theorem 2, the bottom gatekeeper always

<sup>14</sup>Formally,  $s_2$  is the unique solution to  $\int_X (1 - \alpha(x|\underline{\sigma}_1))(1 - \alpha(x|s_2))f(x)dx = 1 - \kappa_1 - \kappa_2$ .

<sup>15</sup>Indeed, since  $\kappa_1 = B_1(\underline{\sigma}_1, \sigma_2) \leq \int_X \alpha(x|\underline{\sigma}_1)f(x)dx$ , the top gatekeeper's equilibrium cutoff is bounded above by some  $s < \sup \Sigma$ . For any  $x_0 > \inf X$ , that implies that his equilibrium acceptance rate  $\alpha_1(x_0)$  is weakly larger than  $\alpha(x_0|s) > 0$  and hence

$$z_1(\alpha_1(x_0)|T_1) \geq \frac{e^{-rT_1}\gamma_1\alpha(x_0|s)}{1 - e^{-rT_1}(1 - \alpha(x_0|s))} \rightarrow \gamma_1 \quad \text{as } T_1 \rightarrow 0.$$

In other words, there exists  $T_1 > 0$  small enough such that  $z_1(\alpha_1(x_0)|T_1) > e^{-rT_2}\gamma_2 \geq z_2(\alpha_2(x_0)|T_2)$  for all  $T'_1 \leq T_1$  no matter the exact equilibrium acceptance odds  $\alpha_1(x_0)$  and  $\alpha_2(x_0)$ . By the definition of the threshold type, this implies  $\hat{x} < x_0$ .



sets a lower cutoff than the top  $\underline{\sigma}_2 < \underline{\sigma}_1$ . In other words, as  $T_1$  goes up, the two cutoffs move closer together.

An expert's total odds of being matched is given by  $\mu(x) = 1 - G(\underline{\sigma}_1|x)G(\underline{\sigma}_2|x)$ . The raise in  $T_1$  affects these odds by

$$d\mu(x) = -d\underline{\sigma}_1 g(\underline{\sigma}_1|x)G(\underline{\sigma}_2|x) - d\underline{\sigma}_2 G(\underline{\sigma}_1|x)g(\underline{\sigma}_2|x).$$

Since capacities always bind in equilibrium, overall matches are not affected by the additional red tape, i.e.  $d\bar{B} = \int_X d\mu(x)f(x)dx = 0$ . This being a weighted sum over terms  $d\mu(x)$  which vary continuously in  $x$ , it follows that at least one of the terms  $d\mu(x_0)$  must be equal to zero. For any higher quality  $x > x_0$ , [Equation \(sMLRP\)](#) implies that

$$\overbrace{-\frac{d\underline{\sigma}_1}{d\underline{\sigma}_2} \frac{g(\underline{\sigma}_1|x)G(\underline{\sigma}_2|x)}{g(\underline{\sigma}_2|x)G(\underline{\sigma}_1|x)}}^{>0} > -\frac{d\underline{\sigma}_1}{d\underline{\sigma}_2} \frac{g(\underline{\sigma}_1|x_0)G(\underline{\sigma}_2|x_0)}{g(\underline{\sigma}_2|x_0)G(\underline{\sigma}_1|x_0)} = \overbrace{\frac{d\mu(x_0)}{d\underline{\sigma}_2 g(\underline{\sigma}_2|x_0)G(\underline{\sigma}_1|x_0)}}^{=0} + 1 = 1$$

and hence, by multiplying both sides with  $d\underline{\sigma}_2 g(\underline{\sigma}_2)G(\underline{\sigma}_1) > 0$ , it follows that  $d\mu(x) > 0$ . By the same argument, it also holds that  $d\mu(x) < 0$  for all  $x < x_0$ .

Together, these two observations imply that  $(x - x_0)d\mu(x) > 0$  for all  $x \in X$ , and hence the total value of all matches increases,

$$d\bar{V} = d\bar{V} - x_0 d\bar{B} = \int_X \underbrace{(x - x_0)d\mu(x)}_{>0} f(x)dx > 0.$$

Part III: Welfare increases in red tape. One can rewrite the welfare measure in [Equation \(1.3\)](#) as

$$W = \gamma_2 \bar{V} + (\gamma_1 - \gamma_2)z_1,$$

where  $z_1$  is the top gatekeeper's equilibrium payoff. By the above arguments, both the total quality of all matches  $\bar{V}$  and the total equilibrium quality of the top gatekeepers matches  $z_1$  are increasing in red tape for low levels of  $\tau_1$ .  $\square$

### A.1.6 Additional proofs for [Section 1.5.2](#)

**Proof of [Theorem 5.A](#).** Consider an alternative strategy  $\underline{\sigma}_1^A < \underline{\sigma}_1 < \underline{\sigma}_1^B$ , along with the corresponding response times  $T_1^A < T_1^B$ . I construct a marginal deviation  $d\underline{\sigma}_1^A$ ,  $dT_1^A$ ,  $d\underline{\sigma}_1^B$  and  $dT_1^B$  that maintains total matches but strictly increases the payoff for the top gatekeeper. Since the baseline equilibrium exhibits red tape, [Equation \(1.2\)](#) implies that  $\int_X (x - \hat{x})g(\underline{\sigma}_1|x)\pi_1(x|\underline{\sigma}_1)f(x)dx = 0$ . The marginal deviation will remove  $dM$  matches in group  $A$  of mean quality below  $\hat{x}$  and add  $dM$  matches in group  $B$  of mean quality above  $\hat{x}$ .

The former claim is easier since  $\underline{\sigma}_1^B > \underline{\sigma}_1 > \underline{\sigma}_2$ , and hence experts propose monotonically. Indeed, by [Lemma A.1.5](#), it follows that  $\int_X (x - \hat{x})g(\underline{\sigma}_1^B|x)\pi_1(x|\underline{\sigma}_1, T_1)f(x)dx > 0$ . Any change

in expert proposal orders may only *increase* match rates *below*  $\hat{x}$  and *decrease* those *above*. This is because the current notation assumes first proposals from experts  $x \geq \hat{x}$  exactly.

$$\int_X (x - \hat{x})g(\underline{\sigma}_1^B|x)\pi_1(x|\underline{\sigma}_1^B, T_1^B)f(x)dx > 0. \quad (\text{A.7})$$

Moreover, since delay is chosen optimally,  $\int_X (x - \hat{x}^B)g(\underline{\sigma}_1^B|x)\pi_1(x|\underline{\sigma}_1^B, T_1^B)f(x)dx \leq 0$  by Equation (1.2). Obviously, this is smaller than Equation (A.7) only if  $\hat{x}^B > \hat{x}$ . Moreover, for any  $d\underline{\sigma}_1^B < 0$ , there exists  $dT_1^B > 0$  that maintains indifference for expert  $\hat{x}^B$  even as his acceptance rate at the top rises (see Equation (A.5)). Choosing  $d\underline{\sigma}_1^B, dT_1^B$  such that total matches increase by  $dM$ , the match rate between the top gatekeeper and  $x$  changes by exactly  $g(\underline{\sigma}_1^B|x)\pi_1(x|\underline{\sigma}_1^B, T_1^B)d\underline{\sigma}_1^B$ . The mean quality of these additional matches is strictly higher than  $\hat{x}$  by Equation (A.7).

As for group  $A$ , the same reasoning implies that

$$\int_X (x - \hat{x})g(\underline{\sigma}_1^A|x)\pi_1(x|\underline{\sigma}_1^A, T_1^A)f(x)dx < 0. \quad (\text{A.8})$$

However, when  $\underline{\sigma}_1^A < \underline{\sigma}_2$ , experts may propose nonmonotonically and as such  $\hat{x}^A$  may not be well defined. Yet, the parametric form of the  $z$ -score implies that for exponential signals, there exist two indifferent experts  $0 \leq \hat{x}_L^A < \hat{x}_H^A$  such that exactly those experts with  $\hat{x}_L^A < x < \hat{x}_H^A$  propose to the bottom gatekeeper first. To construct the marginal change, I consider two different cases:

- (i) If  $\hat{x}_H^A < \hat{x}$ , consider a simple adjustment  $d\underline{\sigma}_1^A > 0$  chosen such that total matches decrease by  $dM$  (i.e. without a change in response times  $T_1^A$ ). Match rates decrease by  $g(\underline{\sigma}_1^A | x)\pi_1(x|\underline{\sigma}_1^A, T_1^A)d\underline{\sigma}_1^A$  at all  $x \notin \{\hat{x}_L^A, \hat{x}_H^A\}$ , and discretely for the indifferent experts with quality lower than  $\hat{x}$ . By Equation (A.8), the mean quality for the lost matches is therefore strictly below  $\hat{x}$ .
- (ii) If  $\hat{x} \leq \hat{x}_H^A$ , red tape is larger than in the original equilibrium,  $T_1^A > T_1$ . Indeed, any weakly shorter response time and a lower cutoff  $\underline{\sigma}_1^A < \underline{\sigma}_1$  raise  $z_1$ -scores throughout, and hence any expert above  $\hat{x}$  would still propose to the top gatekeeper first. As a result, it is possible to consider a marginal change  $d\underline{\sigma}_1^A > 0$  with a *decrease* in response time  $dT_1^A < 0$  that leaves  $\hat{x}_H^A$  indifferent, and such that total matches decrease by  $dM$ .

Furthermore, there exists a feasible response time  $\tilde{T}_1^A > T_1$  that maintains indifference for  $\hat{x}$  under cutoffs  $\underline{\sigma}_1^A$ . For  $\hat{x} < \hat{x}_H^A$ , there exist some experts above  $\hat{x}$  that now prefer the bottom gatekeeper, implying that the actual delays are even larger,  $T_1^A > \tilde{T}_1^A$ . This additional delay drops  $z_1$  scores everywhere, and as a result, some experts below  $\hat{x}$  propose first to the bottom gatekeeper, implying  $\hat{x}_L^A < \hat{x}$ . Since low quality experts

have a lower value for time, the reduced acceptance chances weigh more heavily for  $\hat{x}_L^A$  than the delay reduction, and he starts proposing to the bottom gatekeeper first.<sup>16</sup>

Taken together, it follows that match rates decrease by  $g(\underline{\sigma}_1^A | x)\pi_1(x|\underline{\sigma}_1^A, T_1^A)d\underline{\sigma}_1^A$  at all  $x \neq \hat{x}_L^A$ , and discretely for the indifferent expert of quality  $\hat{x}_L^A < \hat{x}$ . By Equation (A.8), the mean quality for the lost matches is therefore strictly below  $\hat{x}$ .

Having thus designed a marginal deviation that replaces some matches of mean quality lower than  $\hat{x}$  with some of mean quality above  $\hat{x}$ , it follows that  $\underline{\sigma}_1^A < \underline{\sigma}_1 < \underline{\sigma}_1^B$  cannot represent a best response.

Finally, if  $\underline{\sigma}_1 \leq \underline{\sigma}_1^A < \underline{\sigma}_1^B$ , the top gatekeeper necessarily has slack capacity. Indeed, the analysis for group  $B$  above implies that any cutoff  $\underline{\sigma}_1' > \underline{\sigma}_1$  along with the optimal delay  $T_1'$  yields *fewer* proposals since  $\hat{x}' > \hat{x}$ . As a result, the top gatekeeper has fewer total matches in either group. Conversely, for  $\underline{\sigma}_1' < \underline{\sigma}_1$  with  $\hat{x}'_H \leq \hat{x}$ , the analysis for group  $A$  above implies that the top gatekeeper receives *more* proposals and hence *more* total matches. Thus, the only situation where she might receive *fewer* total matches under a *lower* cutoff is when  $\hat{x}'_H \geq \hat{x}$ . In this case, there exists  $\tilde{\sigma}_1 > \underline{\sigma}_1$  along with its optimal response time  $\tilde{T}_1 \in [T_1, T_1']$  such that the matches lost by going from  $(\underline{\sigma}_1', T_1')$  to  $(\underline{\sigma}_1, \tilde{T}_1)$  have a mean quality below  $\hat{x}$ . By further lowering response times slightly to some  $\bar{T}_1 \in [T_1, \tilde{T}_1]$ , the gatekeeper can recover the same amount of matches with higher qualities  $x \in [\hat{x}, \hat{x}'_H]$ . In other words, she has a profitable deviation since she either exceeds capacity or can do achieve higher-quality matches by choosing  $(\underline{\sigma}_1, \bar{T}_1)$ .  $\square$

**Proof of Theorem 5.B.** Focus first on the case with unobservable expert types, fixing the equilibrium behavior of gatekeeper one. Let function  $\nu_2^n : [0, 1 - \kappa_1] \rightarrow \mathbb{R}_+$  describe the payoff  $\nu_2^n(k_2)$  for gatekeeper two under capacity  $k_2$  and response time  $T_2 = \tau_2$ . I will first show that this function is locally convex for  $n$  large enough, and then show how this translates into a profitable deviation that involves specialization.

First, the capacity constraint uniquely pins down the optimal cutoff  $s_2^n$  for the bottom gatekeeper in a neighborhood of  $\underline{\sigma}_2$  as  $M_2^n(\underline{\sigma}_1, s_2^n) = k_2$ . Hence, the implicit function theorem pins down the derivative of  $\nu_2^n$  in a neighborhood of  $\kappa_2$  as

$$\nu_2^{n'}(k_2) = \frac{\partial V_2^n}{\partial s_2^n} \cdot \frac{\partial s_2^n}{\partial k_2} = \frac{\partial V_2^n}{\partial s_2^n} \bigg/ \frac{\partial M_2^n}{\partial s_2^n}.$$

<sup>16</sup>Formally, it can be shown that maintained indifference for  $\hat{x}_H^A$  implies

$$\frac{e^{r(T_1^A + dT_1^A)} - 1}{e^{rT_1^A} - 1} = \frac{\alpha(\hat{x}_H^A | \underline{\sigma}_1^A + d\underline{\sigma}_1^A)}{\alpha(\hat{x}_H^A | \underline{\sigma}_1^A)}.$$

For exponential signals, the latter is equal to  $e^{-d\underline{\sigma}_1^A \phi(\hat{x}_H^A)} > e^{-d\underline{\sigma}_1^A \phi(\hat{x}_L^A)}$ , and hence  $z_1$  decreases at  $\hat{x}_L^A$  under the marginal deviation.

For the second derivative, evaluated at  $\underline{s}_2^n = \underline{\sigma}_2$ , it follows that

$$\nu_2^{n''}(\kappa_2) = \frac{\partial \nu_2^{n'}}{\partial \underline{s}_2^n} \cdot \frac{\partial \underline{s}_2^n}{\partial k_2} \Big|_{\underline{s}_2^n = \underline{\sigma}_2} = \left( \frac{\partial M_2^n}{\partial \underline{\sigma}_2} \frac{\partial^2 V_2^n}{(\partial \underline{\sigma}_2)^2} - \frac{\partial V_2^n}{\partial \underline{\sigma}_2} \frac{\partial^2 M_2^n}{(\partial \underline{\sigma}_2)^2} \right) \Big/ \left( \frac{\partial M_2^n}{\partial \underline{\sigma}_2} \right)^3. \quad (\text{A.9})$$

Since total matches are decreasing in a gatekeeper's own cutoff by [Corollary 1.3.2](#), the function  $\eta_2$  is locally convex if and only if the numerator is negative.

Considering therefore the different components of the numerator, note first that gatekeepers' equilibrium strategies are equal across quality distributions  $F^n$ , and hence the same follows for experts' best response, implying that the marginal expert is equal to  $\hat{x}$  for all  $n$ . Moreover, the existence of red tape implies that  $\hat{x}$  is interior to  $X$  and (at least locally) decreasing in the cutoff  $\underline{\sigma}_2$ . As a result, total matches and the payoff of gatekeeper two locally satisfy

$$\begin{pmatrix} M_2^n(\underline{\sigma}_1, \underline{\sigma}_2) \\ V_2^n(\underline{\sigma}_1, \underline{\sigma}_2) \end{pmatrix} = \int_0^1 \binom{1}{x} \alpha(x|\underline{\sigma}_2) f^n(x) dx - \int_{\hat{x}(\underline{\sigma}_2)}^1 \binom{1}{x} \alpha(x|\underline{\sigma}_2) \alpha(x|\underline{\sigma}_1) f^n(x) dx.$$

Differentiating with respect to  $\underline{\sigma}_2$  and rearranging the individual terms yields a closed form expression for the denominator in [Equation \(A.9\)](#), with limiting behavior

$$\begin{aligned} & \frac{\partial M_2^n}{\partial \underline{\sigma}_2} \frac{\partial^2 V_2^n}{(\partial \underline{\sigma}_2)^2} - \frac{\partial V_2^n}{\partial \underline{\sigma}_2} \frac{\partial^2 M_2^n}{(\partial \underline{\sigma}_2)^2} \\ &= \underbrace{\int g(\underline{\sigma}_2|x) \pi_2(x) f^n(x) dx \int x g'(\underline{\sigma}_2|x) \pi_2(x) f^n(x) dx}_{(\text{direct})} \\ & \quad - \underbrace{\int x g(\underline{\sigma}_2|x) \pi_2(x) f^n(x) dx \int g'(\underline{\sigma}_2|x) \pi_2(x) f^n(x) dx}_{(\text{direct})} \\ & \quad + \underbrace{\alpha_2(\hat{x})^2 \alpha_1(\hat{x})^2 f^n(\hat{x})^2 \left( \frac{\partial \hat{x}}{\partial \underline{\sigma}_2} \right)^3}_{(\text{indirect})} \\ & \quad + \underbrace{\mathcal{O}(f^n(\hat{x})) + \int (x - \hat{x}) g(\underline{\sigma}_2|x) \pi_2(x) f^n(x) dx \cdot \alpha_2(\hat{x}) \alpha_1(\hat{x}) f^{n'}(\hat{x}) \left( \frac{\partial \hat{x}}{\partial \underline{\sigma}_2} \right)^2}_{(\text{mixed})} \end{aligned}$$

The first term measures the *direct* effect, which is conditional on expert proposal order. While it can be shown that this term is always positive, it can also be bounded above by replacing means with extremal values,

$$(\text{direct}) \leq \max_{x, \tilde{x} \in X} g(\underline{\sigma}_2|x) \tilde{x} g'(\underline{\sigma}_2|\tilde{x}) - \min_{x, \tilde{x} \in X} x g(\underline{\sigma}_2|x) g'(\underline{\sigma}_2|\tilde{x}).$$

Since  $g$  is differentiable, this yields an upper bound that is independent of  $n$ .

The second term measures the *indirect* effect, which measures only the impact through changes in expert proposal orders. Since fewer experts first propose to the bottom gatekeeper as she raises her cutoff,  $\partial \hat{x} / \partial \underline{\sigma}_2 < 0$  and hence the entire term is negative and to the order of  $f^n(\hat{x})^2$ .

The final term measures the *mixed* effect, which essentially captures the fact that the two previous effects need not be balanced each. In other words, while total matches remain constant and hence total gains equal total losses, it is possible that some direct gains in the target group are compensated with indirect losses in the nontarget group, or vice versa. Due to the possibility of red tape, the starred term is negative by [Equation \(1.2\)](#), and it is bounded below by  $\min_{x \in X} (x - \hat{x})g(\underline{\sigma}_2|x)\pi_2(x)$ . Eventually, this term therefore either becomes negative (if  $f^n(\hat{x}) > 0$ ) or is dominated by the indirect term.

Either way, the function  $\nu_2^n$  is locally convex for  $n$  large enough. Moreover, a specialization strategy shifts some capacity  $dk$  from the nontarget group to the target group. Using a Taylor series approximation, this increases the total match value in the target group by

$$\nu_2^n(\kappa_2 + dk) - \nu_2^n(\kappa_2) = \nu_2^{n'}(\kappa_2)dk + \nu_2^{n''}(\kappa_2)dk^2 + o(dk^2)$$

and lowering that of the nontarget group by

$$\nu_2^n(\kappa_2 - dk) - \nu_2^n(\kappa_2) = -\nu_2^{n'}(\kappa_2)dk + \nu_2^{n''}(\kappa_2)dk^2 + o(dk^2).$$

Since  $\nu_2^n$  is locally convex, the deviation is profitable for  $dk$  small enough, as

$$\lim_{dk \rightarrow 0} \frac{1}{dk^2} \left( \frac{1}{2}\nu_2^n(\kappa_2 + dk) + \frac{1}{2}\nu_2^n(\kappa_2 - dk) - \nu_2^n(\kappa_2) \right) = \nu_2^{n''}(\kappa_2) + \lim_{dk \rightarrow 0} \frac{o(dk^2)}{dk^2} > 0. \quad \square$$

## A.2 Comparative Statics

In this section, I consider how changes in exogenous parameters affect equilibrium strategies and payoffs. The proofs borrow from properties of the least fixed point and are relegated to the appendix.

First, I show that an increase in capacity affects all experts positively since all gatekeepers become weakly more accessible. The payoff impact for gatekeepers however is ambiguous.

**Lemma A.2.1.** *As capacities increase  $\kappa_j^0 \leq \kappa_j^1$  for all  $j \in J$ , gatekeepers weakly lower their acceptance cutoffs in equilibrium. Equilibrium payoffs increase for all experts.*

*Proof.* The least (and by [Theorem 1](#) unique) fixed point of the best response function  $\Phi : \Sigma^J \rightarrow \Sigma^J$  is defined as

$$\underline{\sigma} = \inf \{ \underline{s} \in \Sigma^J \mid \Phi_j(\underline{s}) \leq \underline{s}_j \ \forall j \in J \}.$$

Since the match rate  $\mu_j(x)$  is decreasing in  $\underline{\sigma}_j$ , the condition  $\Phi_j(\underline{s}) \leq \underline{s}_j$  holds if and only if capacities are slack, i.e.  $M_j(\underline{s}) \leq \kappa_j$ . Moreover, capacities are met under equilibrium

$\underline{\sigma}_j^0$ , hence these cutoffs maintain slack capacities under  $\kappa_1$ , i.e.  $M_j(\underline{\sigma}_j^0) = \kappa_j^0 \leq \kappa_j^1$ . By the definition of the least fixed point, cutoffs are thus lower under  $\kappa^1$ , and since lower cutoffs increase acceptance rates for all experts, their equilibrium payoff increases.  $\square$

Changes in the pool of match participants can be reformulated to fit the framework above: First, the arrival of a new gatekeeper  $j$  is equivalent to a raise in her capacity from zero to  $\kappa_j$  and as such increases experts' payoff. Second, any additional (time-invariant) inflow of experts forces *all* gatekeepers to weakly raise their equilibrium cutoffs by the same logic as above.

Comparative statics for other parameters are more subtle, since they directly influence the *relative* attractiveness of a gatekeeper through her  $z_j$  score. As intuition would have it, when a gatekeeper becomes more attractive, she will also have to be more selective in the new equilibrium.

**Lemma A.2.2.** *Suppose gatekeeper  $j$  becomes more attractive through either decreased response time  $\tilde{T}_j < T_j$  or increased impact  $\tilde{\gamma}_j > \gamma_j$ . In equilibrium, this leads to a raise in her cutoff  $\underline{\sigma}_j \leq \tilde{\sigma}_j$ .*

*Proof.* By the same arguments as in [Theorem 1](#), the best response function remains monotone when component  $j$  is exogenously fixed at level  $\underline{\sigma}_j$ . I refer to this ‘all-but- $j$ ’ best-response function as  $\Phi_{-j} : \Sigma^{J \setminus \{j\}} \rightarrow \Sigma^{J \setminus \{j\}}$ . As above, its least fixed point satisfies

$$\underline{\sigma}'_{-j} = \inf \left\{ s_{-j} \mid \widetilde{M}_{-j}(\underline{\sigma}_j, s_{-j}) \leq \kappa_{-j} \right\}.$$

The change in  $j$ 's external parameters raises score  $z_j(x)$  everywhere, and as a result weakly decreases  $\pi_{-j}(x)$  by [Corollary 1.3.2](#). With fewer proposals at all quality levels, it follows that capacities are now slack,  $M_{-j}(\underline{\sigma}) \leq \kappa_{-j}$ , and hence  $\underline{\sigma}'_{-j} \leq \underline{\sigma}_{-j}$ . This in return implies that there are more overall matches under  $(\underline{\sigma}_j, \underline{\sigma}'_{-j})$  than under  $\underline{\sigma}_j$ , and hence gatekeeper  $j$  weakly exceeds capacity. For the greatest (and unique) fixed point of the entire game,  $\tilde{\sigma} = \sup \left\{ s \mid \widetilde{M}_j(s) \geq \kappa_j \forall j \right\}$ , this thus implies that  $\tilde{\sigma} \geq (\underline{\sigma}_j, \underline{\sigma}'_{-j})$ , implying the stated result.  $\square$

The conclusion of this second statement is weaker since it only speaks to gatekeeper  $j$ 's own equilibrium adjustment. The impact on other gatekeepers is ambiguous, as shows the following example.

**Example 7.** Consider a situation with three gatekeepers

$$(\gamma_1, T_1, \kappa_1) = (1.2, 2, 0.2), \quad (\gamma_2, T_2, \kappa_2) = (1, 2, 0.3), \quad (\gamma_3, T_3, \kappa_3) = (1.1, 2, 0.05)$$

and discount rate  $r = 10\%$ . Experts arrive according to the uniform distribution  $F(x) \equiv 1$ . Signals are distributed exponentially,  $G(\sigma|x) = 1 - x^\sigma$ . In this situation, a raise in the impact of gatekeeper 1 from  $\gamma_1 = 1.2$  to  $\tilde{\gamma}_1 = 5$  causes both gatekeeper 1 *and* 3 to raise their equilibrium cutoff from  $(\underline{\sigma}_1, \underline{\sigma}_2, \underline{\sigma}_3) = (3.19, 1.12, 3.34)$  to  $(\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\sigma}_3) = (3.98, 1.05, 3.62)$ .

Intuitively, *all* experts here propose to gatekeeper 1 *before* gatekeeper 3, even under the lower impact  $\gamma_1$ . Due of this ‘dominance’ (which is illustrative, but not necessary) the surge in impact  $\gamma_1$  merely steers some proposals away from gatekeeper 2 towards gatekeeper 1, but leaves the match distribution of gatekeeper 3 unchanged – at least absent cutoff adjustments. To respect capacity, gatekeeper 1 thus has to become more *selective* and gatekeeper 2 more *accepting*, and the question boils down to whether the increase in  $\underline{\sigma}_1$  or the decrease in  $\underline{\sigma}_2$  has a larger impact on matches made by gatekeeper 3. It turns out that in this instance, the first effect dominates and consequently forces gatekeeper 3 to *raise* her equilibrium cutoff.  $\diamond$

Fortunately, this ambiguity disappears for the case of only two gatekeepers, since at least one opponent has to lower her threshold to maintain total capacity. Still, the joint cutoff adjustment of gatekeepers has ambiguous payoff effects for both sides of the market. While the formal statement of [Lemma 1.3.5](#) is in the main text, the proof follows below.

**Proof of [Lemma 1.3.5](#).** Since overall matches remain equal to total capacity, the two cutoffs need to move in opposite directions, with  $\underline{\sigma}_j$  weakly increasing by [Lemma A.2.2](#). Moreover, absent any cutoff adjustments, the parameter change strictly increases  $z_j(x)$  everywhere. In response, a positive measure of experts near  $\hat{x}$  switch proposal order, thereby upsetting gatekeepers’ individual capacity constraints. By contradiction, it follows that equilibrium adjustments are strict.  $\square$

# Appendix B

## Supplementary Material to Chapter 2

**Proof of Theorem 6.** The first part of the theorem is immediate: When costs are so low that the most able individuals invest even without *any* mentorship, their labor supply never dries out. Under [Hypothesis H2](#) however, no workers get educated unless mentor availability exceeds some positive threshold  $\underline{\Lambda}$ . This ensures that the opposing type has a steady-state mentorship boost of  $\mu(1)$ , which determines labor supply through [Equation \(2.3\)](#). Such a homogeneous steady state is stable since small enough perturbations maintain minority mentor availability below  $\underline{\Lambda}$ .

Let us now turn to mixed steady states  $\mathbf{l} = \mathbf{L} = (\hat{\phi}\hat{L}, (1 - \hat{\phi})\hat{L}) \in \mathbb{R}_{>0}^2$ , where the individual cost-benefit analyses in [Equation \(2.3\)](#) simplify to [Equation \(2.4\)](#). It is easily verified that [Hypothesis H3](#) is *necessary* for a steady supply of minority workers, for otherwise either at least one of the left-side expressions in [Equation \(2.4\)](#) exceeds 1, or total labor participation  $L$  is zero.

As for sufficiency, let  $\underline{\phi} = \inf \{\phi \geq 0 \mid c - \lambda - \alpha\mu(\phi) < 1\}$  be the minimal mentor availability required for the most able individual (if any). Under [Hypothesis H3](#),  $\underline{\phi} \in [0, 0.5)$ . Over the nonempty interval  $[\underline{\phi}, 1 - \underline{\phi}] \cap (0, 1)$ , it is possible to solve each of the conditions in [Equation \(2.4\)](#) individually for  $\hat{L}$ , yielding candidate steady-state labor sizes

$$L^{(1)}(\phi) = \frac{\beta_1}{\phi} x^{-1} (c - 1 - \alpha\mu(\phi)) \quad \text{and} \quad L^{(2)}(\phi) = \frac{\beta_2}{1 - \phi} x^{-1} (c - 1 - \alpha\mu(1 - \phi)). \quad (\text{B.1})$$

A share  $\hat{\phi} \in \Phi$  corresponds to a true steady state if and only if these two quantities are equal,  $L^{(1)}(\hat{\phi}) = L^{(2)}(\hat{\phi})$ . If [Hypothesis H2](#) holds,  $L^{(1)}(\underline{\phi}) = 0 < L^{(2)}(\underline{\phi})$  and  $L^{(1)}(1 - \underline{\phi}) > 0 = L^{(2)}(1 - \underline{\phi})$ , implying a crossing by continuity of all involved functions. Similarly, if [Hypothesis H2](#) is false, then  $\underline{\phi} = 0$  and a crossing exists because  $\lim_{\phi \rightarrow 0} L^{(1)}(\phi) = \infty > \lim_{\phi \rightarrow 0} L^{(2)}(\phi)$  and  $\lim_{\phi \rightarrow 1} L^{(1)}(\phi) < \infty = \lim_{\phi \rightarrow 1} L^{(2)}(\phi)$ .

Finally, we establish the existence of a stable mixed steady state for  $q$  large enough. To this end, we first pin down the approximate location of a mixed steady state, and in a second step show that it must be stable. Formally, let  $\mathbf{L}^0 \in \mathbb{R}_{>0}^2$  be such that  $L_i^0 = \beta_i x^{-1} (c - 1 - \alpha)$ , which is well defined and finite by [Hypothesis H1<sup>+</sup>](#). We refer to the labor composition as



$\phi^0 = \frac{L_1^0}{L_1^0 + L_2^0} \in (0, 1)$ . Fix an arbitrarily small  $0 < \delta < \min\{\phi^0, 1 - \phi^0\}$ , and note that

$$\delta' = \min \left\{ \frac{L_1^0}{\phi^0 - \delta} - \frac{L_1^0}{\phi^0}, \frac{L_2^0}{1 - \phi^0 - \delta} - \frac{L_2^0}{1 - \phi^0} \right\} > 0. \quad (\text{B.2})$$

Furthermore, since  $\mu(\Lambda) \uparrow 1$  pointwise as  $q \rightarrow 1$  for any  $\Lambda > 0$ , there exists  $Q$  large enough such that

$$0 < \frac{L_1^0}{\phi^0 - \delta} - L^{(1)}(\phi^0 - \delta) < \delta' \quad \text{and} \quad 0 < \frac{L_2^0}{1 - \phi^0 - \delta} - L^{(2)}(\phi^0 + \delta) < \delta' \quad \forall q \geq Q. \quad (\text{B.3})$$

As a result, it follows that for those large values of  $q$ ,

$$\begin{aligned} L^{(1)}(\phi^0 - \delta) &= \underbrace{L^{(1)}(\phi^0 - \delta) - \frac{L_1^0}{\phi^0 - \delta}}_{> -\delta'} + \underbrace{\frac{L_1^0}{\phi^0 - \delta} - \frac{L_1^0}{\phi^0}}_{\geq \delta'} + \frac{L_1^0}{\phi^0} > \frac{L_1^0}{\phi^0} = L_1^0 + L_2^0 = \frac{L_2^0}{1 - \phi^0} \\ &= \underbrace{\frac{L_2^0}{1 - \phi^0} - \frac{L_2^0}{1 - (\phi^0 - \delta)}}_{> 0} + \underbrace{\frac{L_2^0}{1 - (\phi^0 - \delta)} - L^{(2)}(\phi^0 - \delta) + L^{(2)}(\phi^0 - \delta)}_{> 0} > L^{(2)}(\phi^0 - \delta) \end{aligned}$$

and

$$\begin{aligned} L^{(1)}(\phi^0 + \delta) &= \underbrace{L^{(1)}(\phi^0 + \delta) - \frac{L_1^0}{\phi^0 + \delta}}_{< 0} + \underbrace{\frac{L_1^0}{\phi^0 + \delta} - \frac{L_1^0}{\phi^0}}_{< 0} + \frac{L_1^0}{\phi^0} < \frac{L_1^0}{\phi^0} = L_1^0 + L_2^0 = \frac{L_2^0}{1 - \phi^0} \\ &= \underbrace{\frac{L_2^0}{1 - \phi^0} - \frac{L_2^0}{1 - (\phi^0 + \delta)}}_{\leq -\delta'} + \underbrace{\frac{L_2^0}{1 - (\phi^0 + \delta)} - L^{(2)}(\phi^0 + \delta) + L^{(2)}(\phi^0 + \delta)}_{< \delta'} < L^{(2)}(\phi^0 + \delta). \end{aligned}$$

By the same argument as above, this implies a crossing  $L^{(1)}(\hat{\phi}) = L^{(2)}(\hat{\phi}) = \hat{L}$ , and therefore the existence of a steady state  $(\hat{\phi}\hat{L}, (1 - \hat{\phi})\hat{L})$  with composition  $\hat{\phi} \in (\phi^0 - \delta, \phi^0 + \delta)$ . By the linearization theorem, this steady state is stable if and only if  $\frac{\partial \dot{\mathbf{L}}}{\partial \mathbf{L}}$  is negative definite. Since  $\dot{\mathbf{L}}(t) = \mathbf{l}(t) - \mathbf{L}(t)$ , this derivative is equal to  $\frac{\partial \mathbf{l}}{\partial \mathbf{L}} - I$ , where  $\frac{\partial \mathbf{l}}{\partial \mathbf{L}}$  can be obtained by the Implicit Function Theorem from Equation (2.3) and  $I$  is the  $2 \times 2$  identity matrix. After simplification, the characteristic polynomial  $\left| \frac{\partial \dot{\mathbf{L}}}{\partial \mathbf{L}} - \gamma I \right|$  at the steady state  $\hat{\mathbf{L}} = (\hat{L}_1, \hat{L}_2) = (\hat{\phi}\hat{L}, (1 - \hat{\phi})\hat{L})$  is proportional to

$$F(\gamma) = \hat{L}^2(1 + \gamma)^2 + \hat{L}(1 + \gamma) \left[ (\hat{\phi}K_2 + (1 - \hat{\phi})K_1) - \gamma(\hat{\phi}K_1 + (1 - \hat{\phi})K_2) \right] - \gamma K_1 K_2$$

where  $K_i = \beta_i \frac{\alpha \mu'(\hat{L}_i/\hat{L})}{x'(\hat{L}_i/\beta_i)} < 0$ . It is a simple exercise in algebra to rewrite this second order polynomial as  $F(\gamma) = a\gamma^2 + b\gamma + c$ , where

$$\begin{aligned} a &= \hat{L}^2 - \hat{L}(\hat{\phi}K_1 + (1 - \hat{\phi})K_2) > 0 \\ b &= 2\hat{L}^2 + \hat{L}(1 - 2\hat{\phi})(K_1 - K_2) - K_1 K_2 \\ c &= \hat{L}^2 + \hat{L}(\hat{\phi}K_2 + (1 - \hat{\phi})K_1). \end{aligned}$$

It can easily be seen that this is an upward sloping quadratic function since  $a > 0$ . The real part of its roots are negative if and only if  $F(0) = c > 0$  and  $F'(0) = b > 0$ , and we will conclude the proof by showing that these hold for  $q$  large enough. Indeed, due to the functional form of  $x$  and  $\mu$ , we can rewrite  $K_1$  as  $-\beta_1 \frac{q}{\lambda} \frac{\alpha(1-\mu(\hat{\phi}))}{x(\hat{\phi}\tilde{L}/\beta_1)}$ , which vanishes as  $q \rightarrow \infty$ ,

$$0 > K_1 \stackrel{(2.4)}{=} -\beta_1 \frac{q}{\lambda} \frac{\alpha(1-\mu(\hat{\phi}))}{c-1-\alpha\mu(\hat{\phi})} > -\beta_1 \frac{q}{\lambda} \frac{\alpha e^{-q\hat{\phi}}}{c-1-\alpha} > -\beta_1 \frac{q}{\lambda} \frac{\alpha e^{-q(\phi^0-\delta)}}{c-1-\alpha} \xrightarrow{q \rightarrow \infty} 0.$$

Similarly,

$$0 > K_2 \stackrel{(2.4)}{=} -\beta_2 \frac{q}{\lambda} \frac{\alpha(1-\mu(1-\hat{\phi}))}{c-1-\alpha\mu(1-\hat{\phi})} > -\beta_2 \frac{q}{\lambda} \frac{\alpha e^{-q(1-\hat{\phi})}}{c-1-\alpha} > -\beta_2 \frac{q}{\lambda} \frac{\alpha e^{-q(1-\phi^0-\delta)}}{c-1-\alpha} \xrightarrow{q \rightarrow \infty} 0.$$

Total equilibrium labor force on the other hand is bounded away from zero for any  $q$ , since

$$\begin{aligned} \hat{L} &= \beta_1 x^{-1}(c-1-\alpha\mu(\hat{\phi})) + \beta_2 x^{-1}(c-1-\alpha\mu(1-\hat{\phi})) \\ &\stackrel{(B.1)}{>} (\phi^0 - \delta)L^1(\phi^0 - \delta) + (1 - \phi^0 - \delta)L^2(\phi^0 + \delta) \\ &\stackrel{(B.3)}{>} \underbrace{L_1^0 - \delta'(\phi^0 - \delta) + L_2^0 - \delta'(1 - \phi^0 - \delta)}_{\text{independent of } q} \stackrel{(B.2)}{>} 0. \end{aligned}$$

Together, these results imply that as  $q$  grows, the term  $\hat{L}^2$  eventually dominates in expressions  $b$  and  $c$ , rendering them positive. Thus, there exists  $\tilde{Q}$  such that there exists a stable steady state near  $\mathbf{L}^0$  for all  $q \geq \max\{Q, \tilde{Q}\}$ .  $\square$

**Proof of Theorem 7.** Consider  $\tilde{L} = \beta_1/(c - \alpha\mu(1) - 1)$ , which is positive by [Hypothesis H6](#). By continuity and the Extreme Value Theorem,  $\pi$  attains its supremum  $\pi(\mathbf{L}^*, \mathbf{L}^*)$  over  $[0, \tilde{L}]^2$ . This supremum is strictly positive and  $\mathbf{L}^* \in \mathbb{R}_+^2$  since  $\pi(\vec{0}, \vec{0}) = 0$  and

$$\left. \frac{\partial}{\partial L_1} \pi((L_1, 0), (L_1, 0)) \right|_{L_1=0} = \lambda + 1 - c + \alpha\mu(1) > 0$$

by [Hypothesis H5](#). Moreover, straightforward substitution<sup>1</sup> shows that  $\pi(\mathbf{L}, \mathbf{L}) < 0$  whenever  $\mathbf{L}$  is outside  $[0, \tilde{L}]^2$ , and hence the local supremum  $\pi(\mathbf{L}^*, \mathbf{L}^*)$  is a global maximum.

Furthermore, the function  $\bar{\pi}_i(L_i, \beta_i) = \beta_i(1 - e^{-\lambda \frac{L_i}{\beta_i}}) + L_i(1 - c + \alpha\mu(L_i/L))$  is strictly supermodular since  $\frac{\partial^2 \bar{\pi}_i}{\partial L_i \partial \beta_i} = \lambda^2 / \beta_i^2 L_i e^{-\lambda \frac{L_i}{\beta_i}} > 0$ . Since  $\pi(\mathbf{L}, \mathbf{L}) = \bar{\pi}_i(L_1, \beta_1) + \bar{\pi}_1(L_2, \beta_2)$ , this therefore implies that the optimal labor force leans towards the dominant type,  $L_1^* \geq L_2^*$ .

In addition, any interior solution satisfies the first order condition

$$\frac{\partial \pi(\mathbf{L}^*)}{\partial L_i} = x\left(\frac{L_i^*}{\beta_i}\right) + 1 - c + \alpha\mu(\phi_i^*) + \alpha\mu'(\phi_i^*)\phi_i^*(1-\phi_i^*) - \alpha\mu'(1-\phi_i^*)(1-\phi_i^*)^2 = 0, \quad (B.4)$$

<sup>1</sup>Indeed,  $\pi(\mathbf{L}, \mathbf{L}) < \sum_{i=1}^2 \beta_i + L_i(1 + \alpha\mu(L_i/(L_1 + L_2)) - c) < 2\beta_1 + 2\tilde{L}(1 + \alpha\mu(1) - c) < 0$ .

where  $\phi_i^* = \frac{L_i^*}{L_1^* + L_2^*}$ .

We can rewrite this equation as

$$\begin{cases} 0 = x \left( \frac{\phi^* L^*}{\beta_1} \right) + 1 - c + \alpha \mu(\phi^*) + \alpha(1 - \phi^*) [\phi^* \mu'(\phi^*) - (1 - \phi^*) \mu'(1 - \phi^*)] \\ 0 = x \left( \frac{(1 - \phi^*) L^*}{\beta_2} \right) + 1 - c + \alpha \mu(1 - \phi^*) - \alpha \phi^* [\phi^* \mu'(\phi^*) - (1 - \phi^*) \mu'(1 - \phi^*)], \end{cases} \quad (\text{B.4''})$$

where  $\phi^* = \frac{L_1^*}{L_1^* + L_2^*}$  and  $L^* = L_1^* + L_2^*$ . In the special case where both pools are of the same size,  $\beta_1 = \beta_2$ , symmetry dictates that  $\phi^* = 0.5$  and the last two terms fall away, making the conditions equivalent to Equation (2.4), and hence implying that the fully balanced steady state maximizes surplus. In all other cases where  $\beta_1 > \beta_2$ , the solution to Equation (B.4'') differs from those to Equation (2.4) whenever  $\phi^* \mu'(\phi^*) - (1 - \phi^*) \mu'(1 - \phi^*) \neq 0$  at the stable steady state.  $\square$

**Proof of Theorem 8.** The first order conditions in Equation (B.4'') give necessary conditions for the size and composition of  $\mathbf{L}^*$ . After multiplying the first equation with  $\phi^*$  and the second with  $(1 - \phi^*)$ , their sum simplifies to

$$0 = \phi^* x \left( \frac{\phi^* L^*}{\beta_1} \right) + (1 - \phi^*) x \left( \frac{(1 - \phi^*) L^*}{\beta_2} \right) + 1 - c + \alpha [\phi^* \mu(\phi^*) + (1 - \phi^*) \mu(1 - \phi^*)]$$

At this point, it is a simple exercise in algebra to verify that the bonuses

$$\Delta_i = c - 1 - x \left( \frac{L_i^*}{\beta_i} \right) - \alpha \mu \left( \frac{L_i^*}{L_1^* + L_2^*} \right)$$

implement  $\mathbf{l} = \mathbf{L}^*$  and satisfy all market clearing equations (2.6) when starting from  $\mathbf{L}^*$ .  $\square$

**Proof of Theorem 9.** Since all educated minority workers find employment, it follows immediately that  $\bar{l}_2 = l_2$ , while the binding quota implies that  $l_1 = \frac{\hat{\phi}}{1 - \hat{\phi}} l_2 \leq \bar{l}_1$ . The partial employment of majority workers reduces their expected lifetime earnings, leading to indifference conditions

$$\begin{cases} c - x \left( \frac{\bar{l}_1}{\beta_1} \right) - \alpha \mu(L_1/\bar{l}) = \frac{\hat{\phi}}{1 - \hat{\phi}} \bar{l}_2 / \bar{l}_1 \\ c - x \left( \frac{\bar{l}_2}{\beta_2} \right) - \alpha \mu(L_2/\bar{l}) = 1. \end{cases}$$

Writing the equations as a function of the total student body  $\bar{l} = \bar{l}_1 + \bar{l}_2$  allows us to apply a similar solution strategy as in Lemma 2.1.1: Indeed, let  $\ell_1^*(\bar{l}, \bar{l}_2)$  be the solution to the first equation for a given student body  $\bar{l}$  and type-2 enrollment  $\bar{l}_2$ . Since  $x' < 0$  and  $\mu' > 0$ , it follows that  $\frac{\partial \ell_1^*}{\partial \bar{l}} < 0$  and  $\frac{\partial \ell_1^*}{\partial \bar{l}_2} > 0$ . Similarly,  $\frac{\partial \ell_2^*}{\partial \bar{l}} < 0$ , where  $\ell_2^*(\bar{l})$  is obtained from the second equation. As such, total student supply  $\ell_1^*(\bar{l}, \ell_2^*(\bar{l})) + \ell_2^*(\bar{l})$  is weakly decreasing in  $\bar{l}$ , implying a single crossing  $\bar{l} = \ell_1^*(\bar{l}, \ell_2^*(\bar{l})) + \ell_2^*(\bar{l})$ .

The second indifference condition is identical to the case without intervention in [Equation \(2.3\)](#), and hence the unregulated student body  $(l_1^0, l_2^0)$  satisfies  $\ell_2^*(l_1^0 + l_2^0) = l_2^0$ . Moreover, since the quota is binding, it also follows that  $\hat{\phi} < \frac{l_1^0}{l_1^0 + l_2^0}$  or, equivalently,  $\frac{\hat{\phi}}{1-\hat{\phi}} l_2^0 / l_1^0 < 1$ . In terms of total student supply, it follows that

$$\ell_1^*(l_1^0 + l_2^0, \ell_2^*(l_1^0 + l_2^0)) + \ell_2^*(l_1^0 + l_2^0) = \ell_1^*(l_1^0 + l_2^0, l_2^0) + l_2^0 < l_1^0 + l_2^0,$$

and hence the quota *reduces* the total student body  $\bar{l} < l_1^0 + l_2^0$ . The resulting improvement in mentoring leads to increased educational investment among type-2 individuals,  $\bar{l}_2 = l_2 = \ell_2^*(\bar{l}) > l_2^0$ , and is compensated by a reduction in type-1 investment,  $\bar{l}_1 = \bar{l} - \bar{l}_2 < l_1^0$ . The dominant type reduces investment *despite* the improved mentoring only because it now suffers from uncertain employment, indicating an over-investment into education.  $\square$