Race, Gender, and the Timing of Preference in Hiring Games

Logan M. Lee and Glen R. Waddell *

August 2018

Preliminary version. Not to be quoted.

Abstract

We model a hiring process in which candidates are evaluated in sequence by two agents of the firm. We introduce the potential for taste-based discrimination and characterize how one agent's private valuation of the candidate indirectly influences the other agent's adjudication of candidates. This influence is often in an offsetting direction and is partially corrective. Yet, this offsetting response can also be large enough that even a high-productivity candidate who is privately favored by one agent is less likely to be hired even when the other agent has no preference over private attributes. In an experimental setting, varying the private values of a second "agent" induces large differences in how male subjects adjudicate female candidates. No such response is evident in female subjects.

Keywords: diversity, hiring, gender, race, discrimination

JEL classification: J1, J7, D8

*Waddell (waddell@uoregon.edu) is a Professor at the University of Oregon and Research Fellow at IZA Bonn. Lee (leelogan@grinnell.edu) is an Assistant Professor at Grinnell College.

1 Introduction

Gender and racial disparities are often quite striking, and efforts to diversify business, political, faculty, and administrative offices are often frustratingly slow in bearing fruit. Our own profession still includes disproportionately few women and members of historically underrepresented racial and ethnic-minority groups, relative both to the overall population and to other academic disciplines (Bayer and Rouse, 2016). "Removing implicit and institutional barriers" is commonly among stated recommendations, and even though identifying subtle implicit barriers is challenging, the rewards to doing so may be large. When potentially obvious instruments of policy (installing Chief Diversity Officers, for example) have no measurable effect on trends in hiring minority groups (Bradley et al., 2018), returning to the individual incentives within hiring processes seems a worthwhile undertaking.

While experimental evidence supports taste-based racial discrimination as a direct contributor to unequal treatment (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2011; Castillo et al., 2013), incomplete information can also give rise to statistical discrimination (Altonji and Pierret, 2001; Farber and Gibbons, 1996; Aigner and Cain, 1977). We approach a mechanism at the intersection of the incomplete information and discrimination arising from decision makers at different places in the hierarchy of a firm having different valuations of non-productive traits. In short, where those differences arise and how they are managed has significant implications for employment and productivity outcomes across race and gender.

We consider a setting in which two agents of a firm participate in a sequential evaluation of a job candidate, we consider the implications of agents anticipating private benefits or costs associated with an observable but non-productive attribute of the candidate. We have in mind the individual's race or gender, for example, in a structure that nicely captures either "bottom-up" or "top-down" efforts to increase the representation of women or racial minorities. Holding the sequence of evaluation constant—an initial screening followed by further consideration if the initial screening goes well—we vary *where* in the sequence and to what degree the candidate's non-productive attribute is valued. Among our results, we show that where pro-diversity interests are stronger at the top of the institution, acting on such preference may be limited in its ability to narrow gaps in outcomes across race or gender, and may even contribute to *increasing* wage

and employment gaps. Thus, in this setting, even preference *for* some non-productive attribute in candidates can be to a candidate's detriment. This has implications for future productivity and the upward mobility of diverse candidates who are successful at the employment stage. Moreover, we show that when those at the top of institutions value these private attributes more than those earlier in the sequence, they are also likely to be insufficiently equipped to incentivize cooperation from those below.

The setting we consider is rich enough to capture the relevant tradeoffs yet sufficiently straightforward that we can speak effectively to policy. We abstract away from the role of committees, for example, and consider only individual agents, two in number, and acting in sequence on behalf of the firm or institution. We assume that the candidate is considered by the second agent (we have in mind the firm's owner, for example, although one could imagine a university administrator also fitting well) only when the first agent (a division manager or department chair, for example) has determined that the candidate is worthy of forwarding in the search. In that way, the process we model captures the typical "up or out" nature of job searches.[1]

Becker (1957) first introduced an economic model of discrimination in which employers had a taste for discrimination, insofar as there was a disamenity to employing minority workers who would have to compensate employers by being more productive at a given wage or being willing to accept a lower wage for identical productivity. Elements of this intuition will remain in our model, although the implications will now depend on where in the sequence such a disamenity is introduced—whether it is introduced "early" or "late." Elements of the longer literature will also be evident in what follows as we reconsider the role of private valuations amid uncertainty around

[1] Green and Laffont (1987) model a two-person decision problem but assumes away a hierarchy of agents. Similarly, Luo (2002) considers collective decision making in a two-person model where agents collaboratively to make decisions. In more-recent work, Guo and Shmaya (2017) and Frankel (2018) consider two-stage hiring processes. Guo and Shmaya (2017) focuses on sender-receiver games in which a sender distributes information in an attempt to influence the actions of a receiver, who receives that information and has access to various other sources of information. Frankel (2018) allows for applicant hiring and considers the role of biases held by the hiring manager who can then make decisions outside of the interests of firms. Frankel (2018) shows that it can be optimal to allow the hiring manager discretion as long as a signal of ability falls above a threshold.

worker productivity (Arrow, 1971; Phelps, 1972; McCall, 1972; Arrow, 1973; Spence, 1973).[2]

In terms of actionable policy, we will speak directly to the implications of directed searches—where private values are arguably a stronger motivating factor at the top of the firm's hierarchy. We will refer to these preferences as "top-down," and demonstrate that in such environments, early decision makers will often take positions that offset the anticipated preferences of later decision makers. In the limit, when the late-arriving preference *for* the personal attribute is large, this "offsetting" effect is sufficient to leave even the high-productivity candidates from the privately preferred group worse off—they face a *lower* probability of employment, not higher.[3] For example, where leadership values female candidates, even highly productive female applicants are harmed by early decision makers protecting their interest against the anticipation of favorable treatment in subsequent rounds. In no way is this due to a disutility associated with hiring a candidate with a particular attribute (e.g., we do not need the first agent to dislike female candidates to find that female candidates can be made worse off when favored by the second agent) but is solely due to agents having incomplete information of candidates' true abilities and the requisite tradeoffs being made at the margin when the early mover anticipates a candidate-favoring bias being introduced by subsequent decision makers. Thus, one might fear that policies designed to encourage the hiring of workers who increase workforce diversity can promote even the opposite outcome if agents of the firm (particularly those acting early in hiring decisions) do not share equally in those interests.

This tension between the first and second decision makers is fundamental. As such, we consider comparative statics around these margins, varying the private values introduced by the first and second agents as we consider the implications on employment and workforce productivity. As private values influence the relative probabilities with which candidates of different abilities are

[2] In other related work, Eriksson and Lagerström (2012) use a resume study in Norway to show candidates who have non-Nordic names, are unemployed, or older receive significantly fewer firm contacts. Kuhn and Shen (2013) find that job postings in China that explicitly seek a certain gender, while suggestive that firms have preferences for particular job-gender matches, only play a significant role in hiring decisions for positions that require relatively little skill. Jacquemet and Yannelis (2012) discuss whether observed bias is due to discrimination against a particular group or favoritism for another group. Other explanations for gender and race gaps include firms benefitting from increased productivity when workforces are homogenous (Breit and Horowitz, 1995), and in-group-favoritism effects (Lewis and Sherman, 2003). Pinkston (2005) introduces the role for differentials in signal variance (e.g., black men have noisier signals of ability than white men) into a model of statistical discrimination. Ewens, Tomlin and Wang (2012) consider separating statistical discrimination from taste-based discrimination and find support for statistical discrimination in rental markets. For a review of the evolution of empirical work on discrimination, see Guryan and Charles (2013).

[3] That "top-down" diversity goals may struggle to increase the number of good candidates of the preferred type is consistent with Chief Diversity Officers having no impact on trends in hiring minority groups (Bradley et al., 2018).

hired, we will also discuss the distributional consequences for subsequent promotion games.

In Section 2, we introduce the model we have in mind, solving the sequential consideration of agents backwards. Throughout, we consider private values of either sign although cases in which candidates are "favored" somewhere in the hiring process may be the more relevant to policy, especially where we demonstrate that this a priori favor can be to their detriment.

In Section 3, we consider a setting in which the second agent in the sequence is somewhat "naive" in forming his expectations of the first agent's action—not expecting that the first agent may respond to the second agent's private incentives. For example, university leadership may reveal that they favor female or minority candidates at the margin and fully expect that departments will not work to oppose these interests. Yet, as long as there is the potential for departments to value those non-productive attributes *differently*, interests can be in conflict. In light of the asymmetries in how early and late decision makers can influence outcomes, we discuss the model's implications for subsequent promotion games and the role of incentive pay.

In Section 4, we provide experimental evidence that, when primed with information about the private interests of a subsequent decision maker, individuals will avoid advancing candidates who should be advanced based on their merits alone. Specifically, we consider a policy relevant scenario where there is preference for female candidates, and the subjects are choosing two candidates from mixed-gender groups of three.

In Section 5, we extend the model to consider promotion, performance pay, and the role of private valuations that are more "bottom up." In Section 6 we offer some additional insights that can be gained with the theoretical model before offering concluding remarks in Section 7.

We do include an appendix section, where we consider the setting in which Agent 2 is "savvy" regarding Agent 1's incentives, and fully anticipates this in his own optimization routine. While we tend to think that those in leadership positions (university deans, for example) may fall short of fully anticipating how others (department committees) respond to "top-down" directives, there is additional intuition offered by considering outcomes in these settings. For example, it is in this setting that we consider whether the second decision maker can incentivize the first in a way that sufficiently aligns their private valuations of the non-productive attribute.

2 Theory

2.1 The setup

We are intent on considering the implications of agents having private values associated with some non-productive attribute of a job candidate as they undertake the hiring responsibilities for the firm. In so doing, we consider a two-stage hiring game in order to speak to the implications of these private values being introduced to the hiring process at different stages. By assumption, Agent 1 considers the candidate first and either rejects the candidate or forwards the candidate to Agent 2 for further consideration. If forwarded, Agent 2 can then reject or hire the candidate. Within such a hierarchy, we then consider private valuations: "bottom-up" preferences (e.g., grass roots efforts to increase racial diversity among co-workers), or "top-down" preferences (e.g., a university administrator's preference to increase the presence of female faculty in STEM fields), or combinations thereof.[4]

As a candidate's productivity is not verifiable, both agents only know that with probability $\alpha \in (0,1)$ a given candidate is highly productive and would therefore be a "good" hire. We quantify the upside to hiring such a candidate as an increase in the firm's value from V_0 to V_g. With probability $(1-\alpha)$ the candidate's productivity is low, and upon hiring would lower the firm's value from V_0 to V_b. In such a case, the firm is always best served by rejecting the candidate, in which case the firm's value would remain at the status-quo level, V_0. (This V_0 can easily be normalized to zero, but we retain for now, thinking that the intuition is made clearer.)

It is uninteresting to consider compensation schemes that do not tie remuneration to agents' actions. That said, these weights are determined outside the model and we simply parameterize these relationships in Agent 1 receiving $\tau_1 \in (0, \tau_2)$ of the value to the firm and Agent 2 receiving $\tau_2 \in (\tau_1, 1)$, such that $\tau_1 + \tau_2 \leq 1$. As agents are moving in strict sequence, consistent with a hierarchy, we think it reasonable to anticipate that $\tau_1 \leq \tau_2$.[5]

We introduce the potential for discrimination and favoritism by allowing for some non-productive but verifiable attribute of the candidate to be privately valued by either or both agents. Given the sequence of actions, we notate any private benefits accruing to Agent 1 from hiring the candidate

[4] STEM: Science, Technology, Engineering, and Mathematics.

[5] For some context regarding the use of incentive pay broadly, see Murphy (2013).

as B_1, and any private benefits accruing to Agent 2 as B_2. To maintain interest and relevance, we will limit agents' private values to those that yield interior solutions.[6] That is, we will limit private values to those that do not have the agents' first-order conditions collapse to "always reject" or "always accept." The model can be solved backwards.

2.2 Agent 2's problem

When the candidate is forwarded to Agent 2 for final consideration, Agent 2 draws an independent signal of the candidate's productivity. The signal, s_2, is drawn from $N(\mu_b, \sigma_b)$ if the candidate is a "bad" type, and from $N(\mu_g, \sigma_g)$ if the candidate is a "good" type, where $\mu_b < \mu_g$. $F_b(\cdot)$ is the CDF of $N(\mu_b, \sigma_b)$ and $F_g(\cdot)$ is the CDF of $N(\mu_g, \sigma_g)$.[7] With such a setup, Agent 2's decision rule can then be summarized in the choice of a reservation signal, $\hat{s}_2$. If the realized signal, s_2, is higher than the chosen reservation signal, $\hat{s}_2$, the candidate is hired. If $s_2 < \hat{s}_2$, the candidate is rejected and no hire is made.

Formally, Agent 2's objective equation can be written as,

$$\begin{aligned} \underset{\hat{s}_2}{\text{Max}}\, V_2(\hat{s}_2) \;=\; & \alpha[F_g(\mathbb{E}_2[\hat{s}_1]) + (1 - F_g(\mathbb{E}_2[\hat{s}_1]))F_g(\hat{s}_2)]\tau_2 V_0 \\ & + \alpha(1 - F_g(\mathbb{E}_2[\hat{s}_1])(1 - F_g(\hat{s}_2))(\tau_2 V_g + B_2) \\ & + (1-\alpha)[F_b(\mathbb{E}_2[\hat{s}_1]) + (1 - F_b(\mathbb{E}_2[\hat{s}_1])F_b(\hat{s}_2)]\tau_2 V_0 \\ & + (1-\alpha)(1 - F_b(\mathbb{E}_2[\hat{s}_1]))(1 - F_b(\hat{s}_2))(\tau_2 V_b + B_2). \end{aligned} \tag{1}$$

As Agent 2 only considers the candidate upon her having successfully navigated Agent 1's evaluation, the probability Agent 2 puts on the candidate being highly productive is updated from the population parameter, α, to reflect Agent 1's evaluation (i.e., that s_1 must have been no smaller than $\hat{s}_1$). Each term in (1) therefore represents the probability weighted outcomes of the hiring

[6] Assuming that $\tau_1 V_b \leq B_1 \leq \tau_1 V_g$, and $\tau_2 V_b \leq B_2 \leq \tau_2 V_g$ effectively limits the set of values where an agent has these dominant strategies to just those where $B_i = \tau_i V_b$ or $B_i = \tau_i V_g$, respectively. More generally, the range of private values over which interesting interactions occur depends on the payoff levels to agents relative to these private values. That is, in the symmetric case, where $B_i > \tau_i V_g$, Agent i will adopt an "always-accept" strategy. Likewise, where $B_i < \tau_i V_b$, Agent i will adopt an "always-reject" strategy. This restriction also implies that $\frac{f_g(s)}{f_b(s)}$ is increasing in s.

[7] Lang and Manove (2011) suggest that employers find it more difficult to evaluate the productivity of black candidates than white candidates. This would imply that non-productive attributes may be correlated with signal noise. Our model can easily encompass this potential by allowing σ_b and σ_g to vary with the candidate's non-productive attribute.

game—the candidate is either a good candidate but not hired (Agent 2 realizes $\tau_2 V_0$), good and hired ($\tau_2 V_g + B_2$), bad and not hired ($\tau_2 V_0$), or bad and hired ($\tau_2 V_b + B_2$). While the true conditional probability depends on Agent 1's reservation signal, $\hat{s}_1$, what matters to characterizing Agent 2's choice is his belief about what Agent 1's reservation signal was in the first stage, which we capture as $\mathbb{E}_2[\hat{s}_1]$.[8]

Given (1), Agent 2's choice of $\hat{s}_2$ solves the first-order condition,

$$\frac{\alpha(1 - F_g(\mathbb{E}_2[\hat{s}_1]))f_g(\hat{s}_2)}{(1-\alpha)(1 - F_b(\mathbb{E}_2[\hat{s}_1]))f_b(\hat{s}_2)} = \frac{\tau_2 V_0 - (\tau_2 V_b + B_2)}{(\tau_2 V_g + B_2) - \tau_2 V_0}. \tag{2}$$

That is, in equilibrium Agent 2's optimal reservation signal, $\hat{s}_2^\star$, equates the ratio of probabilities of committing type-I and type-II errors (i.e., $\alpha(1 - F_g(\mathbb{E}_2[\hat{s}_1]))f_g(\hat{s}_2)$, and $(1-\alpha)(1 - F_b(\mathbb{E}_2[\hat{s}_1])f_b(\hat{s}_2)$, respectively) with the ratio of costs (i.e., $(\tau_2 V_g + B_2) - \tau_2 V_0$, and $\tau_2 V_0 - (\tau_2 V_b + B_2)$).

2.3 Agent 1's problem

In the first stage, Agent 1 draws an independent signal, s_1, of the candidate's productivity to be compared to a chosen reservation signal, $\hat{s}_1$. As above, the candidate's signal of productivity, s_1, is drawn from $N(\mu_b, \sigma_b)$ if the candidate is a "bad" type and from $N(\mu_g, \sigma_g)$ if the candidate is a "good" type. If $s_1 < \hat{s}_1$, the candidate's file is immediately abandoned and no hire is made—Agent 2 never sees the candidate and the resulting firm value is V_0. If $s_1 \geq \hat{s}_1$, the candidate is then subjected to consideration by Agent 2, as described in Equation (2).

Where $R_2(\mathbb{E}_2[\hat{s}_1])$ captures Agent 2's choice of $\hat{s}_2$ given his expectation of $\hat{s}_1$, Agent 1's objective equation can be written,

$$\begin{aligned} \underset{\hat{s}_1}{\text{Max}}\, V_1(\hat{s}_1) \;=\; & \alpha[F_g(\hat{s}_1) + (1 - F_g(\hat{s}_1))F_g(R_2)]\tau_1 V_0 \\ & + \alpha(1 - F_g(\hat{s}_1))(1 - F_g(R_2))(\tau_1 V_g + B_1) \\ & + (1-\alpha)[F_b(\hat{s}_1) + (1 - F_b(\hat{s}_1))F_b(R_2)]\tau_1 V_0 \\ & + (1-\alpha)(1 - F_b(\hat{s}_1))(1 - F_b(R_2))(\tau_1 V_b + B_1). \end{aligned} \tag{3}$$

where we capture in B_1 any private value Agent 1 associates with the candidate's non-productive

[8] Agent 2's expectation of the probability a good candidate cleared Agent 1's reservation is therefore $1 - F_g(\mathbb{E}_2[\hat{s}_1])$, while the expectation of the probability a bad candidate cleared Agent 1's reservation signal is $1 - F_b(\mathbb{E}_2[\hat{s}_1])$.

attribute. In general, Agent 1 chooses $\hat{s}_1$ subject to the first-order condition,

$$\frac{\alpha f_g(\hat{s}_1)(1-F_g(R_2)) + \alpha(1-F_g(\hat{s}_1))f_g(R_2)(\partial R_2/\partial \hat{s}_1)}{(1-\alpha)f_b(\hat{s}_1)(1-F_b(R_2)) + (1-\alpha)(1-F_b(\hat{s}_1))f_b(R_2)(\partial R_2/\partial \hat{s}_1)} = \frac{\tau_1 V_0 - (\tau_1 V_b + B_1)}{(\tau_1 V_g + B_1) - \tau_1 V_0}. \tag{4}$$

As above, Agent 1 chooses his optimal reservation signal, $\hat{s}_1^\star$, to equate the ratio of probabilities of committing type-I and type-II errors with the ratio of costs.[9]

3 When Agent 2 is naive

3.1 Agent behavior

In this section, we begin with the consideration of strictly "top-down" preferences (i.e., $B_2 \neq 0$ while $B_1 = 0$), which is consistent with Agent 1 being interested only in the productivity of the candidate while Agent 2 has private objectives to increase the representation of a races or gender (i.e., $B_2 > 0$).[10]

In making a decision, Agent 2 will obviously have in mind the reservation signal Agent 1 would have had. We model Agent 2's naiveté by setting this expectation, $\mathbb{E}_2[\hat{s}_1]$, equal to what Agent 1 would choose in the absence of any private values (i.e., as if $B_2 = 0$). This is akin to Agent 2 not anticipating that Agent 1 will consider B_2 when choosing $\hat{s}_1$, or update optimally given the expressed interest they've announced. When $\mathbb{E}_2[\hat{s}_1] = \hat{s}_1^\star|_{B_2=0}$, Agent 2's first-order condition in (2) simplifies to

$$\frac{\alpha(1-F_g(\hat{s}_1^\star|_{B_2=0}))f_g(\hat{s}_2)}{(1-\alpha)(1-F_b(\hat{s}_1^\star|_{B_2=0}))f_b(\hat{s}_2)} = \frac{\tau_2 V_0 - (\tau_2 V_b + B_2)}{(\tau_2 V_g + B_2) - \tau_2 V_0}, \tag{5}$$

and $\hat{s}_2^\star$ depends on the expectation of Agent 1's reservation signal, here set to $\hat{s}_1^\star|_{B_2=0}$, which is constant in B_2.

That $\mathbb{E}_2[\hat{s}_1] = \hat{s}_1^\star|_{B_2=0}$ also implies that $\partial R_2(\mathbb{E}_2[\hat{s}_1])/\partial \hat{s}_1 = 0$. As Agent 1 finds neither private cost nor private benefit in the non-productive attribute of the candidate (i.e., $B_1 = 0$), τ_1 drops

[9] This is easy to see in the symmetric case (i.e., $V_b = -V_g$, $V_0 = 0$, and $\alpha = 0.5$), as Agent 2's first-order condition collapses to $f_g(\hat{s}_2) = f_b(\hat{s}_2)$.

[10] In Appendix A we consider the case where Agent 2 is savvy—that is, he correctly anticipates how Agent 1 best responds to $B_2 \neq 0$—and Agent 1 likewise considers Agent 2's best response when choosing $\hat{s}_1$. While this alters the optimal $\hat{s}_1$ and $\hat{s}_2$ profiles for a range of B_2 values, the original result remains—large positive values of B_2 make the "preferred" group less likely to be hired on average.

from the agent's problem, and Agent 1's first-order condition in (4) simplifies to

$$\frac{\alpha f_g(\hat{s}_1)(1 - F_g(R_2(\hat{s}_1|_{B_2=0})))}{(1-\alpha) f_b(\hat{s}_1)(1 - F_b(R_2(\hat{s}_1|_{B_2=0})))} = \frac{V_0 - V_b}{V_g - V_0}. \tag{6}$$

which *will* vary with B_2 through its effect on $R_2(\cdot)$.

In Figure 1 we illustrate the tradeoffs in the sequential screening of candidates by plotting the optimally chosen $\hat{s}_1^\star$ and $\hat{s}_2^\star$ across a range of B_2 between $\tau_2 V_b$ (where the private cost to Agent 2 of hiring someone with this attribute always dominates the value of hiring a "good" worker, and the problem collapses on always reject) and $\tau_2 V_g$ (where the private benefit to Agent 2 of hiring someone with this attribute always dominates the cost of hiring a "bad" worker, and the problem collapses on always reject). For illustrative purposes, we impose *ex ante* symmetry, and abstract away (for now) from the role of incentive pay in agent behavior by setting $\tau_1 = \tau_2 = 0.5$.[11,12]

Where B_2 decreases from zero, and hiring the candidate imposes greater private costs on Agent 2, Agent 2 responds with a higher reservation signal, making it less likely that such a candidate would successfully clear the required standard. In other words, the required productivity required in expectation must be higher in order to offset $B_2 < 0$. While this exposes the firm to higher odds of making a type-I error (i.e., rejecting a good candidate) the perspective of Agent 2 is that the private costs of hiring an individual with the non-productive attribute must be offset by the higher probability that the candidate is a good hire. That Agent 2 is motivated by this private value is clearly costly to the firm. Likewise, then, any *increase* in B_2 from zero is also costly to the firm, as Agent 2 chooses a lower reservation signal in an attempt to increase the probability that the privately-favored candidate is hired, where B_2 would be realized. This in exchange exposes the firm to higher odds of making a type-II error (i.e., hiring a bad candidate).[13]

The shape of Agent 1's choice of $\hat{s}_1^\star$ across B_2 is where we first observe the behavior of con-

[11] Symmetry is defined as $V_b = -V_g$, $V_0 = 0$, and $\alpha = 0.5$. Collectively, the first-order condition for the choice of $\hat{s}_2$ is clear, as $f_g = f_b$ in equilibrium. In characterizing agent behavior, we adopt that $V_b = -4$, $V_g = 4$, $\mu_g = 1$, $\mu_b = -1$, and $\sigma_g = \sigma_b = 1$.

[12] As changes in τ_1 and τ_2 determine the relative weights the private values play in agent decisions (i.e., where τ_i is large, Agent i's incentives are better aligned with the firm's) we will return to consider these margins below.

[13] Figure 1 also reveals two interesting limiting cases, in $B_2 = \tau_2 V_b$ and $B_2 = \tau_2 V_g$, where Agent 2's decision rule collapses on either "never hire" or "always hire." Again, this is in keeping with expectations. Where $B_2 = \tau_2 V_b$, the private cost associated with the non-productive attribute is sufficiently high that there is no possible outcome available (i.e., even $\tau_2 V_g$ is not sufficiently high) that would dominate the status quo of $\tau_2 V_0$ net of B_2. Likewise, where $B_2 = \tau_2 V_g$, the private benefit to the non-productive attribute is sufficiently high that there is no possible outcome available (i.e., even $\tau_2 V_b$ is not sufficiently low) that would dominate the potential that a "bad" hire is made and the firm realizes a value of $\tau_2 V_b$.

sequence. First, as Agent 1 anticipates how $\hat{s}_2^\star$ varies with B_2, Agent 1's first-order condition in (6) implies that a *higher* reservation signal is adopted when B_2 is higher, requiring less uncertainty regarding the candidate's type before forwarding the candidate to Agent 2 where Agent 2 will be favorable toward hiring the candidate.

Proposition 1. *With top-down preferences, for any $|B_2| > 0$ Agent 1's choice of reservation signal acts as a weakly corrective force. That is, Agent 1's mitigating influence on firm value is non-negative.*

Moreover, as B_2 approaches $\tau_2 V_g$ and Agent 2's decision rule collapses to the unproductive act of "always accepting" a candidate who provides the privately valued attribute, Agent 1's decision rule collapses to that which would be chosen by a single decision maker facing the same uncertainty (i.e., $\hat{s}_1^\star = 0$). In effect, while Agent 1's best response to Agent 2 favoring the candidate is corrective and valuable to the firm in expectation, Agent 2's private interest is sufficient to completely dissipate the gains provided to the firm from having the second signal of the candidate's uncertain productivity.[14]

However, this "corrective" ability of Agent 1 is not symmetric around $B_2 = 0$. As the private *costs* to Agent 2 increase and $B_2 < 0$ approaches $\tau_2 V_b$, Agent 2 never hires the candidate and Agent 1's decision is of no consequence to eventual outcomes. The sequential nature of the hiring also limits the influence Agent 1 can have in offsetting $B_2 < 0$ and, in the limit, the firm unavoidably suffers the costs of Agent 2's bias. Again, this private cost results in the complete dissipation of the value to the firm of having multiple signals of productivity. This asymmetry is anticipated, given that both agents must approve a candidate for hire, but rejection can occur with either agent's decision. That fundamental asymmetry, in part, drives that favor is more likely to be offset than is discrimination against a candidate. However, this wedge is larger with late-arriving preferences (i.e., $|B_2 - B_1| > 0$)

3.2 Implications for employment and firm value

In Panel A of Figure 2, we suspend for the moment the role of Agent 1 and plot the associated employment rates associated with Agent 2 acting alone. While any observable attribute would

[14] Note that with symmetry assumed, a single decision maker would solve the first-order condition at $\hat{s} = 0$. In Figure 1, that $\hat{s}_1^\star < 0$ when $B_2 = 0$ is a reflection of the value to the firm of having a second agent—Agent 1 can adopt a lower reservation signal, anticipating that Agent 2's independent draw and evaluation is pending. (While particularly evident at $B_2 = 0$, this is also driving the general result that $\hat{s}_1^* \leq 0$.)

work, for expositional purposes we plot the relative treatments of male and female candidates, with the private value (B_2 in this case) capturing the private value to Agent 2 associated with a female candidate. Clearly, without any offsetting influence of Agent 1, increases in B_2 from zero increase the probability a low-productivity female is hired, and at a faster rate than the increase in probability a high-productivity female is hired. While optimal for Agent 2, this is destructive to firm value as, together, the expected productivity of female workers is falling in B_2. Likewise, as B_2 decreases from zero (and female hires are privately costly) the probability a low-productivity female is hired decreases at a slower rate than does the probability a high-productivity female is hired. This again decreases the value of the firm.

In Panel B of Figure 2, we re-introduce Agent 1. Relative to Agent 2 acting alone, the offsetting and corrective (from the perspective of firm value) influence of Agent 1 is immediately evident. In fact, for both high- and low-productivity candidates, there is now significantly less separation in employment probabilities by gender—this is true across all B_2 other than in the limiting case of $B_2 = \tau_2 V_b$. For B_2 in the vicinity of $\tau_2 V_g$, high-productivity candidates can be strictly worse off than they would be without preference. In this area, some appealing intuition is available.

For example, the effect becomes more pronounced as the fraction of good workers (α) falls. With fewer good candidates in the pool, unbiased agents adopt higher reservations signals. However, where $B_2 > 0$, Agent 2 is less responsive to decreases in α, which Agent 1 best responds to (excessively) by requiring an even-higher reservation signal. This binds on even the most-productive.

Similarly, we find that higher-productivity candidates are differentially worse off when there is more noise in the signals of productivity. With less-informative signals, there is additional benefit to multiple draws from that distribution (and the associated decision), which drives a larger differential between 1) the probability of hiring a good worker based on that worker's type, and 2) Agent 1 best responding by raising the threshold for the "privately preferred" candidates by more than it is increased for non-preferred candidates.

Proposition 2. *With top-down preferences, employment rates among low-ability candidates are strictly increasing in B_2. That is, low-ability candidates are always better off when they can offer employers a privately valued attribute. However, employment rates among high-productivity candidates are not monotonic in B_2—there exists some small private cost ($B_2 < 0$) for which the high-productivity candidate is strictly better off than he would be under a regime in which B_2 is*

large and positive. In a sequential-hiring game, the early decision maker has enough influence on the candidate's prospect that the high-productivity candidate would prefer even mild discrimination in later rounds over having agents in later rounds offer strong favortism.

In Figure 3, we plot the expected value to the firm of a candidate, with and without the influence of Agent 1.[15] Not surprisingly, there is value to the firm in the screen provided by Agent 1, across all $B_2 > V_b$. Given the uncertainty of a candidate's productivity, Agent 1's screen simply enables the hiring of "good" candidates with higher probability. However, more interesting is the asymmetry introduced into the expected outcomes. In the absence of Agent 1, the expected costs to the firm of Agent 2 his private interest are symmetric around $B_2 = 0$. However, when taking an active role in the hiring, Agent 1 is less able to offset Agent 2's inclination to reject female candidates (when $B_2 < 0$) than to offset Agent 2's inclination to hire female candidates (when $B_2 > 0$), hence the asymmetry. In general, in part due to the ability of Agent 1 to unilaterally reject candidates, the expected costs to the firm of agents weighing private interests are higher with top-down discrimination (i.e., for $B_2 < 0$) than they are with top-down favoritism (i.e., for $B_2 > 0$).[16]

4 Empirics

In lieu of somewhat extraordinary data required to fully vet the implications of the above discussion, we do wish to consider the fundamental empirical question around which the above intuitions and results rest—will individuals in the first stage of a hiring game behave differently if they anticipate that a subsequent decision-maker might regard a candidate's private attribute in making a hiring decision? Specifically, we consider whether experimental subjects are more- or less-likely to advance female candidates when they anticipate favor being shown to female candidates in a subsequent decision. Here establishing that experimental subjects are seemingly willing to sabotage female candidates in ways that protect higher-flying male candidates, we will then return to consider extensions of the model.

[15] We normalize to one the expected value to the firm when Agent 2 is naive and there are no private values, $B_1 = B_2 = 0$.

[16] In the limit, as Agent 2's private values decrease, Agent 2 rejects all candidates with the private attribute, regardless of whether Agent 1 is present. In such cases, the expected value to the firm collapses to $V_0 = 0$.

4.1 Design

In an effort to make the experimental setting more easily understood, note also that we will move away from the relatively abstract theory and towards simple, discrete choice. For example, while the firm in our theoretical model may hire any number of applicants, subjects in our experiment will be participating in a hiring process in which a single candidate is chosen. Similarly, agent's in our theoretical model optimally choose thresholds above which a candidate would be advanced for further consideration. To set such a threshold optimally, even in the absence of bias, subjects would need to have an understanding of the means and standard deviations of the high and low productivity pools and be able to use that information to balance type one and type two errors. This is a significant bar for the average experimental subject. Instead, we focus on relatively simple comparisons. Given three possible candidates, which two should be advanced? In the absence of bias, this decision quickly collapses to simply advancing the top two candidates (though in the absence of noisy productivity signals advancing the top two candidates will yield identical outcomes to advancing the best and worst candidates). When bias based on an observable trait is introduced, however, subjects must consider the decision making process of the second agent and adjust their own strategies accordingly. Specifically, if female candidates will receive preference from Agent 2, subjects should be wary of advancing second best female candidates, especially when both the best and worst candidates in a given group of three are male. By advancing the top two candidates in this case, the second-best candidate may be chosen based on the second evaluators preference for females. If instead the subject advances the two male candidates, only expected productivity can be used in determining which candidate will be hired.

A second key way in which the experiment simplifies both the theoretical model and the real world is that expected productivity is based on a single number, total SAT score. Rather than overwhelming subjects with a variety of characteristics that may imply productivity, we simplify the information as much as possible and provide only a name (to signal candidate gender) and SAT score for each candidate with SAT scores in the range 930 to 1230.[17] This simplification also implies a fairly straightforward empirical analysis of subject choices.

[17] All names were drawn from the US Social Security Administration's list of the 200 most-common male and female names given to individuals born in 1990. Any names appearing in the top 1,000 names for both males and females were removed.

All subjects were playing the role of the first agent in a two-stage hiring process. They evaluated thirty sets of three candidates in random order and were asked to chose two of the three candidates to advance. In the initial instructions, subjects were told that of the two candidates they advanced, one would be chosen by a different person. The subjects were told that both they and the other person would receive \$4 if the candidate chosen to compete ended up outperforming another candidate chosen in a similar fashion. The competition was described as one in which "there is no guarantee that the competitor with the highest SAT score will win a competition they enter," but also that "competitors with a higher SAT score are more likely to win than competitors with lower SAT scores."[18] Our fundamental experimental variation comes from the randomly assigned experimental setting—some subjects were told that this second person would earn a \$1 bonus for hiring a female candidate. After each choice, subjects were told which of the two candidates the second person chose to advance into competition. They were not informed of any outcomes of competition, as subjects were paid for a single, randomly chosen round at the end of the experiment.

The behavior of the second agent was computerized. When the second person is not said to receive a bonus, the computer simply chooses the candidate with the highest SAT score. When the second person is said to receive a \$1 bonus for hiring a female, the computer chooses the highest-ranking female, unless the highest-ranking male has an SAT score more than S points higher, where S is drawn randomly (once for each experimental subject) from $\{0, 40, 80, 160\}$.[19] Including zero allows us to observe any effect of the preference frame itself, separately from the degree to which favor may then be applied, and may pick up a notion of what their own perception of preference implies. Though subjects do not observe the degree of preference directly, they can discern this over multiple scenarios as we provided feedback on the decision of the subsequent decision (i.e., which one of the two they forwarded was then chosen).

4.2 Subjects

A growing literature has shown that experiments completed using Amazon's Mechanical Turk (MTurk) yield reliable results similar to those in a typical experimental lab, particularly when

[18] In Figure 7 we replicate the experimental instructions, as seen by subjects, and an example of the sets of three candidates they were asked to evaluate.

[19] 160 is the largest difference in SAT scores observed by subjects within a single set of three candidates, implying that the second decision maker would always advance the highest-ranking female candidate, regardless of SATs.

subject pools are properly restricted (Thomas and Clifford, 2017). In our case, subjects were recruited through MTurk, conditional on having a "Masters" classification and residing in the United States. Subjects were also limited to participating in the experiment a single time. We display their characteristics in Table 1. As should be expected with randomized experiments, subject characteristics are reasonably well balanced with most subject characteristics failing to predict which treatment subjects received. Subjects also completed the experiment in less than 10 minutes, on average, regardless of treatment. In a simple balance test, only GPA predicts assignment to treatment with $p < .05$, with subjects self reporting higher GPA's slightly more likely to be assigned to the treatment group.

In our empirical analysis, we will model deviations from advancing the two highest-ranking candidates by SAT. Without additional information, doing so would assure the highest expected value to the subject, supported further by the instruction to all subjects that "competitors with a higher SAT score are more likely to win than competitors with lower SAT scores." Subjects in the control group advanced the top-two candidates in 88 percent of their scenarios while subjects in the treatment group advanced the top-two candidates in 77 percent of their scenarios. Despite having no incentive to deviate from a top-two strategy, only 50 percent of the control group advanced the top two every time. Among treated subjects, 35 percent advanced the top two subjects in every scenario they were presented. While we attempted to imply variation in the strength of preference by having the subjects learn about the deviation in SAT scores the second decision maker was willing to overlook to exercise preference, the size of the bias had little effect on the fraction of scenarios in which the top-two candidates were advanced.[20]

4.3 Analysis and results

As a baseline specification, we model the behavior of our experimental subjects as

$$\mathbb{1}(\text{AdvancedTopTwo})_{iq} = \beta_0 + \beta_1 \mathbb{1}(\text{FemalePreference}_i) + \gamma_q + \epsilon_{iq} \ , \tag{7}$$

where $\mathbb{1}(\text{AdvancedTopTwo})_{iq}$ equals one if i advances the two candidates with the highest SAT

[20] As we suggested above, the degree to which the second decision maker was wiling to overlook SAT points for gender (i.e., 0 points, 40 points, 80 points, or 160 points) did not significantly predict the fraction of scenarios in which the top-two candidates were advanced in a simple linear regression model ($\beta = 0.00006, p = 0.882$).

scores from among the three candidates randomly observed in question q. We capture any level difference there may be across treatment and control sessions in $\mathbb{1}(\text{FemalePreference}_i)$, and due to random assignment, interpret $\hat{\beta}_1$ as the difference in choice induced by the "preference" frame. As deviations need not be across all scenarios, even given the preference framing, we will also consider variation coming from difference-in-differences parameters—interactions of $\mathbb{1}(\text{FemalePreference}_i)$ and the various positions of male and female candidates in the rank ordering within q. That is, the variation that will identify the key parameters of interest are within q but across i—across different subjects in the experiment who faced the same triple of options (i.e., same gender composition and SAT scores) but had either been told that the subsequent decision would be made by someone interested in hiring a female candidate or told nothing about such a preference. We include γ_q to capture question fixed effects, and we estimate ϵ_{iq} allowing for clustering at the question level. As subjects experience questions in random order, we also control for question-order fixed effects.

In Column (1) of Table 2, we capture the level difference associated with subjects being informed that the second decision will be made by someone who will be paid an additional \$1 if the person hired is female. This difference is not small—when the subjects are told that the second decision will be made by someone who is rewarded for choosing a female, there is a 13.3-percent reduction (11.8 pp) in the probability that the two candidates with the highest and second-highest SAT scores are chosen. As this may confound experimentally induced variation with any unobserved heterogeneity across treatment and control groups, in subsequent columns, we unpack this systematic pattern into its contributors. In Column (2), for example, we see stronger evidence that this difference is experimentally induced, as roughly 65 percent of the average difference is driven by subject behaviour around questions in which a female was among the top two candidates—when a female candidate is among the two-highest SATs, the probability of the top two being chosen decreases by an additional 8.5 pp compared to those in the experimental arm but facing no females in the top two.

In Column (3), we again see the sort of systematic variation that theory implies one would—allowing separate parameters for top first- and second-ranking males and females, respectively, the patterns evident in earlier specifications are clearly driven by second-ranked female candidate, where treated subjects may well fear that preference in a subsequent decision may foreclose on the top-ranking male's opportunity to compete. In Column (4), as we allow for the specific interaction

of a male candidate occupying the top position and a female candidate occupying the second, we see precisely the pattern of decision making predicted—coincident with treatment, it is in these specific opportunities that treated subjects exhibit 16.6-percent lower (14.8 pp) probability of choosing the top-two candidates. Moreover, there is no remaining level difference associated with the positions of either candidate alone—it is precisely in the cells in which a second-ranking female may jeopardize the first-ranked male that our experimental subjects shut down on the female's advancement.

It is clear that where subjects anticipate that the subsequent decision maker is privately motivated to advance female candidates, they act as if they are protecting top-ranking males and sabotaging second-ranking female candidates. Yet, where there is no such opportunity because both second- and third-ranking candidates are female, we might anticipate no such pattern. In Column (5), we push as far as we can with identification, separately identifying within those questions in which there are top-ranking males, but second- *and* third-ranking candidates are female. Indeed, this reveals that when the best candidate is male and both the second and third best candidates are female, the main effect (i.e., the significant reduction in top-two advancement with treatment) is 88.7-percent smaller, with no remaining difference compared to control subjects other than what is being picked up in the level difference of treatment itself ($p < .001$).[21]

In Table 3 we present results separately for male and female experimental subjects, which proves important in any inference one might be inclined to make given our results. Across all specifications, a sample of male subjects reveals similar patterns, with treatment driving a large wedge between how candidates are adjudicated. However, while point estimates follow similar patterns among female subjects, they are small in magnitude and not significantly different from zero.

Specifically, among male subjects adjudicating questions in which there is a first-ranked male and second-ranked female, those who were anticipating that the subsequent decision would be made by one with preference for females (and had opportunity to protect the male candidate) are 22.2-percent less likely to advance the top two candidates than the average control subject. Those without opportunity (given that the third-ranked candidate is also female) are only 6.7 percent less

[21] In some cases, subjects failed to advance any two candidates and instead advanced a single candidate (always the top candidate). We include these observations in our results and code these as cases in which the subject did not advance the top two candidates. Excluding these observations does not change our results in any meaningful way. We also considered the possibility that agents may have been strategically choosing to advance a single candidate as an alternative to avoiding the preference of the second decision maker in the treatment regime. In a model similar to that in Column (5), but on an outcome that captures that only one candidate had been advanced, we find no precisely estimated parameters.

likely to advance the top two. Again, we find no evidence that female experimental subjects are evidencing significant patterns of this sort.

Recall that we provided immediate feedback to subjects on the second decision maker's choices, anticipating that subjects may learn through repeated interactions. While subjects are mildly more responsive to treatment in questions they faced later in the experiment, no significant differences emerge. We also consider the difference in SAT scores between candidates, finding little precision in the estimates.[22]

5 Extensions

In this section, we consider to policy relevant implications of the above model—implications for subsequent promotion, and some thoughts on the role of performance pay in such a setting.

5.0.1 Subsequent promotion games

As $B_2 \neq 0$ induces patterns of hiring that are specific to productivity-by-gender pools of candidates, in any subsequent period, average (within-firm) productivity levels will vary by gender. Even in the absence of private values playing a direct role in promotion decisions, promotion outcomes can be shown to depend on B_2. For example, if $B_2 > 0$ at the hiring decision, the average female in the firm will be of lower productivity than the average male. If subsequent decision makers perceive this difference in productivity, this disparity implies that females will suffer lower promotion probabilities within firms.[23] While the implication of heterogeneous productivity in promotion games has been considered in the literature (Bjerk, 2008), we offer an original source of heterogeneity, active even though the *ex ante* distribution of male and female productivity is common—one driven, somewhat surprisingly, by favoritsm.

[22] Also in unreported analyses, we considered differential effects by both subject age and SAT score, finding only suggestive evidence that subjects with higher SAT scores were more-responsive to treatment.

[23] Of course, if the potential promotion of those with the privately valued attribute continue to be subject to the bias that occurred in the hiring process, outcomes will be affected. In fact, in such a setting, our "hiring" game can itself be recast as a promotion game of sorts.

5.0.2 Performance pay

We next allow for $\tau_1 \leq \tau_2$ as we consider the firm having taken steps to align the incentives differently across the internal hierarchy. In Panel A of Figure 4, we show each agent's reservation signal across B_2 for a range of $\tau_2 \in [.5, 1)$, adjusting τ_1 accordingly, such that $\tau_1 = 1 - \tau_2$. For comparison with the baseline model, the solid lines indicate the $\hat{s}_1^\star$ and $\hat{s}_2^\star$ chosen when $\tau_1 = \tau_2 = 0.5$. Clearly, as τ_2 becomes increasingly large, any bias introduced in $\hat{s}_2^\star$ through $B_2 \neq 0$ (either discrimination or favoritism) is mitigated as Agent 2 cares more about the firm's value relative to his own private value as τ_2 increases. This is seen in the flattening of $\hat{s}_2^\star$ in B_2 in Figure 4. Importantly, the corresponding flattening of Agent 1's optimal $\hat{s}_1^\star$ in B_2 is entirely in response to B_2's influence on $\hat{s}_2^\star$. That is to say, because we have assumed $B_1 = 0$, any $\tau_1 > 0$ achieves unbiased decisions from Agent 1.[24]

In panels B and C, we plot the employment rates for good and bad workers respectively. As expected, increasing τ_2 works to offset biases arising from either $B_2 < 0$ or $B_2 > 0$, and allows for a larger range of these private values over which $\hat{s}_2$ does not collapse to either "always hire" or "never hire" rules.

5.1 The role of Agent 1's private value

As one last consideration before generalizing to both agents valuing the candidate's non-productive attribute, recall the asymmetry in Agent 1's ability to mitigate Agent 2's biases—when Agent 1 foresees Agent 2's bias, Agent 1 plays a corrective role from the firm's perspective. Yet, a naive Agent 2 plays no such role when Agent 1 exercises either favoritism or discrimination. In this way, our model reverts to the Becker (1957) intuition—Agent 2 simply governs over a second signal of productivity and acts unbiasedly. However, for a given candidate, this does imply that the expected outcomes across potential private values can be rank ordered.

Proposition 3. *For a given private value, $W < 0$, the candidate would prefer to be subjected to a regime where $\{B_1, B_2\} = \{0, W\}$ than to a regime where $\{B_1, B_2\} = \{W, 0\}$. That is, if the candidate is to be discriminated against somewhere, she prefers discrimination to fall late in the sequence. Alternatively, for a given private value, $W > 0$, the candidate would prefer to be*

[24] While we do not devote space to $\tau_1 \geq \tau_2$, these scenarios behave as expected. In the limit, where $\tau_2 = 0$, Agent 2 collapses to never hiring members of the non-preferred group for any $B_2 < 0$ and always hiring members of the preferred group for $B_2 > 0$.

subjected to a regime where $\{B_1, B_2\} = \{W, 0\}$ *than to a regime where* $\{B_1, B_2\} = \{0, W\}$*. That is, favoritism is more beneficial if experienced early in the sequence.*

In Figure 5, we allow for $B_1 \neq 0$ and $B_2 \neq 0$, capturing that both agents may value the candidate's non-productive attribute. As before, we plot Agent 2's choice of $\hat{s}_2$, but now with a menu of $\hat{s}_1$ corresponding to values of $B_1 \in (\tau_1 V_b, \tau_1 V_g)$. (As Agent 2 is naive, note that B_1 has no influence on $\hat{s}_2$.) Within the series of plots, Agent 1's decision rule in the strictly "top-down" case (i.e., that corresponding to $B_1 = 0$) can be seen in the solid line.

We illustrate two results in Figure 5. First, as we have assumed that Agent 2 is not best responding to $\hat{s}_1$ at the margin, we document the expected pattern of behavior, that, for any $B_2 \in (\tau_2 V_b, \tau_2 V_g)$, $\hat{s}_1$ is strictly decreasing in B_1. As Agent 1's private value increases, holding constant Agent 2's private value, Agent 1 is less likely to reject those candidates who have the attribute. The less-obvious takeaway from Figure 5, and one we wish to stress, we state as a proposition.

Proposition 4. *For all* B_1*,* $\hat{s}_1^\star$ *is strictly increasing in* B_2*. That is, Agent 1 raises the bar on candidates as Agent 2's private value increases.*

In Figure 6 we plot the *ex post* rates of employment for "good" and "bad" female candidates, assuming that female is the private attribute around which the agents are potentially optimizing. As in Panel B of Figure 2, Figure 6 again captures that employment outcomes are sensitive to B_2, not only as a direct result of Agent 2's private value, but also indirectly through Agent 1's best response to $B_2 \neq 0$. Namely, employment rates among "good" female candidates eventually decline in B_2, reflecting Agent 1's ability to force the rejection of a particular candidate in response to a high B_2. As Agent 1 is less able to force the hiring of a candidate, employment rates among "bad" female candidates again monotonically increase in B_2. Figure 6 also demonstrates an important implication of Agent 2's naiveté. In panels A and B of Figure 6, then, we demonstrate that this pattern remains, across all B_1.

Proposition 5. *Both high- and low-productivity candidates of the preferred type prefer higher* B_1 *to lower* B_1*. That is, in a sequential-hiring game when the late decision maker is naive, candidates weakly benefit from early preference, as late decision makers provide no offsetting or corrective role.*

6 Implications

Before concluding, we note four interesting implications, each of which may motivate additional exploration. First, where a single decision maker discriminates on taste, the average productivity among the "preferred" group decreases. However, as early movers in a sequential decision can take positions offsetting top-down preferences, average *ex post* productivity falls off more slowly among those who are "preferred" *a priori*. For example, with top-down preferences, early decision makers who anticipate excessively favorable treatment of female candidates in subsequent evaluations best respond by increasing the standards they impose on female candidates, which implies that later decision makers will be considering female candidates who are, on average, of higher quality (i.e., able to have cleared the higher standards imposed in early rounds). Therefore, while fewer female candidates advance in the sequence, the average productivity of those who do advance for final consideration is higher. As such, this may leave later decision makers increasingly misinformed of underlying female productivity, thereby reinforcing or strengthening prior beliefs among those in leadership positions. Overall, the influence of late-arriving preference for female candidates will change the mix of low- and high-productivity female employees such that average productivity falls among female employees. This, we presume, introduces a source of downward pressure on female wages in a way that would contribute to the persistence of male-female wage gaps.

Second, the model offers interesting implications in light of existing evidence that resumes with African-American-sounding names receive fewer call backs (Bertrand and Mullainathan, 2004). While such an empirical regularity is consistent with either a single decision maker statistically discriminating or a single decision maker exercising a kind of taste-based discrimination, it is also consistent with the actions of the first of multiple decision makers in a regime where subsequent decision makers are expected to show preference *for* African-American candidates. (We assume that call-back decisions are made by initial screeners and not by those who will ultimately make the hire.) Of course, policy prescriptions across these potential mechanisms will differ significantly.

Third, note that the model we present implies that if preferences for the private attribute are of the top-down variety we describe, we should be concerned that even in regimes where women and racial minorities are valued by leadership, such candidates can be harmed by revealing their identities early if initial screeners merely value those attributes less than leadership. Candidates will

also experience tension, insofar as they do benefit from eventually revealing their identities. (In the model, they would choose to identify strictly between Agent 1 and Agent 2.) "Blind" assessments should arguably be considered in this context, as outcomes are certainly not neutral with respect to the information provided to reviewers. For example, in regimes where preferences for female recruitment are not uniformly held across the firm's hierarchy, pro-minority leadership meets with more success by incorporating blind-recruitment tools in early assessments of job candidates.

Finally, in a setting where late decision makers are savvy enough to anticipate the best responses of early decision makers (see Appendix A), those early-moving agents who would themselves be uninclined to discriminate will *raise* the bar on candidates against whom leadership is inclined to discriminate. Average productivity of female candidates is therefore higher coming out of early stages, moving subsequent priors away from "reject" and toward "accept." Interestingly, where standard models of taste-based discrimination yield heterogeneity in *ex post* productivity by gender and standard models of statistical discrimination yield homogeneity in *ex post* productivity, the sequence of decision making in our setting allows for taste-based discrimination to exist, yet, due to the "corrective" action of an earlier agent of the firm, *not* be evidenced in *ex post* heterogeneity in productivity by gender.

7 Conclusion

In a setting in which two agents of a firm participate in a sequential evaluation of a job candidate, we consider the implications of agents anticipating private benefits or costs associated with an observable but non-productive attribute of the candidate.

We show that private values introduced in one stage of such a game are evident not only in the actions of the agent with those private motivations, but also among agents in other stages of the game. In particular, where preference *for* a personal attribute is introduced late in the sequence, earlier decision makers partially offset this preference by raising the standard they impose on candidates with that attribute. From the firm's perspective, this moves toward first best and we therefore characterize such patterns as partially corrective. In the typical "up-or-out" hiring environment, where earlier decision makers have much more sway in *rejecting* candidates than in *hiring* candidates, the potential response among those who anticipate subsequent favorable treat-

ment still has the potential to subject candidates who are "preferred," on average, to lower odds of employment than they would have experienced had their private attribute not been valued or observable.

In an experimental setting, we vary the conditions under which groups of three candidates are adjudicated by subjects. All subjects see multiple sets of three candidates, with signals of their productivity (i.e., an SAT score) and gender (i.e., a typically male or female name), and must decide which two of the three they wish to forward for further consideration. We vary whether those subjects were informed that the subsequent consideration would be done by someone with the private incentive to hire females. Without knowledge of that private incentive, the dominant strategy is to forward the two candidates with the highest signals of productivity. However, in the treatment arm, we demonstrate strong willingness among male subjects to protect high-flying male candidates by passing over the best female candidates.

References

Aigner, Dennis J and Glen G Cain, "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, 1977, pp. 175–187.

Altonji, Joseph G and Charles R Pierret, "Employer Learning and Statistical Discrimination," *The Quarterly Journal of Economics*, 2001, *116* (1), 313–350.

Arrow, Kenneth, "Some Models of Racial Discrimination in the Labor Market," 1971.

_ , "The Theory of Discrimination," *Discrimination in Labor Markets*, 1973, *3* (10), 3–33.

Bayer, Amanda and Cecilia Elena Rouse, "Diversity in the Economics Profession: A New Attack on an Old Problem," *Journal of Economic Perspectives*, 2016, *40* (4), 221–242.

Becker, Gary S, *The Economics of Discrimination*, University of Chicago press, 1957.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, September 2004, *94* (4), 991–1013.

Bjerk, David James, "Glass Ceilings or Sticky Floors? Statistical Discrimination in a Dynamic Model of Hiring and Promotion," *The Economic Journal*, 2008, *118* (530), 961–982.

Bradley, Steven W., James R. Garven, Wilson W. Law, and James E. West, "The Impact of Chief Diversity Officers on Diverse Faculty Hiring," *NBER Working Paper 24969*, 2018.

Breit, William and John B. Horowitz, "Discrimination and Diversity: Market and Non-Market Settings," *Public Choice*, 1995, *84*, 63–75.

Carlsson, Magnus and Dan-Olof Rooth, "Revealing Taste Based Discrimination in Hiring: A Correspondence Testing Experiment with Geographic Variation," *IZA Discussion Paper*, 2011, (6153).

Castillo, Marco, Ragan Petie, Maximo Torero, and Lise Vesterlund, "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination," *Journal of Public Economics*, 2013, *99*, pp. 35–48.

Eriksson, Stefan and Jonas Lagerström, "Detecting Discrimination in the Hiring Process: Evidence From an Internet-Based Search Channel," *Empirical Economics*, 2012, *43* (2), 537–563.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang, "Statistical Discrimination or Prejudice? A Large Sample Field Experiment," *Review of Economics and Statistics*, 2012, (0).

Farber, Henry S. and Robert Gibbons, "Learning and Wage Dynamics," *Quarterly Journal of Economics*, 1996, *111* (4), 1007–47.

Frankel, Alex, "Selecting Applicants," *working paper*, 2018.

Green, Jerry R. and Jean-Jacques Laffont, "Posterior Implementability in a Two-Person Decision Problem," *Econometrica*, 1987, *55* (1), pp. 69–94.

Guo, Yingni and Eran Shmaya, "The Interval Structure of Optimal Disclosure," *working paper*, 2017.

Guryan, Jonathan and Kerwin Kofi Charles, "Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots," *The Economic Journal*, 2013.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," *The Quarterly Journal of Economics*, 2015.

Jacquemet, Nicolas and Constantine Yannelis, "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market," *Labour Economics*, 2012.

Kuhn, Peter and Kailing Shen, "Gender Discrimination in Job Ads: Evidence from China," *The Quarterly Journal of Economics*, 2013, *128* (1), 287–336.

Lang, Kevin and Michael Manove, "Education and Labor Market Discrimination," *The American Economic Review*, 2011, *101* (4), 1467–1496.

Lewis, Amy C. and Steven J Sherman, "Hiring You Makes Me Look Bad: Social-Identity Based Reversals of the Ingroup Favoritism Effect," *Organizational Behavior and Human Decision Processes*, 2003, pp. 262–276.

Luo, Guo Ying, "Collective Decision-Making and Heterogeneity in Tastes," *Journal of Business & Economic Statistics*, 2002, *20* (2), pp. 213–226.

McCall, John J, "The Simple Mathematics of Information, Job Search, and Prejudice," *Racial Discrimination in Economic Life, Lexington Books*, 1972, pp. 205–224.

Murphy, Kevin J., "Executive Compensation: Where We Are, and How We Got There," in G. Constantinides, M. Harris, and R. Stulz, eds., *Handbook of the Economics of Finance*, Elsevier Science North Holland, Elsevier, 2013.

Phelps, Edmund S, "The Statistical Theory of Racism and Sexism," *The American Economic Review*, 1972, pp. 659–661.

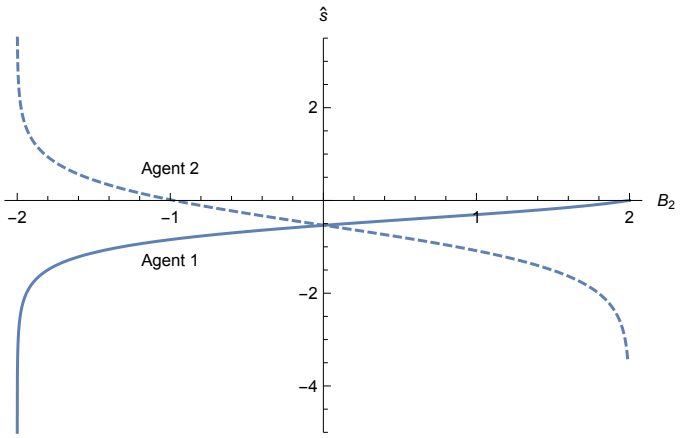
Pinkston, Joshua C, "A Test of Screening Discrimination with Employer Learning," *Industrial & Labor Relations Review*, 2005, *59*, 267.

Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, pp. 355–374.

Thomas, Kyle A and Scott Clifford, "Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments," *Computers in Human Behavior*, 2017, *77*, 184–197.

8 Figures

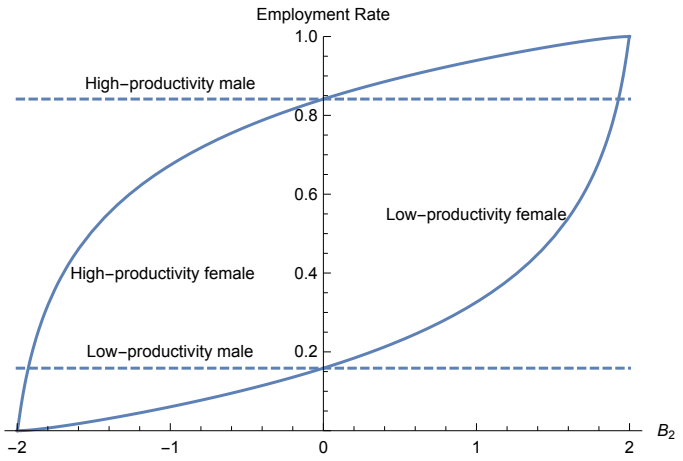
Figure 1: Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2)



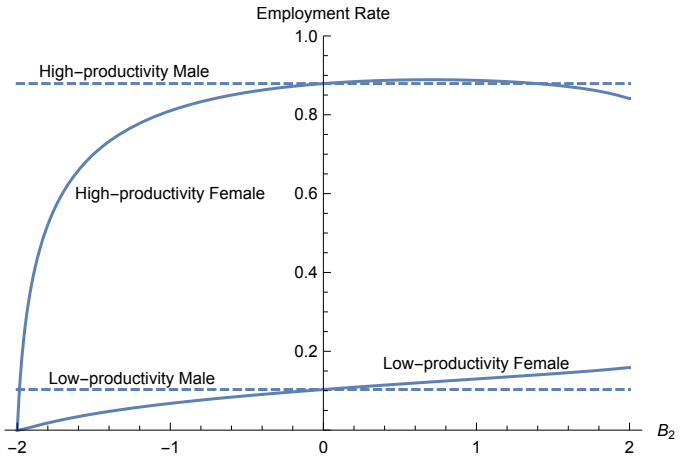
Notes: Each curve shows the optimal reservation signal of quality, above which a candidate of the "preferred" type will be advanced for further consideration by Agent 1 or hired by Agent 2. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 2. We further assume Agent 2 does not anticipate that Agent 1 will act to offset Agent 2's bias.

Figure 2: Employment probabilities with top-down preferences ($B_1 = 0$, as we vary B_2)

Panel A: No screening provided by Agent 1

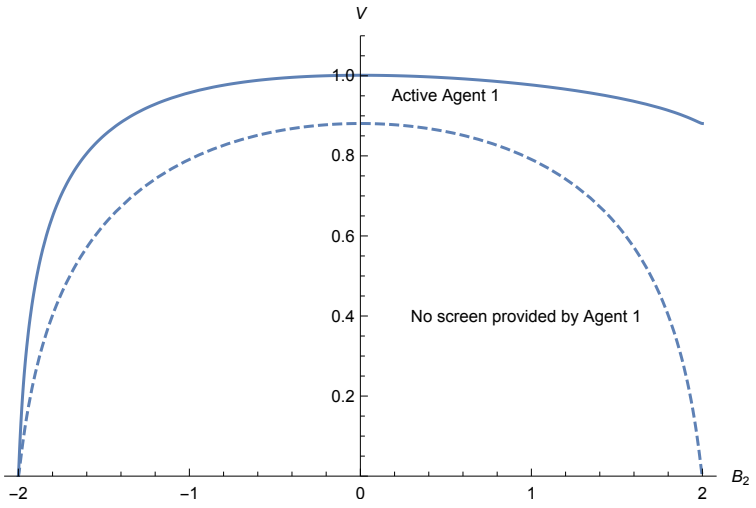


Panel B: Agent 1 screens candidates prior to Agent 2



Notes: Each curve indicates the probability that candidates from the indicated group will be hired by the firm. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 2. In Panel A, Agent 2 is acting alone with no screen provided by Agent 1. In Panel B, both Agent 1 and Agent 2 must approve of a candidate in order for that candidate to be hired.

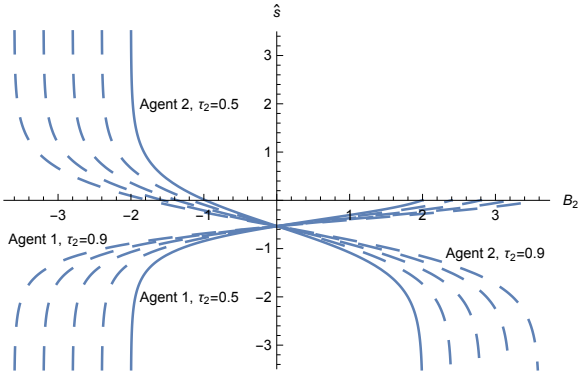
Figure 3: Firm value with top-down preferences ($B_1 = 0$, as we vary B_2)



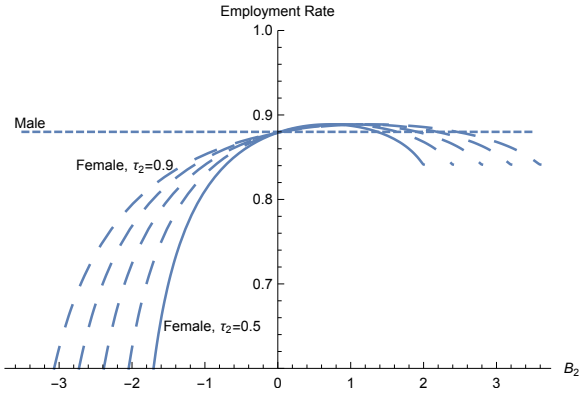
Notes: Each curve indicates the expected value to the firm of considering a randomly chosen candidate. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 2. The solid line indicates value where both agents actively participate in deciding whether to hire candidates while the dashed line indicates the value of Agent 2 acting alone.

Figure 4: Reservation signals and employment rates across τ_2, with top-down preferences ($B_1 = 0$, as we vary B_2)

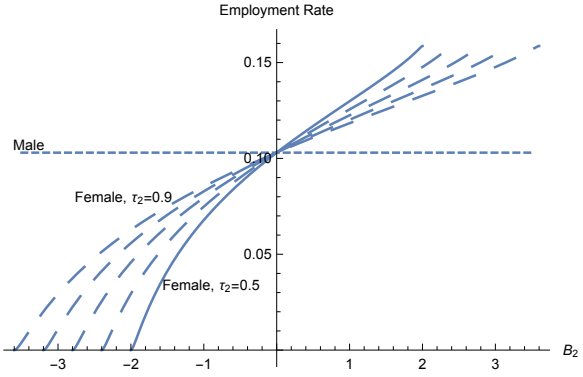
Panel A: Reservation signals



Panel B: Employment of "good" candidates

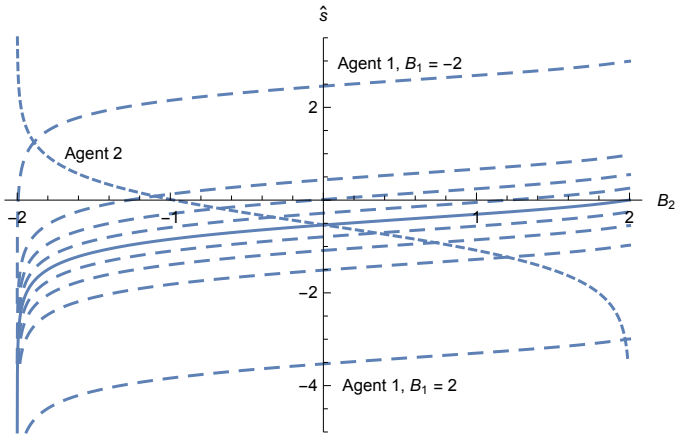


Panel C: Employment of "bad" candidates



Notes: All panels present results allowing the fraction of the benefits (costs) of hiring good (bad) candidates earned by Agent 2 (τ_2) to vary. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 2. Solid lines replicate the case presented in earlier figures and assume that Agent 2 earns half of the value of each hired candidate. Panel A indicates optimal minimum quality signals above which candidates will be advanced for further consideration (Agent 1) or hired (Agent 2). Panel B displays employment rates for "good" candidates and Panel C displays employment rates for "bad" candidates.

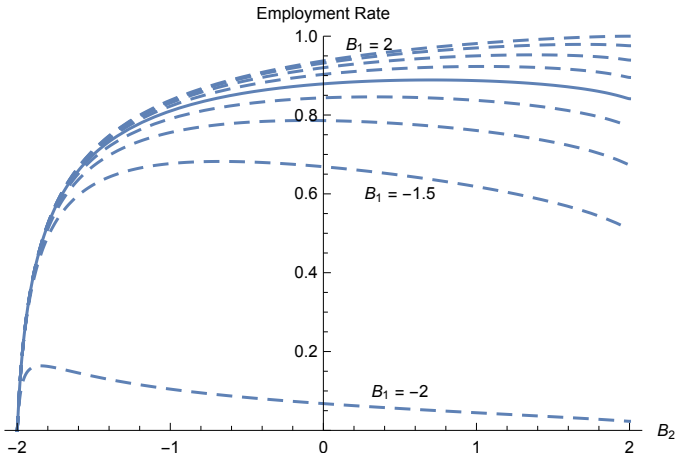
Figure 5: Reservation signals with bottom-up preferences ($B_2 = 0$, as we vary B_1)



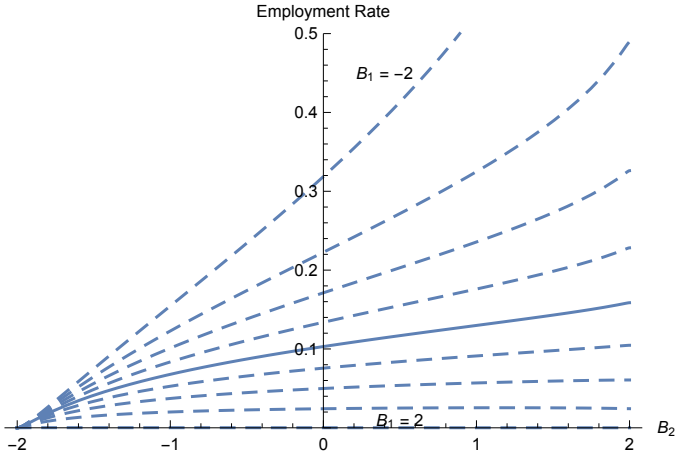
Notes: Each curve shows the optimal reservation signal of quality, above which a candidate of the "preferred" type will be advanced for further consideration by Agent 1 or hired by Agent 2. We plot results across a variety of potential valuations of this non-productive trait by both Agent 1 and Agent 2. Because Agent 2 does not anticipate Agent 1's behavior in this setting, Agent 2's minimum quality signal is unaffected by Agent 1's minimum quality signal.

Figure 6: Employment probabilities among female (i.e., the "preferred") candidates

Panel A: High-productivity female candidates



Panel B: Low-productivity female candidates



Notes: Panel A displays employment rates for "good" candidates and Panel B displays employment rates for "bad" candidates. We plot results across a variety of potential valuations of this non-productive trait by both Agent 1 and Agent 2.

Figure 7: Experiment Examples

Panel A: Instructions

Please read the following instructions carefully. You will not be able to return to these instructions once you move on.

You will see lists of 3 potential competitors and their SAT scores, which are out of 1600. You will be eliminating 1 from each list of potential competitors, leaving 2 potential competitors for a different person to consider.

This other person will then eliminate 1 more of these competitors, leaving only 1. This last remaining competitor will then compete in a task against another competitor who was chosen in a similar way, by two other people.

While there is no guarantee that the competitor with the highest SAT score will win a competition they enter, competitors with higher SAT scores are more likely to win than competitors with lower SAT scores.

You will receive $4 if the competitor you helped to chose wins the competition they enter. The person who chose the 1 competitor to actually enter into competition from the list of 2 you forwarded, will also receive $4 if the competitor wins. This other person will also receive $1 if they choose a woman to compete, even if that woman loses in competition. You will not receive this $1.

When you are finished with the game, one of the lists you saw will be chosen randomly to determine which you actually receive payment for.

Panel B: Sample question

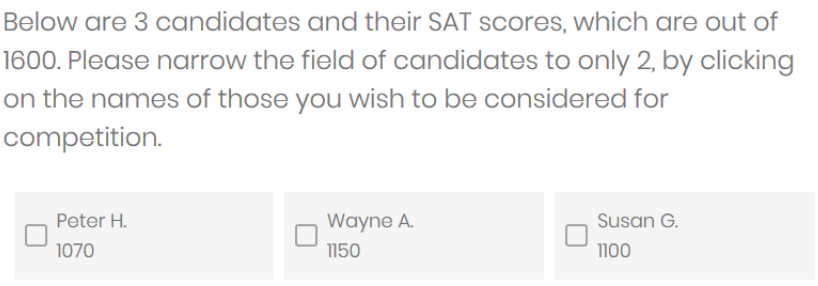


Table 1: Summary Statistics

	(1) Control	(2) Treated	(3) Difference
Male	0.457	0.570	0.114
	(0.504)	(0.497)	(0.085)
Age	38	38.51	0.514
	(9.468)	(10.22)	(0.164)
White	0.804	0.782	-0.023
	(0.401)	(0.415)	(0.068)
Black	0.109	0.0563	-0.052
	(0.315)	(0.231)	(0.050)
Asian	0.0435	0.106	0.062
	(0.206)	(0.308)	(0.040)
Hispanic	0.0217	0.0211	-0.001
	(0.147)	(0.144)	(0.025)
Other Race	0.0217	0.00704	-0.015
	(0.147)	(0.0839)	(0.023)
Income $\leq$ \$20,000	0.370	0.218	-0.151*
	(0.488)	(0.415)	(0.080)
Income \$20,001 - \$40,000	0.239	0.345	0.106
	(0.431)	(0.477)	(0.075)
Income \$40,001 - \$60,000	0.217	0.155	-0.062
	(0.417)	(0.363)	(0.068)
Income \$60,001 - \$80,000	0.0870	0.106	0.019
	(0.285)	(0.308)	(0.049)
Income $>$ \$80,001	0.0870	0.148	0.061
	(0.285)	(0.356)	(0.051)
GPA	3.305	3.480	0.176**
	(0.499)	(0.476)	(0.085)
SAT Score	1335.3	1299.0	-36.3
	(130.5)	(161.4)	(35.4)
ACT Score	26.71	26.63	-0.08
	(4.536)	(3.953)	(1.81)
Minutes To Complete	9.43	9.10	-0.33
	(9.43)	(6.58)	(1.10)
Observations	46	142	188

Notes: Observations are at the subject level. Columns 1 and 2 report standard deviations in parentheses. Column (3) presents robust standard errors. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: How inclined were subjects to forward the top-two candidates across treatment and control?

	(1)	(2)	(3)	(4)	(5)
2nd decision-maker paid for female	-0.118***	-0.045***	-0.034*	-0.060***	-0.054***
	(0.011)	(0.014)	(0.017)	(0.011)	(0.011)
× Female among top two		-0.085***			
		(0.018)			
× First if M			-0.030	0.015	
			(0.020)	(0.018)	
× Second is F			-0.079***	-0.014	
			(0.018)	(0.013)	
× First is M × Second is F				-0.088***	-0.097***
				(0.022)	(0.014)
× Third is F					-0.005
					(0.016)
× First is M × Second is F × Third is F					0.086***
					(0.019)
Mean (control)	0.890	0.890	0.890	0.890	0.890
Observations	4,980	4,980	4,980	4,980	4,980
Question FE	Yes	Yes	Yes	Yes	Yes
Question-order FE	Yes	Yes	Yes	Yes	Yes

Notes: Observations are at the subject-by-question level. In all columns, the outcome variable is whether the two candidates with the highest SAT scores were advanced. In the case of a tie in the scores of the second- and third-best candidates, subjects were counted as advancing the top-two candidates if the top candidate and either of the tied candidates were advanced. Standard errors in parentheses, allowing for clustering at the question level. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Were male and female subjects differently inclined to forward the top-two candidates?

	Male subjects			Female subjects		
	(1)	(2)	(3)	(4)	(5)	(6)
2nd decision-maker paid for female	-0.173***	-0.100***	-0.083***	-0.061***	-0.025	-0.033
	(0.014)	(0.013)	(0.019)	(0.011)	(0.020)	(0.027)
× First is M × Second is F		-0.136***	-0.137***		-0.028	-0.043
		(0.032)	(0.021)		(0.047)	(0.030)
× Third is F			-0.016			0.008
			(0.022)			(0.034)
× First is M × Second is F × Third is F			0.153***			0.006
			(0.026)			(0.038)
Mean (control)	0.922	0.922	0.922	0.866	0.866	0.866
Observations	2,742	2,742	2,742	2,238	2,238	2,238
Question FE	Yes	Yes	Yes	Yes	Yes	Yes
Question-order FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Observations are at the subject-by-question level. In all columns, the outcome variable is whether the two candidates with the highest SAT scores were advanced. In the case of a tie in the scores of the second- and third-best candidates, subjects were counted as advancing the top-two candidates if the top candidate and either of the tied candidates were advanced. Standard errors in parentheses, allowing for clustering at the question level. *** p<0.01, ** p<0.05, * p<0.1.

A Appendix: When Agent 2 is savvy

A.1 Agent behavior

In this section we relax the earlier assumption that Agent 2 is naive (i.e., does not recognize how Agent 1 best responds to $B_2 \neq 0$) and, instead, allow both agents to choose reservation signals while fully anticipating the effect that choice will have on the other agent's choice. While we are granting much more forethought and consideration to Agent 2 than may be evidenced in the field, this case fully bounds the possible scenarios relevant to policy and provides a richer understanding of the potential implications of private values in hiring games.

In Figure A1, we return to consider "top-down" preferences (i.e., $B_1 = 0$) across a range of $B_2 \in (\tau_2 V_b, \tau_2 V_g)$, but allow Agent 2 to recognize that Agent 1 will adjust $\hat{s}_1$ in response to B_2. First, note that when $B_2 = 0$, both $\hat{s}_1^\star$ and $\hat{s}_2^\star$ are as they were in the case with a naive Agent 2. (This is expected, as one model nests the other when private values are absent.) Likewise, when $B_2 > 0$, the general patterns of behavior are similar to that in the naive-owner case. Yet, where $B_2 < 0$ and Agent 2 correctly anticipates $\hat{s}_1^\star$, both $\hat{s}_1^\star$ and $\hat{s}_2^\star$ behave differently in B_2 (than was the case with naiveté, in Figure 1). In particular, Agent 1's reservation signal is no longer monotonically increasing through $B_2 \in (\tau_2 V_b, \tau_2 V_g)$. To contrast, $\hat{s}_1^\star$ is now U-shaped, decreasing in B_2 for all $B_2 < 0$ in this range.

Proposition 6. *With top-down preferences, when Agent 2 is savvy in setting expectations of Agent 1's reservation signal, $\hat{s}_1^\star$ is monotonically decreasing in $B_2 \in (\tau_2 V_b, 0)$. (As when Agent 2 is naive, when Agent 2 is savvy $\hat{s}_1^\star$ is monotonically increasing in $B_2 \in (0, \tau_2 V_g)$.)*

The intuition for this result is again found in Agent 1's inability to fully offset prejudicial bias that arises late in the hiring sequence—while Agent 1 can secure a candidate's rejection, he cannot secure a candidate's hire. When Agent 2 anticipates a higher $\hat{s}_1$, he best responds by increasing $\hat{s}_2^\star$ all the more, which ultimately decreases employment rates among those presenting the privately costly attribute. By increasing $\hat{s}_1^\star$ as Agent 2 is more inclined to discriminate (i.e., as B_2 decreases from zero), Agent 1 is able to induce a lower $\hat{s}_2^\star$ than in the naive case. In essence, where Agent 2 is naive and Agent 1 then has no ability to influence Agent 2's decision, his decision rule was motivated solely by the potential to offset Agent 2's bias at the margin. Now, where Agent 2 is aware that $\hat{s}_1$ responds to B_2, Agent 1's choice of $\hat{s}_1$ influences $\hat{s}_2^\star$ at the margin. By raising his standard on candidates in the first period, Agent 1 lowers the marginal benefit to Agent 2 increasing $\hat{s}_2^\star$ in the second period, thereby allowing the firm to better exploit the gains available through the second signal of productivity. We learn by this that prejudicial bias introduced late in a sequential-hiring game can motivate what looks like a prejudicial bias in earlier rounds; a preemptive bias-correction, of a sort. In this way taste-based discrimination introduced late in a sequence can yield a sort of statistical discrimination earlier in the sequence. However, in this setting, Agent 1 is not responding to a perceived difference in the average productivity of female candidates—as would be the case in standard models of statistical discrimination—but in recognizing that subsequent decision makers will lean away from an unbiased assessment of productivity, treats female candidates differently as a corrective action.

A.2 Implications for employment and firm value

In Panel A of Figure A2, we again plot employment rates—the patterns are remarkably similar to those in the naive case (see Figure 2). With Agent 2 now savvy, both high- and low-productivity

females are less likely to be hired for $B_2 < 0$, but there are nonlinearities in the effect of $B_2 > 0$ on employment probabilities for high-productivity females. In particular, we again see that at high values of $B_2 > 0$, high-productivity females are less likely to be hired than are high-productivity males.

In Panel B of Figure A2 we plot the expected value to the firm of considering a candidate for the savvy and naive cases. While the firm's expected value is invariant to the assumption of naiveté when $B_2 = 0$, slight differences emerge at other values of B_2. In general, the firm suffers more from Agent 2's privately motivated decisions when Agent 2 is savvy; Agent 1 offers less of a corrective influence in such cases. The exception to this rule is for extreme discrimination (i.e., B_2 approaching V_b), where Agent 1's higher standard enables the firm to escape Agent 2's "always reject" regime.

A.3 The role of Agent 1's private value

In Panel A of Figure A3, for various values of B_1, we plot the rates at which high-productivity female candidates are hired across B_2. (Recall that we use the hiring of female candidates as a placeholder of sorts in the figures, which more-broadly apply to any observable non-productive attribute for which there may be private consideration.) The bold line captures the parameterization already represented in Figure A2. Around this line, however, we see the interesting asymmetry of employment rates. For example, where B_2 is large and negative and Agent 2 is increasingly inclined toward adopting a "never hire" position, Agent 1 has no ability to influence employment regardless of his inclination to do so (i.e., for any B_1). Thus, for all B_1, employment rates converge to zero as B_2 decreases to $\tau_2 V_b$. As B_2 increases from $\tau_2 V_b$, employment rates fan out across B_1, with rates increasing faster in B_2 for higher values of B_1. This, again, reflects Agent 1's ability to "force" rejections (e.g., when B_1 is low), while being quite unable to force hires—even in the limit (as B_1 increases to $\tau_1 V_g$), employment is still very much dependent on Agent 2's private value (B_2).

In Panel B of Figure A3 we plot the expected value to the firm of a female candidate. That the expected value is highest when $B_1 = B_2 = 0$ again reflects that any privately motivated interest, in either agent, is costly to the firm. Moreover, it is interesting to note that for all B_2, firm value is maximized when $B_1 = 0$. That is, in the sequential-hiring game, the full value to having multiple signals drawn and evaluated is only exploited when the first agent is free from bias. Any departure from this not only costs the firm directly (through Agent 1 choosing a standard that depends on B_1), but indirectly costs the firm through Agent 1's influence on Agent 2's decision (even when $B_2 = 0$).

The timing of preference—whether introduced with Agent 1 or Agent 2—yields striking differences in agents' optimal thresholds. In Figure A4, we impose bottom-up preferences (i.e, $B_2 = 0$) and plot agents' optimal thresholds (Panel A) and associated employment probabilities (Panel B) across B_1. Most notable, with bottom-up preferences, Agent 2's optimal threshold is monotonically increasing in B_1. This is different from the patterns evident with "top-down" preferences (recall Figure A1), where the agent without private preference appears to "buy" more-lenient treatment from the agent who finds the candidate's non-productive attribute privately costly.

The importance of the timing of bias is also seen in Panel B of Figure A4, where we plot associated employment probabilities by productivity. With discrimination, the timing of the introduction of private values is of little consequence to employment; either agent can unilaterally dismiss candidates. As no single agent can unilaterally hire a candidate, preference for a candidate's non-productive attribute yields different patterns of behavior. With bottom-up preferences, both good and bad female candidates are more likely to be hired than male candidates, for all B_1. This

contrasts with top-down preferences (see Panel A of Figure A2) where strong preference on the part of Agent 2 ultimately leaves good female candidates less likely to be hired.

A.4 Can Agent 2 incentivize Agent 1's cooperation?

Given the similarity in employment outcomes when we assume Agent 2 is savvy, we forgo additional discussion of subsequent hiring and promotion games and the implications of performance pay in this environment. Yet, unique to the environment in which Agent 2 fully anticipates Agent 1's best response to $B_2 \neq 0$ (which, loosely speaking, is to take corrective action and mitigate Agent 2 acting on his private valuations), it is interesting to consider the potential for a transfer, from Agent 2 to Agent 1, to incentivize Agent 1's cooperation.[25]

Here we consider one important extension to the model—a potential transfer, from the firm (i.e., Agent 2, as the residual claimant) to Agent 1, attached to the hiring of a candidate presenting a particular non-productive attribute. We ask, then, whether there are private values $\{B_1, B_2\}$ for which Agent 2 will choose to reward Agent 1 for hiring such a candidate.[26]

Such practice appears in academic markets, for example, where payments would typically be made, by college-level administrators to departments, conditional on hiring a candidate who presents with a non-productive attribute, such as a minority race of gender. We parameterize this payment with ρ, through which we allow Agent 2 to transfer $\rho > 0$ from the firm to Agent 1, conditional on hiring a candidate with a particular (non-productive but verifiable) attribute. Agent 2's objective can therefore be written as,

$$\begin{aligned}\underset{\hat{s}_2,\rho}{\text{Max}}\, V_2(\hat{s}_2) \;=\;& \alpha[F_g(\mathbb{E}_2[\hat{s}_1]) + (1 - F_g(\mathbb{E}_2[\hat{s}_1]))F_g(\hat{s}_2)]\tau_2 V_0 \\ &+\alpha(1 - F_g(\mathbb{E}_2[\hat{s}_1])(1 - F_g(\hat{s}_2))(\tau_2(V_g - \rho) + B_2) \\ &+(1-\alpha)[F_b(\mathbb{E}_2[\hat{s}_1]) + (1 - F_b(\mathbb{E}_2[\hat{s}_1])F_b(\hat{s}_2)]\tau_2 V_0 \\ &+(1-\alpha)(1 - F_b(\mathbb{E}_2[\hat{s}_1]))(1 - F_b(\hat{s}_2))(\tau_2(V_b - \rho) + B_2),\end{aligned} \tag{8}$$

where the payment reflects a reduction in firm value by the amount ρ upon hiring. Similarly, as Agent 1 receives ρ, his objective equation becomes,

$$\begin{aligned}\underset{\hat{s}_1}{\text{Max}}\, V_1(\hat{s}_1) \;=\;& \alpha[F_g(\hat{s}_1) + (1 - F_g(\hat{s}_1))F_g(R_2)]\tau_1 V_0 \\ &+\alpha(1 - F_g(\hat{s}_1))(1 - F_g(R_2))(\tau_1(V_g - \rho) + B_1 + \rho) \\ &+(1-\alpha)[F_b(\hat{s}_1) + (1 - F_b(\hat{s}_1))F_b(R_2)]\tau_1 V_0 \\ &+(1-\alpha)(1 - F_b(\hat{s}_1))(1 - F_b(R_2))(\tau_1(V_b - \rho) + B_1 + \rho).\end{aligned} \tag{9}$$

In giving away part of the firm, the private cost to Agent 2 is merely his share of the direct reduction in firm value, $\tau_2\rho$. On this margin, then, any increase in ρ is less costly to Agent 2 when τ_2 is small. Regardless, however, Agent 2 benefits by any such payment only to the extent that it moves Agent 1 in his preferred direction. Since Agent 1 also pays a share of the cost of $\rho > 0$ (in terms of firm

[25] We do not discuss the feasibility of such a payment in the "naive" case, as Agent 2 recognizing the need to "correct" Agent 1's action seems a prerequisite to explaining the use and effect of such payments.

[26] US labor law forbids deductions from employee pay without serious violations of workplace rules. As such, we do not consider whether there are values for which Agent 2 would tax Agent 1 for hiring a candidate with a particular non-productive attribute. Regardless, the sequential nature of the hiring process limits Agent 2's ability to require payment from Agent 1 for hiring a candidate, as Agent 1 can always avoid such penalties by raising the required standard for hire. Agent 1 still solves the first-order condition for $\hat{s}_1$, of course, so while Agent 1 will not collapse to an "always reject" position immediately, in the limit, $\hat{s}_1^\star$ approaches "always reject."

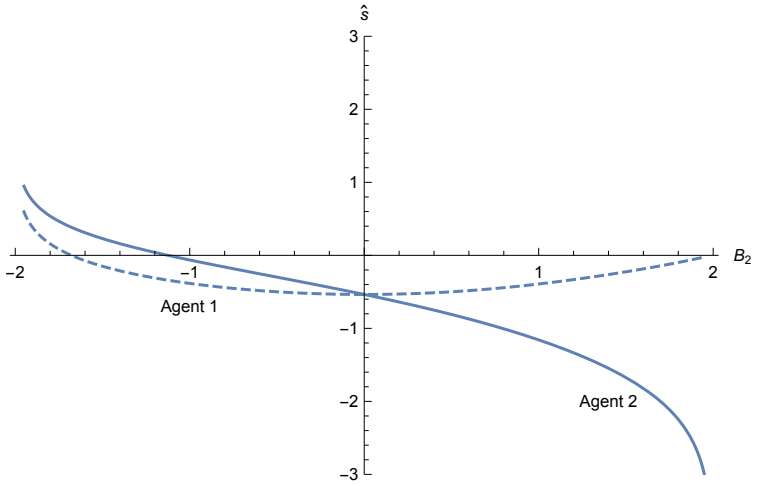
value, $\tau_1\rho$), awarding $\rho > 0$ to Agent 1 is more powerful when τ_1 is small. Thus, only for small τ_1 and τ_2 can Agent 2 benefit from a non-zero transfer of $\rho > 0$ from the firm to Agent 1.

In many cases, however, Agent 2 finds $\rho^\star = 0$ to be optimal. This implies that the additional dollar that would be used to influence $\hat{s}_1^\star$ generates less than a dollar's worth of return in noise reduction and increased probability a candidate will be hired. Intuitively, Agent 2 is most likely to choose a non-zero ρ in cases where B_2 is large. In the extreme case, where $B_2 \to \tau_2 V_g$, we have shown (in Figure A1) that Agent 1 acts as though he were the only screen ($\hat{s}_1^\star = 0$) while Agent 2 collapses to always hiring candidates that make it through the first screen. This leads to a significant increase in the number of low-productivity workers hired relative to the number of high-productivity workers hired and limits the payoffs to all parties. By choosing $\rho > 0 > B_2$, Agent 2 incentivizes Agent 1 to lower his chosen threshold, bringing $\hat{s}_1^\star$ more in line with $\hat{s}_2^\star$ and increasing the average productivity of workers hired.

We can also consider the optimal choice of ρ from the firm's perspective. Given the existence of some discrimination, the firm benefits from the maximum possible screening that can be offered by the two agents, which occurs where $\hat{s}_1^\star = \hat{s}_2^\star$. As such, the optimal ρ from the firm's perspective can be solved as $\rho = \frac{1}{2}((\tau_2 - \tau_1)V_X + B_2 - B_1)$. At this point, each agent has identical incentives and their chosen thresholds are identical.[27]

[27] The firm may also choose to move away from a sequential hiring process using two agents and instead adopt a model that uses test scores instead of personal judgement in at least one stage of the process (Hoffman et al., 2015).

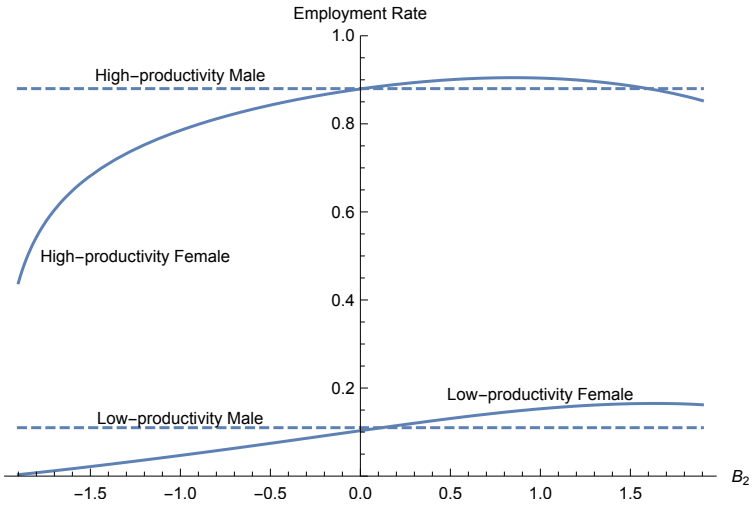
Figure A1: Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2



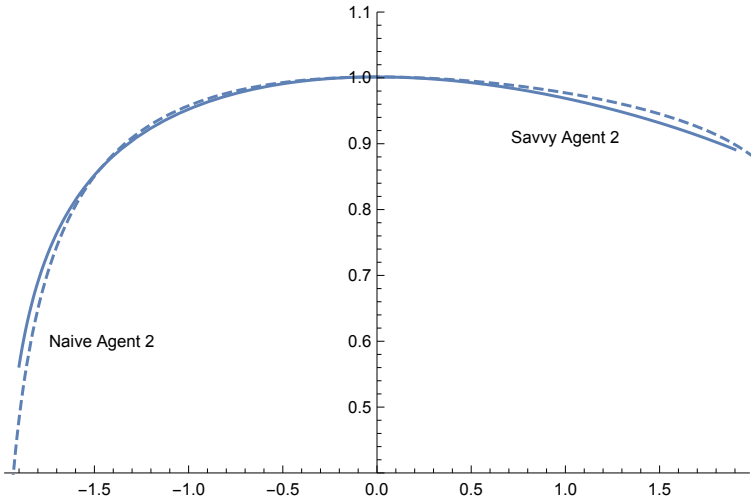
Notes: Each curve shows the optimal reservation signal of quality, above which a candidate of the "preferred" type will be advanced for further consideration by Agent 1 or hired by Agent 2. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot result across a variety of potential valuations of this non-productive trait by Agent 2. We further assume that both agents fully predict the other agent's behavior.

Figure A2: Employment probabilities and firm value with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2

Panel A: Employment probabilities



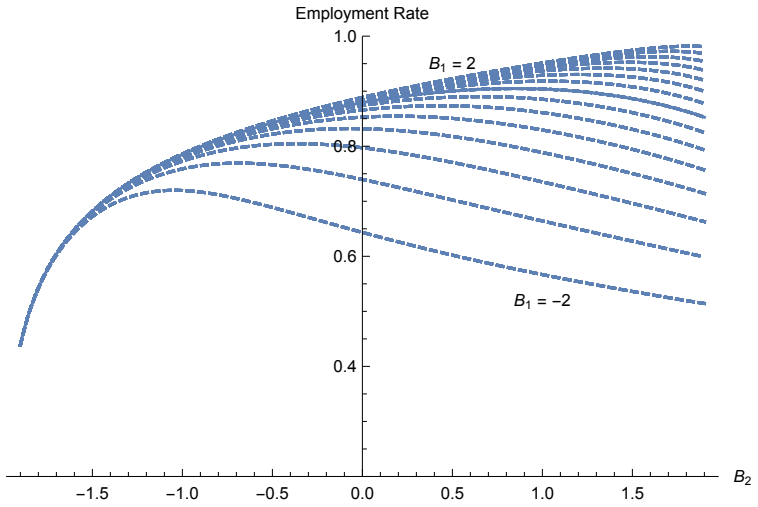
Panel B: Firm value



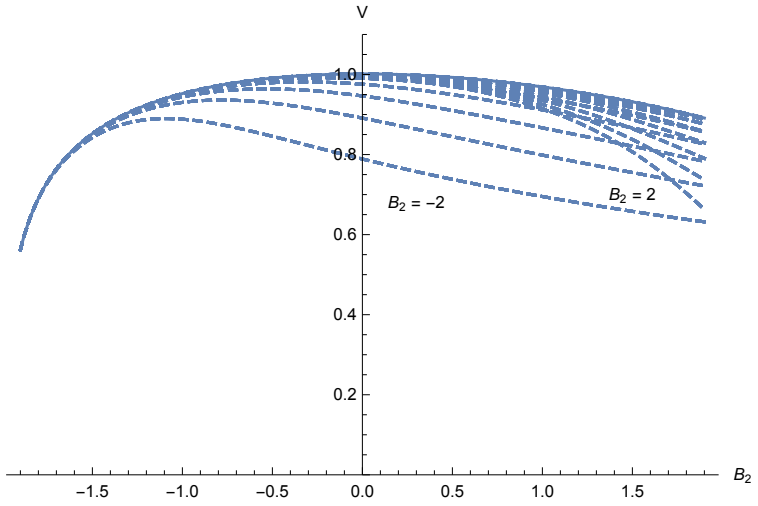
Notes: Panel A indicates the probability that candidates from the indicated group will be hired by the firm. Panel B indicates the expected value to the firm of considering a randomly chosen candidate. We assume in this case that Agent 1 places no value on non-productive traits ($B_1 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 2. We further assume that both agents fully predict the other agent's behavior.

Figure A3: Employment probabilities and firm value when Agent 2 is savvy

Panel A: Employment probabilities among "good" female candidates



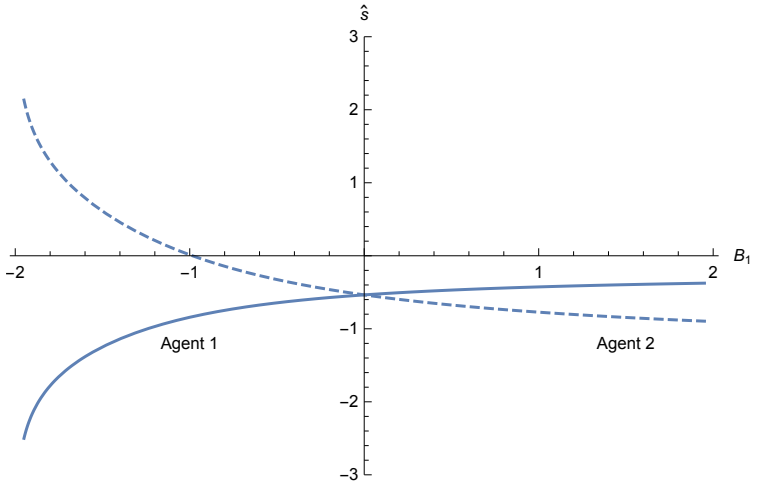
Panel B: Expected firm value in assessing a privately valued candidate



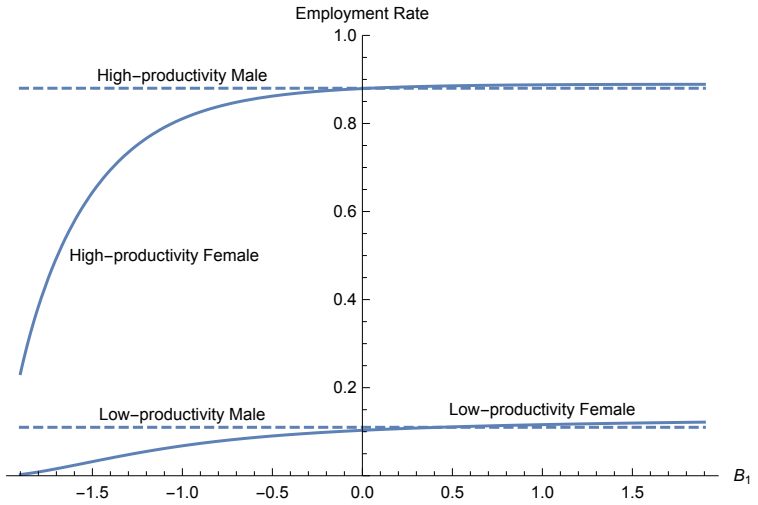
Notes: Panel A indicates the probability that good candidates with a particular non-productive trait will be hired by the firm. Panel B indicates the expected value to the firm of considering a randomly chosen candidate. We plot results across a variety of potential valuations of this non-productive trait by both Agent 1 and Agent 2. We assume that both agents fully predict the other agent's behavior.

Figure A4: Reservation signals and employment probabilities with bottom-up preferences ($B_2 = 0$, as we vary B_1) and a savvy Agent 2

Panel A: Reservation signals



Panel B: Employment probabilities



Notes: Panel A shows the optimal reservation signal of quality, above which a candidate of the "preferred" type will be advanced for further consideration by Agent 1 or hired by Agent 2. Panel B indicates the probability that candidates from the indicated group will be hired by the firm. We assume in this case that Agent 2 places no value on non-productive traits ($B_2 = 0$) and plot results across a variety of potential valuations of this non-productive trait by Agent 1. We further assume that both agents fully predict the other agent's behavior.