Diversity and the Timing of Preference in Hiring Games

Logan M. Lee and Glen R. Waddell *

March 2019

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, characterizing how one agent's interest in enhancing diversity can indirectly influence the other agent's decision. Where there is an unequal interest in diversity across the two decision makers, this can be sufficiently offsetting that even highly productive candidates who also enhance diversity are less likely to be hired. In an experimental setting, varying the interest the second agent has in a candidate's gender induces large differences in how male experimental 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. The data used in this article can be obtained by contacting Logan M. Lee. Neither author has any conflicts of interests to disclose. IRB approval for the project has been obtained.

1 Introduction

Gender and racial disparities in labor-market outcomes are often quite striking, and efforts to diversify business, political, faculty, and administrative offices are often frustratingly slow in bearing fruit. Bayer and Rouse (2016) write that 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. A number of recommendations for increasing diversity have been proposed including "Removing implicit and institutional barriers." While identifying these subtle implicit barriers is challenging, the rewards to doing so may be large. Given that even the installation of Chief Diversity Officers has 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 consider a mechanism at the intersection of incomplete information and discrimination, stemming from decision makers at different places in the hierarchy of a firm having different objectives. Specifically, we consider situations in which there are returns to the firm from enhancing diversity but these returns are not internalized by all decision makers. In short, where those differences arise has significant implications for the employment and productivity of candidates that offer diversity.

Consider, for example, a setting in which two agents of a firm participate in the sequential evaluation of a job candidate. In this setting, agents might differently anticipate benefits associated with an observable attribute of the candidate, we have in mind the individual's race or gender, for example, in a structure that nicely captures 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 consider *where* in the sequence interests remain narrowly defined around finding productive individuals and where in the sequence interests are broader, and include firm-wide considerations such as diversity. 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* diversity enhancing attributes can be to the detriment of candidate's who offer diversity. 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 diversity more than those earlier in the sequence, they have difficulty incentivizing 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, acting in sequence on behalf of a 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 has determined that the candidate is worthy of forwarding in the search. (We think of the early movers as division managers or department chairs, for example.) In that way, the process we model captures the typical "up or out" nature of job searches.[1]

In terms of actionable policy, we will speak directly to the implications of directed

[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.

searches—where valuations beyond individual employee productivities 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 high-productivity candidates from the diversity enhancing group worse off. That is, here, diversity enhancing candidates face a *lower* probability of employment, not higher.[2] For example, where leadership values female candidates, 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 disutility associated with hiring a candidate with a particular attribute. More simply, 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. Instead, the results is solely due to agents having incomplete information of candidates' true abilities and the tradeoffs being made at the margin when the early mover anticipates a favorable bias being introduced by subsequent decision makers. Thus, one might fear that policies designed to encourage the hiring of employees who increase workforce diversity can promote 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-maker is fundamental. As such, we consider variation in the relative value agents place on diversity as we consider the implications on employment and workforce productivity. As interests in diversity influence the relative probabilities with which candidates of different abilities are hired, we also discuss the distributional consequences for subsequent promotion games.

In Section 2, we introduce the model, solving the sequential consideration of agents backwards. Throughout, we consider cases where diverse candidates are directly discriminated

[2] 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).

against, although cases in which diverse candidates are favored somewhere in the hiring process we see as the more-relevant to policy, especially where we demonstrate that this *a priori* favor can be to the detriment of diversity enhancing candidates.

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 broader 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 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 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 diversity interests.

In Section 5, we provide experimental evidence that, when primed with information about the 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 subjects are asked to choose two candidates to advance for further consideration from mixed-gender groups of three. In Section 6 we offer some additional insights as we conclude.

2 Theory

2.1 The setup

We consider the implications of agents having different values or responsibilities associated with managing broader diversity interests as they undertake the hiring responsibilities for the firm.[3] In so doing, we consider a two-stage hiring game in order to speak to the implications of these 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 the placement of these interests: "bottom-up" preferences (e.g., grass roots efforts to increase racial diversity among co-workers), "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 high-productivity (H) candidate as an increase

[3] Becker (1957) first introduced an economic model in which employers had a taste for discrimination. In Becker's model, workers possessing an undesirable trait have to compensate employers by being more productive at a given wage or by being willing to accept a lower wage for equal 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." (Though our consideration is around where in a sequential hiring process diversity is valued, the math of the problem does allow for disutility.) Elements of the longer literature will also be evident in what follows as we reconsider the role of private valuations amid uncertainty around worker productivity (Arrow, 1971; Phelps, 1972; McCall, 1972; Arrow, 1973; Spence, 1973). 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).

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

in the firm's value from V_0 to V_H. With probability $(1-\alpha)$ the candidate's productivity is low (L), and upon hiring—a "bad" hire—would lower the firm's value from V_0 to V_L.[5] The firm's interest is always in rejecting the low-quality candidates, which leaves the firm's value at the status-quo level, V_0.[6]

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 $\tau_1 \leq \tau_2$.[7]

We introduce the potential for diversity to be valued by allowing for some verifiable attribute of the candidate to be valued by either or both agents. Given the sequence of actions, we notate any benefits accruing to Agent 1 from hiring the candidate as B_1, and any benefit to diversity accruing to Agent 2 as B_2.[8] To maintain interest and relevance, we will limit agents' interests to those that yield interior solutions.[9] That is, we will limit these 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.

[5] Note that α, V_H, V_L, and V_0 are not influenced by the personal attributes of a diversity enhancing candidate. When these traits are made more tangible (e.g., race, gender), it is possible that differential access to resources and a variety of other factors could cause the terms to carry a trait-specific designation. To the extent that both agents equally value this additional productivity and may still have preferences that are not related to productivity, the main results of our model are unchanged except that direct comparisons over the employment and promotion of favored and non-favored groups is more difficult.

[6] This V_0 can easily be normalized to zero, but we retain for now, thinking that the intuition is made clearer.

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

[8] We remain agnostic about the nature of these diversity benefits and rely on B_1 and B_2 to capture movement in what might constitute the relevant benefits. Some of these benefits may effect the total value of the firm (if, for example, diverse candidates increase the productivity of their coworkers) while others may only accrue to these decision makers privately.

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

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 $N(\mu_L, \sigma_L)$ if the candidate is a low-productivity type and would therefore be a "bad" hire, and from $N(\mu_H, \sigma_H)$ if the candidate is a high-productivity type and would therefore be a "good" hire, where $\mu_L < \mu_H$. $f_L(\cdot)$ is the CDF of $N(\mu_L, \sigma_L)$ and $f_H(\cdot)$ is the CDF of $N(\mu_H, \sigma_H)$.[10] 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,

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

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 game—the candidate is either an H type but not hired (Agent 2 realizes $\tau_2 V_0$), an H type and hired ($\tau_2 V_H + B_2$), an L type not hired ($\tau_2 V_0$), or an L type but hired ($\tau_2 V_L + B_2$). 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

[10] 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 personal attributes may be correlated with signal noise. Our model can easily encompass this potential by allowing σ_L and σ_H to vary with the candidate's personal attribute.

reservation signal was in the first stage, which we capture as $\mathbb{E}_2[\hat{s}_1]$.[11]

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

$$\frac{\alpha(1-f_H(\mathbb{E}_2[\hat{s}_1]))f_H(\hat{s}_2)}{(1-\alpha)(1-f_L(\mathbb{E}_2[\hat{s}_1]))f_L(\hat{s}_2)} = \frac{\tau_2 V_0 - (\tau_2 V_L + B_2)}{(\tau_2 V_H + B_2) - \tau_2 V_0}. \tag{2}$$

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., $\alpha(1-f_H(\mathbb{E}_2[\hat{s}_1]))f_H(\hat{s}_2)$, and $(1-\alpha)(1-f_L(\mathbb{E}_2[\hat{s}_1])f_L(\hat{s}_2)$, respectively) with the ratio of costs (i.e., $(\tau_2 V_H + B_2) - \tau_2 V_0$, and $\tau_2 V_0 - (\tau_2 V_L + B_2)$).

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 $N(\mu_L, \sigma_L)$ if the candidate is an L type and from $N(\mu_H, \sigma_H)$ if the candidate is an H 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,

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

[11] Agent 2's expectation of the probability a high-productivity candidate (an H type) cleared Agent 1's reservation is therefore $1 - f_H(\mathbb{E}_2[\hat{s}_1])$, while the expectation of the probability a low-productivity candidate (an L type) cleared Agent 1's reservation signal is $1 - f_L(\mathbb{E}_2[\hat{s}_1])$.

where we capture in B_1 any value Agent 1 associates with the candidate's personal attribute. In general, Agent 1 chooses $\hat{s}_1$ subject to the first-order condition,

$$\frac{\alpha f_H(\hat{s}_1)(1-f_H(R_2))+\alpha(1-f_H(\hat{s}_1))f_H(R_2)(\partial R_2/\partial \hat{s}_1)}{(1-\alpha)f_L(\hat{s}_1)(1-f_L(R_2))+(1-\alpha)(1-f_L(\hat{s}_1))f_L(R_2)(\partial R_2/\partial \hat{s}_1)} = \frac{\tau_1 V_0-(\tau_1 V_L+B_1)}{(\tau_1 V_H+B_1)-\tau_1 V_0}. \tag{4}$$

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.[12]

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 $B_1 = 0$). That is, we assume that Agent 1 is interested only in the individual productivity of the candidate while Agent 2 has the objective of more-broadly increasing the representation of a race or gender (i.e., $B_2 > 0$). 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 value to personal attributes (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

$$\frac{\alpha(1-f_H(\hat{s}_1^\star|_{B_2=0}))f_H(\hat{s}_2)}{(1-\alpha)(1-f_L(\hat{s}_1^\star|_{B_2=0}))f_L(\hat{s}_2)} = \frac{\tau_2 V_0-(\tau_2 V_L+B_2)}{(\tau_2 V_H+B_2)-\tau_2 V_0}, \tag{5}$$

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's interests do

[12] This is easy to see in the symmetric case (i.e., $V_L = -V_H$, $V_0 = 0$, and $\alpha = 0.5$), as Agent 2's first-order condition collapses to $f_H(\hat{s}_2) = f_L(\hat{s}_2)$.

not include the broader diversity interests of the firm (i.e., $B_1 = 0$), τ_1 drops from the agent's problem, and Agent 1's first-order condition in (4) simplifies to

$$\frac{\alpha f_H(\hat{s}_1)(1 - f_H(R_2(\hat{s}_1|_{B_2=0})))}{(1-\alpha) f_L(\hat{s}_1)(1 - f_L(R_2(\hat{s}_1|_{B_2=0})))} = \frac{V_0 - V_L}{V_H - V_0}. \tag{6}$$

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 $\tau_2 V_L$ (where the disutility to Agent 2 of increasing diversity always dominates the value of hiring a high-productivity employee, and the problem collapses on "always reject") and $\tau_2 V_H$ (where the utility to Agent 2 of increasing diversity always dominates the cost of hiring a low-productivity employee, and the problem collapses on "always hire"). For illustrative purposes, we impose *ex ante* symmetry, and abstract away (for now) from the role of incentive pay in agent behavior by setting $\tau_1 = \tau_2 = 0.5$.[13,14]

Where B_2 decreases from zero, and hiring the candidate is costly for 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 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 an H type) the perspective of Agent 2 is that the costs of diversity must be offset by the higher probability that the candidate is an H type. That Agent 2 is motivated by this private value is clearly costly to the firm. As B_2 *increases* from zero, Agent 2 chooses a lower reservation signal in an attempt to increase the probability that the diversity-enhancing 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 an L

[13] Symmetry is defined as $V_L = -V_H$, $V_0 = 0$, and $\alpha = 0.5$. Collectively, the first-order condition for the choice of $\hat{s}_2$ is clear, as $f_H = f_L$ in equilibrium. In characterizing agent behavior, we adopt that $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$.

[14] As changes in τ_1 and τ_2 determine the relative weights the diversity interests 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.

type).[15]

The shape of Agent 1's choice of $\hat{s}_1^\star$ across B_2 is where we first observe the behavior of consequence. 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 more certainty that the candidate is an H type, and therefore a good hire, before forwarding the candidate to Agent 2. This occurs because Agent 1 anticipates that Agent 2 will hire the candidate even if the quality signal Agent 2 receives is relatively low.

Proposition 1. *With top-down preferences, for any $|B_2| > 0$ Agent 1's choice of reservation signal acts as a weakly offsetting force. That is, Agent 1's mitigates the favorable treatment diverse candidates would otherwise receive.*

Moreover, as B_2 approaches $\tau_2 V_H$ and Agent 2's decision rule collapses to the unproductive act of "always accepting" a candidate who increases diversity, Agent 1's decision rule collapses to that which would be chosen by a single decision maker. In effect, Agent 1's efforts to mitigate Agent 2's interest are sufficient to completely dissipate the gains provided to the firm from having the second signal of the candidate's uncertain productivity.[16]

Agent 1's behavior in cases where $B_2 > 0$ has two interesting implications. First, we note that African-American-sounding names receiving fewer call backs in resume studies (Bertrand and Mullainathan, 2004), while obviously evidencing a certain discrimination against applicants thought to be African American, is also consistent with the callback decision resting with someone early in a sequence of decisions who is herself unbiased, though anticipates subsequent decision makers showing preference *for* African-American applicants.

[15] Figure 1 also reveals two interesting limiting cases, in $B_2 = \tau_2 V_L$ and $B_2 = \tau_2 V_H$, where Agent 2's decision rule collapses on either "never hire" or "always hire." Again, this is in keeping with expectations. Where $B_2 = \tau_2 V_L$, any private cost associated with diversity is sufficiently high that there is no possible outcome (i.e., even $\tau_2 V_H$ is not sufficiently high) that would dominate the status quo of $\tau_2 V_0$ net of B_2. Likewise, where $B_2 = \tau_2 V_H$, the benefit to diversity is sufficiently high that there is no possible outcome (i.e., even $\tau_2 V_L$ is not sufficiently low) that would dominate the potential that an L type is hired and the firm realizes a value of $\tau_2 V_L$.

[16] 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 $\hat{s}_1^* \leq 0$).

This, of course, is a subtle but very real distinction when it comes to the design of policy. Second, by raising the minimum threshold for advancement for candidates that offer diversity, Agent 1 increases the average productivity of diverse candidates seen by Agent 2, which is likely to influence Agent 2's perception of the average quality of diverse candidates. In a larger game, one could consider the propagation of this misinformation and its effect on subsequent decision making.

This influence of Agent 1 is not symmetric around $B_2 = 0$. As $B_2 < 0$ approaches $\tau_2 V_L$, Agent 2 always rejects the candidate and Agent 1's decision is of no consequence to eventual outcomes. In the limit, the firm unavoidably suffers the costs of Agent 2's bias. 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 average employee productivity

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 work, for expositional purposes we plot the relative treatments of male and female candidates, with the diversity interest (B_2 in this case) capturing Agent 2's perception of the value associated with hiring a female candidate, all else equal. Clearly, with productivity uncertain, and no offsetting influence of Agent 1, increases in B_2 from zero increase the probability a female is hired. On the other hand, the rate at which low low-productivity female candidates are hired increases faster than the probability high-productivity female candidates are hired. The differential rates imply that the average productivity of female employees is falling in B_2. Likewise, as B_2 decreases from zero (and Agent 2 sees disutility in the hiring of diversity enhancing employees) 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 reduces the average productivity of female employees.

In Panel B of Figure 2, we re-introduce Agent 1. Relative to Agent 2 acting alone, the offsetting 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 $B_2 = \tau_2 V_L$. As B_2 approaches $\tau_2 V_H$, high-productivity diverse candidates can be strictly worse off than they would be without preference.

Proposition 2. *With top-down preferences, employment rates are strictly increasing in B_2 among low-ability candidates who also offer the firm the ability to increase diversity.*

Proposition 3. *Employment rates among high-productivity candidates who also offer the firm the ability to increase diversity are not monotonic in B_2. In a sequential-hiring game, there exists some small $B_2 < 0$ for which the high-productivity candidate who also offers the firm the ability to increase diversity is strictly better off than she would be under a regime in which B_2 is large and positive.*

In essence, the early decision maker has enough influence on the candidate's prospects that the high-productivity candidate who also offers the firm the ability to increase diversity would prefer even mild discrimination in later rounds over having agents in later rounds offer strong favoritism. This effect becomes more pronounced as the fraction (α) of highly productive employees falls. With fewer H types 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, the benefits to having multiple signals of productivity increase for both high productivity candidates and the firm.

To the extent that Agent 2 represents the diversity interests of the firm as a whole, firms are best served by "blinding" Agent 1 to the candidates' diverse attributes. This mechanism

prevents Agent 1 from offsetting the perceived favoritism of Agent 2 while still allowing Agent 2 to consider the value of the diversity each candidate offers. Interestingly, where $B_2 > 0$, this is also optimal from the candidate's perspective. Given the choice, if $B_2 > 0$ diverse candidates will choose not to reveal their diversity during the initial screen by Agent 1, benefiting from the revelation that they offer diversity subsequent to Agent 1 and prior to Agent 2. The overlap of optimal procedure from the perspective of both the firm and diverse candidates suggests this may be a promising policy opportunity.

In Figure 3, we plot the average productivity of hired diverse candidates, with and without the influence of Agent 1.[17] Not surprisingly, the screen provided by Agent 1 increases average employee productivity, across all $B_2 > V_L$. Given the uncertainty of a candidate's productivity, Agent 1's screen simply enables the hiring of highly productive candidates—"good" hires—with higher probability.

However, more interesting is the asymmetry introduced into the expected outcomes. In the absence of Agent 1, reductions in average productivity 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 he is able to offset Agent 2's inclination to *hire* female candidates (when $B_2 > 0$).[18]

Proposition 4. *Due to the ability of Agent 1 to unilaterally reject candidates, average productivity among hired diverse candidates will fall more with top-down discrimination (i.e., $B_2 < 0$) than with top-down favoritism (i.e., $B_2 > 0$).*

3.3 Subsequent promotion games

As $B_2 \neq 0$ induces patterns of hiring that are specific to diverse candidates, in any subsequent period, average (within-firm) productivity levels will vary by the diverse trait. Even in the

[17] We normalize to one average productivity when Agent 2 is naive and there are no private values, $B_1 = B_2 = 0$.

[18] 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 $V_0 = 0$.

absence of diversity considerations playing a direct role in promotion decisions, promotion outcomes can be shown to depend on B_2. For example, if female is the diverse trait and $B_2 > 0$ 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.[19] While the implication of heterogeneous productivity in promotion games has been considered in the literature (Bjerk, 2008), we offer an original source of heterogeneity driven, somewhat surprisingly, by a desire to increase firm diversity.

3.4 The role of Agent 1's diversity interest

Thus far in the analysis, we have assumed that $B_1 = 0$. In Figure 4, we allow for $B_1 \neq 0$ and $B_2 \neq 0$, allowing both agents to have internalized the value on diversity. 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 $B_1 \in (\tau_1 V_L, \tau_1 V_H)$. For now, we continue to assume that Agent 2 is naive, which implies 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. In Panel A of Figure 4, we document the expected pattern of behavior—for any $B_2 \in (\tau_2 V_L, \tau_2 V_H)$, $\hat{s}_1$ is strictly decreasing in B_1. As Agent 1's interest in promoting diversity increases (holding constant Agent 2's), Agent 1 is less likely to reject those candidates who enhance diversity. The less-obvious takeaway from Panel A we state as a proposition:

Proposition 5. *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 interest in diversity increases.*

This highlights the complexity of the sequential decision—even when Agent 1 values diversity, his response to Agent 2 valuing diversity more is to raise the threshold imposed on candidates offering opportunities for enhancing diversity.

[19] Of course, if the potential promotion of females 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.

In Panels B and C of Figure 4 we plot the *ex post* rates of employment for high- and low-productivity female candidates, assuming that female is the personal attribute around which the agents are potentially optimizing. As in Panel B of Figure 2, Panels B and C of Figure 4 again capture that employment outcomes are sensitive to B_2, not only as a direct result of Agent 2's interest in diversity, but also indirectly through Agent 1's best response to $B_2 \neq 0$. Namely, employment rates among high-productivity 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 high-productivity female candidates again monotonically increase in B_2. Figure 4 also demonstrates an important implication of Agent 2's naiveté. Both high- and low-productivity candidates who offer the firm the ability to increase diversity 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 role.

4 When Agent 2 is savvy

The lack of a response from Agent 2 where $B_1 \neq 0$ suggests that it is worthwhile to relax the assumed naiveté of Agent 2 and allow her to correctly anticipate (an fully respond to) $\hat{s}_1$. In this section we 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.

4.1 Agent behavior

In Figure 5, we return to consider "top-down" preferences (i.e., $B_1 = 0$) across a range of $B_2 \in (\tau_2 V_L, \tau_2 V_H)$, 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 there are no diversity considerations.) 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 than was the case with naiveté, in Figure 1. In particular, Agent 1's reservation signal is no longer monotonically increasing through $B_2 \in (\tau_2 V_L, \tau_2 V_H)$. Instead, $\hat{s}_1^\star$ 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 $B_2 \in (\tau_2 V_L, 0)$. (As when Agent 2 is naive, when Agent 2 is savvy $\hat{s}_1^\star$ is monotonically increasing in $B_2 \in (0, \tau_2 V_H)$.)*

The intuition for this result is again found in Agent 1's inability to fully offset discrimination 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 lower $\hat{s}_1$, he responds by increasing $\hat{s}_2^\star$ all the more, which ultimately decreases employment rates among those presenting the diversity enhancing 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 diversity interests 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 $\hat{s}_2^\star$, thereby allowing the firm to better exploit the gains available through the second signal of productivity. We learn by this that prejudicial interests introduced late in a sequential-hiring game can motivate what looks like prejudicial interests 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. Interestingly, Agent 1's behavior implies that the average productivity of female candidates will be higher coming out of early stages, potentially moving subsequent priors away from "reject" and toward "accept."

4.2 Implications for employment and firm value

In Panel A of Figure 6, 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 6 we plot the average productivity of hired diverse candidates for the savvy and naive cases. While average productivity is invariant to the assumption of naiveté when $B_2 = 0$, slight differences emerge at other values of B_2. In general, productivity falls more from Agent 2's diversity interests when Agent 2 is savvy; Agent 1 offers a less-offsetting influence in such cases. The exception to this rule is for extreme discrimination (i.e., B_2 approaching V_L), where Agent 1's higher standard enables the firm to escape Agent 2's "always reject" regime.

4.3 The role of Agent 1's private value

In Panel A of Figure 7, 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 interest.) The bold line captures the parameterization already represented in Figure 6. 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 $\tau_2 V_L$. As B_2 increases from $\tau_2 V_L$, 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 $\tau_1 V_H$), employment is still very much dependent on Agent 2's interest in diversity (B_2).

In Panel B of Figure 7 we plot the average productivity of hired female candidates. That the expected value is highest when $B_1 = B_2 = 0$ again reflects that any diversity related interest, in either agent, causes a less efficient evaluation process and increases the probability that a low-productivity female candidate more than the increase in probability of hiring a high-productivity female candidate. 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 solely attuned to the individual hiring decision, and not directly motivated by broader concerns for diversity.

The timing of preference—whether introduced with Agent 1 or Agent 2—yields striking differences in agents' optimal thresholds. In Figure 8, 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 5), where the agent without diversity related preference appears to "buy" more-lenient treatment from the agent who finds the

candidate's personal attribute privately costly.

The importance of the timing of bias is also seen in Panel B of Figure 8, where we plot associated employment probabilities by productivity. With discrimination, the timing of the introduction of an awareness of a candidate's personal attributes is of little consequence to employment; either agent can unilaterally dismiss candidates. As no single agent can unilaterally hire a candidate, preference over a candidate's personal attribute yields different patterns of behavior. With bottom-up preferences, both high- and low-productivity 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 6), where strong preference on the part of Agent 2 ultimately leaves highly productive female candidates less likely to be hired.

4.4 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 response to $B_2 \neq 0$, it is interesting to consider the potential for a transfer, from Agent 2 to Agent 1, to incentivize Agent 1's cooperation.[20]

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 diversity enhancing personal attribute. We ask, then, whether there are any $\{B_1, B_2\}$ for which Agent 2 will choose to reward Agent 1 for hiring such a candidate.[21]

Such practice appears in academic markets, for example, where payments would typically

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

[21] 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 personal 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."

be made by college-level administrators to departments conditional on hiring a candidate who presents with a desirable personal attribute, such as a minority race or 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,

$$\begin{aligned}
\underset{\hat{s}_2,\rho}{\text{Max}}\, V_2(\hat{s}_2) \;=\;& \alpha[f_H(\mathbb{E}_2[\hat{s}_1]) + (1 - f_H(\mathbb{E}_2[\hat{s}_1]))f_H(\hat{s}_2)]\tau_2 V_0 \\
&+\alpha(1 - f_H(\mathbb{E}_2[\hat{s}_1])(1 - f_H(\hat{s}_2))(\tau_2(V_H - \rho) + B_2) \\
&+(1-\alpha)[f_L(\mathbb{E}_2[\hat{s}_1]) + (1 - f_L(\mathbb{E}_2[\hat{s}_1])f_L(\hat{s}_2)]\tau_2 V_0 \\
&+(1-\alpha)(1 - f_L(\mathbb{E}_2[\hat{s}_1]))(1 - f_L(\hat{s}_2))(\tau_2(V_L - \rho) + B_2),
\end{aligned} \tag{7}$$

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,

$$\begin{aligned}
\underset{\hat{s}_1}{\text{Max}}\, V_1(\hat{s}_1) \;=\;& \alpha[f_H(\hat{s}_1) + (1 - f_H(\hat{s}_1))f_H(R_2)]\tau_1 V_0 \\
&+\alpha(1 - f_H(\hat{s}_1))(1 - f_H(R_2))(\tau_1(V_H - \rho) + B_1 + \rho) \\
&+(1-\alpha)[f_L(\hat{s}_1) + (1 - f_L(\hat{s}_1))f_L(R_2)]\tau_1 V_0 \\
&+(1-\alpha)(1 - f_L(\hat{s}_1))(1 - f_L(R_2))(\tau_1(V_L - \rho) + B_1 + \rho).
\end{aligned} \tag{8}$$

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 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 $B_2 \to \tau_2 V_H$, we have shown (in Figure 5) 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 employees hired relative to the number of high-productivity employees 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 employees 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.[22]

5 Empirics

Here, we design experimental conditions to consider the fundamental empirical question around which the above results rest—will individuals in the first stage of a hiring game behave differently if they anticipate that a subsequent decision-maker will value a personal attribute in the candidate 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.

[22] 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).

5.1 Design

In an effort to make the experimental setting less cumbersome, we will move away from relatively abstract theory and toward simple, discrete choices. 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 diversity considerations, subjects would need 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-I and type-II errors. Instead, we focus on relatively simple comparisons—given three possible candidates, which two should be advanced? In the absence of diversity considerations, 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 diversity considerations are introduced into Agent 2's objective function, 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 evaluator's 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.[23] This simplification also implies a fairly straightforward empirical analysis of subject choices.

All subjects were playing the role of the first agent in a two-stage hiring process. They evaluated thirty sets of three candidates and in each set 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."[24] Our fundamental experimental variation comes from the randomly assigned setting. Some subjects were told that the 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 was not said to receive a bonus, the computer simply chose the candidate with the highest SAT score. When the second person was said to receive a \$1 bonus for hiring a female, the computer chose the highest-ranking female, unless the highest-ranking male has an SAT score more than S points higher, where S was drawn randomly (once for each experimental subject) from $\{0, 40, 80, 160\}$.[25] Including zero allows us to observe any effect of the preference frame itself, separately from the degree to which favor may then be applied. Though subjects do

[23] 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.

[24] In Figure 9 we replicate the experimental instructions, as seen by subjects, and an example of the sets of three candidates they were asked to evaluate.

[25] 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 SAT score.

not observe the degree of preference directly, they can discern this over multiple scenarios as we provided feedback on the subsequent decision (i.e., which one of the two candidates they forwarded was then chosen).

With a three-choose-two design, there are eight configurations of gender and ordinal (SAT) rank. Clearly, in both control and treatment arms, if a subject faces all female or all male triples, there is no role for gender in the decision and subjects would be expected to advance the top-two candidates. As such, we spend fewer resources on increasing precision in this dimension. Likewise, where subjects are not anticipating that there will be a preference for female in the subsequent decision, all eight combinations of three candidates should similarly yield identical results. As subjects realize the highest expected return when they advance the candidates with the two-highest SAT scores, we expect that they will advance the top two regardless of gender.

Where we do have interest in increasing precision is around the identifying variation available within the eight possible configurations. Namely, we are most interested in the experimental variation across decisions made when facing a "Male-Female-Male" ranking of candidates, having the potential to evidence a systematic increase (among treated subjects) in the advance of a third-ranking male over a second-ranking female. For example, if a subject was asked to choose two from among Dave (SAT 1200), Emily (SAT 1150), and Adam (SAT 1100), theory would have to explain the advancement of Adam over Emily.[26] Of the thirty sets of three candidates, twenty followed the Male-Female-Male pattern.[27] All

[26] The model's prediction could also evidence in a second version of this decision—in Male-Female-Female configurations—where a subject could choose to advance the male and bottom-female candidates hoping that the score-gap between the top male and bottom female candidates would be sufficient to dissuade Agent 2 from choosing the female candidate.

[27] The remaining ten are Male-Male-Female (2), Male-Female-Female (1), Female-Male-Male (2), Female-Male-Female (3), and Female-Female-Male (2). In the treatment arm, when the top-two candidates are both male (i.e., Male-Male-Female), subjects can simply advance the top-two candidates and avoid any exercised preference from Agent 2 altogether. In all cases where the top candidate is female (Female-Female-Male, Female-Male-Female, Female-Male-Male), the preference of Agent 2 should only serve to reinforce the selection of the candidate the subject believes to be the best. Cases in which the highest SAT is a female do raise the interesting theoretical possibility that subjects could choose to advance candidates with lower SAT scores to "punish" the anticipated bias of Agent 2. This potential is clearest in the Female-Male-Male cases, where subjects could advance the bottom two males and ensure that Agent 2 did not earn the $1 reward. In practice we observed 0 instances of this behavior.

subjects see the thirty sets in a random order with the exception of the first-two sets, which were always Male-Female-Male and acted to inform subjects of the number of SAT points Agent 2 would be willing to give up in order to advance a lower-scoring female candidate.

5.2 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 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 ($p < .05$) assignment to treatment, with subjects self reporting higher GPA's slightly more likely to be assigned to the treatment group.

In our empirical analysis, we 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.[28]

5.3 Analysis and results

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

$$\mathbb{1}(\text{AdvancedTopTwo})_{iq} = \beta_0 + \beta_1 \mathbb{1}(\text{FemalePreference}_i) + \gamma_q + \epsilon_{iq} \ , \tag{9}$$

where $\mathbb{1}(\text{AdvancedTopTwo})_{iq}$ equals one if subject i advances the two candidates with the highest SAT scores from among the three candidates randomly observed in question q.[29] We capture any level difference 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, 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. Our model thus estimates differences in the choices of subjects within a given question (i.e., same gender composition, names, and SAT scores).based on whether they had told that the subsequent decision would be made by someone interested in hiring a female candidate. 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

[28] 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 ($\beta = 0.00006, p = 0.882$).

[29] Recall, most sets of candidates (20 out of 30) included a top male candidate, second-best female candidate, and third-best male candidate. In these scenarios, advancing the top and bottom candidates (by far the most-common deviation from a top-two strategy) amounted to intentionally avoiding the advancement of a female candidate.

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 behavior around questions in which a female was among the top two candidates. That is, 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.[30]

In Column (3), we again see the sort of systematic variation that theory implies one would. By allowing separate parameters for first and second ranking males and females, respectively, the patterns evident in earlier specifications are clearly driven by second-ranked female candidates. 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. It is in these specific opportunities that treated subjects are 16.6-percent (14.8 pp) less likely to choose the top-two candidates. Moreover, there is no remaining level difference associated with the positions of either candidate alone. In the sets in which a second-ranking female may jeopardize the first-ranked male, our experimental subjects shut down the female's advancement.

[30] 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.

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 ($p < .001$).[31]

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 significantly influencing which candidates are advanced in the expected direction. 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 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 likely to advance the top two. Again, we find no evidence that female experimental subjects are following similar patterns.

There are at-least three possible explanations for the gender divide. First, female subjects

[31] This result also suggests that subjects were unwilling to risk a third best female candidate being hired in an effort to expand the score gap between the top male candidate and the other option advanced to Agent 2.

may be more prosocial and willing to allow Agent 2 to earn the bonus dollar when available. Second, to the extent that subjects interpreted the sequential-hiring process as implying a hierarchy, female subjects may have been less willing to explicitly counteract the preferences of the Agent 2. Finally, female subjects may have had a personal preferences for candidates with whom they shared a gender, leading them them to be less willing to offset anticipated bias.[32]

6 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 having diversity enhancing interests as they adjudicate candidates. We show that the introduction of these interests in one stage of such a game are evident not only in the actions of the agent with those 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 can partially offset this preference by raising the standard they impose on candidates with that attribute. In the typical "up-or-out" hiring environment, where earlier decision makers have much more sway in *rejecting* candidates than in *hiring* candidates, the response among those who anticipate that subsequent treatment will be favorable still has the potential to subject candidates who are preferred, on average, to lower odds of employment than they would have experienced had their personal attributes not been valued or observable.

We note four interesting implications, each of which may motivate additional exploration. First, the model offers a new explanation for existing evidence that resumes with African-American-sounding names receive fewer call backs (Bertrand and Mullainathan,

[32] 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. 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.

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 sequential decision responding to subsequent decision makers showing preference *for* African-American candidates. Of course, policy prescriptions across these potential mechanisms will differ significantly.

Second, note that the model we present implies that if preferences for the personal 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 value those personal 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 the adjudication by 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-diversity leadership will meet with more success by incorporating blind-recruitment tools in early assessments of job candidates.

Third, compared to a single agent acting alone, where diversity interests tradeoff with productivity, when the decision is made by two agents in sequence, average productivity falls off less. Moreover, while fewer female candidates advance in the sequence, the average productivity of those who do advance for final consideration is higher. This may leave later decision makers increasingly misinformed of the underlying distribution of female productivity, thereby reinforcing or strengthening prior beliefs among those in leadership positions.

Finally, as both agents must approve a candidate for hire, while rejection can occur with either agent's unilateral decision, there is a fundamental asymmetry in this offset—it more effectively offsets pro-diversity interests than it does offset pro discrimination interests, for example. This suggests that efforts to increase diversity through hiring may be slower, for

example, than would efforts to limit that diversity be. Especially so when those interests are late arriving.

In an experimental setting, we test the fundamental empirical question on which the theory rests—whether subjects act to offset the anticipated preference for diversity among other decision makers. We vary the conditions under which groups of three candidates are evaluated 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 male or female name), and must decide which two of the three they wish to forward for further consideration. We vary whether experimental subjects are informed that the subsequent consideration would be done by someone with the incentive to hire females. Without knowledge of that 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 experimental subjects to protect the best male candidates by terminating the candidacy of 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.

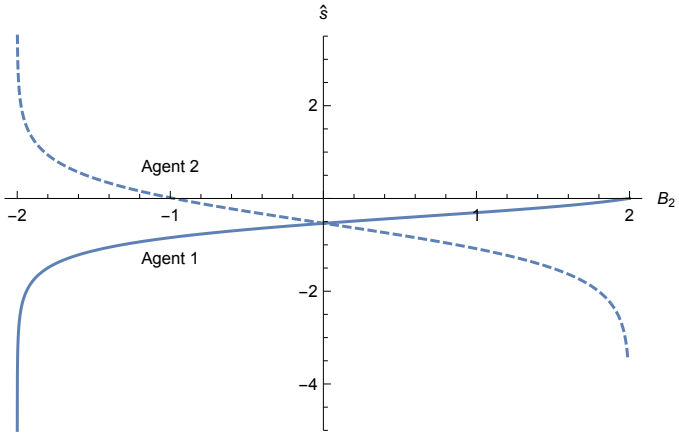
Pinkston, Joshua C, "A Test of Screening Discrimination with Employer Learning," *Industrial & Labor Relations Review*, 2005, *59*, 267.

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.

7 Figures

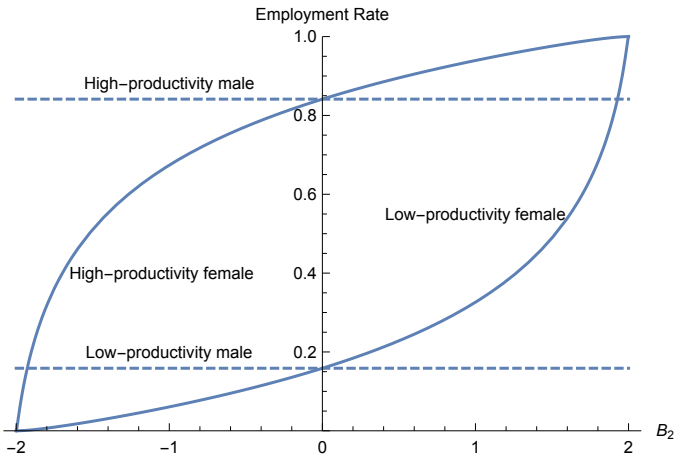
Figure 1: Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2)



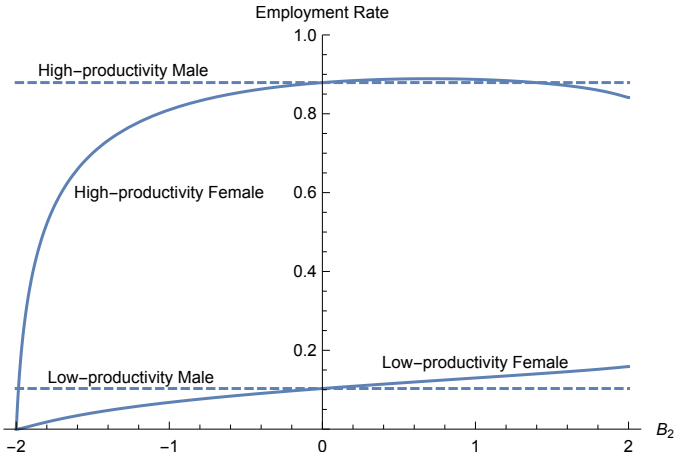
Notes: Each curve shows the optimal reservation signal of quality, above which a diversity enhancing candidate will be advanced for further consideration by Agent 1 or hired by Agent 2. We assume in this case that Agent 1 places no value on diversity enhancing attributes ($B_1 = 0$) and plot results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

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

Panel A: No screening provided by Agent 1

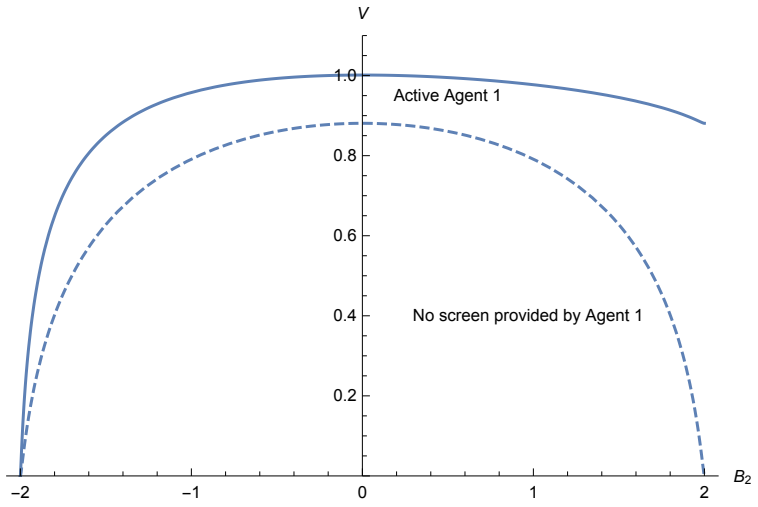


Panel B: Agent 1 screens candidates prior to Agent 2



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 diversity enhancing attributes ($B_1 = 0$) and plot results across a variety of potential valuations of this personal attribute 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.

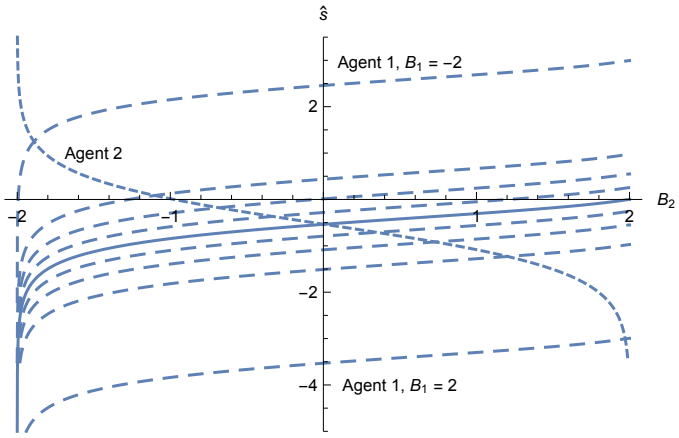
Figure 3: Average employee productivity with top-down preferences ($B_1 = 0$, as we vary B_2)



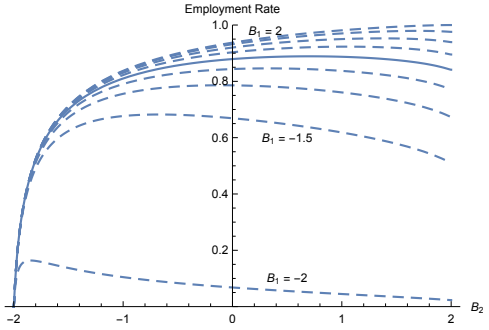
Notes: Each curve indicates the expected value to the firm when considering a randomly chosen candidate. We assume in this case that Agent 1 places places no value on diversity enhancing attributes ($B_1 = 0$) and plot results across a variety of potential valuations of this personal attribute by Agent 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 B_1 and B_2

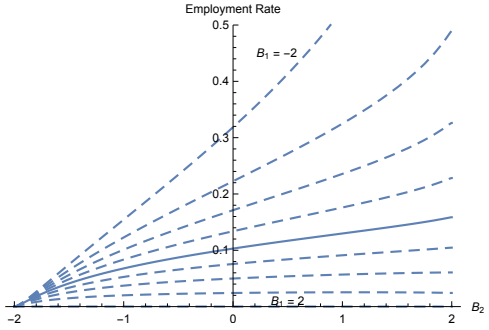
Panel A: Reservation signals



Panel B: Employment of high-productivity diversity-enhancing candidates

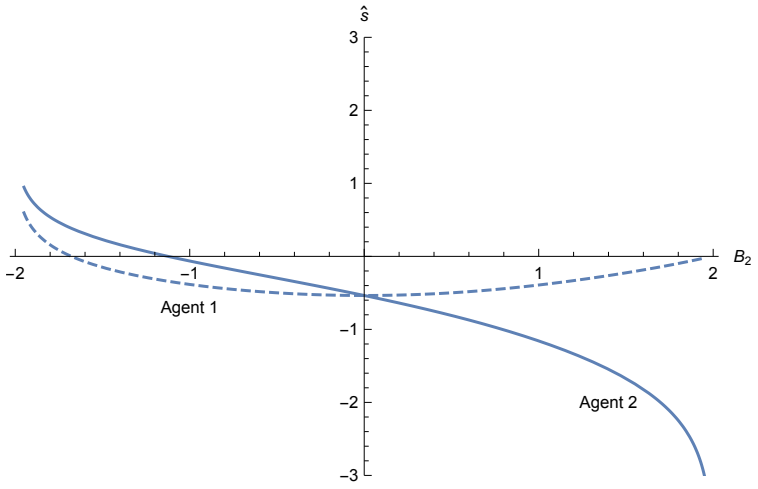


Panel C: Employment of low-productivity diversity-enhancing candidates



Notes: Each curve in Panel A shows the optimal reservation signal of quality, above which a diversity enhancing candidate will be advanced for further consideration by Agent 1 or hired by Agent 2. We plot results across a variety of potential valuations of a diversity enhancing attribute, for 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. Panel B displays employment rates for H types and Panel C displays employment rates for L types hires.

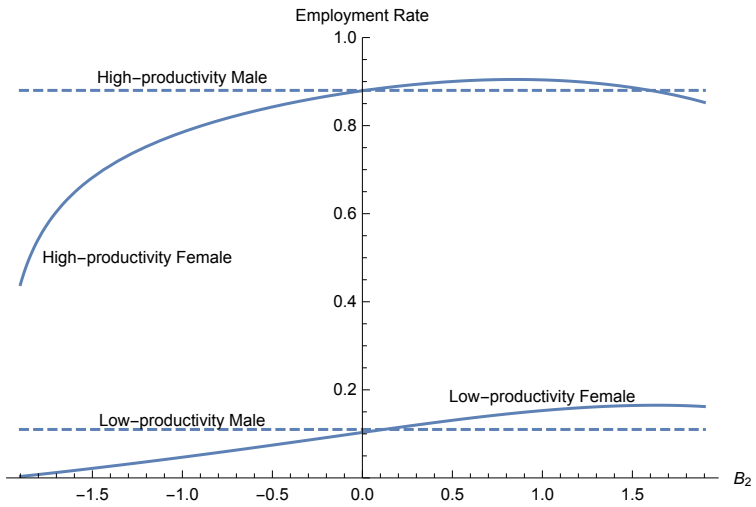
Figure 5: Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2



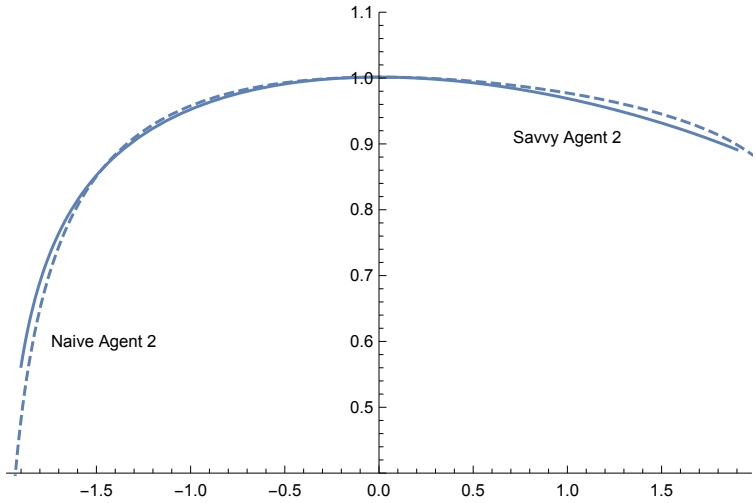
Notes: Each curve shows the optimal reservation signal of quality, above which a candidate who can enhance diversity will be advanced for further consideration by Agent 1 or hired by Agent 2. We assume in this case that Agent 1 places no value on diversity enhancing attributes ($B_1 = 0$) and plot result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

Figure 6: Employment probabilities and employee productivity with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2

Panel A: Employment probabilities



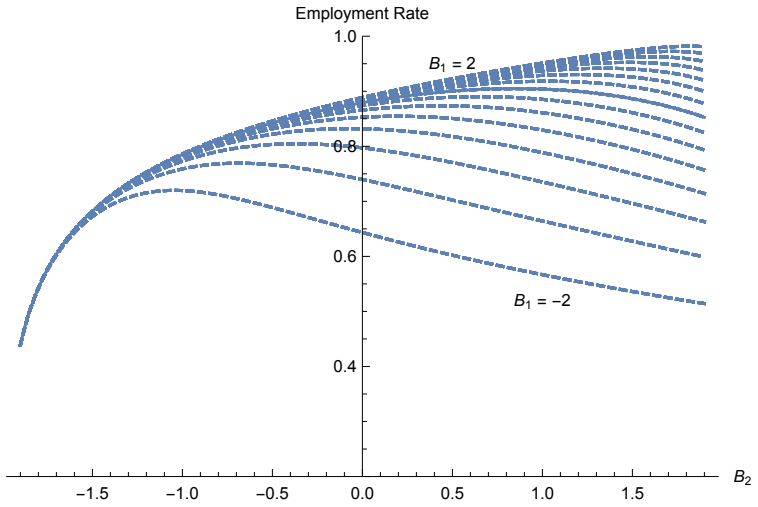
Panel B: Firm value



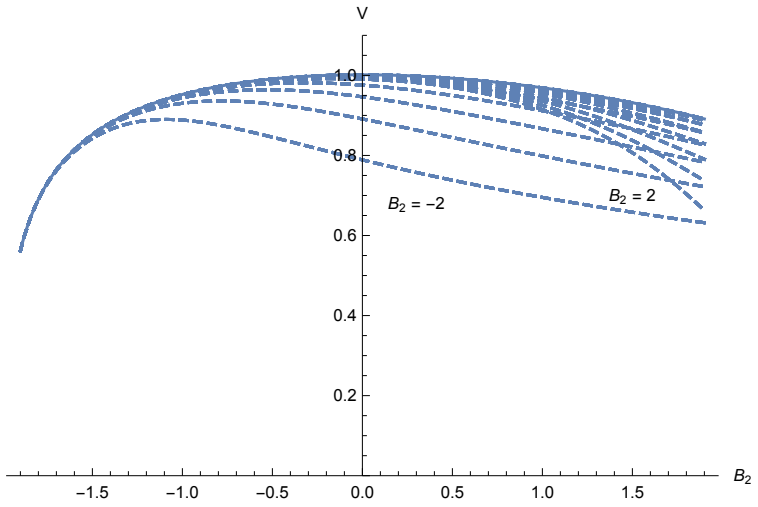
Notes: Panel A indicates the probability that candidates from the indicated group will be hired by the firm. Panel B indicates the expected value to the firm from considering a randomly chosen candidate. We assume in this case that Agent 1 places no value on diversity enhancing attributes ($B_1 = 0$)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

Figure 7: Employment probabilities and employee productivity when Agent 2 is savvy

Panel A: Employment probabilities among highly productive female candidates



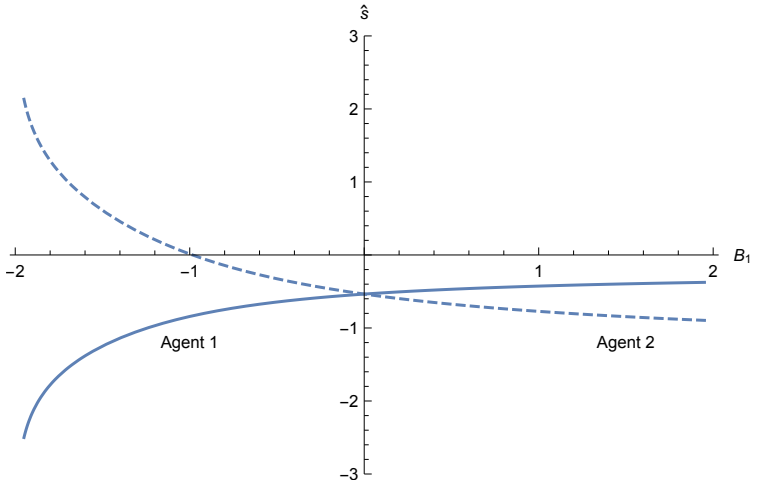
Panel B: Average employee productivity across the internalized value of diversity



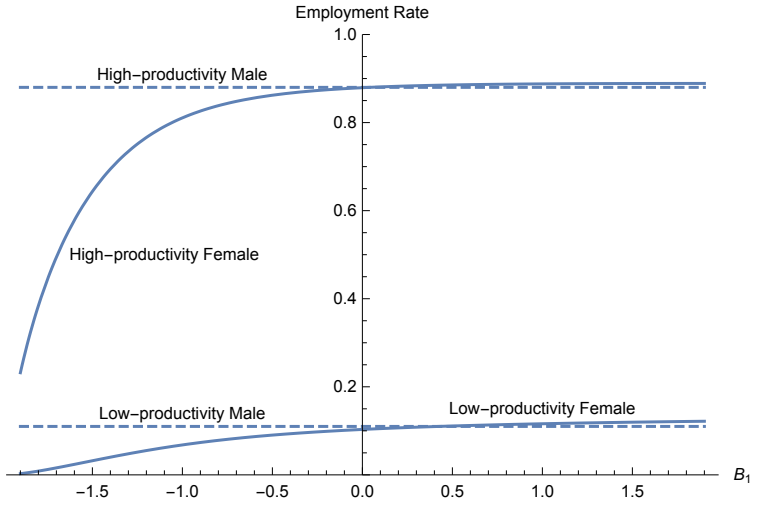
Notes: Panel A indicates the probability that an H type with a particular diversity enhancing attributes will be hired by the firm. Panel B indicates the expected value to the firm from considering a randomly chosen candidate. We plot results across a variety of potential valuations of this diversity enhancing attributes, by both Agent 1 and Agent 2. We assume that both agents fully predict the other agent's behavior.

Figure 8: 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



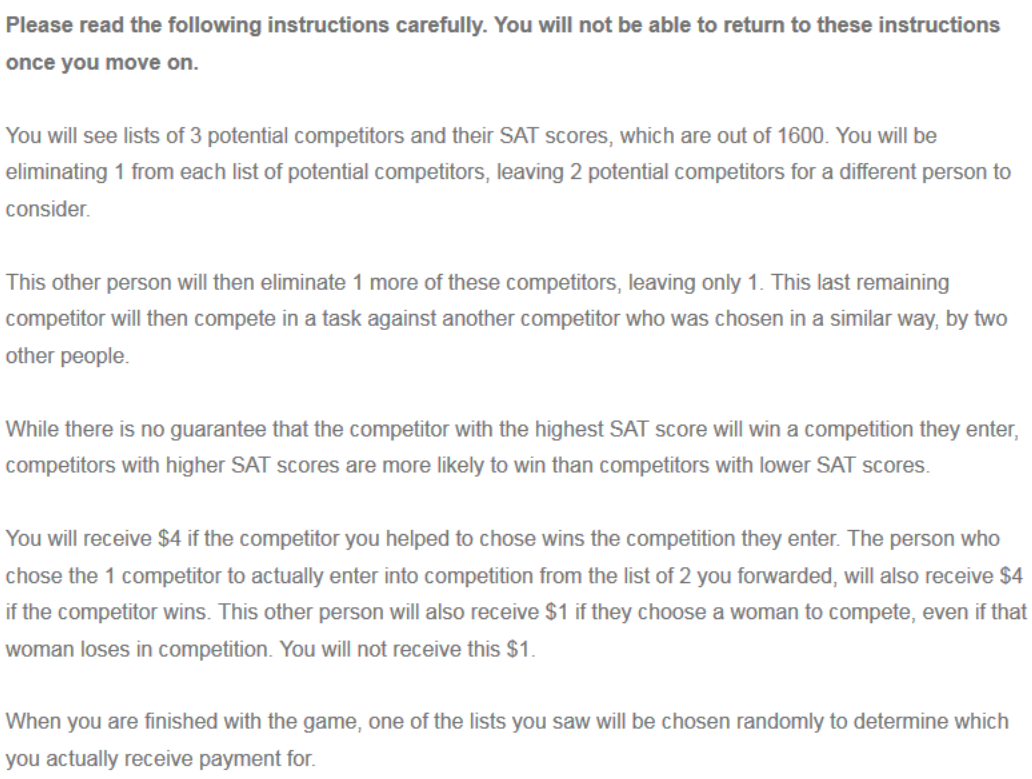
Panel B: Employment probabilities



Notes: Panel A shows the optimal reservation signal of quality, above which a candidate who offers a diversity enhancing attribute 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 diversity enhancing attributes ($B_2 = 0$) and plot results across a variety of potential valuations of diversity enhancing attributes by Agent 1. We further assume that both agents fully predict the other agent's behavior.

Figure 9: 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

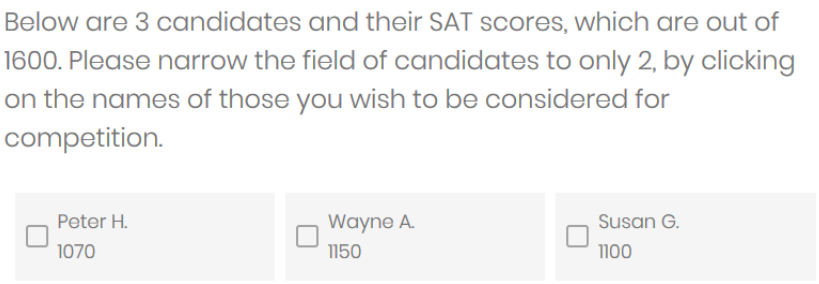


Table 1: Summary Statistics

	(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: Were experiment subjects differentially inclined to forward the top-two candidates when they anticipated that female candidates would subsequently be shown favor?

	(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 experiment 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.