Journal of Economic Behavior and Organization 184 (2021) 432–459

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Diversity and the timing of preference in hiring decisions☆

Logan M. Lee[a], Glen R. Waddell[b,c,*]

[a] Grinnell College, 1115 8th Ave, Grinnell, IA 50112, United States
[b] University of Oregon, 1285 University of Oregon, Eugene, OR 97405, United States
[c] IZA Bonn Germany

ARTICLE INFO

Article history:
Received 20 December 2019
Revised 11 November 2020
Accepted 15 November 2020
Available online 4 March 2021

JEL classification:
J7
D8

Keywords:
Diversity
Hiring
Gender
Race
Discrimination

ABSTRACT

We consider a hiring procedure in which candidates are evaluated in sequence by two agents of the firm. We illustrate how one agent's interest in enhancing diversity can indirectly influence the other agent's hiring decisions. 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, we first establish that incentivizing subjects to choose females (males) induces them into choosing females (males). Importantly, then, we establish that when subjects who screen candidates in an earlier stage know about this pending incentive they systematically avoid forwarding females (males) when they jeopardize the candidacy of higher-ranking male (female) candidates.

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. Frankly, our own profession still includes disproportionately few women and members of historically underrepresented racial and ethnic-minority groups (Bayer and Rouse, 2016). While "removing implicit and institutional barriers" is a common directive, and the rewards are potentially large, identifying these often-subtle barriers can be challenging

While experimental evidence supports taste-based racial discrimination as a direct contributor to unequal treatment (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2012; 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).[1] We consider a mechanism at the intersection of incomplete information and discrimination, stemming from decision makers at different levels in a hierarchy of a firm having different objectives. Specifically, we consider situations in which there are returns to

☆ 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. We thank Bill Harbaugh, Chris Ellis, Bruce McGough, Derek Neal, Paul Oyer, Jason Query, Nick Sly, and Jon Thompson for productive interactions, and seminar participants at the Midwestern Economic Association Annual Meetings. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author at: University of Oregon, 1285 University of Oregon, Eugene, OR 97405, United States.
E-mail addresses: leelogan@grinnell.edu (L.M. Lee), waddell@uoregon.edu (G.R. Waddell).

[1] Exley and Kessler (2019) also finds that women are less likely than men to engage in self promotion during the hiring process—this may further exacerbate differences in labor-market outcomes.

https://doi.org/10.1016/j.jebo.2020.11.014
0167-2681/

the firm from enhancing diversity but these returns are not internalized by all decision makers. In short, *where* in a hierarchy those differences arise will have 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 gender or race, for example, in a structure in which efforts are devoted to increasing the representation of women and 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 differences in the private interests of decision makers give rise to unequal treatment. 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 inequality. Thus, we learn, even preference *for* diversity enhancing attributes can be to the detriment of candidates 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.

These results have significant implications for firms and academic institutions, attempting to diversify their workforces. Firms have increasingly made public pledges stating their commitments to advancing diversity and inclusion in the workplace (McGrit, 2017). According to its website, The "CEO Action for Diversity & Inclusion" commitment has more than 1000 signatories including more than 300 of the Fortune 500 companies. Yet, despite these commitments, fewer than one percent of all CEOs in the Fortune 500 are black and fewer than eight percent of all CEOs in the Fortune 500 are women (Hinchiffe, 2020). While racial and gender disparities are possibly more visible in leadership, they persist broadly across education levels and job types (Gould, 2019; Hegewisch and Tesfaselassie, 2019). One reason for the lack of progress despite public commitments may simply be that the firms making these commitments are not genuinely interested in diversity and inclusion (Ahmed, 2007). We offer another possibility. Our results suggest that a top-down commitment to increasing diversity, even when well intentioned, may not only fail but could make disparities *larger* when the interest in diversifying is not equally shared at lower levels of the hierarchy. One policy prescription that comes from the model, for example, is not to focus on ensuring C-level executives understand the value of diversity, but instead to make sure mid-level managers—those in earlier rounds of applicant screening— share those convictions.

Below, we consider a theoretical setting 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 only considered by a second agent (we have in mind the firm's owner, for example, although one could imagine a university administrator also fitting well) when a first agent (for example, a division manager or department chair) has determined that the candidate is worthy of forwarding in the search. In that way, the process we model captures the typical "up-or-out" nature of job searches.[2]

In terms of actionable policy, we will speak directly to the implications of directed 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, diversity enhancing candidates can face a *lower* probability of employment, not higher.[3] For example, where leadership values female candidates, highly productive female applicants are harmed by the early decision maker's interest in protecting their preferred male candidate against the favorable treatment they anticipate being given to female candidates in subsequent rounds. In no way is this due to any disutility associated with hiring a candidate with a particular attribute—more fundamentally, we do not need the first-moving agent to dislike female candidates at all to find that female candidates can be made worse off when favored by the second-moving agent. Instead, the result is solely due to agents having incomplete information of candidates' true abilities and *difference* in how agents privately value diversity that drives the tradeoffs being made when the early mover anticipates a favorable preference being introduced in a subsequent evaluation. 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—the model gets its true import from the *relative* private values associated with diversity, and one need not consider that there are "biased" and "unbiased" agents, but merely that agents in the firm may value gender, or race, or any other attribute

[2] There is a rich literature related to our problem. For example, Green and Laffont (1987) models 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 collaborate to make decisions. In more-recent work, Guo and Shmaya (2019) and Frankel (2019) consider two-stage hiring processes. Guo and Shmaya (2019) 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 (2019) 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 (2019) shows that it can be optimal to allow the hiring manager discretion as long as a signal of ability falls above a threshold.

[3] Note that "top-down" diversity goals struggling 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).

differently. As such, we will consider variation in the relative values agents place on diversity, considering implications on employment and workforce productivity. As interests in diversity influence the relative probabilities that candidates with different abilities are hired, we also discuss the distributional consequences for subsequent promotion games.

In Section 2 we add some formality to the setting we have in mind, solving the sequential consideration of agents backwards. In so doing, we allow for cases where diverse candidates are directly discriminated against, although we think it is more relevant to policy to focus on diverse candidates being favored somewhere in the hiring process. This relevance is especially salient as we demonstrate that *a priori* favor can be to the detriment of diversity enhancing candidates.

We then write down the model in two ways. 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 well 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. We change this in Section 4, where we allow the second decision maker to be "savvy" regarding the incentives the first faces. While we tend to think that those in leadership positions may fall short of fully anticipating how subordinates 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 explicitly consider whether the second mover might overcome the influence of the first mover either by adjusting their own threshold or by incentivizing the first in a way that aligns their interest in diversity.

In Section 5 we provide experimental evidence showing that, when informed about the interests of a subsequent decision maker, subjects tend to avoid advancing candidates who would be advanced in the absence of gender considerations. Specifically, experimental subjects consider mixed-gender groups of three and are asked to choose two candidates to advance for additional consideration by someone else. Consistent with the model, when subjects expect that a subsequent player is incentivized to choose female (male) candidates, the "first movers" act in a way that is consistent with protecting top-ranked male (female) candidates from this threat—they forgo the advancement of *middle*-ranked female (male) candidates in favor of *bottom*-ranked candidates. In Section 6 we offer some additional insights as we conclude.

2. A motivation

2.1. The setup

We consider the implications of agents differentially internalizing, or having different responsibilities associated with, diversity interests as they undertake the hiring responsibilities for the firm.[4] In so doing, we consider a two-stage hiring game in order to speak to the implications of these values existing at different stages in the hiring process. By assumption, "Agent 1" considers the candidate first and either rejects the candidate or forwards the candidate to "Agent 2" for further consideration. We model Agent 1 as choosing a threshold, $\hat{s}_1$, and advancing all candidates who generate a signal of productivity, s_1, that falls above this threshold value. There is no ability for Agent 1 to influence the decision thereafter so, if forwarded to Agent 2, Agent 2 can then reject or hire the candidate unilaterally. Like Agent 1, Agent 2's decision is based on the candidate's productivity signal, s_2, and Agent 2's threshold value, $\hat{s}_2$, for that 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.[5]

A candidate's productivity is not verifiable, and both agents only know that a given candidate is highly productive with probability $\alpha \in (0, 1)$ and would therefore be a "good" hire. We quantify the upside to hiring such candidate (H) as an increase 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—this would be a "bad" hire—would lower the firm's value from V_0 to V_L. The firm's interest is always in rejecting

[4] Becker (1957) first introduces 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) uses 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) finds 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) discusses 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 et al. (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).

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

low-productivity candidates, which leaves the firm's value at the status-quo level, V_0. Note that α, V_H, V_L, and V_0 are not influenced by the personal attributes of the candidate.[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 as 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 $\tau_1 + \tau_2 \leq 1$. Consistent with a hierarchy, and agents moving in strict sequence, it is reasonable to anticipate that $\tau_1 \leq \tau_2$. We therefore impose these relative magnitudes throughout, which leaves Agent 1 having a smaller stake in the firm's value than Agent 2.[7]

We introduce the potential for diversity interests to be internalized by decision makers by allowing for some verifiable attribute of the candidate to enter directly into the agents' objective equations. Given the sequence of actions, we notate any private benefits accruing to Agent 1 from hiring the candidate as B_1, and any private benefits to diversity accruing to Agent 2 as B_2.[8] Of course, productivity may well relate directly to diversity within the firm, which we absorb in V_H and V_L. We explicitly leave B_1 and B_2 to represent only the private valuations of diversity perceived by agents 1 and 2, respectively. As such, we are in no way precluding a firm-level interest in workplace diversity—key is that we add to the model the potential for there to be private interests that *vary across agents* in the firm. To maintain 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, which we walk through below.

2.2. Agent 2's problem

When the candidate is forwarded to Agent 2 for a second and 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 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,

$$\begin{aligned}
\max_{\hat{s}_2} V_2(\hat{s}_2) = {} & \overbrace{\alpha[F_H(\mathbb{E}_2[\hat{s}_1]) + (1 - F_H(\mathbb{E}_2[\hat{s}_1]))F_H(\hat{s}_2)]}^{\text{Probability candidate is an } H \text{ type and Agent 2 rejects}} \cdot \tau_2 V_0 \\
& + \overbrace{\alpha(1 - F_H(\mathbb{E}_2[\hat{s}_1]))(1 - F_H(\hat{s}_2))}^{\text{Probability candidate is an } H \text{ type and Agent 2 hires}} \cdot (\tau_2 V_H + B_2) \\
& + \overbrace{(1-\alpha)[F_L(\mathbb{E}_2[\hat{s}_1]) + (1 - F_L(\mathbb{E}_2[\hat{s}_1]))F_L(\hat{s}_2))]}^{\text{Probability candidate is an } L \text{ type and Agent 2 rejects}} \cdot \tau_2 V_0 \\
& + \overbrace{(1-\alpha)(1 - F_L(\mathbb{E}_2[\hat{s}_1]))(1 - F_L(\hat{s}_2))}^{\text{Probability candidate is an } L \text{ type and Agent 2 hires}} \cdot (\tau_2 V_L + B_2)\,. \qquad (1)
\end{aligned}$$

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

[6] If access to resources differs by personal attributes (e.g., gender, race) these terms could carry trait-specific designations. However, to the extent that both agents experience these additional productivity benefits similarly, the main results of our model are unchanged.

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

[8] We remain agnostic about the nature of these benefits and rely on B_1 and B_2 to capture anything that might constitute the private benefits associated with the candidate's personal attributes.

[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.

[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. For ease of exposition, we assume $\sigma_H = \sigma_L$.

[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 $1 - F_L(\mathbb{E}_2[\hat{s}_1])$.

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 Eq. (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,

$$\begin{aligned}\underset{\hat{s}_1}{\text{Max}}\, V_1(\hat{s}_1) = {} & \overbrace{\alpha[F_H(\hat{s}_1) + (1-F_H(\hat{s}_1))F_H(R_2(\mathbb{E}_2[\hat{s}_1]))]}^{\text{Probability candidate is an } H \text{ type and Agent 1 rejects}} \cdot \tau_1 V_0 \\ & + \overbrace{\alpha(1-F_H(\hat{s}_1))(1-F_H(R_2(\mathbb{E}_2[\hat{s}_1])))}^{\text{Probability candidate is an } H \text{ type and Agent 1 forwards}} \cdot (\tau_1 V_H + B_1) \\ & + \overbrace{(1-\alpha)[F_L(\hat{s}_1) + (1-F_L(\hat{s}_1))F_L(R_2(\mathbb{E}_2[\hat{s}_1]))]}^{\text{Probability candidate is an } L \text{ type and Agent 1 rejects}} \cdot \tau_1 V_0 \\ & + \overbrace{(1-\alpha)(1-F_L(\hat{s}_1))(1-F_L(R_2(\mathbb{E}_2[\hat{s}_1])))}^{\text{Probability candidate is an } L \text{ type and Agent 1 forwards}} \cdot (\tau_1 V_L + B_1)\end{aligned} \tag{3}$$

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,

$$\frac{\alpha f_H(\hat{s}_1)(1-F_H(R_2(\mathbb{E}_2[\hat{s}_1]))) + \alpha(1-F_H(\hat{s}_1))f_H(R_2(\mathbb{E}_2[\hat{s}_1]))(\partial R_2(\mathbb{E}_2[\hat{s}_1])/\partial \hat{s}_1)}{(1-\alpha)f_L(\hat{s}_1)(1-F_L(R_2(\mathbb{E}_2[\hat{s}_1]))) + (1-\alpha)(1-F_L(\hat{s}_1))f_L(R_2(\mathbb{E}_2[\hat{s}_1]))(\partial R_2(\mathbb{E}_2[\hat{s}_1])/\partial \hat{s}_1)} = \frac{\tau_1 V_0 - (\tau_1 V_L + B_1)}{(\tau_1 V_H + B_1) - \tau_1 V_0}\,. \tag{4}$$

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 normalize Agent 1's diversity interest to 0 and consider the differences in Agent 2's *relative* interest in diversity. While we allow for $B_2 < 0$—and in the figures we capture the implications—our main focus is on cases where Agent 2 values the attribute, consistent with having the objective of increasing the representation of a race or gender (i.e., $B_2 > 0$). In making a decision, Agent 2 will have in mind the reservation signal Agent 1 would have chosen, but (here, at least) not that $\hat{s}_1$ could have been chosen by Agent 1 with $B_2 \neq 0$ in mind. 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 Agent 2 has 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.

[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 $F_H(\hat{s}_2) = F_L(\hat{s}_2)$.

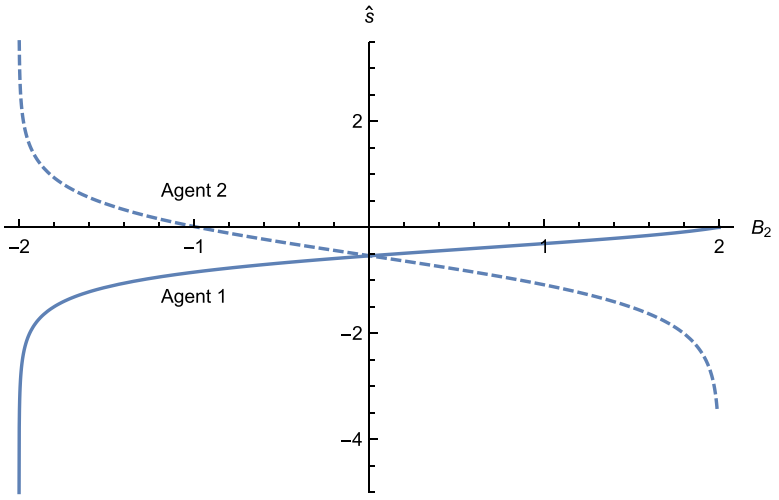

Fig. 1. Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2) *Notes:* Each curve shows the optimal reservation signal of productivity, above which a diversity enhancing candidate will be advanced for further consideration by Agent 1 or hired by Agent 2. We assume in this case that 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid Agent 2's decision collapsing on either "always advance" or "always reject.".

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 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 Fig. 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 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 productivity signal Agent 2 receives is relatively low. *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 mitigates the favorable treatment diverse candidates would otherwise receive.*

[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 the figures, we characterize agent behavior having adopted $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$.

[14] 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.

[15] Fig. 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 $\tau_2 V_L$.

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, noting that African-American-sounding names receiving fewer call backs in resume studies (Bertrand and Mullainathan, 2004) is evidencing a certain discrimination against applicants thought to be African American, it is as consistent with the callback decision resting with someone early in a sequence of decisions who does not herself value the applicant's race, but anticipates subsequent decision makers showing preference *for* African-American applicants. 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 productivity 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 outcomes. In the limit, the firm unavoidably suffers the costs of Agent 2's relative bias against hiring the candidate. This asymmetry is anticipated—both agents must approve a candidate for hire, but rejection can occur with either agent's decision to reject. That asymmetry is also fundamental–favor for a candidate 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 Fig. 2 we suspend for the moment the role of Agent 1 and plot the 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 B_2 capturing Agent 2's private 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-productivity female candidates are hired increases faster than the rate at which 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 Fig. 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. That is, with top-down preferences, among low-ability candidates who also offer the firm the ability to increase diversity, employment rates are strictly increasing in B_2. Moreover, employment rates are not monotonic in B_2 among high-productivity candidates who likewise offer the firm the ability to increase diversity. In a sequential-hiring game, then, there can be 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 was large and positive. This counterintuitive result occurs because, for large positive values of B_2, Agent1's decision to advance is the only hurdle to be cleared by the candidate. Collapsing on only one signal of productivity and a second decision maker who always advances leaves additional noise in the process, which helps low-productivity females but hurts high-productivity females.

The result implies that 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 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, 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 high-productivity candidates and for the firm.

In terms of policy, to the extent that Agent 2 is more aligned with 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 $B_2 > 0$ and $B_1 = 0$, 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

[16] Note that with symmetry assumed, a single decision maker would solve the first-order condition at $\hat{s} = 0$. In Fig. 1, that $\hat{s}_1^* < 0$ 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 $\hat{s}_1^* \leq 0$).

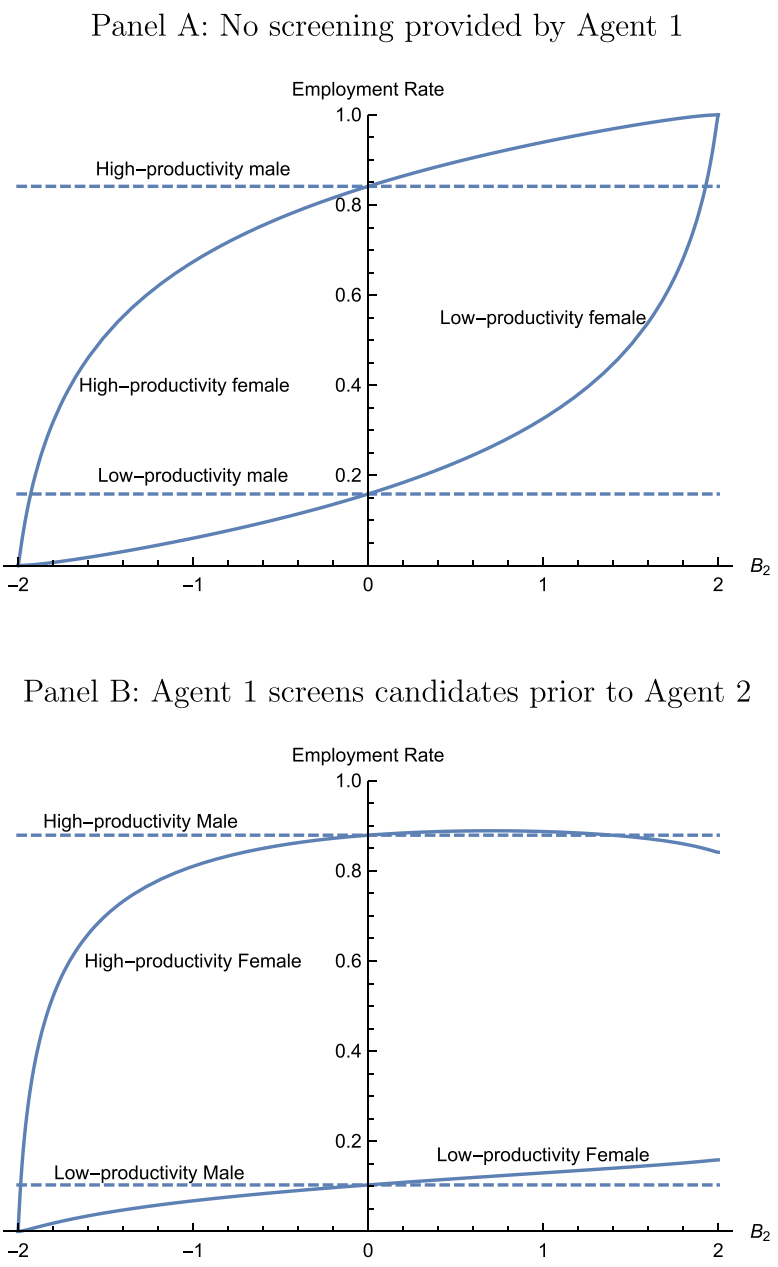

Fig. 2. Employment probabilities with top-down preferences ($B_1 = 0$, as we vary B_2) *Notes:* Each curve indicates the probability that candidates from the indicated group will be hired by the firm. We assume in this case that Agent 1 places no value on diversity enhancing attributes ($B_1 = 0$) and 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid Agent 2's decision collapsing on either "always advance" or "always reject." .

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 Fig. 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.

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

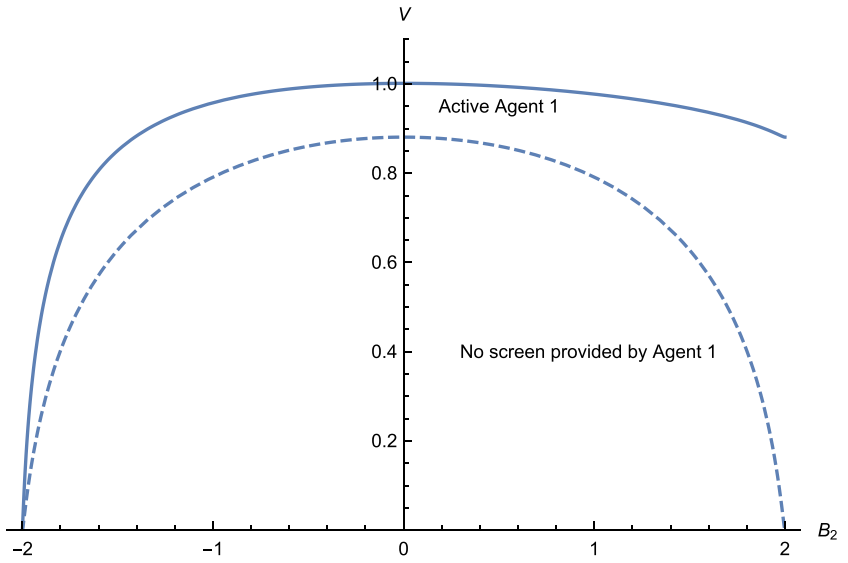

Fig. 3. Average employee productivity with top-down preferences ($B_1 = 0$, as we vary B_2) *Notes:* Each curve indicates the expected value to the firm when considering a randomly chosen candidate. We assume in this case that Agent 1 places no value on diversity enhancing attributes ($B_1 = 0$) and 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid Agent 2's decision collapsing on either "always advance" or "always reject." We normalize to one average productivity when $B_1 = B_2 = 0$.

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] *Due to the ability of Agent 1 to unilaterally reject candidates, average productivity among hired diverse candidates will be lower 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 diversity traits. Even in the absence of diversity considerations playing a direct role in promotion decisions, promotion outcomes can be shown to depend on B_2. For example, if those overseeing subsequent promotions perceive this B_2-induced difference in productivity, 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 that is driven, somewhat surprisingly, by a desire to *increase* firm diversity.

3.4. The role of Agent 1's diversity interest

Thus far, we have assumed that $B_1 = 0$. In Fig. 4 we allow for $B_1 \neq 0$ and $B_2 \neq 0$, allowing both agents to have private value of diversity. As before, we plot Agent 2's choice of $\hat{s}_2$, but now with a menu of $\hat{s}_1$ that correspond to different 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, then, Agent 1's decision rule in the strictly "top-down" case (i.e., where $B_1 = 0$) can be seen in the solid line. In Panel A of Fig. 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 is that *for all B_1, $\hat{s}_1^*$ 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.

[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 $V_0 = 0$.

[19] Of course, if the potential promotion of females continues to be subject to the B_1 and (especially) B_2 values that occurred in the hiring process, outcomes will be influenced. In fact, in such a setting, our "hiring" game can itself be recast as a promotion game of sorts.

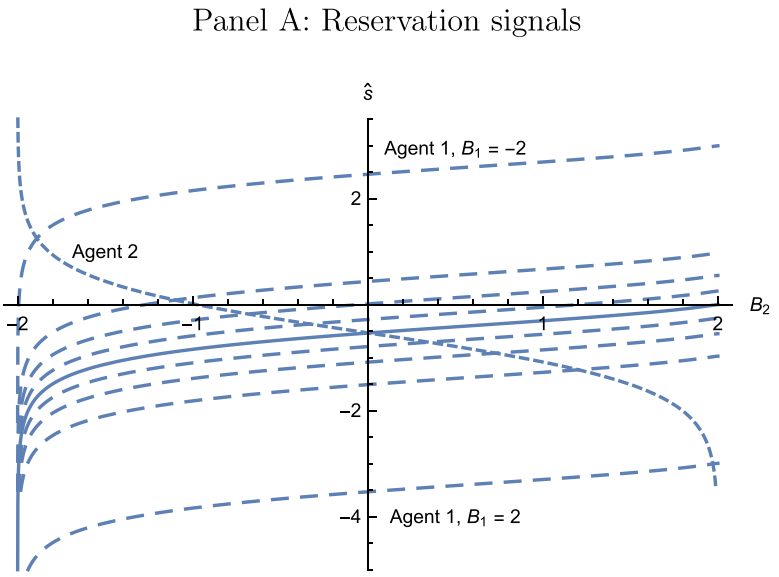

Fig. 4. Reservation signals and employment rates across B_1 and B_2 *Notes:* Each curve in Panel A shows the optimal reservation signal of productivity, above which a diversity enhancing candidate will be advanced for further consideration by Agent 1 or hired by Agent 2. We plot results across a variety 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-productivity signal is unaffected by Agent 1's minimum-productivity signal. Panel B displays employment rates for H types and Panel C displays employment rates for L types hires. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_1 to $\tau_1 V_L \leq B_1 \leq \tau_1 V_H$, and B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid either Agent's decision collapsing on either "always advance" or "always reject.".

In Panels B and C of Fig. 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 Fig. 2, Panels B and C of Fig. 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. In Fig. 4 we also demonstrate 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.

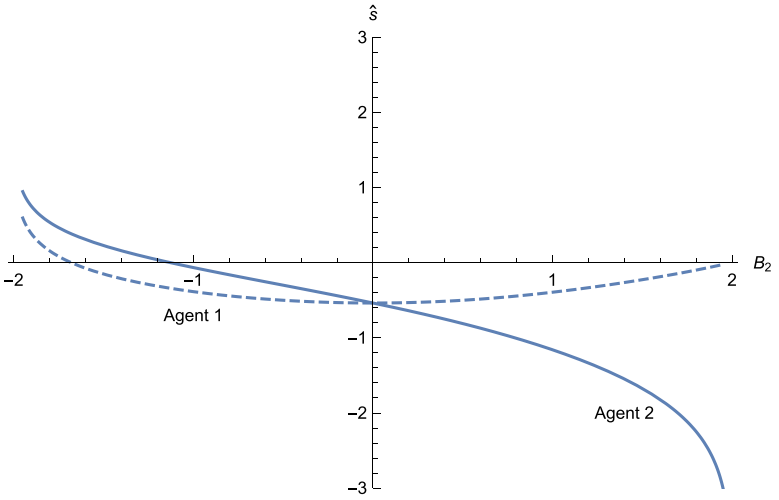

Fig. 5. Reservation signals with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2 *Notes:* Each curve shows the optimal reservation signal of productivity, above which a candidate who can enhance diversity will be advanced for further consideration by Agent 1 or hired by Agent 2. 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid Agent 2's decision collapsing on either "always advance" or "always reject.".

One final consideration in cases where both $B_1 > 0$ and $B_2 > 0$ is that blind first-stage assessments will often be detrimental to female candidates. To the extent that agents throughout the firm place value on candidates being female, a blind first-stage evaluation will reduce employment of female candidates. This effect will be particularly large where $B_1 \geq B_2$. This concern is consistent with the broad empirical finding that blind early assessments can hurt minority candidates, especially when firms are allowed to self-select into an anonymous resume system (Behaghel et al., 2015). Similarly, female and minority candidates have been shown to suffer where blind evaluations have caused a change the relative weight given to various signals of productivity as well as the way these signals were interpreted (Krause et al., 2012; Behaghel et al., 2015).[20]

Blind candidate assessments have been considered in other settings as a policy that can help to mitigate discrimination against certain groups of candidates. Goldin and Rouse (2000) finds that a change that prevented juries from observing gender in orchestra auditions increased the probability that females were advanced and ultimately hired. Similarly, Åslund and Skans (2012) finds that anonymous application procedures increased the probability that both women and people of non-Western origin were advanced to the interview stage and were ultimately hired. In the context of our model, these results could be evidence of early stage decision makers actively offsetting the expected favorable treatment minority candidates will receive if advanced for further consideration.

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 and respond to $\hat{s}_1$. We therefore return to the general first-order conditions in Eqs. (2) and (4). That is, we allow both agents to choose reservation signals while fully anticipating the effect that choice has 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 Fig. 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 Fig. 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 $B_2 < 0$. *With top-down preferences, when Agent 2 is savvy in setting expectations of Agent 1's reservation*

[20] Arguably, the "ban the box" papers also relate on this margin Agan and Starr (2017); Doleac and Hansen (2020).

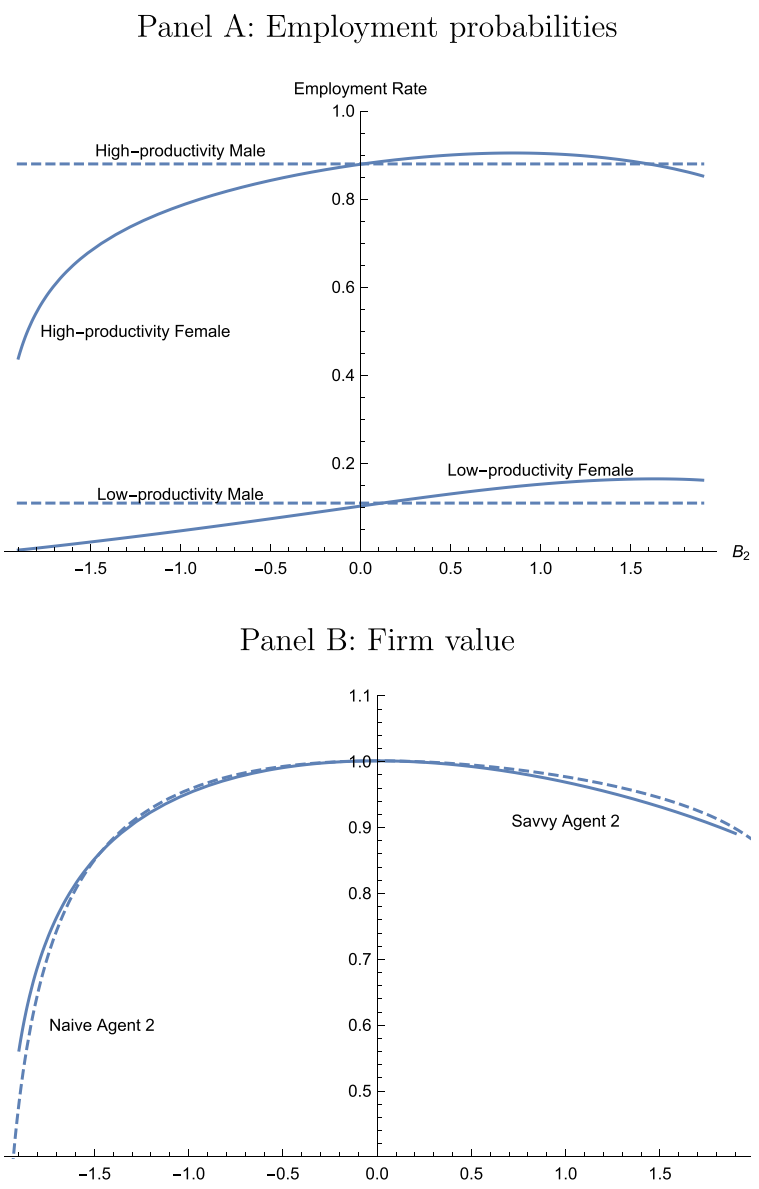

Fig. 6. Employment probabilities and employee productivity with top-down preferences ($B_1 = 0$, as we vary B_2) and a savvy Agent 2 *Notes:* Panel A indicates the probability that candidates from the indicated group will be hired by the firm. Panel B indicates the expected value to the firm from considering 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid Agent 2's decision collapsing on either "always advance" or "always reject." We normalize to one average productivity when $B_1 = B_2 = 0$.

signal, $\hat{s}_1^\star$ is monotonically decreasing in $B_2 \in (\tau_2 V_L, 0)$. (Whether Agent 2 is naive or 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 likewise 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—this 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, Agent 1's decision is 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 correction, of a sort. In this way, taste-based discrimination introduced *late* in a sequence can yield a sort of statistical discrimination *early* 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, by recognizing that subsequent decision makers will not base their decision entirely on expected productivity, Agent 1 will treat 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 Fig. 6 we again plot employment rates—the patterns are remarkably similar to those in the naive case (see Fig. 2). With Agent 2 now savvy, both high- and low-productivity females are less likely to be hired for $B_2 < 0$, while 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. The intuition in this result is similar to the naive case—any increase in $\hat{s}_1^*$ decreases the likelihood that a female candidate is advanced past the first stage of the evaluation. Thus, for high-productivity females, that Agent 1 adopts a higher standard is not fully offset by the decrease in $\hat{s}_2^*$ because high-productivity females were already likely to clear Agent 2's evaluation, even where $B_2 = 0$. The opposite is true for low-productivity females—they benefit disproportionately from the decrease in $\hat{s}_2^*$ because in the limit they need only generate one sufficiently high signal (when being evaluated by Agent 1) in order to be hired by the firm—this is more likely than generating two above-average signals.

In Panel B of Fig. 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's influence is less-offsetting 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 AAgent 1's private value

In Panel A of Fig. 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 Fig. 6. Around this line, however, we see an interesting asymmetry in how employment rates vary with agents' diversity interests. 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 Fig. 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, private concerns for diversity.

The timing of preference—whether introduced with Agent 1 or Agent 2—yields striking differences in agents' optimal thresholds. In Fig. 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 Fig. 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 implications of these private values, and where they show up, is also seen in Panel B of Fig. 8, where we plot associated employment probabilities by productivity. Where discrimination against a candidate's personal attributes is introduced—with Agent 1 or Agent 2—is of little consequence to employment, as either agent can unilaterally dismiss candidates. Yet, as no single agent can unilaterally hire a candidate, preference for a candidate's personal attribute yields different patterns of outcome. 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 Fig. 6), where strong preference on the part of Agent 2 ultimately leaves highly productive female candidates less likely to be hired.[21]

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 evaluate a candidate differently if they anticipate that a subsequent decision-maker will value a personal attribute in that candidate in making the ultimate hiring decision? As in the theory section, we focus our discussion around private values related to gender (though it could easily be recast with respect to race, for example). Specifically, then, we consider three environments—in them, the second decision maker is incentivized to choose a female candidate, a male candidate, or neither (i.e., a control condition). Despite the gender framing, the experiment is not meant to ascertain the gender preferences of subjects at all. Rather, the experiment is intended to determine whether there are systematic regularities in decisions that imply a willingness in early decision makers to take positions that ultimately *offset* the anticipated preference of subsequent decision makers—if so, an important mechanism to integrate into our understanding going forward.

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, agents in our theoretical model optimally choose thresholds above which a candidate would be advanced for further consideration. However, to set such a threshold (even in the absence of diversity considerations) requires subjects to have an understanding of the distribution of high- and low-productivity pools and be able to use that information to balance type-I and type-II errors.

To not overwhelm subjects with a variety of characteristics that imply diversity and productivity, the experiment simplifies both the theoretical model and the real world by providing subjects with a name (to imply gender) and SAT score for each candidate (to imply productivity).[22] This simplification also implies a fairly straightforward empirical analysis of subject choices.[23]

Similarly, while the agents in our theoretical model may hire any number of applicants, subjects in our experiment participate in a hiring process in which a single candidate is chosen. In particular, we characterize the "Agent 1" problem as a relatively simple comparison—given three possible candidates, which two should be advanced?

In the absence of diversity considerations, this decision quickly collapses to advancing the top-two candidates. However, when diversity considerations are introduced into the second decision maker's objective function—the one making the "two-choose-one" choice—early-moving subjects might adjust their own strategies accordingly. For example, if female candidates will receive preference when being evaluated by the second decision maker, subjects might be wary of advancing middle-ranked female candidates, especially when both the top and bottom candidates in a given group of three are male. By advancing the top-two candidates in such a scenario, the first decision maker exposes their choice of candidates to the potential that the second decision-maker's preference for females leads to the middle-ranked candidate being chosen. If, instead, the first decision maker advances the two male candidates, gender cannot be used to determine which candidate is ultimately chosen. This is the sort of systematic choice we will be measuring experimentally.

One final and related difference between the theoretical model and the experiment is that subjects in the experiments are evaluating sets of candidates jointly—in the theory, candidates were evaluated sequentially. Existing literature has demonstrated that when presented with two attributes, the weighting of each attribute will depend on whether candidates are evaluated jointly or individually, which can cause different candidates to be selected in joint and individual evaluations of the same candidate pool (Hsee, 1996; Bazerman et al., 1998; Shaffer and Arkes, 2009). In particular, Bohnet et al. (2016) finds that when candidate attributes are presented side by side (rather than one at a time) gender bias is all-but eliminated.[24] Given this, the joint evaluation of candidates in our experimental setting might be expected to work *against* finding a role for gender. In our experiment, subjects face "side-by-side" comparisons yet still exhibit gender "bias"

[21] In Appendix A we consider strategies that could be employed to mitigate the efficiency losses we have documented, including potential transfers that would attempt to align the interests of agents.

[22] All names were drawn from the US Social Security Administration's list of the 100 most-common male and female names given to individuals born in 1990. Any names appearing in the 1000 most-common names for both males and females were removed. SAT scores were drawn from a range 960 to 1210.

[23] While we imagine hiring decisions often implicating multiple signals of a candidate's ability, we limit the dimensionality to one in our environment so there is little question about the rank ordering of candidates in the absence of gender-related strategy—we need a true test, in essence, and more noise in what experimental subjects might value is not informative here.

[24] Li and Hsee (2019) offers an explanation for this by showing that attributes that should clearly be taken into account but that can be difficult to evaluate without comparison to other values are given particular focus in joint evaluations. On the other hand, attributes that are easy to evaluate but that are less clearly important to the candidates evaluation are given extra weight in individual evaluations. In our context, SAT score is obviously something that subjects should take into account as their payment is directly based on it, but it is also likely difficult to evaluate without comparison to others as the subjects were not informed of the mean SAT score among candidates. Gender on the other hand is relatively easy to evaluate, but the extent to which it should be taken into account is unclear for subjects acting as the first agent because they are unsure of the extent to which subjects acting as the second agent may choose to pass over higher SAT score candidates in order to secure a bonus payment for hiring candidates of a particular gender.

when the decision is in anticipation of a pending gender preference. That we do find a role for gender, then, suggests that differences in the private valuations of candidate attributes (gender being just one) can remain an important determinant of employment outcomes, even when candidate attributes are presented side by side.

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.[25]

All subjects were assigned to participate as either Agent 1 or as Agent 2. With our interest in evaluating the potential for Agent 1 to act as the theory predicts (e.g., forgoing high-ability female candidates who could threaten the candidacy of a highly ranked male candidates) we have most subjects (140) participate in the role of Agent 1. We display subject characteristics in Table 1. As should be expected with randomized experiments, subject characteristics are well balanced with no subject characteristics predicting whether subjects were assigned to treatment or control groups. Among subjects acting in the role of Agent 1, those in the female-preferred treatment spent 2.4 minutes (33 percent, $p = .129$) longer to complete the tasks than those in the control group. On the other hand, subjects in the male-preferred treatment spent only thirty seconds (7 percent, $p = .678$) longer than those in the control group. While none of these differences are statistically significant, they are consistent with subjects experiencing more dissonance around treatments in which female candidates are favored.

5.2.1. Subjects playing "Agent 1" roles

Agent 1 subjects evaluated 30 sets of three candidates and in each set were asked to chose two of the three candidates to advance. In the instructions, subjects were told that the two candidates they advanced would "be seen by another person, who will choose only one of them." The subjects were also told that this final candidate would ultimately be compared to another candidate who was chosen in a similar fashion, and that they and the other person would both receive $4 if the final candidate they chose had a higher SAT score than this other candidate.[26] To assist with comprehension, subjects acting as Agent 1 were then shown an example.[27]

Our experimental variation comes from the randomly assigned setting within which experimental subjects perform—some subjects (50) were told that the subsequent decision would be made by someone who earns a $3 bonus for hiring a female candidate, some (48) were told that the subsequent decision would be made by someone who earns a $3 bonus for hiring a male candidate, and some (42) were told nothing beyond that the subsequent decision would be made by someone else.

With a three-choose-two design, there are eight possible combinations of gender and ordinal (SAT) rank—we did not allow for ties in SAT, so in all combinations subjects would be able to distinctly order candidates by SAT. As subjects realize the highest expected return when they advance the two candidates with the highest SAT scores, we expect "control" subjects to advance the top-two candidates. However, they play an important role nonetheless, as one could imagine baseline preferences being important to distinguish (e.g., to forward one candidate from each gender). Of the eight possible scenarios, however, we have particular interest in two of them—the two that most-clearly confront the theoretical propositions we wish to evaluate.

First, where the second agent has been given an incentive to choose female candidates, for example, there is an interesting evaluation of the theory when the first agent faces a set of two male candidates and one female candidate, and the female is in the middle on observable productive attributes—this we will notate as $M_1F_2M_3$, with subscripts indicating the candidate's ordinal rank with respect to SAT. Specifically, a gender-neutral evaluation of candidates would imply that subjects choose M_1 and F_2. Likewise, a general preference for forwarding one of each gender would evidence in M_1 and F_2

[25] 174 subjects participated in the experiment—all were recruited and completed the experiment online using Amazon's MTurk service. As the purpose of the experiment is to learn about "Agent 1" choices—whether subjects respond to the pending favor to be given to a candidate of a particular gender, as we predict above—we up-weighted the collection of "Agent 1" data. Specifically, we recruited 140 subjects as Agent 1 in the experiment while the remainder (34) acted as Agent 2. Subjects were randomly assigned to the control, to the male-preferred treatment, or to the female-preferred treatment, and experienced questions in random order. At the end of the experiment, Agent 1 subjects were randomly matched with an Agent 2 subject (within the same treatment arm) to determine the outcome and the associated payment. (All subjects were paid for a randomly chosen round at the end of the experiment.) Agent 2 subjects spent an average of five minutes (298 seconds) completing the experiment and earned an average of $2.11 in addition to their $1 show up fee. Agent 1 subjects spent an average of eight minutes (491 seconds) completing the experiment and earned an average of $1.80 in addition to their $1 show up fee.

[26] The instructions for settings in which Agent 2 was incentivized to hire a female candidate can be seen in Appendix B, in Panel A of Fig. B1. The instructions for settings in which Agent 2 was incentivized to hire a male candidates are comparable. For those in the control group, instructions simply omitted the phrases "This other person will receive $3 for choosing a female candidate. You will not receive this $3." and "Regardless of the candidate's SAT score, they will keep the $3 for choosing a female candidate."

[27] A representative example can be found in Panel B of Fig. B1 for settings in which Agent 2 was incentivized to hire a female candidate. In the example scenario, after choosing two of the three candidates to advance for further consideration, subjects were informed about how the subsequent two-choose-one choice had been made. A representative example of this feedback, which was based on the actual decisions made by subjects who participated in the experiment as Agent 2, can be found in Panel C of Fig. B1.

Table 1
Summary statistics.

		Agent 2 rewarded for choosing a female candidate		Agent 2 rewarded for choosing a male candidate	
	Control (1)	Treated (2)	Difference (3)	Treated (4)	Difference (5)
Female	0.620	0.595	0.025	0.458	-0.137
	[0.490]	[0.497]	(0.103)	[0.504]	(0.106)
Age	42.48	44.04	1.56	42.15	-0.330
	[10.28]	[11.80]	(2.30)	[12.74]	(2.429)
White	0.881	0.860	-0.021	0.813	-0.068
	[0.328]	[0.351]	(0.071)	[0.394]	(0.076)
Black	0.0476	0.0800	0.032	0.0625	0.015
	[0.216]	[0.274]	(0.051)	[0.245]	(0.049)
Asian	0.0238	0.0200	-0.004	0.0625	0.039
	[0.154]	[0.141]	(0.031)	[0.245]	(0.043)
Hispanic	0.0238	0.000	-0.024	0.0417	0.018
	[0.154]	0.000	(0.024)	[0.202]	(0.038)
Other race	0.0238	0.0400	0.016	0.000	-0.024
	[0.154]	[0.198]	(0.037)	0.000	(0.024)
Income $\leq$ \$20,000	0.167	0.240	0.073	0.188	0.021
	[0.377]	[0.431]	(0.084)	[0.394]	(0.081)
Income \$20,001 - \$40,000	0.357	0.360	0.003	0.313	-0.045
	[0.485]	[0.485]	(0.101)	[0.468]	(0.101)
Income \$40,001 - \$60,000	0.190	0.180	-0.010	0.229	0.039
	[0.397]	[0.388]	(0.082)	[0.425]	(0.087)
Income \$60,001 - \$80,000	0.190	0.180	-0.010	0.104	-0.086
	[0.397]	[0.388]	(0.082)	[0.309]	(0.076)
Income $>$ \$80,001	0.0952	0.0400	-0.055	0.167	0.071
	[0.297]	[0.198]	(0.054)	[0.377]	(0.071)
GPA	3.285	3.458	0.173	3.442	0.157
	[0.739]	[0.438]	(0.132)	[0.427]	(0.131)
SAT score	1271.6	1291.6	20.02	1237.3	-34.23
	[150.6]	[143.1]	(37.56)	[177.8]	(42.40)
Minutes to complete	7.163	9.560	2.397	7.660	0.497
	[7.074]	[8.328]	(1.606)	[3.378]	(1.195)
Subjects	42	50		48	

Notes: Observations are at the subject level and include all subjects who participated as Agent 1. Columns 1, 2, and 4 report standard deviations in square brackets. Columns 3 and 5 estimate the different between the indicated treatment group and the control group and present robust standard errors in parentheses. Subjects were asked to report ACT score if they had not taken the SAT. All reported ACT scores have been converted to SAT scores. Even after this conversion, 46 subjects did not have a reported SAT score (14 missing in the control group, 16 missing in the male preference group, and 16 missing in the female treatment group). Some subjects did not report a GPA (two in the control group, two in the male preference group, and three in the female treatment group). *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

being chosen. Thus, it would be informative (and consistent with the theory) to find subjects skip middling females in favor of forwarding two male candidates when the second decision is being made by someone who prefers female candidates—a treatment-induced shift toward choosing M_1 and M_3. Likewise, a systematic increase in the likelihood a subject chooses F_1 and F_3 from the triple of $F_1M_2F_3$ when anticipating the second decision being made by someone who prefers *male* candidates would be consistent with the theoretical model presented above.

The model's prediction can also evidence in a second scenario, and in a way that is even more-difficult to explain away with a general preference for forwarding one of each gender, for example. Specifically, with a bonus paid to the second decision maker for choosing a female candidate, seeing an $M_1F_2F_3$ scenario leaves Agent 1 subjects having to choose at-least one female candidate. Like the scenario above, a preference for forwarding one of each gender should, all else equal, evidence in M_1 and F_2 being chosen. Thus, to see a systematic shift toward advancing the top male and the *bottom*-ranked female candidate—choosing M_1 and F_3—is consistent with subjects anticipating that increasing the score-gap between the top male and bottom-ranked female candidate increases the costs to Agent 2 exercising a preference for choosing a female candidate. The comparable scenario for the male treatment would be an $F_1M_2M_3$ scenario, which we also upweight.

In the experiment, each subject saw 30 sets of three candidates in random order.[28] As is typical in experiments, when the experimental subject has completed all tasks, one was randomly chosen for payment. Specifically, we choose one set of

[28] "Agent 1" subjects in the female-preference treatment chose from fifteen $M_1F_2M_3$ scenarios (up-weighted to increase precision around this key test of the theory), five $M_1F_2F_3$ scenarios, five $M_1M_2F_3$ scenarios, three $F_1M_2M_3$ scenarios, three $F_1M_2F_3$ scenarios, three $F_1F_2M_3$ scenarios, and one $F_1F_2F_3$ scenario. Subjects in the male-preference treatment saw the same distribution of threesomes, with genders reversed.

Table 2
Agent 2 behavior: Do Agent 2 subjects choose the top-ranked candidate (in two-choose-one decisions) when incentivized to choose a female or male?.

	Top-ranked candidate advanced			
	Agent 2 rewarded for choosing a female candidate		Agent 2 rewarded for choosing a male candidate	
	(1)	(2)	(3)	(4)
Treated = 1	-0.161***	-0.080**	-0.179***	0.063
	(0.033)	(0.035)	(0.034)	(0.052)
Treated = 1 × Top candidate not preferred gender × Bottom candidate preferred gender		-0.224***		-0.693**
		(0.072)		(0.247)
Top candidate not preferred gender) × Bottom candidate preferred gender		0.044		-0.067
		(0.048)		(0.041)
Top candidate not preferred gender		-0.021		-0.018
		(0.025)		(0.031)
Treated = 1 × Top candidate not preferred gender		-0.006		-0.003
		(0.035)		(0.041)
Bottom candidate preferred gender		0.060**		0.084***
		(0.024)		(0.023)
Treated = 1 × Bottom candidate is preferred gender		0.000		-0.033
		(0.038)		(0.043)
Mean (control)	0.933	0.933	0.933	0.933
Impact (%)	-17.2		-19.2	
Observations	646	646	645	645

Notes: In all columns, the dependent variable is equal to one if subjects choose the top-ranked candidate, and zero otherwise. Observations are at the subject-by-question level. In all columns, the control group is all subjects who were in neither female- or male-treatment arms. Robust standard errors are reported in parentheses. Reported mean values are calculated only among subjects in the control group. *** p<0.01, ** p<0.05, * p<0.1.

three candidates at random, matching the two chosen by Agent 1 to one of the "Agent 2" subjects in the same experimental setting, and payment was determined by the outcome of that two-choose-one choice. We describe this choice below.

5.2.2. Subjects playing "Agent 2" roles

Each subject assigned to an Agent 2 role evaluated 45 sets of two candidates, with the instruction to choose one to be compared to another candidate who was chosen in a similar fashion.[29] Agent 2 subjects were told that the SAT score of the candidate they chose would be compared to the SAT score of another candidate chosen in a similar way, and that they would earn $4 if the SAT score of the candidate they chose was higher. Subjects who were randomly assigned to the "female-bonus" treatment were told that if the chosen candidate was female then they would earn an *additional* $3. Likewise, subjects who were randomly assigned to the "male-bonus" treatment were told that if the chosen candidate was male then they would earn an additional $3. All subjects were paid for a randomly chosen round at the end of the experiment.[30]

In Table 2 we test the extent to which the $3 incentive motivated subjects toward choosing the second-ranked candidate over the first-ranked candidate, which we assume would be the tendency in the control group. In columns (1) and (3) we simply capture the average differences between treatment and control subjects—those incentivized to choose female candidates were 17.2-percent less likely to advance the first-ranked candidate (16.1 pp) relative to subjects who were not given any gender incentives, while those incentivized to choose male candidates were 19.1-percent less likely to advance the first-ranked candidate (17.9 pp). In columns (2) and (4) we confirm that the scenarios driving this difference are cases where the first-ranked candidate will not earn the subject any additional payment but the second-ranked candidate will earn the subject the $3. In these cases, subjects incentivized to choose female candidates were 24-percent (22.4 pp) less likely to advance the first-ranked candidate while subjects incentivized to choose male candidates were 74-percent (69.3 pp) less likely to advance the first-ranked candidate. Treatment has induced Agent 2 subjects toward choosing the gender for which they are directly rewarded—while we do not perform a belief elicitation stage directly, this is consistent with Agent 1 subjects in the treatment groups anticipating that gender-specific payments to Agent 2 subjects will translate into Agent 2 showing preference for those genders in the subsequent two-choose-one decisions.

[29] The 30 sets of three candidates seen by subjects who were assigned to Agent 1 roles create 90 possible sets of two candidates—we divided these 90 possible sets of candidates into two sets of 45 to keep the length of the experiment for Agent 2 subjects similar to that of Agent 1 subjects.

[30] The instructions for Agent 2 subjects in the male-preference treatment are reproduced in Fig. B2 in Appendix B. Similarly, the example question seen by these subjects is reproduced in Panel B of Fig. B2. Agent 2 subjects were not given feedback on their choice in the example question.

Table 3
Agent 1 behavior: Does anticipating gender preference move subjects away from advancing the top-two candidates (in three-choose-two decisions)?

	Top-two candidates advanced			
	Agent 2 rewarded for choosing a female candidate		Agent 2 rewarded for choosing a male candidate	
	(1)	(2)	(3)	(4)
Treated = 1	-0.184***	-0.141***	-0.186***	-0.131***
	(0.040)	(0.038)	(0.044)	(0.049)
Treated = 1 × Top candidate not preferred gender	-0.060**			-0.087**
		(0.028)		(0.034)
Top candidate not preferred gender		-0.021		-0.021
		(0.020)		(0.020)
Mean (control)	0.958	0.958	0.958	0.958
Impact (%)	-19.2		-19.4	
Observations	2756	2756	2696	2696

Notes: In all columns, the dependent variable is equal to one if subjects choose the top-two candidates, and zero otherwise. Observations are at the subject-by-question level. In all columns, the control group is all subjects who were in neither female- or male-treatment arms. Standard errors allow for clustering at the subject level and are reported in parentheses. Reported mean values are calculated only among subjects in the control group. *** p<0.01, ** p<0.05, * p<0.1.

5.3. Analysis and results

We now turn to consider our fundamental interest, on which the theory rests, whether individuals in the first stage of a hiring game behave differently when they anticipate that a subsequent decision-maker will value a personal attribute in the candidate in making a hiring decision.

In our empirical analysis, we model deviations from advancing the top-two candidates (ranked by SAT), as doing so would assure the highest expected value to the subject (in the role of Agent 1) in the absence of gender-neutral decisions. As a baseline specification, then, we consider modeling the behavior of our experimental subjects as

$$\mathbb{1}(\text{Top two advanced}_{iq} = 1) = \beta_0 + \beta_1 \mathbb{1}(\text{Treated}_i = 1) + \epsilon_{iq} \tag{7}$$

where $\mathbb{1}(\text{Top two advanced}_{iq} = 1)$ equals one if subject i advances the two candidates with the highest SAT scores from among the three candidates in question q.[31] We capture any level difference across treatment and control sessions in $\mathbb{1}(\text{Treated}_i = 1)$ and, due to random assignment, interpret $\hat{\beta}_1$ as the difference in choice induced by treatment. We estimate ϵ_{iq} allowing for clustering at the subject level.

In Table 3 we first report the level differences associated with subjects being informed that the second decision—the pending two-choose-one decision—will be made by someone who is paid an additional $3 if a female is chosen (Column 1) or if a male is chosen (Column 3). Each of these is compared to control observations where subjects were not anticipating any such preference. Neither of these differences are small. When subjects are told that the second decision will be made by someone who is rewarded for choosing a female, there is a 19-percent reduction (18.4 pp) in the probability that Agent 1 subjects choose the two candidates with the highest and second-highest SAT scores. Likewise, anticipating that the second decision will be made by someone with preference for a male, we find a 19-percent reduction (18.6 pp) in the probability that subjects choose the top-two candidates.

In columns (2) and (4) of Table 3 we consider variation coming from the interactions of $\mathbb{1}(\text{Treated}_i = 1)$ and the gender of the top-ranked candidate within each three-way comparison. We single out these, in particular, as variation in the gender of the top-ranked candidate guides whether the interests of Agent 1 and Agent 2 are aligned within the treated groups. For example, when Agent 1 anticipates that Agent 2 will prefer female candidates *and the top-ranked candidate is female*, Agent 1 should see no misalignment of incentives at all. To the contrary, it is when the top-ranked candidate is male that Agent 1 might envision a misalignment, and act to protect their own interest. In Table 3 we code this as "Top candidate is the 'other' gender," allowing for both the interaction and any level effects.

While the systematic move away from advancing the top-two candidates evident in (1) and (3) is itself a validation of the theory, in columns (2) and (4) we see even-stronger evidence that our experimental variation is moving subjects toward the theoretical prediction. Namely, when the top-ranked candidate is male but the subsequent decision will be made in a way that prefers female candidates (Column 2) experimental subjects are 21.5-percent less inclined to choose the top-two candidates (20.1 pp). Patterns are similar when the genders are flipped—when the top-ranked candidate is female but the subsequent decision will be made in a way that prefers male candidates, subjects are 23.3-percent less inclined to choose

[31] Recall, half of the sets of candidates in the female-preference setting included a top-ranked male candidate, a middle-ranked female candidate, and a bottom-ranked 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.

Table 4
Agent 1 behavior: Do subjects conform to theory?

	Agent 2 rewarded for choosing a female candidate (1)	Agent 2 rewarded for choosing a male candidate (2)
Panel A: *Do subjects skip over middling candidates to protect high flyers?*		
Sample restricted to:	$M_1F_2M_3$	$F_1M_2F_3$
$Y = 1$ if subject choses:	M_1M_3	F_1F_3
Treated = 1	0.123***	0.126**
	(0.038)	(0.051)
Mean (control)	0.007	0.037
Impact (%)	1757	341
Observations	772	752
Panel B: *Do subjects forward lower-ranked candidates than they need to?*		
Sample restricted to:	$M_1F_2F_3$	$F_1M_2M_3$
$Y = 1$ if subject choses:	M_1F_3	F_1M_3
Treated = 1	0.042*	0.069***
	(0.023)	(0.023)
Mean (control)	0.018	0.006
Impact (%)	233	1150
Observations	417	407

Notes: Observations are at the subject-by-question level. In all columns, the control group is all subjects who were in neither female- or male-treatment arms. Standard errors allow for clustering at the subject level and are reported in parentheses. Reported mean values are calculated only among subjects in the control group. *** p<0.01, ** p<0.05, * p<0.1.

Table 5
Do male and female subjects behave similarly with respect to the advancement of the top-two candidates?

	Top-two candidates advanced			
	Agent 2 rewarded for choosing a female candidate		Agent 2 rewarded for choosing a male candidate	
	(1)	(2)	(3)	(4)
Panel A: *Female Subjects*				
Treated = 1	-0.187***	-0.144***	-0.099*	-0.022
	(0.052)	(0.040)	(0.056)	(0.056)
Treated = 1 × Top candidate not preferred gender		-0.040		0.040
		(0.029)		(0.029)
Top candidate is the "other" gender		-0.052		-0.124**
		(0.040)		(0.048)
Mean (control)	0.96	0.96	0.96	0.96
Impact (%)	-19.5		-10.3	
Observations	1678	1678	1408	1408
Panel B: *Male Subjects*				
Treated = 1	-0.177**	-0.131*	-0.257***	-0.228***
	(0.066)	(0.077)	(0.064)	(0.071)
Treated = 1 × Top candidate not preferred gender		0.012		-0.012
		(0.036)		(0.036)
Top candidate is the "other" gender		-0.072		-0.039
		(0.045)		(0.052)
Mean (control)	0.95	0.95	0.95	0.95
Impact (%)	-18.6		-27.1	
Observations	1048	1048	1258	1258

Notes: In all columns, the dependent variable is equal to one if subjects choose the top-two candidates, and zero otherwise. Observations are at the subject-by-question level. In all columns, the control group is all subjects who were in neither female- or male-treatment arms. Standard errors allow for clustering at the subject level and are reported in parentheses. Reported mean values are calculated only among subjects in the control group. *** p<0.01, ** p<0.05, * p<0.1.

the top-two candidates (21.8 pp). Our experimental variation is clearly moving subjects toward an awareness of gender generally, and a willingness to re-order candidates systematically away from the simple strategy of forwarding the top-two candidates in a three-choose-two choice. In Table 3, then, we identify that the tendency to not choose a "top-two" strategy is stronger in the scenarios where the incentives of Agent 1 and Agent 2 subjects are misaligned—that is, where the gender of the top-ranked candidate and the gender that Agent 2 receives payment for are not the same.

Table 6
Do male and female subjects conform similarly to theory?

	Agent 2 rewarded for choosing a female candidate		Agent 2 rewarded for choosing a male candidate	
Panel A: *Do subjects skip over middling candidates to protect high flyers?*				
Sample restricted to:	$M_1F_2M_3$		$F_1M_2F_3$	
$Y = 1$ if subject chose:	M_1M_3		F_1F_3	
	(1)	(2)	(3)	(4)
Treated = 1	0.123***	0.163**	0.126**	0.149
	(0.038)	(0.070)	(0.051)	(0.095)
Treated = 1 × Female = 1		-0.067		-0.082
		(0.082)		(0.109)
Female = 1		0.014		-0.075
		(0.013)		(0.070)
Mean (control)	0.007	0.007	0.037	0.037
Impact (%)	1757		341	
Observations	772	772	752	752
Panel B: *Do subjects forward lower-ranked candidates than they need to?*				
Sample restricted to:	$M_1F_2F_3$		$F_1M_2M_3$	
$Y = 1$ if subject chose:	M_1F_3		F_1M_3	
	(1)	(2)	(3)	(4)
Treated = 1	0.042*	0.061	0.069***	0.093**
	(0.023)	(0.045)	(0.023)	(0.040)
Treated = 1 × Female = 1		-0.030		-0.056
		(0.052)		(0.043)
Female = 1		0.007		-0.016
		(0.024)		(0.015)
Mean (control)	0.018	0.018	0.006	0.006
Impact (%)	233		1150	
Observations	417	417	407	407

Notes: Observations are at the subject-by-question level. In all columns, the control group is all subjects who were in neither female- or male-treatment arms. Standard errors allow for clustering at the subject level and are reported in parentheses. Reported mean values are calculated only among subjects in the control group. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

In Table 4 we drill down on specific scenarios in which the theoretical predictions unambiguously suggests deviations from a "top-two" strategy. For treated subjects who anticipate the subsequent choice will be made by one who prefers female candidates, facing the three-choose-two problem of a top-ranked male, a middle-ranked female, and a bottom-ranked male—this we notate as $M_1F_2M_3$—puts their private interest against the anticipated interest of "Agent 2." The theoretical prediction is for treated subjects to forward the two male candidates. That is, Agent 1 may protect the top-ranked male by stepping over the F_2 candidate and only forwarding the two male candidates. Indeed, in Column (1) of Panel A we find treated subjects roughly eighteen times more likely to forward the two male candidates than are control subjects, who only choose M_1M_3 one percent of the time. Likewise, in the "male" treatment, Agent 1 facing a triple of $F_1M_2F_3$ might be inclined to protect their top pick from the anticipated favor to be given to male subjects in subsequent decisions—in Column (2) of Panel A we indeed see F_1F_3 chosen more than three times as often than in the treated group. Though the low base probabilities in the control group (0.007 and 0.037, respectively) contribute to larger relative increases, the impact is much larger in the "female" treatment, which suggests that individuals may be more likely to skip over a middling female and elevate a lower-ranked male, than they are to skip over a middling male and elevate a lower-ranked female.

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 the middle-ranking female candidates. And though the magnitude of response is smaller, we see a similar direction of response in protecting the top-ranking female candidates from middle-ranking male candidates when the subsequent decision maker is privately motivated to advance male candidates.

Yet, where there is no such opportunity to avoid choosing the favored gender we might anticipate no such pattern. In Panel B of Table 4 we therefore explore a second set of scenarios in which there is potential conflict in the interests of Agent 1 and Agent 2—$M_1F_2F_3$ for the "female" treatment and $F_1M_2M_3$ for the "male" treatment—they differ from those of Panel A insofar as a female candidate (male candidate) must now be included in the choice of Agent 1. A strict reading of the theory might have one anticipate that when facing the triple of $M_1F_2F_3$ in a "female" treatment, treated subjects would be found advancing M_1F_3, or when facing the triple of $F_1M_2M_3$ in a "male" treatment, treated subjects would be found advancing F_1M_3. In Panel B of Table 4 we see exactly this—treated subjects are twice as likely to advance M_1F_3 in response to female preference (though we lack precision in that point estimate) and more than eleven times more likely to advance F_1M_3 in

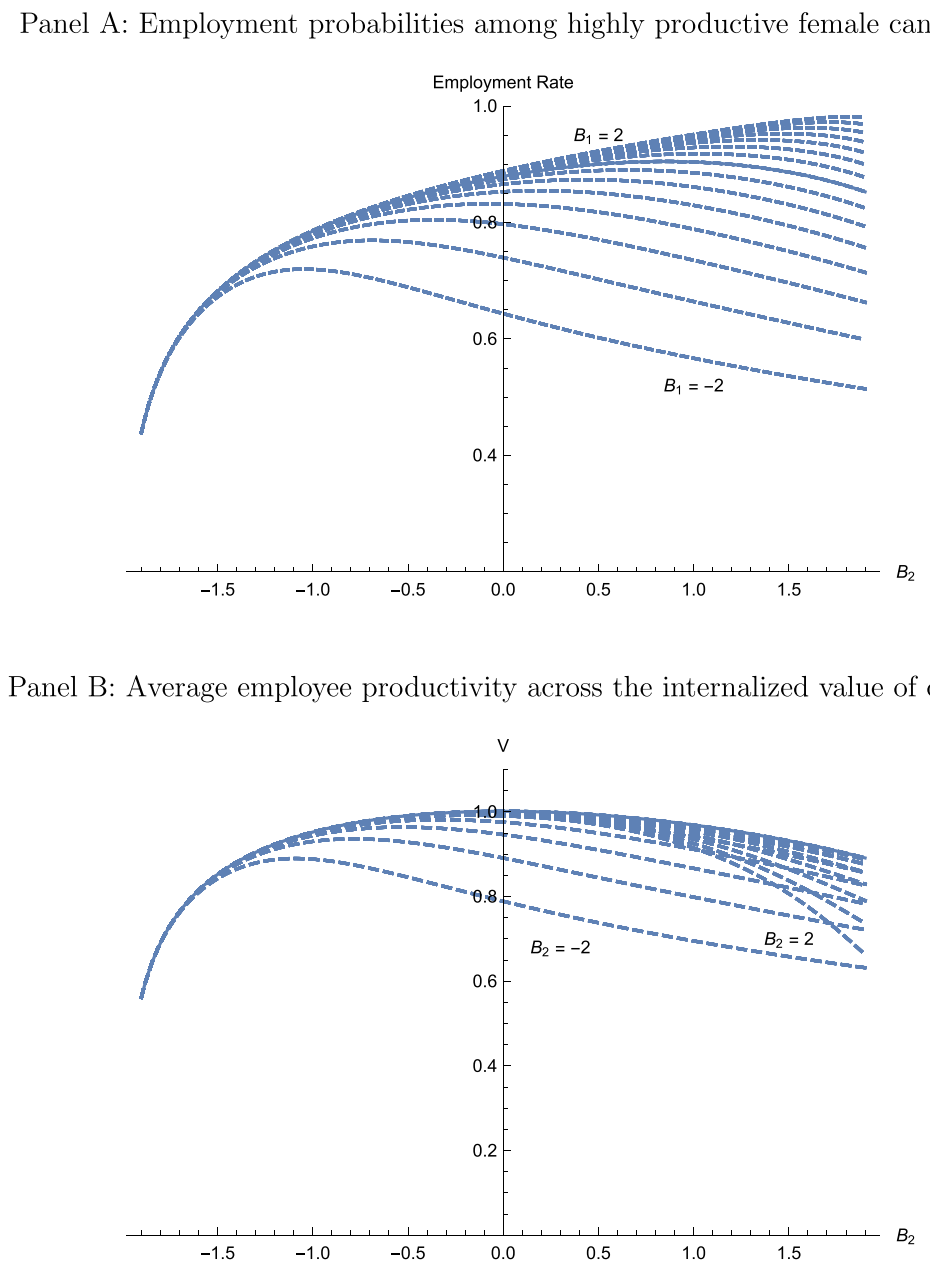

Fig. 7. Employment probabilities and employee productivity when Agent 2 is savvy *Notes:* Panel A indicates the probability that an H type with a particular diversity enhancing attributes will be hired by the firm. Panel B indicates the expected value to the firm from considering a randomly chosen candidate. 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_1 to $\tau_1 V_L \leq B_1 \leq \tau_1 V_H$, and B_2 to $\tau_2 V_L \leq B_2 \leq \tau_2 V_H$ in order to avoid either agent's decision collapsing on either "always advance" or "always reject." We normalize to one average productivity when $B_1 = B_2 = 0$.

response to male preference. This is consistent with subjects anticipating that increasing the score-gap between the two candidates increases the costs to Agent 2 exercising a preference for choosing a specific gender.[32]

[32] A traditional interpretation of experimenter-demand effects (Zizzo, 2010) would have Agent 1 subjects infer from the payment to Agent 2 subjects associated with choosing female candidates as a sign that they are to choose female candidates. This is inconsistent with the patterns evident in our data, as Agent 1 systematically avoids females (males) that could jeopardize males (females), but collapses on high-ability females (males) *when it is in their interest to do so*. We do not, therefore, imagine that the results we present are driven by experimenter-demand effects.

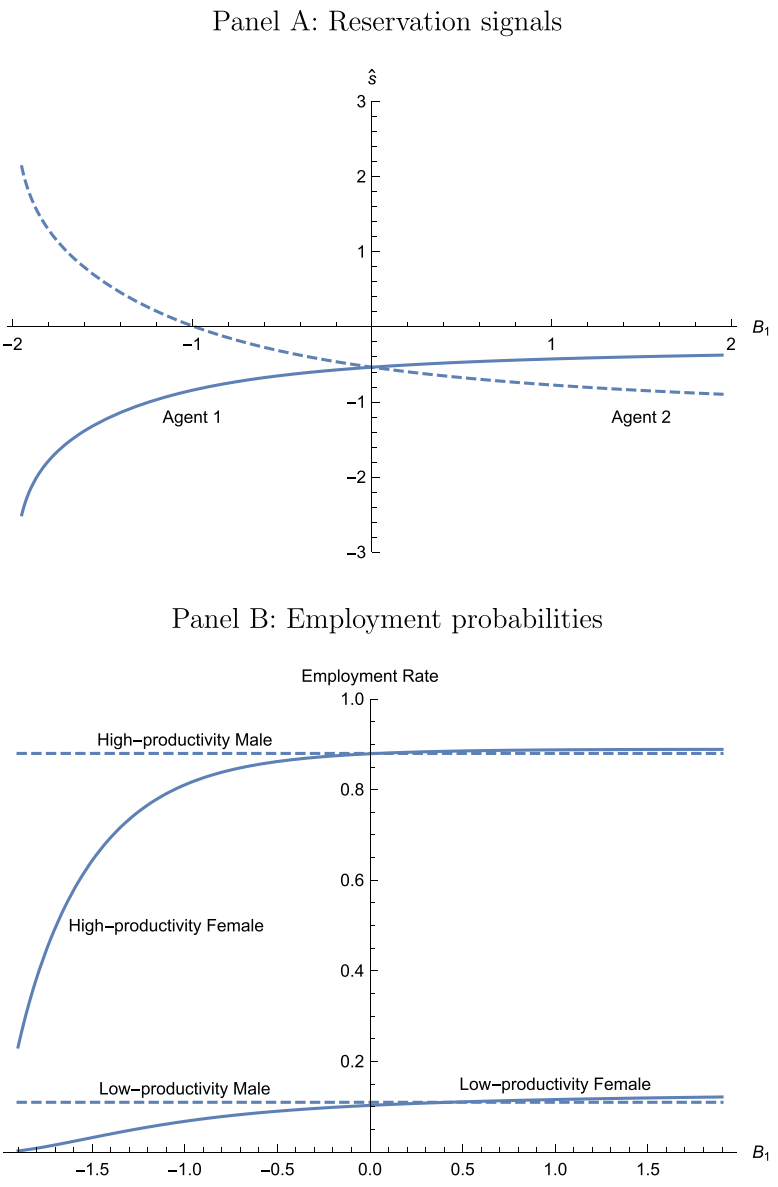

Fig. 8. Reservation signals and employment probabilities with bottom-up preferences ($B_2 = 0$, as we vary B_1) and a savvy Agent 2 *Notes:* Panel A shows the optimal reservation signal of productivity, above which a candidate who offers a diversity enhancing attribute will be advanced for further consideration 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. We assume here that $\tau_1 = \tau_2 = 0.5$, $\alpha = 0.5$, $V_L = -4$, $V_H = 4$, $\mu_H = 1$, $\mu_L = -1$, and $\sigma_H = \sigma_L = 1$. In addition, we limit the range of B_1 to $\tau_1 V_L \leq B_1 \leq \tau_1 V_H$ in order to avoid Agent 1's decision collapsing on either "always advance" or "always reject.".

5.4. Do male and female subjects behave differently?

Given the focus on gender throughout this paper, one may reasonably wonder whether male and female experimental subjects responded differently to treatment either generally or in particular scenarios. In Tables 5 and 6 we consider the role of subject gender both in terms of treatment and in the specific scenarios of interest. Specifically, in Table 5 we replicate Table 3 but stratify subjects by gender.[33] While the smaller number of observations limits the precision of Table 5, we find suggestive evidence that male and female subjects are reacting differently to treatment. Specifically, when the subsequent decision will be made by someone who is paid for choosing a female candidate, female experimental subjects appear more

[33] One subject reported their gender as "other" in the survey. They are not included in either Table 5 or 6.

responsive to treatment than are male subjects. Similarly, male subjects seem to be relatively more responsive to the anticipation that the subsequent decision will be made by someone paid for choosing a male candidate. Also of note, Column (4) of Panel A suggests that female subjects in the male-bonus arm of treatment are not responsive to treatment but are less likely to advance the top-two candidates if the top candidate is female. This may suggest that female subjects are anticipating favorable treatment of male candidates in the second stage even when the second agent has no monetary incentive to favor a male candidate.

In Table 6 we replicate earlier models for the specific scenarios we first reported in Table 4, but add additional columns in which we test whether female and male subjects respond differently within those specific scenarios. We find that male and female subjects behaved similarly in these key scenarios of interest although the consistent negative coefficients on the $\mathbb{1}(\text{Treated} = 1)$ x $\mathbb{1}(\text{Female} = 1)$ terms are again reflect that female subjects are somewhat less responsive to treatment.

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.

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 scores) 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 is made by someone who gets paid for choosing female or male candidates—the control groups are told nothing of the interest of the subsequent decision maker. Without knowledge of that incentive, the dominant strategy is to forward the two candidates with the highest signals of productivity. However, in both female- and male-preference treatments, we see sizable departures from this strategy—roughly 18- and 21-percent reductions in the likelihood of choosing that top-two strategy, respectively.

Moreover, we demonstrate a strong willingness among experimental subjects to protect top-ranked male candidates by skipping over middle-ranked female candidates—forcing the subsequent decision maker to choose between the top- and bottom-ranked male, as though they are removing the temptation for the second decision maker to choose the middle-ranked female over the top-ranked male. We see similar patterns under male preference—though the magnitude is far smaller—where subjects protect the top-ranked female candidates by terminating the candidacies of middle-ranked male candidates in favor of bottom-ranked females.

We also demonstrate a willingness among experimental subjects to protect the top-ranked male candidates by terminating the candidacies of the middle-ranked female candidates in favor of bottom-ranked females—forcing the subsequent decision maker to choose between the top male and a less-able female than was actually available. We see similar patterns under male preference—the protection of the top-ranked female candidates by terminating the candidacies of the middle-ranked male candidates in favor of bottom-ranked males. We also find suggestive evidence of differences across subject gender. Male experimental subjects appear more responsive in treatments in which male candidates preferred, while female experimental subjects appear more responsive in treatments in which female candidates preferred.

We conclude by noting 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, 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. This strategy is likely to be particularly effective when combined with recruitment materials

that explicitly signal interest in diversity (Flory et al., 2019). Alternatively, firms could rely more heavily on algorithmic assessments of productivity in the early stages of the review process. Even if these programs were not explicitly designed to favor particular groups, they would also not be expected to actively offset the expected preferences later in the process. Finally, firms could invest in ensuring that diversity goals were shared throughout the firm's hierarchy. By ensuring that all parties involved in the hiring process are invested in the firm's diversity goals, those goals can be achieved without penalizing high-productivity workers who also advance diversity objectives.

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. This suggests that efforts to increase diversity through hiring may be slower, and less effective, than would be efforts to limit that diversity. This difference will be particularly pronounced when those interests are late arriving within the candidate evaluation process.

7. Figures and tables

Declaration of Competing Interest

Authors declare that they have nothing to disclose–They have no relevant or material financial interests that relate to the research described in this paper. IRB approval was obtained from Grinnell College. They are aware of and abide by the principle of providing our original code and data.

Appendix A. Theoretical appendix

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.[A1]

Here we consider one important extension to the model—a potential transfer attached to the hiring of a candidate presenting a diversity enhancing personal attribute (from the firm, or Agent 2 as the residual claimant, to Agent 1). 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.[A2]

Such practice appears in academic markets, for example, where payments would typically be made by college-level administrators to departments conditional on hiring a candidate who presents with a 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. Specifically, Agent 2's objective can be written as,

$$\begin{aligned}\underset{\hat{s}_2,\rho}{\text{Max}}\, V_2(\hat{s}_2) = {} & \alpha[F_H(\mathbb{E}_2[\hat{s}_1]) + (1 - F_H(\mathbb{E}_2[\hat{s}_1]))F_H(\hat{s}_2)]\tau_2 V_0 \\ & + \alpha(1 - F_H(\mathbb{E}_2[\hat{s}_1])(1 - F_H(\hat{s}_2))(\tau_2(V_H - \rho) + B_2) \\ & + (1-\alpha)[F_L(\mathbb{E}_2[\hat{s}_1]) + (1 - F_L(\mathbb{E}_2[\hat{s}_1])F_L(\hat{s}_2)]\tau_2 V_0 \\ & + (1-\alpha)(1 - F_L(\mathbb{E}_2[\hat{s}_1]))(1 - F_L(\hat{s}_2))(\tau_2(V_L - \rho) + B_2) \end{aligned} \tag{A1}$$

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,

$$\begin{aligned}\underset{\hat{s}_1}{\text{Max}}\, V_1(\hat{s}_1) = {} & \alpha[F_H(\hat{s}_1) + (1 - F_H(\hat{s}_1))F_H(R_2)]\tau_1 V_0 \\ & + \alpha(1 - F_H(\hat{s}_1))(1 - F_H(R_2))(\tau_1(V_H - \rho) + B_1 + \rho) \\ & + (1-\alpha)[F_L(\hat{s}_1) + (1 - F_L(\hat{s}_1))F_L(R_2)]\tau_1 V_0 \\ & + (1-\alpha)(1 - F_L(\hat{s}_1))(1 - F_L(R_2))(\tau_1(V_L - \rho) + B_1 + \rho)\,. \end{aligned} \tag{A2}$$

[A1] 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.

[A2] 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, $\hat{s}_1^*$ approaches "always reject."

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, 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 Fig. 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.[A3]

We can also consider the optimal choice of ρ from the firm's perspective. Given the existence of some differential in the private values of the two decision makers (i.e., B_1 and B_2), 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, agents have similar incentives and their chosen thresholds are likewise equivalent.[A4]

An alternative approach, arguably, could be for the firm to hire an intermediary of sorts, similar to Cowgill and Perkowski (2020). While they address the interest of potential employees in a way that's outside of the scope of our model, we simply assume that employees always desire employment, they do demonstrate the agency concerns remaining in intermediary-type relationships, as the intermediary matchmaking firms place greater weight on the preferences of the firm than on the preferences of the employees.

Another possible solution would be for the firm to engage in a campaign to ensure that diversity goals received buy-in throughout the firm's hierarchy, possibly through changes in corporate leadership, or company culture, for example. To the extent that the firm can align both B_1 and B_2 with their own diversity goals, real progress can be made without sacrificing worker productivity.

Appendix B. Experimental appendix

[A3] There is evidence that factors other than direct incentives may be effective in shaping preferences (or a willingness to act on preference). For example, Bursztyn et al. (2020) shows that men in Saudi Arabia think of themselves as more supportive of women working than their neighbors. Interestingly, increases in female job applications followed the revelation of neighbors' true preferences.

[A4] 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 or some other algorithmic assessment instead of personal judgement in at least one stage of the process (Hoffman et al., 2015).

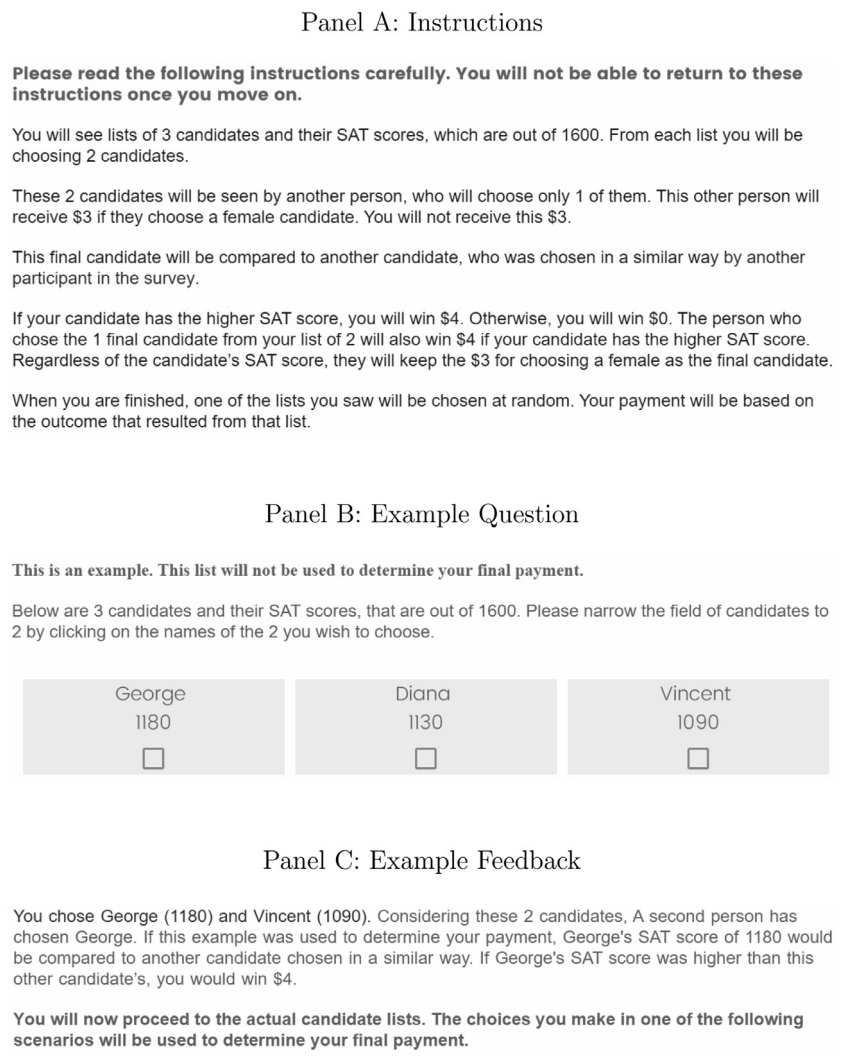

Fig. B1. Agent 1 instructions when "Agent 2" is paid to hire female candidates.

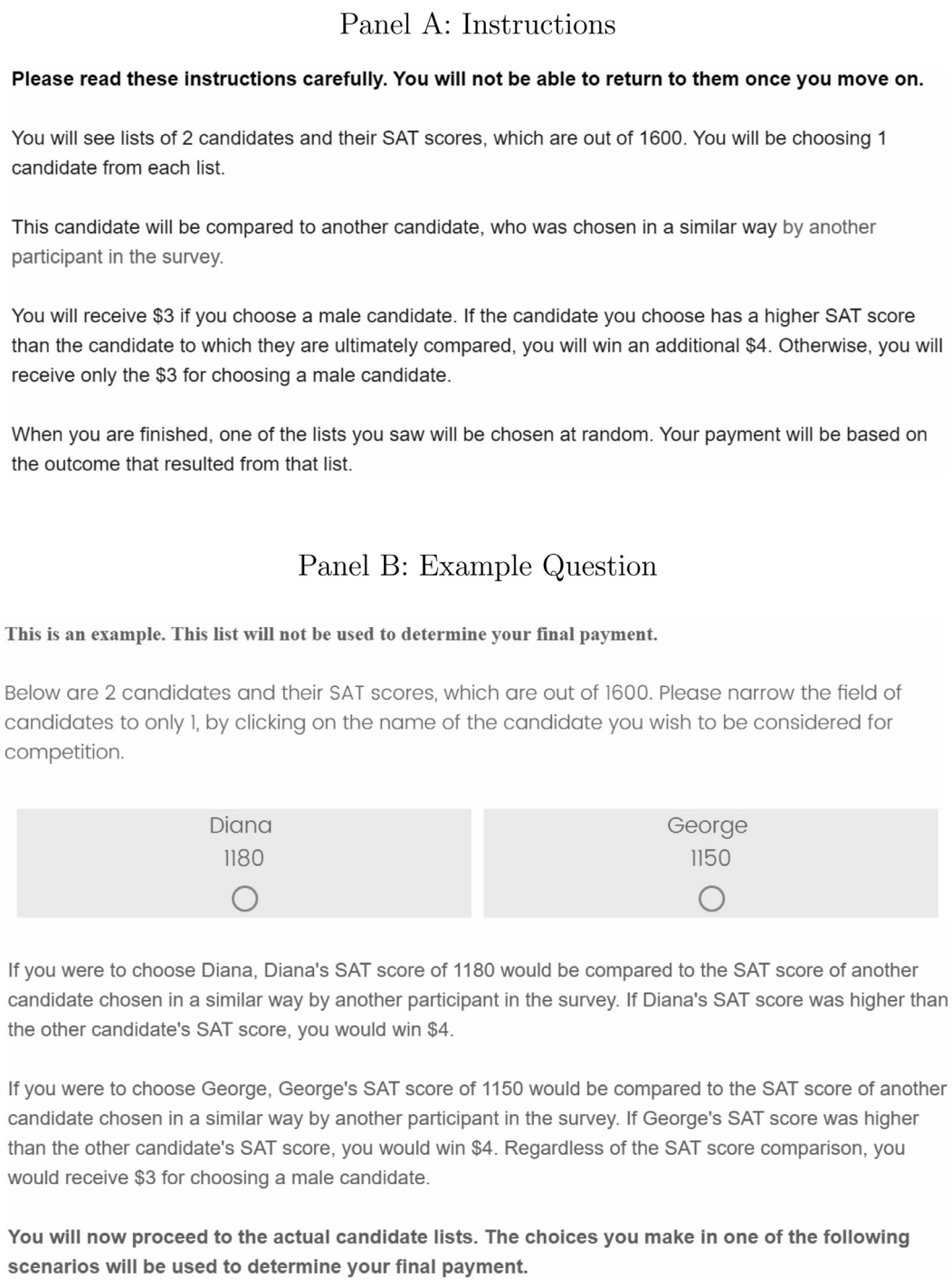

Fig. B2. Experiment example: Subjects acting as Agent 2.

References

Agan, A., Starr, S., 2017. Ban the box, criminal records, and racial discrimination: a field experiment. Q. J. Econ. 133 (1), 191–235.
Ahmed, S., 2007. 'You end up doing the document rather than doing the doing': diversity, race equality and the politics of documentation. Ethn. Racial Stud. 30 (4), 590–609.
Aigner, D.J., Cain, G.G., 1977. Statistical theories of discrimination in labor markets. Ind. Labor Relat. Rev. 175–187.

Altonji, J.G., Pierret, C.R., 2001. Employer learning and statistical discrimination. Q. J. Econ. 116 (1), 313–350.
Arrow, K., 1971. Some models of racial discrimination in the labor market.
Arrow, K., 1973. The theory of discrimination. Discrim. Labor Mark. 3 (10), 3–33.
Åslund, O., Skans, O.N., 2012. Do anonymous job application procedures level the playing field? ILR Rev. 65 (1), 82–107.
Bayer, A., Rouse, C.E., 2016. Diversity in the economics profession: a new attack on an old problem. J. Econ. Perspect. 40 (4), 221–242.
Bazerman, M.H., Tenbrunsel, A.E., Wade-Benzoni, K., 1998. Negotiating with yourself and losing: making decisions with competing internal preferences. Acad. Manag. Rev. 23 (2), 225–241.
Becker, G.S., 1957. The Economics of Discrimination. University of Chicago press.
Behaghel, L., Crépon, B., Le Barbanchon, T., 2015. Unintended effects of anonymous resumes. Am. Econ. J. Appl. Econ. 7 (3), 1–27.
Bertrand, M., Mullainathan, S., 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. Am. Econ. Rev. 94 (4), 991–1013.
Bjerk, D.J., 2008. Glass ceilings or sticky floors? statistical discrimination in a dynamic model of hiring and promotion. Econ. J. 118 (530), 961–982.
Bohnet, I., Van Geen, A., Bazerman, M., 2016. When performance trumps gender bias: joint vs. separate evaluation. Manage. Sci. 62 (5), 1225–1234.
Bradley, S. W., Garven, J. R., Law, W. W., West, J. E., 2018. The impact of chief diversity officers on diverse faculty hiring. NBER Working Paper 24969.
Breit, W., Horowitz, J.B., 1995. Discrimination and diversity: market and non-market settings. Public Choice 84, 63–75.
Bursztyn, L., González, A.L., Yanagizawa-Drott, D., 2020. Misperceived social norms: women working outside the home in saudi arabia. Am. Econ. Rev. 110 (10), 2997–3029.
Carlsson, M., Rooth, D.-O., 2012. Revealing taste-based discrimination in hiring: a correspondence testing experiment with geographic variation. Appl. Econ. Lett. 19 (18), 1861–1864.
Castillo, M., Petie, R., Torero, M., Vesterlund, L., 2013. Gender differences in bargaining outcomes: a field experiment on discrimination. J. Public Econ. 99, pp.35–48.
Cowgill, B., Perkowski, P., 2020. Agency and workplace diversity: evidence from a two-sided audit. working paper.
Doleac, J., Hansen, B., 2020. The unintended consequences of "ban the box"': statistical discrimination and employment outcomes when criminal histories are hidden. J. Labor Econ..
Eriksson, S., Lagerström, J., 2012. Detecting discrimination in the hiring process: evidence from an internet-based search channel. Empir. Econ. 43 (2), 537–563.
Ewens, M., Tomlin, B., Wang, L.C., 2012. Statistical discrimination or prejudice? a large sample field experiment. Rev. Econ. Stat. (0).
Exley, C. L., Kessler, J. B., 2019. The gender gap in self-promotion. NBER Working Paper.
Farber, H.S., Gibbons, R., 1996. Learning and wage dynamics. Q. J. Econ. 111 (4), 1007–1047.
Flory, J., Leibbrandt, A., Rott, C., Stoddard, O., 2019. Increasing workplace diversity: evidence from a recruiting experiment at a fortune 500 company. J. Hum. Resour..
Frankel, A., 2019. Selecting applicants. working paper.
Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. Am. Econ. Rev. 90 (4), pp.715–741.
Gould, E., 2019. Stark black-white divide in wages is widening further. Econ. Policy Inst.
Green, J.R., Laffont, J.-J., 1987. Posterior implementability in a two-person decision problem. Econometrica 55 (1), pp.69–94.
Guo, Y., Shmaya, E., 2019. The interval structure of optimal disclosure. Econometrica 87 (2), 653–675.
Guryan, J., Charles, K.K., 2013. Taste-based or statistical discrimination: the economics of discrimination returns to its roots. Econ. J..
Hegewisch, A., Tesfaselassie, A., 2019. The gender wage gap by occupation 2018. Inst. Women's Policy Res.
Hinchiffe, E., 2020. The number of female ceo's in the fortune 500 hits an all-time record. Fortune.
Hoffman, M., Kahn, L.B., Li, D., 2015. Discretion in hiring. Q. J. Econ.
Hsee, C.K., 1996. The evaluability hypothesis: an explanation for preference reversals between joint and separate evaluations of alternatives. Organ. Behav. Hum. Decis. Process. 67 (3).
Jacquemet, N., Yannelis, C., 2012. Indiscriminate discrimination: a correspondence test for ethnic homophily in the chicago labor market. Labour Econ..
Krause, A., Rinne, U., Zimmermann, K.F., 2012. Anonymous job applications of fresh Ph.D. economists. Econ. Lett. 117 (2), 441–444.
Kuhn, P., Shen, K., 2013. Gender discrimination in job ads: evidence from china. Q. J. Econ. 128 (1), 287–336.
Lang, K., Manove, M., 2011. Education and labor market discrimination. Am. Econ. Rev. 101 (4), 1467–1496.
Lewis, A.C., Sherman, S.J., 2003. Hiring you makes me look bad: social-identity based reversals of the ingroup favoritism effect. Organ. Behav. Hum. Decis. Process. 262–276.
Li, X., Hsee, C.K., 2019. Beyond preference reversal: distinguishing justifiability from evaluability in joint versus single evaluations. Organ. Behav. Hum. Decis. Process. 153, 63–74.
Luo, G.Y., 2002. Collective decision-making and heterogeneity in tastes. J. Bus. Econ. Stat. 20 (2), pp.213–226.
McCall, J.J., 1972. The simple mathematics of information, job search, and prejudice. In: Racial Discrimination in Economic Life, Lexington Books, pp. 205–224.
McGrit, E., 2017. 175 Ceo's join forces for diversity and inclusion. Fortune.
Murphy, K.J., 2013. Executive compensation: Where we are, and how we got there. In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of the Economics of Finance. Elsevier Science North Holland, Elsevier.
Phelps, E.S., 1972. The statistical theory of racism and sexism. Am. Econ. Rev. 659–661.
Pinkston, J.C., 2005. A test of screening discrimination with employer learning. Ind. Labor Relat. Rev. 59, 267.
Shaffer, V.A., Arkes, H.R., 2009. Preference reversals in evaluations of cash versus non-cash incentives. J. Econ. Psychol. 30 (6), 859–872.
Spence, M., 1973. Job market signaling. Q. J. Econ. 355–374.
Thomas, K.A., Clifford, S., 2017. Validity and mechanical turk: an assessment of exclusion methods and interactive experiments. Comput. Human. Behav. 77, 184–197.
Zizzo, D.J., 2010. Experimenter demand effects in economic experiments. Exp. Econ. 13 (1), 75–98.