Inscribing Diversity Policies in Algorithmic Hiring Systems: Theory and Empirics

Prasanna Parasurama* Panos Ipeirotis

Abstract

We study the downstream effects of diversity policies inscribed as algorithmic fairness constraints in Human+AI hiring systems. We present, solve, and empirically estimate a 2-stage hiring model consisting of: (1) a screening algorithm, which selects and shortlists candidates from a pool of applicants, and (2) an unbiased hiring manager, who hires from the shortlist. We consider the *equal selection* fairness constraint (i.e., a diversity policy), which constrains the screening algorithm to shortlist an equal number of men and women. We solve this model analytically and show that under certain conditions, even when both the algorithm and the hiring manager are unbiased, the fairness constraint can be ineffective in increasing the diversity of the hires. The *more* correlated the screening algorithm's and the hiring manager's assessment criteria are, the *less* effective the equal selection constraint becomes in increasing workforce diversity. Based on our theoretical findings, we propose a screening algorithm that would increase the effectiveness of the equal selection constraint. We empirically test our theoretical predictions using hiring data from eight IT firms and show via counterfactual policy simulation that the equal selection constraint in the shortlist can only improve the gender diversity of hires by a modest amount and not up to parity. We benchmark these results against our proposed algorithmic design and other commonly used fairness constraints and show the effectiveness of the proposed design in increasing workforce diversity.

*pparasurama@emory.edu

1 Introduction

In recent years, many firms have adopted various hiring diversity policies to increase the diversity of their workforce (Shi et al. 2018). One such policy is to diversify the shortlist/interview pool, sometimes called a "soft" affirmative action policy. These "soft" affirmative action policies aim to increase the share of minority candidates in the *initial* interview pool, without imposing any "hard" quota on the final hiring decision. This is in contrast with "hard" affirmative action policies such as hiring quotas, which are explicitly prohibited under the US employment law (Civil Rights Act of 1974) (Schuck 2002). An example of a soft affirmative action policy is the NFL's Rooney Rule[1], a hiring policy that requires the leagues to interview at least one ethnic minority for the head coach position. Various large tech firms including Patreon[2], Pinterest[3], and Facebook[4] have adopted similar policies in an effort to increase the racial and gender diversity of their workforce.

As hiring becomes increasingly aided by algorithms, inscribing such diversity policies in the form of algorithmic fairness constraints is also becoming a common practice. For example, in 2018, *LinkedIn Recruiter*[5] deployed a fairness-aware ranking algorithm to increase the gender diversity of candidates shown to recruiters (Geyik et al. 2019). However, showing recruiters a more diverse set of candidates does not necessarily translate to more diverse candidates being contacted, interviewed, or hired. Based on LinkedIn's own report, it is unclear whether improving gender representation in ranking improved *outcomes* (such as LinkedIn messages, interview requests) for underrepresented candidates (Geyik et al. 2019). After all, these algorithms do not work in isolation; ultimately, a human decision-maker such as a recruiter or hiring manager makes the final decision based on the algorithm's recommendations.

Prior lab studies have shown that the effectiveness of such fairness constraints depends vastly on the job type (Sühr et al. 2021; Peng et al. 2019). When these fairness constraints or diversity policies do not work, the conventional wisdom is to attribute the ineffectiveness to the bias of human decision-makers. Indeed, human bias can negate any fairness constraints in an algorithm, but there is little understanding of what *other* factors contribute to the effectiveness of fairness constraints.

[1] https://operations.nfl.com/inside-football-ops/diversity-inclusion/the-rooney-rule/
[2] https://www.slideshare.net/TarynArnold/patreon-culture-deck-april-2017
[3] https://newsroom.pinterest.com/en/post/our-plan-for-a-more-diverse-pinterest
[4] https://shorturl.at/bk1Xs
[5] LinkedIn's sourcing and talent search tool for recruiters.

To identify these factors in a structured manner, we first develop a 2-stage hiring model consisting of an algorithmic and human component. In the first stage of the model, a screening algorithm screens and shortlists candidates from a pool of applicants. There are more male than female applicants[6] (typical of firms that employ hiring diversity policies); we assume that both males and females have the same underlying quality distribution. To improve the gender diversity of the workforce, the algorithm has an *equal selection* constraint (i.e., a diversity policy) such that it shortlists an equal number of men and women. In the second stage, a hiring manager interviews the shortlisted candidates and hires the best candidate based on her own assessment.

We solve this model analytically and show that under certain conditions the equal selection constraint in the shortlist can be ineffective in increasing the gender diversity of hires even when the hiring manager is gender-unbiased. Importantly, we show that the effectiveness of the constraint decreases as the correlation between the algorithm's and the hiring manager's assessment criteria increases—i.e., the better the screening algorithm learns the hiring manager's assessment criteria, the less effective the fairness constraint becomes. Based on our theoretical findings, we propose a screening algorithm that is designed to be complementary to the hiring manager's assessment criteria, which can improve the effectiveness of the equal selection constraint in increasing workforce diversity.

We empirically estimate the model parameters and test the theoretical predictions on real-world hiring data from eight technology firms (799k applicants, and 3.6k job postings). We perform counterfactual policy simulation to understand the effectiveness of equal selection and show that: (1) In line with theoretical predictions, the equal selection constraint in the shortlist does not always lead to increased gender diversity in hires and, in some instances, does not affect the outcome. (2) The effectiveness of the constraint varies widely across job types, due to the difference in correlation in assessment criteria between the screening algorithm and the hiring manager across different jobs. We benchmark these results against our proposed complementary algorithmic design and other commonly used fairness constraints and show the effectiveness of the proposed design in increasing workforce diversity.

We make two main contributions to the literature on algorithmic fairness in pipelines. First, we

[6]For ease of presentation, we use the male vs. female hiring example, instead of framing it in "majority" and "minority" classes.

make a theoretical contribution that characterizes the effectiveness of the equal selection constraint in a hiring pipeline setting and show that the correlation in assessment criteria is a key determinant of the constraint's effectiveness. Second, we empirically estimate the effectiveness of the equal selection constraint in a hiring pipeline setting and propose an algorithmic design that can improve the constraint's effectiveness.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 introduces the theoretical hiring model. Section 4 presents the theoretical results. Section 5 proposes an algorithmic design to improve the effectiveness of algorithmic fairness constraints. Section 6 introduces the data and empirical modeling techniques. Section 7 presents the empirical results. Section 8 discusses the managerial implications of the results and concludes.

2 Related Work

This paper is related to the algorithmic fairness literature, which studies the design and evaluation of algorithms aimed to mitigate bias and improve fairness in algorithmic decision-making (Dwork, Hardt, et al. 2012; Zemel et al. 2013; Hardt et al. 2016; Zafar, Valera, Gomez Rodriguez, et al. 2017; Zafar, Valera, Rodriguez, et al. 2017; Geyik et al. 2019; Blum et al. 2022). In this literature, two broad notions of fairness exist: *individual fairness*, which requires that similar individuals are treated similarly by the algorithm; and *group fairness*, which requires that some statistic of interest is on average equal across groups along the lines of *protected attributes*.[7] Within group fairness, different definitions of fairness exist, such as demographic (or statistical) parity, equal selection, equal false-positive rates, equal false-negative rates, equal odds, equal accuracy rates, and equal positive predictive values across groups (see Table 7 for precise definitions and Mitchell et al. (2021) for a review). Except for in trivial cases, it is impossible to simultaneously satisfy all fairness criteria (Chouldechova 2017; Kleinberg, Mullainathan, et al. 2016), so the choice of fairness criteria depends on the context and is often informed by laws, policies, and desired outcomes.

Fairness constraints are not only used to mitigate any potential bias in the algorithm but can also be used as a tool to inscribe diversity policies that proactively correct for pre-existing societal

[7]Protected attributes are attributes that are protected under the law against discrimination. U.S. federal law prohibits employment discrimination based on race, gender, religion, national origin, age, disability, sexual orientation, and pregnancy.

and systemic bias. For example, in the hiring context, prior studies have shown that women preclude themselves from applying to male-dominated jobs because they anticipate discrimination in the hiring process (Storvik and Schøne 2008; Brands and Fernandez-Mateo 2017; Bapna et al. 2021). To address such pre-existing disparities, firms have adopted hiring diversity policies that increase or equalize the representation of minorities in the shortlist (Shi et al. 2018).[8] As hiring becomes increasingly aided by algorithms, these diversity policies can be inscribed in the form of algorithmic fairness constraints. Of particular interest is the *equal selection* fairness constraint (Khalili et al. 2021; Jiang et al. 2023), which requires positive outcomes to be equal across groups regardless of the proportions in the baseline population. For example, in algorithmic hiring, an equal selection constraint might require that the screening algorithm shortlists an equal number of men and women, regardless of the proportion of women in the applicant pool.[9]

Although these constraints guarantee fairness on algorithmic outputs, when these outputs are used as inputs in downstream decisions, the overall effects of these constraints in either mitigating bias or increasing diversity are not guaranteed. An emerging line of literature studies the efficacy of algorithmic fairness constraints in "pipelines"—i.e., settings where decisions are made sequentially. Bower et al. (2017) analyzes the equal opportunity constraint in a pipeline setting and shows that individually fair algorithms, when assembled sequentially, do not necessarily guarantee fair final outcomes with respect to equal opportunity. Similarly, Dwork and Ilvento (2019) analyzes the individual fairness constraint and conditional parity constraints in composition settings and show that individually fair algorithms, when composed together, do not necessarily guarantee fair final outcomes. Blum et al. (2022) propose a fair algorithm that satisfies the equality of opportunity constraint across the entire selection pipeline. Our main contribution to this algorithmic fairness and fair pipelines literature is that we study the *equal selection* constraint in a hiring pipeline setting, where decisions are made sequentially. We propose an algorithmic design to increase the effectiveness of the equal selection constraint and demonstrate its effectiveness using empirical hiring

[8]For example, the diversity hiring policies implemented in high-tech firms such as Facebook, Pinterest, Patreon, and *LinkedIn Recruiter's* ranking algorithm (Geyik et al. 2019) all seek to increase the representation of minorities in the shortlist.

[9]This is in contrast to *demographic parity*, another common fairness constraint in the algorithmic hiring setting, which requires the proportion of positive outcomes across groups to be equal to the proportions in a baseline population (Raghavan et al. 2020). For example, in algorithmic screening, demographic parity may require that the proportion of women on the shortlist be equal to the proportion of women in the applicant pool. Whereas demographic parity ensures that bias is not introduced in the hiring process, it does not correct for pre-existing disparities.

data.

Outside the algorithmic fairness literature, our work is also related to a number of theoretical papers that study bias and fairness in hiring settings. Kleinberg and Raghavan (2018) provide a theoretical hiring model in the presence of implicit bias and show that the Rooney Rule can increase the proportion of minority hires while also increasing the payoff of the decision-maker (see also Celis et al. (2021)). Fershtman and Pavan (2021) present a model to study the effect of "soft" affirmative action policies that increase the proportion of minority candidates in the candidate pool. Lee and Waddell (2021) study a 2-stage hiring setting with agents with different levels of interest in diversity and show that this difference can lower the likelihood of highly qualified candidates being hired even when they enhance diversity. Our contribution to this theoretical hiring literature is that we explicitly model the correlation in assessment criteria between the screener and the hiring manager, which we show to be a key determinant of the effectiveness of a common diversity policy.

3 Model Setup

Consider a hiring context in which a firm seeks to hire someone for a job. There are n_a applicants for the job and each applicant is characterized by their group membership $g \in \{m, f\}$, where m $(male)$ is the majority group and f $(female)$ is the minority group. The proportion p_a of the n_a applicants is female ($p_a < 0.5$), and $1 - p_a$ is male. Each candidate is also characterized by their true quality Q, which is unobserved at the time of hiring but revealed once hired (for example, job performance).

The hiring process consists of two stages. In the first stage, a screening algorithm estimates a quality score Q^S (a noisy estimate of the candidate's true quality Q) for each candidate, and shortlists candidates whose Q^S scores exceed the shortlist threshold τ^S—i.e., $y^S = \mathbb{1}\{Q^S > \tau^S\}$. Let p_s be the proportion of women in the shortlist. In the second stage, a hiring manager interviews the shortlisted candidates, comes up with her own estimate of quality Q^H, and hires candidates whose Q^H scores exceed the hiring threshold τ^H—i.e., $y^H = \mathbb{1}\{Q^H > \tau^H\}$. Let p_h be the proportion of women in the hired pool.

The firm prefers to maximize both the quality and the gender diversity of the hires, while still

Table 1: Table of notations

Symbol	Definition
n_a	Number of applicants
n_s	Number of candidates in the shortlist
n_h	Number of hires
p_a	Proportion of women in the applicant pool
p_s	Proportion of women in the shortlist
p_h	Proportion of women in the hired pool/finalist
Q	Candidate's true quality
Q^S	Screening score (screener's estimate of true quality)
y^S	Screening outcome
Q^H	Hiring manager score (hiring manager's estimate of true quality)
y^H	Hiring outcome
τ^S	Screening threshold
τ^H	Hiring manager threshold
θ	Correlation between Q^S and Q^H
θ^S	Correlation between Q^S and Q
θ^H	Correlation between Q^H and Q
δ	Gender difference in correlation $\theta_m - \theta_f$

Notes: Capital letters denote random variables, small letters denote variables, Greek letters denote parameters to estimate, $\hat{x}$ denotes the estimate of x. Generally, subscripts denote some subset of candidates. For example $Q_{s,m}$ denotes the quality (Q) of the shortlisted (s) male (m) candidates.

delegating the final hiring decision to the hiring manager, who is assumed to only maximize quality.[10] Putting a constraint on the hiring manager to hire an equal number of men and women would trivially increase the gender diversity of hires; however, such a constraint on the hiring manager would be considered a hiring quota, which is prohibited under US Employment Law (Title VII, Civil Rights Act of 1974).[11] As such, any firm-level policies intended to increase the diversity of hires must be inscribed in the screening algorithm.

We consider the *equal selection* constraint, which constrains the screening algorithm to shortlist an equal number of men and women.[12] Under equal selection, the algorithm may use different

[10] A high-profile example of this is Facebook's recruiting policy (Huet 2017). In 2016, Facebook implemented a point system for recruiters to source and recruit diverse candidates. Under this system, Facebook recruiters received 1 point for every new hire, and an additional bonus point if the new hire is diverse. Final hiring decisions were delegated to hiring managers, but they did not receive extra incentives to hire diverse candidates.

[11] This is the reason many diversity-focused hiring policies (e.g., Rooney Rule, Facebook's hiring policy (Huet 2017), LinkedIn's screening algorithm (Geyik et al. 2019)) target the initial screening decision rather than the final hiring decision.

[12] The main focus of this paper is on the equal selection constraint since it is a hiring policy that is widely used in practice. For example, the Rooney Rule, the diversity hiring policies implemented in high-tech firms such as Facebook, Pinterest, Patreon, and *LinkedIn Recruiter's* ranking algorithm (Geyik et al. 2019) all take some form of equal selection. In the empirical section, we benchmark the results against other fairness constraints, including

Table 2: Stages of the hiring model

Stage	Constraint	Selection Rule
(1)	Equal Selection $\mathbb{P}(g = f \mid y^S = 1) = \mathbb{P}(g = m \mid y^S = 1)$	$y^S = \begin{cases} 1 & \text{if } Q^S > \tau_m^S, g = m \\ 1 & \text{if } Q^S > \tau_f^S, g = f \\ 0 & \text{otherwise} \end{cases}$
(2)	None	$y^H = \begin{cases} 1 & \text{if } Q^H > \tau^H \\ 0 & \text{otherwise} \end{cases}$

Notes: y^S and y^H are binary indicators of selection in the screening and hiring stages, respectively. In the screening stage, the thresholds τ_m^S and τ_f^S are gender-specific and are set such that the equal selection constraint is satisfied (i.e., an equal number of men and women are shortlisted). In the hiring stage, there is no constraint, and the threshold τ^H is the same for both men and women.

shortlist thresholds for men (τ_m^S) and women (τ_f^S), such that an equal number of men and women are shortlisted. In the second stage, the hiring manager uses a common hire threshold τ^H for both men and women. The selection rule for each stage is summarized in Table 2.

We jointly model the quality scores Q^S, Q^H, Q as random variables drawn from a multivariate Gaussian distribution[13], with location parameter $\boldsymbol{\mu} = [\,\mu^Q \;\; \mu^{Q^S} \;\; \mu^{Q^H}\,]$ and covariance $\Sigma = \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}$, where Σ is positive semi-definite. Male and female applicants can have different quality distributions. Without loss of generality, we fix the location parameter for men at $\boldsymbol{\mu}_m = 0$. Table 3 summarizes what each parameter in the model represents, and what assumptions we make about them.

$$
\begin{aligned}
(Q_m, Q_m^S, Q_m^H) &\sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta_m^S & \theta_m^H \\ \theta_m^S & 1 & \theta_m \\ \theta_m^H & \theta_m & 1 \end{bmatrix} \right) \\
(Q_f, Q_f^S, Q_f^H) &\sim \mathcal{N}\left(\begin{bmatrix} \alpha & \alpha + \beta^S & \alpha + \beta^H \end{bmatrix}, \begin{bmatrix} 1 & \theta_f^S & \theta_f^H \\ \theta_f^S & 1 & \theta_f \\ \theta_f^H & \theta_f & 1 \end{bmatrix} \right)
\end{aligned}
\tag{3.1}
$$

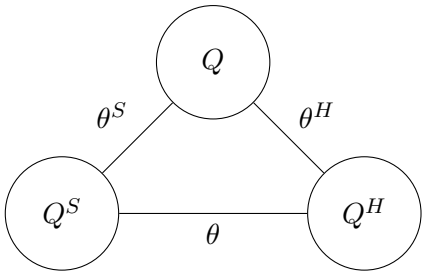
demographic parity, conditional demographic parity, equalized odds (equal opportunity), and error rate parity.

[13] In Appendix C, we empirically show that the Gaussian distribution captures the dependency structure between the scores well.

Table 3: Model parameters, definitions, and assumptions

Parameter	Definition	Assumption
θ^S	Correlation between Q and Q^S; measure of how good the screening algorithm is at predicting true quality	$\theta^S \in [0,1]$
θ^H	Correlation between Q and Q^H; measure of how good the hiring manager is at predicting true quality	$\theta^H \in [0,1]$
θ	Correlation between Q^S and Q^H; degree to which the screening algorithm and the hiring manager agree in their quality assessment	$\theta \in [0,1]$
$\delta \coloneqq \theta_f - \theta_m$	Gender difference in assessment criteria between the screener and the hiring manager	$\delta = 0$ (for now)
$\delta^S \coloneqq \theta_f^S - \theta_m^S$	Predictive gender bias of screening scores	$\delta^S = 0$
$\delta^H \coloneqq \theta_f^H - \theta_m^H$	Predictive gender bias of the hiring manager scores	$\delta^H = 0$
α	Mean quality difference between men and women; positive α implies women have higher mean quality than men	$\alpha = 0$. We extend the model in the Online Appendix A.4
β^S	Systematic gender bias of screening scores	$\beta^S = 0$
β^H	Systematic gender bias of hiring manager scores	$\beta^H = 0$

Figure 1: The correlation structure between Q, Q^S and Q^H



4 Theoretical results

In this section, we analyze how the gender diversity of hires (p_h) and the expected quality of hires ($E[Q_h]$) varies as a function of the firm's design parameters. Not all model parameters are design parameters that can be controlled by the firm. For a given candidate, Q is fixed, and estimation of Q^H is delegated to the hiring manager, which fixes θ^H. The firm has control over the screening

algorithm, and thus how Q^S is estimated. Therefore, the design parameters that the firm can control are θ^S (i.e., how good the screening algorithm is in predicting true quality), and θ (how similar the screening algorithm is compared to the hiring manager in assessing quality).

We show three main results: (1) Under equal selection, gender diversity of hires decreases with the correlation parameter θ, (2) Conditional on θ^S and θ^H, the expected quality of hires decreases with θ. (3) Gender diversity of hires decreases with gender difference in assessment criteria between the screener and hiring manager, δ. All proofs are provided in Appendix A.

Proposition 1. *Under equal selection, the female proportion of hires decreases with the correlation between screening and hiring manager scores, θ. This result is independent of θ^S and θ^H.*

Corollary. *When screening and hiring manager scores are perfectly uncorrelated, the equal selection constraint effectively balances the gender proportion of hires. In contrast, equal selection has no effect on the gender proportion of hires when the scores are perfectly correlated.*

Proposition 2. *Conditional on θ^S and θ^H, the expected quality of hires decreases with the correlation between screening and hiring manager score, θ.*

So far, we have assumed that the correlation in assessment criteria between the screener and hiring manager is the same for both men and women. We now extend this model to allow for different assessment criteria using different correlation parameters, θ_m, and θ_f, for male and female candidates, respectively. We parametrize this difference using δ, where:

$$\delta := \theta_m - \theta_f \tag{4.1}$$

Proposition 3. *The female proportion of hires decreases with the gender difference in the correlation parameter, δ. This result is independent of θ^S and θ^H.*

5 Implications for screening algorithm selection and design

The theoretical analyses provide insights for the selection and design of the optimal screening algorithm that maximizes both the expected quality of hires and the effectiveness of the equal selection constraint (i.e., diversity). As we have shown: (1) the expected quality of hires is increasing

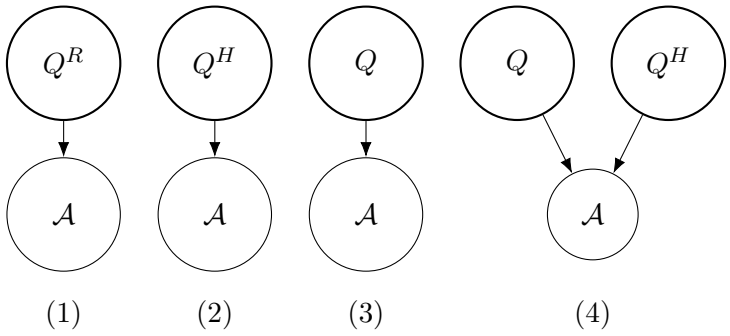
with θ^S, (2) the proportion of female hires decreases with θ under equal selection, and (3) conditional on θ^S, the expected quality of hire decreases with θ. This implies that to maximize hire quality *and* diversity, the firm should select a screening algorithm that has high θ^S and low θ. In other words, the firm should select a screening algorithm that is good at predicting true quality, but different from the hiring manager's assessment—i.e., the screening algorithm should be *complementary* to the hiring manager.

Consider, for example, the case where the firm is selecting a screening algorithm from two different vendors. Both algorithms have the same performance in predicting true quality—i.e., $\theta^S_{\mathcal{A}_1} = \theta^S_{\mathcal{A}_2}$, but different correlations with the hiring manager's scores—i.e., $\theta_{\mathcal{A}_1} \neq \theta_{\mathcal{A}_2}$. This can happen, for example, when the two algorithms use different feature sets to predict candidate quality. Which algorithm should the firm choose? The firm should choose whichever algorithm has the lower correlation with the hiring manager scores. Even though both algorithms have the same performance in predicting true quality, the one with the lower θ will have less redundant information about true quality, leading to a higher expected quality of hires. More importantly, the algorithm with the lower θ will also be more effective in increasing workforce diversity under equal selection.

5.1 Simultaneously maximizing θ^S and minimizing θ with fixed information

In the above example, two algorithms have independent θ^S and θ. We now consider the case, where the firm designs its own screening algorithm with fixed information (e.g., training the screening algorithm using resumes), where there may be a tradeoff between θ^S and θ. To show how the firm can optimize for both diversity and quality of hires by adjusting θ and θ^S, we first consider all the available design options for training the screening algorithm.

Figure 2: Target variable options for training the screening algorithm



(1) The first option is to train the screening algorithm with the historical recruiter's scores/decisions Q^R as the target variable (i.e., decisions of human screeners that used to do the job of the algorithmic screener). The advantage of this approach is that the number of observations to train the algorithm will be large and uncensored (i.e., the firm observes all applicants who applied and the recruiter's decision for each applicant). The disadvantages are: (a) Q^R is only a proxy of Q, so θ^S may not be maximized directly, and (b) the firm will have no control over the effectiveness of the equal selection constraint, as it partly depends on how similar the recruiter's and hiring manager's assessment criteria are (i.e., the firm will have no control over θ).

(2) The second option is to train the algorithm with the hiring manager's scores/decisions Q^H as the target variable. The equal selection constraint will not be very effective in this case, since the algorithm is designed to be similar to the hiring manager (i.e., θ is maximized by design).

(3) The third option is to train the algorithm with true quality Q (e.g., job performance) as the target variable. Although this will maximize the screening performance in predicting quality (i.e., maximizes θ^S), the firm will have no control over the effectiveness of the equal selection constraint, since it will have no control over θ.

(4) The last option is to combine the above approaches and train the algorithm with both true quality Q and the hiring manager's scores/decisions Q^H as the target variables. As we have shown, to optimize both the expected quality of hire and the effectiveness of equal selection, the screening algorithm should be complementary to the hiring manager—i.e., the screening algorithm should be good at predicting true quality, but different from the hiring manager's assessment. One way to achieve this is through multi-objective learning (e.g., with adversarial learning. See Zhang et al. (2018)), where the model is trained to learn screening scores that are good at predicting Q but poor at predicting Q^H. The model trades off the two objectives to learn a scoring function that maximizes θ^S and minimizes θ. Indeed, because the information is fixed, the θ^S cannot be adjusted independently of θ. Once θ^S is maximized, making the screening algorithm dissimilar to the hiring manager (i.e., minimizing θ) necessitates removing information that is predictive of hiring manager scores. Insofar that the hiring manager scores are predictive of true quality (which is a function of θ^H), θ will be weakly decreasing with θ^S once θ^S is maximized.

To formalize this, we fix the conditional entropy of true quality given the screening score and

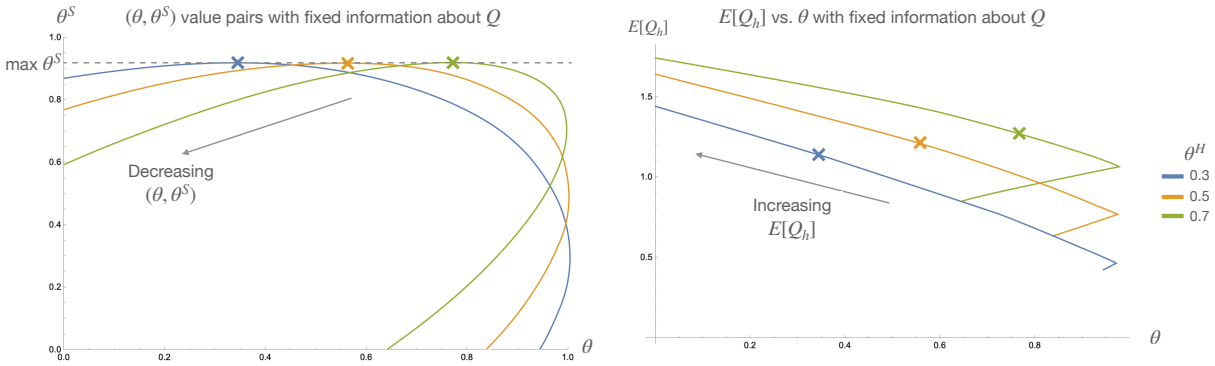
hiring manager score, $H(Q|Q^S, Q^H)$.

$$H(Q|Q^S, Q^H) = \frac{1}{2} \cdot \log\left(\frac{2e\pi \cdot Det\left(\begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}\right)}{1-\theta}\right) \tag{5.1}$$

The conditional entropy is a measure of information about true quality given the screening score and hiring manager score. The left panel in Figure 3 plots the (θ, θ^S) value pairs that yield the same conditional entropy, $H(Q|Q^S, Q^H)$, while holding θ^H fixed. Once θ^S is maximized ("x" marker), decreasing θ further (moving leftward) comes at the cost of also decreasing θ^S.

Next, we can ask how simultaneously changing (θ^S, θ) values affects the expected quality of hires, $E[Q_h]$ while holding the conditional entropy of Q fixed. The right panel plots $E[Q_h]$ as a function of θ using the equal-information (θ, θ^S) value pairs. Decreasing θ affects the expected quality of hires via two channels: (1) there is a direct effect of θ, where decreasing θ increases $E[Q_h]$, and (2) there is an indirect effect via θ^S, where decreasing θ also decreases θ^S, which in turn decreases $E[Q_h]$. Interestingly, the net effect of decreasing θ still increases $E[Q_h]$ even though θ^S is simultaneously decreasing—meaning that the direct increase in $E[Q_h]$ due to the decrease in θ offsets the indirect decrease in $E[Q_h]$ due to decreasing θ^S.

Figure 3: Simultaneously optimizing θ^S and θ with fixed information about Q



Notes: The left panel plots the equal-information (θ, θ^S) value pairs that yield the same conditional entropy $H(Q|Q^S, Q^H) = 0.5$. The "x" marks the point where θ^S is maximized. The right panel plots the expected quality of hire, $E[Q_h]$, as a function of θ using the equal-information (θ, θ^S) value pairs. θ^H is held fixed at $\theta^H \in \{0.3, 0.5, 0.7\}$.

5.2 Directly minimizing gender difference in Q^H scores in the shortlist

Another design option is to *directly* minimize the gender difference in Q^H scores in the shortlist rather than minimizing it through θ. One way to implement this would be to train two predictors—one for Q and another for Q^H. The screening algorithm then shortlists candidates that have the highest predicted Q scores, while minimizing the gender difference in Q^H scores in the shortlist—for example, by shortlisting a pair of male and female candidates that have high predicted Q scores but roughly the same predicted Q^H scores, so both men and women are equally likely to be hired once shortlisted. Compared to an algorithm that only uses predicted Q to shortlist candidates, this algorithm will have a lower θ (i.e., relatively more complementary), however, it will not be minimized to 0. Therefore, this algorithm will not maximize the expected quality of hires to the fullest extent (since θ is not minimized to 0), but will still maximize diversity under the equal selection constraint while being much simpler to implement.

6 Empirical modeling

Theoretical analyses in the previous sections show how the effectiveness of the equal selection constraint depends on key parameters, which in turn inform us how to select and design the optimal screening algorithm. Although we have shown that it is optimal to train the screening algorithm on both true quality and the hiring manager's estimate of quality, it's not typically what's done in practice. Typically, screening algorithms are trained on the historical decisions of human screeners/recruiters because the training data is abundant and readily available. For example, *LinkedIn Recruiter's* recommendation algorithm is trained on the recruiter's decisions (Geyik et al. 2019). Similarly, firms may choose to train their screening algorithm on the recruiter's decisions rather than true quality (e.g., job performance) if the training data for the latter is scarce.[14] If a firm were to train the screening algorithm on the recruiter's decisions, how effective the equal representation constraint would be in increasing diversity is an empirical question. The effectiveness of equal selection depends on the values of the parameters discussed thus far, which in turn depends on how screeners and hiring managers make hiring decisions in the real world, how correlated their

[14]Firms observe recruiter decisions for all applicants, whereas they observe job performance only for candidates that were hired. Within our data, we observe that the hire rate is between 1-5% across firms, meaning that the training data to predict recruiter decisions will be 20-100x larger than training data to predict job performance.

assessment criteria are (θ), and how much the assessment criteria differs between men and women (δ)—all of which are empirical questions.

In this section, we introduce real hiring data into the model to estimate the model parameters for each job and use these parameters to estimate the effectiveness of the equal selection constraint across jobs using counterfactual policy simulation. We then benchmark these results against our proposed complementary screening algorithm and other commonly used fairness criteria.

6.1 Data

We use Applicant Tracking System (ATS) data from eight technology firms based in the U.S. These firms are clients of an HR analytics software provider, who provided us with the aggregated ATS data. The ATS keeps a detailed record of all the applicants, their characteristics (gender, years of experience, etc.), the applicant's resume, the job posting to which they applied and the corresponding business unit, and the application outcome (0/1) in each stage of the hiring funnel. We have data on 799k applicants (60% male, 40% female) across 3,608 job postings. We categorize each job posting into the following categories based on the business unit of the job. We only consider external applicants (i.e., outside applicants who applied for a job).

Table 4: Number of applicants and job postings by job category

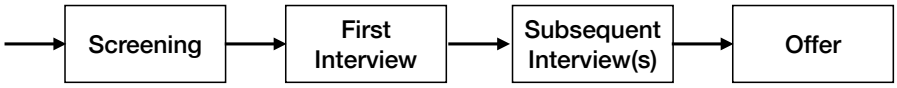
Job Category	N Applicants	N Jobs
Engineering & Technical	214,943	1,178
Product & Design	130,669	534
Sales & Marketing	92,559	391
Legal & PR	75,955	332
Other	70,864	53
Finance & Accounting	69,536	308
Biz Dev & Operations	51,523	299
HR	48,122	246
Customer Service & Acct Management	42,199	238
Overall	799,108	3,608

Each firm may have slightly different hiring processes, but because we only consider external applicants, they all fit the following canonical hiring process: Screening, First Interview, Subsequent Interviews, and Offer.[15] The first stage is the screening stage, where the screener reviews the

[15]Although this multistaged hiring process deviates from our 2-stage hiring model, it is sufficient to only examine

applications received. If candidates pass the screening stage, they move on to the first interview. If they clear that stage, they move on to subsequent interviews and finally to the offer stage.

Figure 4: Hiring funnel



For an average job, 233 applicants apply, of which 36 pass the initial screening and are shortlisted for the first interview. After the first interview, 7 finalists move on to subsequent interviews, of which 2 receive an offer (there can be multiple vacancies per job posting).

6.2 Parameter estimation

Estimation of screening and hiring manager scores. To estimate parameters θ and δ, we first need the screening and hiring manager scores for each candidate.[16] In the hiring data, we only observe the binary decisions of the screener and the hiring manager. We can exploit the variation in the screener and hiring manager's binary decisions and recover the latent screening and hiring manager scores, respectively. Formally, consider the binary choice decision $y_{i,j}$ faced by the screener/hiring manager when assessing candidate i for job j.

$$y_{i,j} = \begin{cases} 1 & q > \tau_j \\ 0 & otherwise \end{cases} \tag{6.1}$$

$$\mathbb{P}(y_{i,j} = 1) = \hat{p}_{i,j} = f(X_i, X_j, \epsilon_{i,j}) \tag{6.2}$$

the following two stages: (1) the stage with the equal selection constraint (Screening) and (2) the stage after the equal selection constraint (First Interview). To see why this is the case, consider an example where the applicant pool is 70% male and 30% female. Under equal selection, the proportion of male/female candidates after the screening stage will be 50/50. In Interview 1, the hiring manager can "undo" some of this constraint resulting in 60/40, for example. As long as the selection process in the subsequent stages is unbiased (which is the assumption we make in the theoretical model), the same proportion of 60/40 will be maintained in the rest of the hiring funnel.

[16]The reason we cannot estimate the parameters with just the binary decisions is that we only observe hiring manager decisions for applicants who passed screening. This means we won't have any variation in screening decisions (all 1s). Instead, we need scores from the screener and hiring manager for *all* applicants (even those that did not pass screening) to calculate the correlation between the two.

The decision-maker makes a positive decision $y_{i,j} = 1$ if the applicant's quality score exceeds the corresponding threshold, or makes a negative decision otherwise. The decision is a function of applicant characteristics X_i, job characteristics X_j as well as some idiosyncratic error $\epsilon_{i,j}$. In the hiring data, we observe the binary decisions $y_{i,j}$, applicant characteristics X_i (resume text), and job characteristics X_j (job description text) for each applicant-job pair. We can use this information and train two ML models, one for the screener and one for the hiring manager, to flexibly estimate the decision function f.

Since we only observe hiring managers decisions for candidates that were shortlisted, we correct for selection bias by re-weighting the observations by the inverse of the probability of being shortlisted ($\frac{1}{\mathbb{P}(y^S_{i,j}=1)}$). We discuss these models in detail in Section 6.3.

To map the probability estimates to quality scores, we use a Gaussian copula transformation.[17] Let $\hat{p}^S_{i,j}$ be the predicted probability of being shortlisted for candidate i in job posting j, and let $\hat{\boldsymbol{p}}^S_j$ be the list of predicted probabilities for all candidates in job posting j. The screening score $\hat{q}^S_{i,j}$, and similarly, hiring manager score $\hat{q}^H_{i,j}$ for candidate i in job posting j are given by:

$$\hat{q}^S_{i,j} = Quantile(\hat{p}^S_{i,j}, \hat{\boldsymbol{p}}^S_j) \tag{6.3}$$

$$\hat{q}^H_{i,j} = Quantile(\hat{p}^H_{i,j}, \hat{\boldsymbol{p}}^H_j) \tag{6.4}$$

Estimation of θ. We estimate the correlation parameter $\hat{\theta}_j$ for job posting j using the Spearman rank correlation coefficient.[18] We calculate the average by taking a weighted sum of the parameters

[17] Gaussian copulas are multivariate Gaussian distributions, whose marginals are uniformly distributed. They offer a flexible way to disentangle multivariate Gaussian distribution as a product of uniform marginal distributions and a Gaussian copula that "couples" them (See Joe (2014) and Nelsen (2007) for a reference on copulas). Formally, the joint empirical distribution of quality scores $(\hat{q}, \hat{q}^S, \hat{q}^H)$ has CDF $F_{\hat{q},\hat{q}^S,\hat{q}^H}(x, y, z; \Sigma) = C(F_{\hat{q}}(x), F_{\hat{q}^S}(y), F_{\hat{q}^H}(z))$. Here, C is the 3-dimensional Gaussian copula, $C(u, v, k) = \Phi(\Phi^{-1}(u), \Phi^{-1}(v), \Phi^{-1}(k))$, and Φ is the CDF of a multivariate Gaussian distribution. This transformation ensures that we stay close to the theoretical model, which assumes that the quality scores have a multivariate normal distribution. We show that the Gaussian copula has good goodness-of-fit measures on our empirical data compared to other copulas. See Appendix C.

[18] We are using Spearman correlation because we are performing a copula transformation on the scores. Spearman rank correlation is robust to increasing transformations and is more commonly used in practice than Pearson correlation (Xiao and Zhou 2019). Our results are similar if we use Pearson correlation.

(weighted by the size of applicant pool n_a) across job postings in the hold-out test set.

$$\hat{\theta}_j = Spearman(\hat{\boldsymbol{q}}_j^S, \hat{\boldsymbol{q}}_j^H) \tag{6.5}$$

$$\hat{\theta}_{avg} = \sum_j \frac{n_{a,j}}{n_a} \hat{\theta}_j \tag{6.6}$$

Estimation of δ. To estimate δ, we calculate $\hat{\theta}$ for male and female applicants separately for each job posting j in the hold-out test set and take the difference.

$$\hat{\delta}_j = \hat{\theta}_{j,m} - \hat{\theta}_{j,f} \tag{6.7}$$

$$\hat{\delta}_{avg} = \sum_j \frac{n_{a,j}}{n_a} \hat{\delta}_j \tag{6.8}$$

6.3 ML models for screener and hiring manager

To train our resume screening and hiring manager models, we use a state-of-the-art deep learning model for text classification called BigBird[19] (Zaheer et al. 2020). BigBird uses a variant of the popular BERT-style transformer architecture and is optimized for long documents[20] (Vaswani et al. 2017; Devlin et al. 2019).

For input data, we concatenate the resume text with job characteristics (company name, job name, business unit, employment type, location, skills, and keywords), and feed the concatenated text as a single document to the model. We get the company name, job name, business unit, and location directly from the ATS. For skills and keywords, we use a dictionary of skills that was created in a separate analysis by aggregating all the skills and keywords listed in the skills section of LinkedIn profiles. We then concatenate this text with the resume text to create a single document.

Since all the parameters of interest are at the job level, we split the dataset by stratifying on job postings. We randomly take 80% of the job postings for the training set, 10% for the evaluation set, and 10% for the hold-out test set. All the parameters are estimated on the hold-out test set. This ensures that the parameters are evaluated on a test set that the model has never seen before.

[19] https://github.com/google-research/bigbird

[20] The classic BERT architecture uses a self-attention mechanism that scales quadratically with the number of tokens in the document. This makes using BERT for long documents such as resumes infeasible due to memory and computational footprint. BigBird overcomes this by using a sparse attention mechanism that scales linearly with the number of tokens in the document.

We follow Sun et al. (2019) for initial hyperparameters and fine-tune them by making small adjustments. The following parameters yield the best results based on the area under ROC criteria on the evaluation set: `Epochs=3, Batch Size=14, Learning Rate=2e-5, Weight Decay=2e-5`.

For the screening model, the AUC score is 0.83, and there is no difference in ROC curves between male and female candidates. For the hiring manager model, the AUC score is 0.68, and there is no difference in ROC curves between genders. We provide more details on the predictive performance of the models in Appendix B.

Inverse propensity weighting of observations. For the screening model, we use the full dataset. After stratifying on job postings, the size of the training/evaluation set is 725,351, and the hold-out test set is 73,757. For the hiring manager model, we can only use the subset of candidates that were shortlisted (we don't observe interview outcomes for the candidates that did not get shortlisted). To correct for this selection issue, we re-weight the observations by the inverse of the probability of being shortlisted—i.e., we give more weight to the candidates that have a low estimated probability of being shortlisted, and less weight to candidates that have a high estimated probability of being shortlisted. So long there is noise/variation in the screening decisions, the hiring manager model should recover the unbiased hiring manager scores (see Cowgill (2020)). The resulting size of the training/evaluation set for the hiring manager model is 106,419, and the hold-out test set (for performance evaluation) is 11,357. Note that we use the full hold-out test set including candidates that were not shortlisted for the estimation of hiring manager scores $\hat{q}^H$ and model parameters θ and δ.

7 Empirical results

This section reports the results of the empirical models. The following results are estimated on the hold-out test set.

7.1 Parameter estimates

We estimate the model parameters (θ, δ) for each job posting separately and plot the distribution in Figure 5. We also aggregate the parameters at the job category level by taking the weighted

average across job postings and report the results in Table 5. Note that these parameter estimates are calculated on the hold-out test set.

The average estimate of the correlation parameter, $\hat{\theta}$ is 0.43, with a high level of heterogeneity across jobs. When comparing estimates across job categories, an interesting pattern emerges: the correlation tends to be higher in technical jobs that require "hard skills" (e.g., in Finance & Accounting, Engineering & Technical) and lower in non-technical jobs that require "soft skills" (e.g., in HR, Legal & PR). A plausible explanation for this pattern is that it is easier to screen for hard skills using resumes than it is to screen for soft skills, which are better assessed in interviews. So, in technical jobs (e.g., software engineering) when both the screener and the hiring manager are aligned on the assessment criteria involving some hard skills (e.g., python proficiency), the screener is *able* to screen for those hard skills using the resume (e.g., experience and projects involving python, open source contributions, etc.), making the correlation with the hiring manager criteria higher. In non-technical jobs (e.g., HR) on the other hand, even when the screener and hiring manager are aligned on the same assessment criteria involving some soft skills (e.g., good communication skills), it is harder for the screener to screen for those soft skills from just the resume and may resort to attending to other proxies such as college major. This lowers the correlation in assessment criteria between the screener and the hiring manager.

Turning to the gender difference in correlation, δ, the average estimate is -0.007, [21] implying that the screening algorithm, on average, has the same assessment criteria with respect to the hiring manager for both men and women. However, there is heterogeneity across jobs with estimates ranging between -0.4 to 0.2.

[21] Note that the "Other" category has a high estimate of δ, but this is likely due to the small sample size (only 214 applicants in this category in the hold-out test set).

Figure 5: Distribution of parameter estimates across job postings

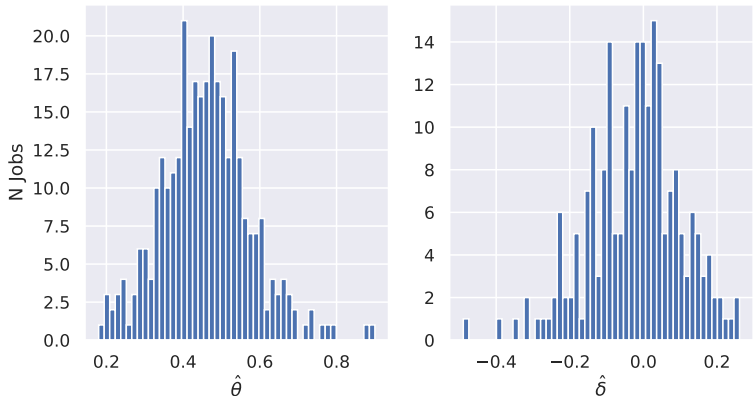


Table 5: Average parameter estimates

Job Category	$\hat{\theta}$	$\hat{\delta}$
Finance & Accounting	0.493	-0.019
Engineering & Technical	0.449	-0.009
Sales & Marketing	0.442	0.003
Product & Design	0.441	0.031
Customer Service & Acct Management	0.436	-0.02
Biz Dev & Operations	0.413	-0.014
HR	0.403	-0.074
Legal & PR	0.348	-0.016
Other	0.245	-0.332
Average	0.434	-0.007

Notes: This table reports the average parameter estimates for each job category. We estimate the parameters at the job posting level and aggregate it up to the job category level for all jobs in the hold-out test set.

7.2 Effectiveness of the equal selection constraint

We estimate the effect of equal selection on the gender proportion of hires with counterfactual policy simulations. We first estimate the job-specific model parameters for each job posting in the hold-out test set where female applicants are underrepresented ($p_a < 0.5$). We set the job-specific shortlist and hiring manager thresholds based on the actual size of the shortlist and finalist observed in the data.[22] For each job, we impute the parameters into the theoretical model and calculate the female

[22]We conceive the thresholds as exogenous variables. For example, the firm may have a limited budget to interview candidates and can only afford to interview a certain number of candidates. The shortlist and hiring manager threshold is therefore set based on the observed size of the shortlist and finalist respectively.

proportion of hires with and without the equal selection constraint. We then aggregate these results up to the job category level and report the results in Table 6. On average, there were 31% women in the applicant pool. Without the equal selection constraint, this proportion remains constant for the rest of the hiring funnel. With the equal selection constraint, the proportion of women in the shortlist increases to 50% (by design), but drops back down to 41% in the finalist. While the fairness constraint increases the proportion of women in the finalist compared to the applicant pool, the effect is modest and does not reach parity.

We also find that the effectiveness of the equal selection constraint varies widely across job categories due to differences in model parameters across jobs. In Engineering & Technical jobs, for example, the equal selection constraint only increases the proportion of women in the finalist to 36%.

Table 6: Estimated effectiveness of the equal selection constraint

Job Category	Parity Constraint	p_a	p_s	p_h
Biz Dev & Operations	False	0.38	0.38	0.38
	True	0.38	0.50	0.45
Customer Service & Acct Management	False	0.38	0.38	0.38
	True	0.38	0.50	0.47
Engineering & Technical	False	0.23	0.23	0.23
	True	0.23	0.50	0.36
Finance & Accounting	False	0.35	0.35	0.35
	True	0.35	0.50	0.45
HR	False	0.41	0.41	0.41
	True	0.41	0.50	0.46
Legal & PR	False	0.35	0.35	0.35
	True	0.35	0.50	0.44
Product & Design	False	0.33	0.33	0.33
	True	0.33	0.50	0.43
Sales & Marketing	False	0.36	0.36	0.36
	True	0.36	0.50	0.44
Overall	False	0.31	0.31	0.31
	True	0.31	0.50	0.41

Notes: This table reports the proportion of women in the applicant pool p_a, shortlist p_s and hired pool p_h — with and without the equal selection constraint. p_a is observed in the data. p_s and p_h are estimated by first estimating the job-specific model parameters $(\hat{\theta}_j, \hat{\delta}_j)$, and imputing the model parameters into the theoretical model. We estimate at the job posting level and aggregate it up to the job category level for all jobs with $p_a < 0.5$ in the hold-out test set.

7.3 Benchmarking against other screening algorithms and fairness constraints

We now turn to benchmarking the effectiveness of the equal selection constraint against other common fairness criteria and our proposed complementary screening algorithm. We consider seven different screening algorithms.

1. No Constraint algorithm is the baseline where no fairness constraint is imposed.
2. Equal Selection algorithm constrains the proportion of men and women in the shortlist to be equal.
3. Demographic Parity algorithm constrains the proportion of men and women in the shortlist to be the same as the proportion of men and women in the applicant pool.
4. Error Rate Parity algorithm has an equal error rate constraint—i.e., the error rates are equal between men and women. We use fairlearn's implementation of the algorithm.[23]
5. Equalized Odds algorithm has an equalized odds constraint—i.e., the true positive and false positive rates are equal between men and women. We use fairlearn's implementation of the algorithm.[24]
6. Equal Selection min Q^S Diff. algorithm has the equal selection constraint *and* minimizes the expected difference in screening scores between shortlisted men and women. To implement this, we first rank order the male and female candidates based on screening score $\hat{q}^S$. For each male candidate, we try to find the closest female candidate who has a score within $\epsilon = 0.01$—i.e., $\hat{q}^S \pm \epsilon$. If we find a match, we shortlist the pair. If we don't find a match, we move on to the next male candidate and repeat the process until the observed shortlist size is reached. If we run out of candidates in this process, we increase ϵ until we reach the shortlist size observed in the data.
7. Complementary Equal Selection algorithm has the equal selection constraint *and* minimizes the expected difference in hiring manager scores between shortlisted men and women. The implementation is the same as the Equal Selection min Q^S Diff. algorithm but minimizes the gender difference in hiring manager scores $\hat{q}^H$.

[23] https://fairlearn.org/v0.10/user_guide/mitigation/reductions.html#error-rate-parity
[24] https://fairlearn.org/v0.10/user_guide/mitigation/reductions.html#equalized-odds

Table 7: Screening algorithms

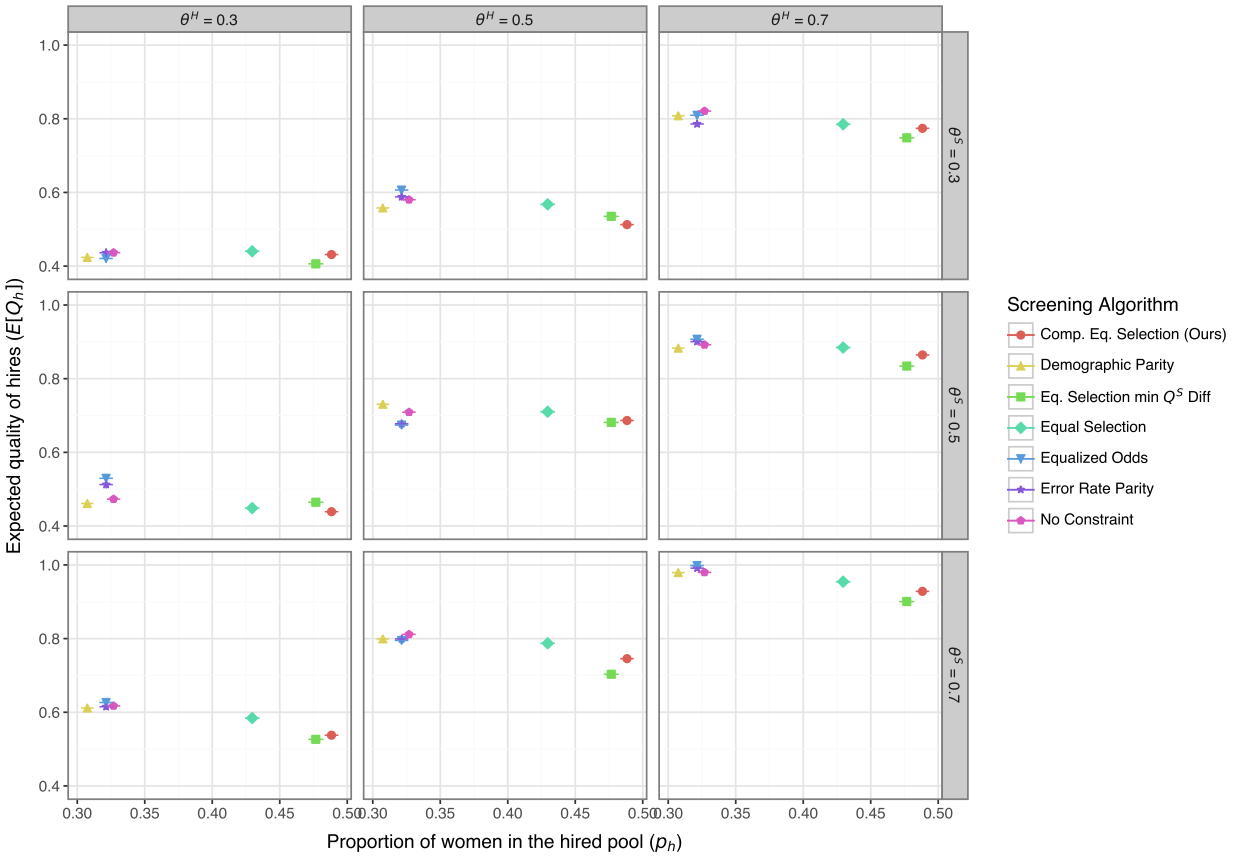
Screening Algorithm	**Constraint**
No Constraint	None
Equal Selection	$\mathbb{P}(g = f \mid \hat{y}^S = 1) = \mathbb{P}(g = m \mid \hat{y}^S = 1)$
Demographic Parity	$\mathbb{P}(\hat{y}^S \mid g = f) = \mathbb{P}(\hat{y}^S \mid g = m)$
Error Rate Parity	$\mathbb{P}(\hat{y}^S \neq y^S \mid g = f) = \mathbb{P}(\hat{y}^S \neq y^S \mid g = m)$
Equalized Odds	$\mathbb{P}(\hat{y}^S = 1 \mid y^S, g = f) = \mathbb{P}(\hat{y}^S = 1 \mid y^S, g = m),$ $y^S \in \{0, 1\}$
Equal Selection min Q^S Diff.	$\mathbb{P}(g = f \mid \hat{y}^S = 1) = \mathbb{P}(g = m \mid \hat{y}^S = 1)$ $\min \mathbb{E}[\hat{q}_s^S \mid g = f] - \mathbb{E}[\hat{q}_s^S \mid g = m]$
Complementary Equal Selection	$\mathbb{P}(g = f \mid \hat{y}^S = 1) = \mathbb{P}(g = m \mid \hat{y}^S = 1)$ $\min \mathbb{E}[\hat{q}_s^H \mid g = f] - \mathbb{E}[\hat{q}_s^H \mid g = m]$

Notes: This table summarizes the screening algorithms used for benchmarking. $\hat{y}^S$ is the predicted screening outcome, and y^S is the true screening outcome observed in the data. $\hat{q}_s^S$ is the predicted screening score of the shortlisted candidates, and $\hat{q}_s^H$ is the predicted hiring manager score of the shortlisted candidates. The primary objective of all the algorithms is to shortlist candidates with the highest screening score.

To understand how each of the screening algorithms affects the diversity and the expected quality of hires, we perform agent-based hiring simulation experiments, where each of the screening and hiring manager ML models serve as the agents who make screening and interview decisions. Because we are using agent-based simulation, these results are not necessarily subject to the assumptions of the theoretical model (same quality distribution among men and women, unbiased hiring manager, distributional assumptions, etc.) Although we are interested in measuring the expected quality of hires, we do not observe the true quality of the applicants in the data. To overcome this, we generate semi-synthetic true quality scores for each applicant for different θ^S and θ^H values, while fixing the observed θ value in each job.[25] We run the simulation at the job posting level for all jobs in the hold-out test set where women are underrepresented in the applicant pool ($p_a < 0.5$), and report the results in Figure 6.

[25] For each job we observe the vectors $\boldsymbol{q}_j^S$ and $\boldsymbol{q}_j^H$, which fixes θ. To generate $\boldsymbol{q}$, we first generate a random vector. We then orthogonalize it with respect to $\boldsymbol{q}_j^S$ and $\boldsymbol{q}_j^H$. We then transform the vector for a given value for θ^S and θ^H. This produces a random vector $\boldsymbol{q}_j$ that has the defined correlation structure $\Sigma = \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}$

Figure 6: Quality and diversity of hires using different screening algorithms



Notes: This figure plots the expected quality of hires (y-axis) and the proportion of women in the hired pool (x-axis) for different screening algorithms using agent-based hiring simulation experiments. We use semi-synthetic data for true quality Q. Each grid in the facet corresponds to a different θ^S and θ^H value. Error bars in both the x and y-axis represent bootstrapped 95% confidence intervals. Error bars are not visible on the y-axis because they are narrow.

There is a high level of variation in the diversity of hires across different screeners. The DEMOGRAPHIC PARITY screener achieves the lowest diversity of hires, and it is lower than the NO CONSTRAINT screener. This suggests that, in the training data, female applicants have a higher likelihood of being shortlisted than male applicants (we show that this is indeed the case with regression analysis in Appendix C.2), and the demographic parity constraint removes this disparity, resulting in a higher proportion of men in the shortlist.

The ERROR RATE PARITY and EQUALIZED ODDS screeners are clustered around the NO CONSTRAINT screener. That the diversity of hires does not increase with these error-rate constraints is not surprising, since these constraints correct for any differences in error rates between men and women. As such, they can increase the diversity of hires insofar that the NO CONSTRAINT

model suffers from disparate error rates. However, as shown in Appendix B, these models do not exhibit differences in error rates between men and women—the ROC curves for men and women are identical. This leaves no room for error rate constraints to increase the diversity of hires.

The Equal Selection screener increases the diversity of hires but not up to parity, as already seen in the previous section. The Equal Selection min Q^S Diff screener further increases the diversity of hires.

But this still falls short of our proposed Complementary Equal Selection screener, which achieves the highest level of diversity compared to all other screeners. The reason the Complementary screener outperforms the Equal Selection min Q^S Diff screener is that the former more directly minimizes the difference in hiring manager scores, which is what determines who is hired given the shortlist. Minimizing the gender difference in screening scores does not guarantee that the difference in hiring manager scores will be minimized when $\delta \neq 0$.

Next turning to the quality of hires, comparing across the grids, we see that the expected quality of hires increases with θ^S and θ^H, as expected. Within a particular grid, there is still variation in the expected quality of hires, but the variation is much smaller. In general, our proposed Complementary Equal Selection screener achieves lower expected quality of hires than the No Constraint screener, which is to be expected since there is some cost to the fairness constraint. However, this cost is minimal, especially when θ^S and θ^H are low. Compared to the Equal Selection min Q^S Diff screener, the proposed Complementary Equal Selection screener achieves a higher expected quality of hires in most cases.

8 Discussion and conclusion

This paper studies the effectiveness of diversity policies inscribed as algorithmic fairness constraints in Human+AI hiring systems. We first develop a theoretical model of the hiring process and show that the effectiveness of a common diversity policy (i.e., equal selection in the shortlist) depends on some seemingly inconsequential but important parameters such as the correlation between the screener and the hiring manager's assessment criteria. We recover these parameters using hiring data from IT firms and study the effectiveness of equal selection using counterfactual policy simulation. The results show that equal selection in the shortlist will have a modest effect in increasing the

gender diversity of the hires, but not up to parity. We also find that the effectiveness of the equal selection constraint will vary significantly across job types. Based on our findings, we propose a screening algorithm that is complementary to the hiring manager's assessment criteria and show that it is more effective in increasing the diversity of hires compared to other common fairness constraints. We now discuss the managerial and algorithmic design implications and suggest some paths forward for future work.

First, gender parity in the shortlist does not necessarily lead to gender parity in finalists/hires, even with unbiased hiring managers. The lack of gender parity in hires, when there is gender parity in the shortlist, is not necessarily an indication that the hiring manager is biased. Unless this is understood and anticipated by all stakeholders, the equal selection constraint runs the risk of making otherwise unbiased hiring managers *appear* biased since they seem to "undo" the constraint in the interview stage.

Second, the effectiveness of equal selection depends on the job. When there is variation in the effectiveness across jobs, it is not necessarily an indication that the hiring managers are more or less biased in certain jobs. As we have shown, equal selection is less effective in Engineering and Technical jobs, because the correlation between the screener and the hiring manager's assessment criteria is higher in these jobs, not necessarily because the hiring managers are more biased. Ironically, the very jobs in which women are underrepresented are the ones in which the equal selection constraint is least effective.

Third, the gender difference in the correlation between the screener and the hiring manager's assessment criteria also affects the effectiveness of the equal selection constraint. Screening algorithms are typically audited gender differences in predictive accuracy. As we've shown, in a multi-stage hiring context, it is not sufficient for the screening algorithm to have similar predictive accuracy across gender groups (i.e., $\theta^S_m = \theta^S_f$), but it should also have a similar correlation with the hiring manager's assessment criteria ($\delta = 0$).

Lastly, as the correlation between the screener's and the hiring manager's assessment criteria increases, both the effectiveness of the equal selection constraint and the expected quality of hires decrease. This means that the better a screening algorithm learns the hiring manager's estimate of candidate quality, the less effective the equal selection constraint will be and the lower the expected quality of hires will be. A design implication is that the screening algorithm's assessment should be

complementary to the hiring manager's assessment rather than similar.

Limitations and future work

One of the key insights of this paper is that the screening algorithm should be *complementary* to the hiring manager's assessment rather than similar. The complementary nature of the screening algorithm could cause organizational challenges. Hiring managers may view AI screening algorithms as a means to automate part of their work, in which case they would want the AI to be as similar as possible to their own assessment. Prior work has studied similar tensions in the algorithmic hiring setting (van den Broek et al. 2021). Future work could study Human-AI management when the AI is designed to be explicitly dissimilar but complementary to the human.

A limitation of our model is that the hiring manager does not update her beliefs based on the implementation of the equal selection constraint. Future work could study whether the introduction of algorithmic fairness constraints *induces* bias in hiring managers who were previously unbiased. The psychology and management literature has documented that in the presence of affirmative action programs (AAPs), the majority groups stigmatize AAP hires and view them as less competent (Heilman et al. 1997; Leslie et al. 2013). These negative views are extended to minorities even if they are not hired under AAPs through stereotyping (Heilman et al. 1997). Since algorithmic fairness constraints can be perceived as variations of AAPs, future work could study whether the introduction of fair algorithms induces bias in hiring managers who were previously unbiased.

References

Bapna, Sofia, Alan Benson, and Russell Funk (Oct. 2021). *Rejection Communication and Women's Job-Search Persistence*. SSRN Scholarly Paper. Rochester, NY.

Blum, Avrim, Kevin Stangl, and Ali Vakilian (June 20, 2022). "Multi Stage Screening: Enforcing Fairness and Maximizing Efficiency in a Pre-Existing Pipeline". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, pp. 1178–1193.

Bower, Amanda et al. (July 2017). "Fair Pipelines". In: *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. arXiv. arXiv: 1707.00391 [cs, stat].

Brands, Raina A. and Isabel Fernandez-Mateo (Sept. 1, 2017). "Leaning Out: How Negative Recruitment Experiences Shape Women's Decisions to Compete for Executive Roles". In: *Administrative Science Quarterly* 62.3, pp. 405–442.

Celis, L. Elisa et al. (Mar. 3, 2021). "The Effect of the Rooney Rule on Implicit Bias in the Long Term". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 678–689.

Chouldechova, Alexandra (June 1, 2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2, pp. 153–163.

Clemen, Robert T. and Robert L. Winkler (Apr. 1985). "Limits for the Precision and Value of Information from Dependent Sources". In: *Operations Research* 33.2, pp. 427–442.

Cowgill, Bo (2020). "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Re´sume´ Screening".

Devlin, Jacob et al. (May 24, 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv: 1810.04805 [cs].

Dwork, Cynthia, Moritz Hardt, et al. (Jan. 8, 2012). "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. New York, NY, USA: Association for Computing Machinery, pp. 214–226.

Dwork, Cynthia and Christina Ilvento (2019). "Fairness Under Composition". In: *LIPIcs, Volume 124, ITCS 2019* 124, 33:1–33:20. arXiv: 1806.06122 [cs, stat].

Fershtman, Daniel and Alessandro Pavan (Mar. 1, 2021). ""Soft" Affirmative Action and Minority Recruitment". In: *American Economic Review: Insights* 3.1, pp. 1–18.

Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi (Apr. 30, 2019). "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". arXiv: 1905.01989 [cs].

Hardt, Moritz, Eric Price, and Nathan Srebro (Oct. 7, 2016). "Equality of Opportunity in Supervised Learning". arXiv: 1610.02413 [cs].

Heilman, Madeline E., Caryn J. Block, and Peter Stathatos (June 1, 1997). "The Affirmative Action Stigma Of Incompetence: Effects Of Performance Information Ambiguity". In: *Academy of Management Journal* 40.3, pp. 603–625.

Huet, Ellen (Jan. 10, 2017). "Facebook's Hiring Process Hinders Its Effort to Create a Diverse Workforce". In.

Jiang, Xiangyu, Yucong Dai, and Yongkai Wu (June 2023). "Fair Selection through Kernel Density Estimation". In: *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Joe, Harry (June 26, 2014). *Dependence Modeling with Copulas*. CRC Press. 483 pp. Google Books: 09ThAwAAQBAJ.

Khalili, Mohammad Mahdi, Xueru Zhang, and Mahed Abroshan (2021). "Fair Sequential Selection Using Supervised Learning Models". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 28144–28155.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (Nov. 17, 2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores". arXiv: 1609.05807 [cs, stat].

Kleinberg, Jon and Manish Raghavan (Jan. 4, 2018). "Selection Problems in the Presence of Implicit Bias". arXiv: 1801.03533 [cs, stat].

Lee, Logan M. and Glen R. Waddell (Apr. 1, 2021). "Diversity and the Timing of Preference in Hiring Decisions". In: *Journal of Economic Behavior & Organization* 184, pp. 432–459.

Leslie, Lisa M., David M. Mayer, and David A. Kravitz (July 23, 2013). "The Stigma of Affirmative Action: A Stereotyping-Based Theory and Meta-Analytic Test of the Consequences for Performance". In: *Academy of Management Journal* 57.4, pp. 964–989.

Mitchell, Shira et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163.

Nelsen, Roger B. (June 10, 2007). *An Introduction to Copulas*. Springer Science & Business Media. 277 pp. Google Books: yexFAAAAQBAJ.

Peng, Andi et al. (Oct. 28, 2019). "What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (1), pp. 125–134.

Raghavan, Manish et al. (Jan. 27, 2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, pp. 469–481.

Schuck, Peter H. (2002). "Affirmative Action: Past, Present, and Future". In: *Yale Law & Policy Review* 20.1, pp. 1–96.

Shi, Wei et al. (2018). "The Adoption of Chief Diversity Officers among S&P 500 Firms: Institutional, Resource Dependence, and Upper Echelons Accounts". In: *Human Resource Management* 57.1, pp. 83–96.

Storvik, Aagoth Elise and Pål Schøne (2008). "In Search of the Glass Ceiling: Gender and Recruitment to Management in Norway's State Bureaucracy1". In: *The British Journal of Sociology* 59.4, pp. 729–755.

Sühr, Tom, Sophie Hilgard, and Himabindu Lakkaraju (July 21, 2021). "Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. New York, NY, USA: Association for Computing Machinery, pp. 989–999.

Sun, Chi et al. (2019). "How to Fine-Tune BERT for Text Classification?" In: *Chinese Computational Linguistics*. Ed. by Maosong Sun et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 194–206.

Van den Broek, Elmira, Anastasia Sergeeva, and Marleen Huysman (Sept. 2021). "When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring". In: *MIS Quarterly* 45.3, pp. 1557–1580.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., pp. 5998–6008.

Xiao, Qing and Shaowu Zhou (Apr. 2019). "Matching a Correlation Coefficient by a Gaussian Copula". In: *Communications in Statistics - Theory and Methods* 48.7, pp. 1728–1747.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, et al. (Apr. 3, 2017). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW

'17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, pp. 1171–1180.
Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, et al. (Nov. 28, 2017). "From Parity to Preference-based Notions of Fairness in Classification". arXiv: 1707.00010 [cs, stat].
Zaheer, Manzil et al. (2020). "Big Bird: Transformers for Longer Sequences". In: Neural Information Processing Systems (NeurIPS).
Zemel, Rich et al. (May 26, 2013). "Learning Fair Representations". In: *International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 325–333.
Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (Dec. 27, 2018). "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New York, NY, USA: Association for Computing Machinery, pp. 335–340.

Appendix

A Proofs

Preliminaries

The quality scores (Q, Q^S, Q^H) have a multivariate Gaussian distribution.

$$(Q, Q^S, Q^H) \sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}\right) \tag{A.1}$$

Let $F_{QQ^SQ^H}$ be the CDF of the quality scores, and $f_{QQ^SQ^H}$ be the PDF.

A.1 Proof of Proposition 1

Proposition. *Under equal selection, the female proportion of hires decreases with the correlation between screening and hiring manager scores, θ.*

Proof. We are interested in finding the proportion of women in the hired pool, p_h, as a function of the correlation parameter θ.

$$p_h = \frac{Pr(female\ is\ hired)}{Pr(female\ is\ hired) + Pr(male\ is\ hired)} \tag{A.2}$$

No Constraint. With no equal selection constraint in the screening algorithm, both men and women have the same shortlist threshold $\tau^S = \tau_m^S = \tau_f^S$. The probability that a female candidate is hired will be equal to the proportion of women in the applicant pool, and the probability that a male candidate is hired will be equal to the proportion of men in the applicant pool.

$$\begin{aligned} p_h &= \frac{p_a}{p_a + (1 - p_a)} \\ &= p_a \end{aligned} \tag{A.3}$$

The proportion of women in the hired pool is therefore the same as the proportion of women in the applicant pool.

Equal Selection. We first provide an intuition for the result, and then give the formal proof. Under equal selection, the female shortlist threshold (τ_f^S) is adjusted such that an equal number of women and men are shortlisted. When the screening and hiring manager scores are perfectly correlated ($\theta = 1$), $Q^S = Q^H$. This happens, for example, when the screener and the hiring manager use the same assessment criteria. Say, they both use years of experience as an assessment of quality. Since the screening algorithm shortlists the male and female candidates with the highest screening scores in each group, and there are more men than women in the applicant pool, the shortlist threshold for women will be lower compared to men ($\tau_f^S < \tau_m^S$). This implies that the expected quality of the shortlisted women will be lower than the expected quality of the shortlisted men. So, even though there are an equal number of male and female candidates in the shortlist, the shortlisted female candidate will be less likely to get hired compared to the male candidate. This probability is exactly equal to the proportion of females in the applicant pool, p_a.

On the other extreme, consider the case when the two scores are perfectly uncorrelated ($\theta = 0$). This happens when the screening algorithm and the hiring manager use orthogonal assessment criteria. Say, the screening algorithm uses years of experience, and the hiring manager uses GPA (assuming they are indeed uncorrelated). After the screening stage, the shortlisted female candidates will have less experience in expectation than the male candidates. But since the hiring manager only considers college GPA, shortlisted male and female candidates will have the same GPA in expectation (since the underlying quality distributions are the same). Therefore, when $\theta = 0$, the probability that a female is hired equals $\frac{1}{2}$.

Formally, we first solve for female shortlist threshold τ_f^S.

$$\begin{aligned} Pr(g = f|y^S = 1) = Pr(g = m|y^S = 1) \implies \\ 1 - F_{Q^S}(\tau_m^S) \cdot (1 - p_a) = 1 - F_{Q^S}(\tau_f^S) \cdot p_a \end{aligned} \tag{A.4}$$

Solving for τ_f^S:

$$\tau_f^S = F_{Q^S}^{-1}(1 - (1 - p_a) \cdot (1 - F_{Q^S}(\tau_m^S))/p_a) \tag{A.5}$$

Next, the probability that a female candidate is hired is:

$$\begin{aligned} Pr(\textit{female is hired}) &= p_a \cdot Pr(Q^H > \tau^H \mid Q^S > \tau_f^S) \\ &= p_a \int_{\tau_f^S}^{\infty} (1 - F_{Q^H|Q^S}(\tau^H)) \cdot f_{Q^S}(q^S)\, \mathrm{d}q^S \end{aligned} \tag{A.6}$$

where the conditional distribution of the hiring manager score given the screener score is:

$$F_{Q^H|Q^S} = \Phi_{\mathcal{N}}\left(\frac{\tau^H - \theta q^S}{\sqrt{1 - \theta^2}}\right) \tag{A.7}$$

Similarly, for a male candidate:

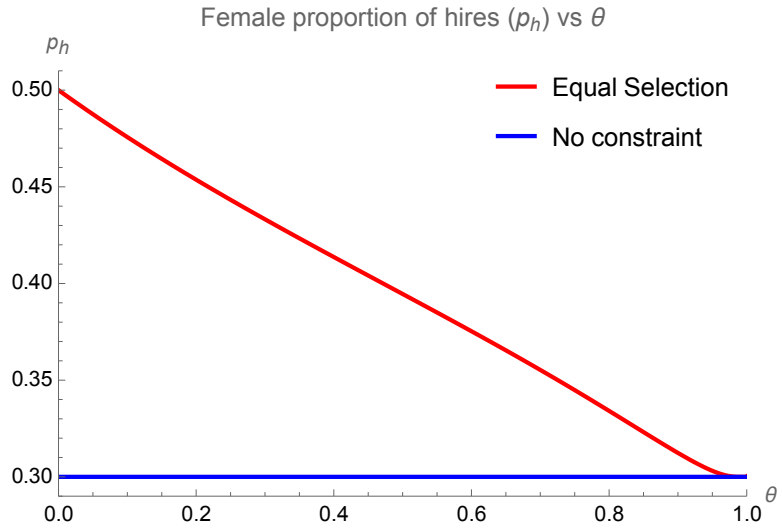
$$\begin{aligned} Pr(\textit{male is hired}) &= (1 - p_a) \cdot Pr(Q^H > \tau^H \mid Q^S > \tau_m^S) \\ &= (1 - p_a) \int_{\tau_m^S}^{\infty} (1 - F_{Q^H|Q^S}(\tau^H)) \cdot f_{Q^S}(q^S)\, \mathrm{d}q^S \end{aligned} \tag{A.8}$$

Plugging in the above expressions into A.2, we get:

$$p_h(\theta) = \frac{p_a \int_{\tau_f^S}^{\infty} (1 - F_{Q^H|Q^S}(\tau^H)) \cdot f_{Q^S}(q^S)\, \mathrm{d}q^S}{p_a \int_{\tau_f^S}^{\infty} (1 - F_{Q^H|Q^S}(\tau^H)) \cdot f_{Q^S}(q^S)\, \mathrm{d}q^S + (1 - p_a) \int_{\tau_m^S}^{\infty} (1 - F_{Q^H|Q^S}(\tau^H)) \cdot f_{Q^S}(q^S)\, \mathrm{d}q^S} \tag{A.9}$$

The above expression simplifies to $p_h(1) = p_a$ when $\theta = 1$ and $p_h(0) = 0.5$, when $\theta = 0$ (Proposition 0). The expression does not otherwise have a closed-form solution. So, we solve for p_h numerically and plot it as a function of θ.

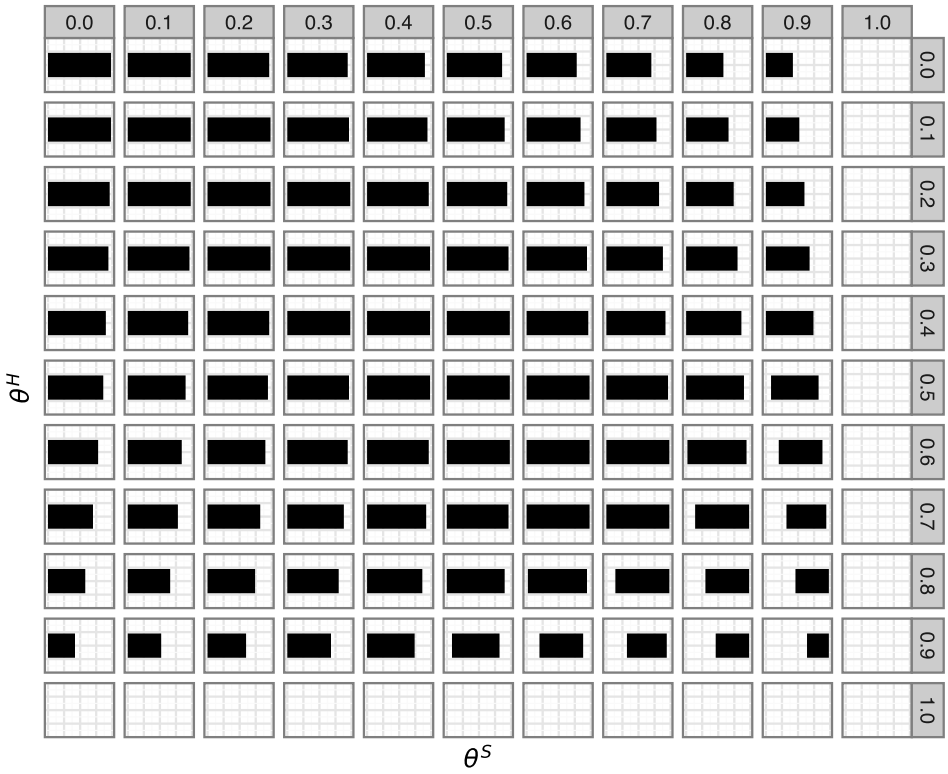
Figure 7: Proportion of women in the hired pool (p_h) vs. correlation parameter (θ)



Notes: This figure plots the female proportion of hires, p_h, as a function of the correlation parameter, θ. The proportion of women in the applicant pool is fixed at $p_a = 0.3$. This result does not depend on the θ^S and θ^H parameters.

This result does not depend on the θ^S and θ^H parameters. However, because Σ is positive semi-definite, how much θ can change depends on the values of θ^S and θ^H. Below, we plot the feasible region of θ for different θ^S and θ^H values.

Figure 8: Feasible regions of θ for different θ^S and θ^H values



Notes: This figure plots the feasible region of θ for different θ^S and θ^H values. Each row corresponds to a different θ^H value, and each column corresponds to a different θ^S value. The black bar in each cell is the feasible region of θ. The feasible region is the set of all θ values that satisfy the condition that the covariance matrix is positive semi-definite.

□

A.2 Proof for Proposition 2

Proposition. *Conditional on θ^S and θ^H, the expected quality of hires decreases with the correlation between screening and hiring manager score, θ.*

Proof. The 2-stage hiring process is an aggregation of noisy signals from the screener and the hiring manager. When aggregating noisy signals, uncorrelated signals reveal more information about the true quality of the candidate than correlated signals, conditional on individual signal informativeness (see Clemen and Winkler (1985)). Hence, conditional on the informativeness of individual signals—i.e., conditional on θ^S and θ^H—the expected quality of hire decreases with θ.

Formally, we are interested in finding the expected quality of hires, $E[Q_h]$, as a function of θ and θ^S.

No Constraint. The expected quality of hires is:

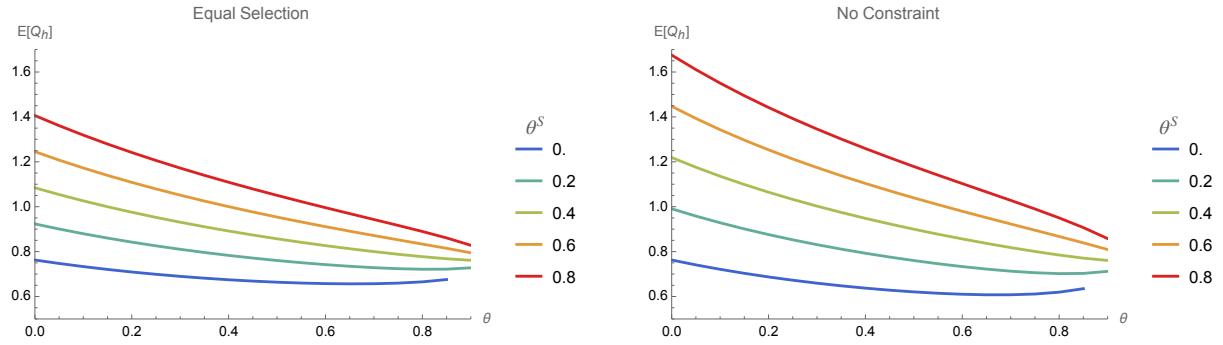
$$\begin{aligned} E[Q_h] &= E[Q \mid Q^S > \tau^S, Q^H > \tau^H] \\ &= \int_{-\infty}^{\infty} q \cdot f_{Q|Q^S>\tau^S, Q^H>\tau^H}(q)\, \mathrm{d}q \\ &= \frac{\int_{\tau^S}^{\infty}\int_{\tau^H}^{\infty}\int_{-\infty}^{\infty} q \cdot f_{Q,Q^S,Q^H}(q, q^S, q^H)\, \mathrm{d}q\, \mathrm{d}q^S\, \mathrm{d}q^H}{Pr(Q^S > \tau^S, Q^H > \tau^H)} \\ &= \frac{\int_{\tau^S}^{\infty}\int_{\tau^H}^{\infty}\int_{-\infty}^{\infty} q \cdot f_{Q,Q^S,Q^H}(q, q^S, q^H)\, \mathrm{d}q\, \mathrm{d}q^S\, \mathrm{d}q^H}{\int_{\tau^S}^{\infty}\int_{\tau^H}^{\infty}\int_{-\infty}^{\infty} f_{Q,Q^S,Q^H}(q, q^S, q^H)\, \mathrm{d}q\, \mathrm{d}q^S\, \mathrm{d}q^H} \end{aligned} \tag{A.10}$$

Equal Selection. Under equal selection, we calculate the expected quality of hire separately for male and female candidates, with their respective shortlist thresholds τ_m^S and τ_f^S. The overall expected quality of hires is the weighted average of the two, where the weight is the proportion of each gender in the hired pool given by A.9.

$$E[Q_h] = p_h \cdot E[Q_{h,f}] + (1 - p_h) \cdot E[Q_{h,m}] \tag{A.11}$$

The above expression does not have a closed-form solution, so we solve for it numerically and plot the results as a function of θ and θ^S.

Figure 9: Expected quality of hire ($E[Q_h]$) vs. θ, θ^S



Notes: This figure plots the expected quality of hires, $E[Q_h]$, as a function of the correlation parameter, θ for different θ^S values. The rest of the parameters are fixed at $\theta^H = 0.5, p_a = 0.3, \delta = 0$.

□

A.3 Proof for Proposition 3

Proposition. *The female proportion of hires decreases with the gender difference in the correlation parameter, δ.*

Proof. The proof for Proposition 3 is similar to the proof for Proposition 1. The only difference is that male and female candidates have different quality score distributions.

For male candidates, the quality score distribution is the same as before.

$$(Q_m, Q_m^S, Q_m^H) \sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix} \right) \tag{A.12}$$

For female candidates, there is an additional parameter δ that captures the gender difference in correlation between the screener and the hiring manager.

$$(Q_f, Q_f^S, Q_f^H) \sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta - \delta \\ \theta^H & \theta - \delta & 1 \end{bmatrix} \right) \tag{A.13}$$

Given these distributions, the probability that a female candidate will be hired is:

$$\begin{aligned} Pr(\textit{female is hired}) &= p_a \cdot Pr(Q_f^H > \tau^H \mid Q_f^S > \tau_f^S) \\ &= p_a \int_{\tau_f^S}^{\infty} (1 - F_{Q^H|Q^S,f}(\tau^H)) \cdot f_{Q^S,f}(q^S)\, \mathrm{d}q^S \end{aligned} \tag{A.14}$$

where the conditional distribution of the hiring manager score given the screener score is:

$$F_{Q^H|Q^S,f} = \Phi_{\mathcal{N}}\left(\frac{\tau^H - (\theta - \delta) q^S}{\sqrt{1 - (\theta - \delta)^2}} \right) \tag{A.15}$$

Similarly, for a male candidate:

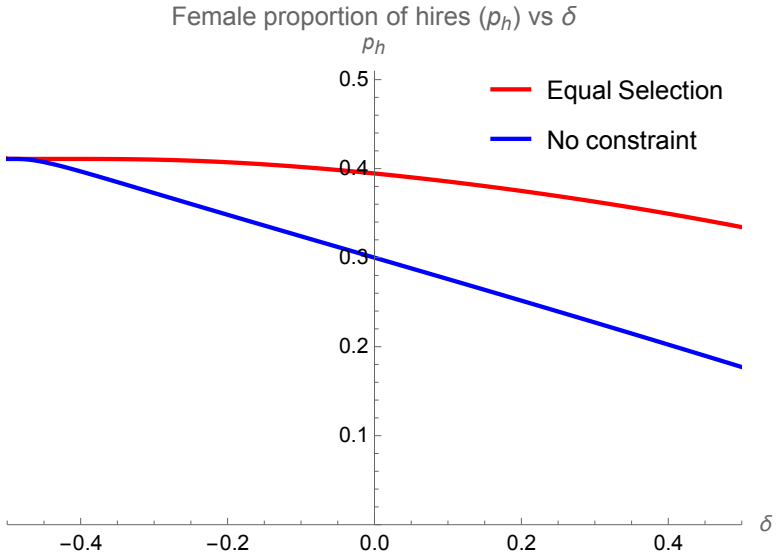
$$\begin{aligned} Pr(\textit{male is hired}) &= (1 - p_a) \cdot Pr(Q_m^H > \tau^H \mid Q_m^S > \tau_m^S) \\ &= (1 - p_a) \int_{\tau_m^S}^{\infty} (1 - F_{Q^H|Q^S,m}(\tau^H)) \cdot f_{Q^S,m}(q^S)\, \mathrm{d}q^S \end{aligned} \tag{A.16}$$

where the conditional distribution of the hiring manager score given the screener score is:

$$F_{Q^H|Q^S,m} = \Phi_{\mathcal{N}}\left(\frac{\tau^H - \theta q^S}{\sqrt{1-\theta^2}}\right) \tag{A.17}$$

We substitute these expressions into A.2 and solve for p_h as a function of δ.

Figure 10: Proportion of women in the hired pool (p_h) vs. correlation parameter (δ)



Notes: This figure plots the female proportion of hires, p_h, as a function of the correlation parameter, θ. The proportion of women in the applicant pool is $p_a = 0.3$. This result does not depend on the θ^S and θ^H parameters.

□

A.4 Difference in quality between men and women

So far we have considered the case where male and female applicants are equally qualified. We now consider the case where female applicants can be more/less qualified than men on average. We parametrize this difference using the location parameter α, as follows:

$$\begin{aligned}
(Q_m, Q_m^S, Q_m^H) &\sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}\right) \\
(Q_f, Q_f^S, Q_f^H) &\sim \mathcal{N}\left(\begin{bmatrix} \alpha & \alpha & \alpha \end{bmatrix}, \begin{bmatrix} 1 & \theta^S & \theta^H \\ \theta^S & 1 & \theta \\ \theta^H & \theta & 1 \end{bmatrix}\right)
\end{aligned} \tag{A.18}$$

A positive α implies that women are more qualified on average than men, and a negative α implies the opposite.

The conditional distribution of the hiring manager score given the screener score is:

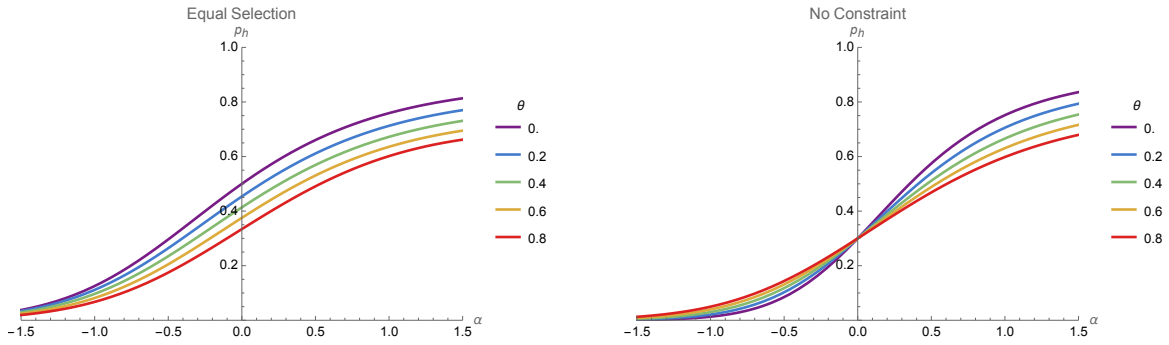
$$F_{Q^H|Q^S} \sim \mathcal{N}(\alpha + (Q^S - \alpha)\theta, \sqrt{1-\theta^2}) \tag{A.19}$$

We repeat the same analysis as in the previous proofs but with the additional parameter α. We plot the proportion of women in the hired pool p_h as a function of α and θ in Figure 11.

When women are more qualified than men ($\alpha > 0$), the equal selection constraint becomes redundant. Higher mean quality compensates for the lower proportion of women in the applicant pool. This implies that the proportion of women in the hired pool p_h increases with α. Therefore, the equal selection constraint becomes redundant.

When women are less qualified than men ($\alpha < 0$), equal selection becomes even less effective.

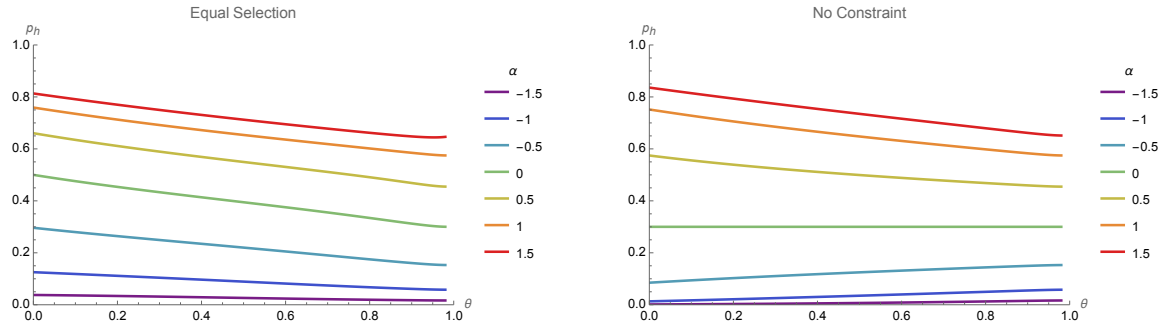
Figure 11: Female proportion of hires (p_h) vs. quality difference parameter (α)



Notes: This figure plots the female proportion of hires, p_h, as a function of the quality difference parameter, α. The proportion of women in the applicant pool is $p_a = 0.3$. This result does not depend on the θ^S and θ^H parameters.

Interestingly, without the equal selection constraint, the female proportion of hires decreases with θ when women have higher mean quality ($\alpha > 0$), and vice versa when women have lower mean quality.

Figure 12: Female proportion of hires (p_h) vs. correlation parameter (θ) for different α values



Notes: This figure plots the female proportion of hires, p_h, as a function of the correlation parameter, θ for different values of α. The proportion of women in the applicant pool is $p_a = 0.3$. This result does not depend on the θ^S and θ^H parameters.

B Additional details on the ML models

B.1 Predictive performance

We measure the predictive performance of the ML models using the Area Under ROC curve (AUC) criteria[26] on the hold-out test set and report the results in Table 8.

For the screening model, the overall AUC score is 0.83, and there is no difference in AUC scores between the male and female candidates. We also find that there is some heterogeneity in performance across job types, as reported in Table 9.

For the hiring manager model, the predictive performance is lower compared to the screening model since the hiring manager has more information from the interview, which we do not observe. Nonetheless, the predictive performance based on just resume characteristics is still reasonably high at 0.68, and there is no difference between genders. Note that the hiring manager model is evaluated on a subset of applicants in the hold-out test set who were, in fact, shortlisted.

[26]AUC is a widely-used measure for predictive performance for classification models since it is agnostic to both imbalanced classes and classification thresholds. The score ranges from 0.5 to 1, where 0.5 corresponds to a random classifier, and 1 corresponds to a perfect classifier.

Table 8: Predictive model performance by gender on hold-out test set

	Screening		Hiring Manager	
Group	AUC	Support	AUC	Support
Female	0.83	31,364	0.68	4,679
Male	0.83	42,393	0.68	6,678
Overall	0.83	73,757	0.68	11,357

Notes: This table reports the predictive performance of the screening and hiring manager classification models on the hold-out test set broken down by male and female candidates.

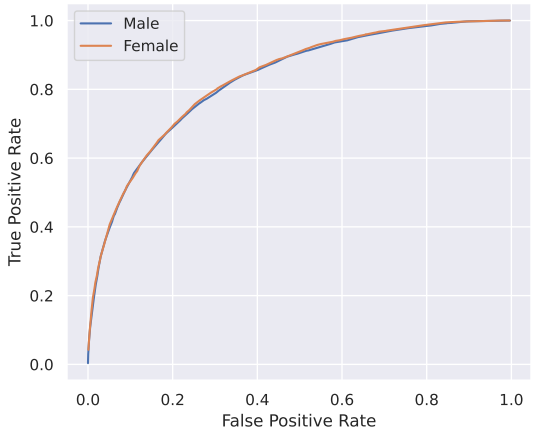
Table 9: Predictive model performance by job category on hold-out test set

	Screening		Hiring Manager	
Job Category	AUC	Support	AUC	Support
Legal & PR	0.86	7,337	0.65	926
Product & Design	0.85	12,519	0.66	1,863
Sales & Marketing	0.85	15,169	0.67	2,120
Other	0.83	214	0.59	25
Engineering & Technical	0.82	16,506	0.66	3,315
Finance & Accounting	0.82	7,026	0.67	886
Biz Dev & Operations	0.81	4,836	0.65	628
HR	0.79	3,355	0.69	452
Customer Service & Acct Management	0.78	6,734	0.7	1,112
Overall	0.83	73,757	0.68	11,357

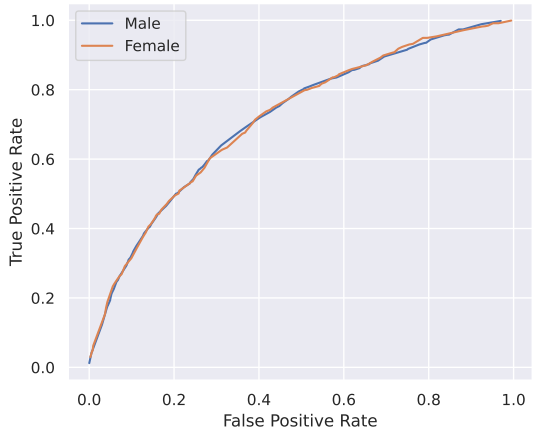
Notes: This table reports the predictive performance of the screening and hiring manager classification models on the hold-out test set. We estimate the metrics at the job posting level and aggregate up to the job category level.

Figure 13: ROC Curves for Screening and Hiring Manager Models

(a) Screening Model ROC Curve



(b) Hiring Manager Model ROC Curve



C Additional empirical analyses

C.1 Measures of observable quality differences between men and women

In this section we empirically assess differences in observable quality measures between men and women. To do so, we first identify four measures of observable quality: job-resume skill similarity, years of experience, attended a top 100 school, and educational attainment. We operationalize these measures as follows:

- **Job-Resume skill similarity:** We measure the average cosine similarity between skills listed in the job description and the applicant's resume. To get the cosine similarity, we first tokenize the job description and resume text. We then filter the tokens to extract only skills-related tokens (e.g., `python`, `data_analysis`, `project_management`) using a dictionary of skills[27]. We then get the vector representation of each skill token using a custom word2vec model trained on resumes, and take the average cosine similarity between the job description and resume skill vectors.

- **Years of experience:** We get the applicant's years of experience from the ATS.

[27]This dictionary was created using the skills section of LinkedIn profiles in a separate analysis.

- **Attended a top 100 school:** We create a binary variable indicating if the applicant attended a top 100 school based on the undergraduate institution listed in the resume. We use U.S. News and World Report's ranking of top 100 schools as the reference.

- **Educational attainment:** We create binary variables indicating if the applicant has a bachelor's, master's, or doctorate degree based on the highest degree listed in the resume.

For each of these outcomes, we estimate a linear regression model with job posting fixed effects

$$y_{ij} = \beta_{Female} \cdot Female_i + \alpha_j + \epsilon_{ij}$$

where y_{ij} is the observable quality measure for applicant i applying to job j, $Female_i$ is a binary variable indicating if the applicant is female, α_j is the job posting fixed effect, and ϵ_{ij} is the error term.

We report the coefficients and percentage differences below. Compared to male applicants, female applicants have roughly the same job-resume skill similarity, fewer years of experience, are more likely to have attended a top 100 school, more likely to have a bachelor's or master's degree, and less likely to have a doctorate degree.

Table 10: Regression coefficients of observable quality measures

Variable	β_{Female}	% Difference	p-value
Job-Resume skill similarity	0.003	0.51%	< 0.001
Yrs exp	-0.53	-6.1%	< 0.001
Attended top 100 school	0.013	4.66%	< 0.001
Has bachelor's degree	0.015	1.79%	< 0.001
Has master's degree	0.03	7.16%	< 0.001
Has doctorate	-0.004	-6.98%	< 0.001

Notes: This table shows the estimated regression coefficients and percentage differences of various observable quality measures. Job-Resume Similarity is the cosine similarity between the job description and the resume. Yrs Exp is the number of years of experience. Attended Top 100 School is a binary variable indicating if the candidate attended a top 100 school. Has Bachelor's Degree, Has Master's Degree, and Has Doctorate are binary variables indicating if the candidate has a bachelor's, master's, or doctorate degree, respectively.

C.2 Regression estimates on the likelihood of being shortlisted

Table 11: Likelihood of being shortlisted, OLS estimates

Dependent Variable:	Shortlisted (1=YES)
Variables	
Male	-0.0128***
	(0.0012)
Yrs Exp	0.0012***
	(0.0002)
Job Resume Similarity	0.3918***
	(0.0107)
Fixed-effects	
Job Posting	Yes
School Rank	Yes
Degree	Yes
Fit statistics	
Observations	595,246

Clustered (Job Posting) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: This table shows the OLS estimates of the likelihood of being shortlisted. Each observation corresponds to an application. Male applicants are more likely to be shortlisted than female applicants after controlling for job-resume skill similarity, years of experience, education, and job posting.

C.3 Goodness of fit using different copulas

Below we provide Kolmogorov-Smirnov (KS) statistics of empirical quality scores q^S and q^H fit against commonly used copulas. Lower KS statistics indicate a better fit. The Gaussian copula has the 2nd best fit after the Frank copula.

Table 12: Copula goodness-of-fit measures

Copula	KS-Statistic	p-value
Gaussian	2.62	< 0.0001
Gumbel	6.25	< 0.0001
Frank	2.29	< 0.0001
Clayton	6.07	< 0.0001
Joe	11.84	< 0.0001
AMH	5.37	< 0.0001

Notes: This table shows the goodness-of-fit measures of empirical quality scores q^S and q^H fit against various copulas.