I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17004

Delegation in Hiring: Evidence from a Two-Sided Audit

Bo Cowgill
Patryk Perkowski

MAY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17004

Delegation in Hiring: Evidence from a Two-Sided Audit

Bo Cowgill
Columbia Business School and IZA

Patryk Perkowski
Yeshiva University

MAY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Delegation in Hiring: Evidence from a Two-Sided Audit*

Firms increasingly delegate job screening to third-party recruiters, who must not only satisfy employers' demand for different types of candidates, but also manage yield by anticipating candidates' likelihood of accepting offers. We study how recruiters balance these objectives in a novel, two-sided field experiment. Our results suggest that candidates' behavior towards employers is very correlated, but that employers' hiring behavior is more idiosyncratic. Workers discriminate using the race and gender of the employer's leaders more than employers discriminate against the candidate's race and gender. Black and female candidates face particularly high uncertainty, as their callback rates vary widely across employers. Callback decisions place about two thirds weight on employer's expected behavior and one third on yield management. We conclude by discussing the accuracy of recruiter beliefs and how they impact labor market sorting.

JEL Classification: M51, C93, J71
Keywords: hiring, recruiting, discrimination, field experiments

Corresponding author:
Bo Cowgill
Columbia Business School
665 W 130th St
New York, NY 10027
USA
E-mail: bo.cowgill.work@gmail.com

* The authors thank participants at the NBER Summer Institute (2018 & 2019), Behavioral Decision Research in Management conference (BDRM), the Conference on Digital Experimentation (CODE) at MIT, Wharton's People and Organizations Conference, and Columbia Business School. We thank Dan Wang for contributions to the original draft. We also thank Amanda Agan, Alan Benson, Peter Cappelli, Fabrizio Dell'Acqua, Laura Gee, Rem Koning, Jean Oh and Orie Shelef for helpful comments and feedback. We also thank Hailey Brace, Nadine Fares and Matthew Fondy for excellent research assistance. Cowgill thanks the Kauffman Foundation for supporting this research. An earlier version of this paper was circulated under the title "Agency and Workplace Diversity: Evidence from a Two-Sided Audit." Edited by Juanna Joensen.

1 Introduction

Modern employers often delegate key parts of employee screening to third-party intermediaries. Previously limited to executive search, this practice is widespread for rank-and-file openings. A recent survey by Korn Ferry indicates that 40% of U.S. firms have delegated all or part of their hiring process to third party intermediaries. Using data from the U.S. Bureau of Labor Statistics (BLS) and the U.S. Economic Census, we document rapid growth in outsourced recruiting since 2000: the number of outsourced recruiters has grown up to ten times faster than U.S. employment. As Peter Cappelli (2019) writes, "the recruiting and hiring function has been eviscerated" in modern firms.[1]

This paper studies the origin and nature of third-party recruiting practices. We focus on the two-sided matchmaking aspects of this work: Recruiters must align employer requirements with candidate availability, effectively balancing client demand with worker preferences. Previous work has shown employer discrimination against candidates (Bertrand and Mullainathan, 2004; Charles and Guryan, 2008; Bertrand and Duflo, 2017; Gaddis, 2018), as well as candidate discrimination against employers or managers (Stoll et al., 2004; Giuliano et al., 2009; Chakraborty et al., 2018; Doerrenberg et al., 2020; Ayalew et al., 2019; Abel, 2019).

In this paper, we explore how recruiters absorb and re-route the pressures from either side of the market when selecting candidates for a job opening. We introduce a novel field experimental approach called *two-sided audits* for studying this topic and related questions at the intersection of labor supply and demand. In our design, we hire a recruiting workforce to evaluate job applications on behalf of clients. Like most audit studies, the job applications they review are similar to real resumes, but randomized (and thus fictitious). However, unlike traditional audit studies, the employers' characteristics (i.e., the recruiters' clients) were simultaneously randomized. We call this approach *two-sided* because the

[1] https://hbr.org/2019/05/your-approach-to-hiring-is-all-wrong

recruiters face randomized treatments on both sides of the market.

In our experiment, we manipulate biographical details – including race, gender, education and professional background – of applicants as well as the hiring manager assigned to interview the applicants (after callback decisions are made). For each job candidate, we ask recruiters to report not only a callback decision, but also how both sides (employer and candidate) are likely to react to the match (conditional on the callback). This includes how likely the employer will extend a job offer, and how likely the candidate will be to accept it.

Our design permits an in-depth look at recruiter decision-making and some of the underpinnings of labor market sorting. By examining how the candidate and manager characteristics impact recruiters' beliefs about both sides, we can better understand how labor markets integrate the actions of employees and employers. Our study helps understand how recruiters balance employee and employer expectations when matching candidates and openings.

We have three main results. First, we find that recruiters expect job candidates to care more about the race and gender of the employer's leaders than employers care about the race and gender of candidates. Company executives who are female or black are expected to face the strongest discrimination from job candidates. We find that recruiters expect different job candidates to react similarly to the set of potential job opportunities. By contrast, they expect different employers to react idiosyncratically to the set of different candidates (i.e., each employer has a distinct ordering of candidates). On average, recruiters act as if employers are more neutral to race and gender than candidates are.

Second, we find robust evidence that recruiters' choices are highly match-specific. The same candidate's outcomes vary widely depending on the hiring manager they are assigned. Black and female candidates face particularly high variability. We can statistically detect match-specific effects for all non-blinded candidate and manager characteristics we study. While there is robust evidence of match-specific effects, we find little evidence that recruiters facilitate homophily (for example, by pairing workers with demographically-similar managers), as we had expected.

Third, we examine how recruiters synthesize these beliefs into callback decisions. We find that recruiters place about 2/3rds weight on their expectations about the employer's behavior, and about 1/3rd on the candidate's side. Our finding that expectations about candidate behavior influence callback outcomes contrasts with typical interpretations of audit studies as reflecting solely employer behavior. However, we do find that employer responses are weighted more heavily, despite the widespread use of incentive contracts that reward yield.

As a byproduct of these results, many candidates received callbacks despite a low expected probability of accepting the employer's job. In our setting, elite university and large company job applicants benefit from this practice: although these candidates are not more likely to result in a hire, they are much more likely to receive an interview from recruiters. Indeed towards the end of our paper we present survey evidence suggesting that recruiters have relatively accurate beliefs regarding behavior, but nonetheless make different callback decisions than hiring managers would. We find suggestive evidence that reputational incentives compel recruiters to impress employers by delivering employer-approved candidates at the expense of yield management (choosing candidates more likely to accept). Although this would appear to be a waste of the employer's time, we cannot make strong claims about whether this behavior is optimal (either for the recruiter's private interests or employers and job candidates). It is possible that employers or recruiters benefit from interviewing these candidates through some other mechanism (besides the opportunity to hire them).

Related Literature and Contributions. Our paper contributes to three complementary literatures. The first is about labor market discrimination. A large literature spanning multiple social science disciplines has used audit studies to experimentally test for discrimination by employers (see Gaddis (2018) for an overview), especially for race and gender (see, for example, Bertrand and Mullainathan (2004) and Neumark et al. (1996)). A smaller literature studies worker discrimination against managers (Stoll et al., 2004; Giuliano et al., 2009; Chakraborty et al., 2018; Doerrenberg et al., 2020; Ayalew et al., 2019; Abel, 2019). We examine both types of discrimination simultaneously, and find that recruiters believe that job candidates discriminate using race and gender more than employers do.

Second, we contribute to the methodological literature on hiring and selection. Several recent papers have introduced new tools for studying discrimination and hiring, such as Kessler et al.'s 2019 "incentivized resume rating." Kline and Walters (2021) show how to extend the traditional audit toolkit to detect illegal discrimination by specific employers. Our paper complements these papers by providing a new extension to the audit toolkit.

Lastly, this paper contributes to understanding firm hiring practices. Advances in IT have shaped how companies screen and select workers. For example, digitization and job platforms have also led to outbound recruiting, whereby firms seek out candidates directly rather than waiting for them to apply (Carrillo-Tudela et al., 2015; Black et al., 2022; Kim and Pergler, 2022). Additionally, advances in machine learning and A.I. have led some firms to screen workers using algorithms (Kuncel et al. 2014; Chalfin et al. 2016; Cowgill 2020; Li et al. 2021; Hunkenschroer and Luetge 2022; Perkowski 2023). An often-mentioned benefit of hiring algorithms is the lower cost of screening, compared to the higher cost of using employee time to manually review resumes. In this paper, we document the rise of an alternative screening approach (outsourcing to (human) third-party screeners) that is also often justified on cost-savings grounds. In related work, Kohlhepp and Aleksenko (2021) formalizes a model whereby delegated recruitment leads to distortions in the hiring process.

The rest of this paper proceeds as follows. In Section 2, we present motivating data and facts about the growth and practice of outsourced recruiting among employers. Section 3 describes our experimental setting and intervention. Sections 4 and 5 contain empirical results and discussion, respectively. Section 6 concludes.

2 Institutional Setting

In this section we describe the role of outsourced recruiters, the scope of their responsibilities, incentives, and business models. The workers in our experiment are in the BLS occupational category code #13-1071 ("Human resource specialists"). According to the BLS, their primary job is to "recruit, screen, interview, and place workers" on behalf of clients (either within the same firm or externally). Delegated recruiting happens both through outsourcing firms, as well as through temporarily-employed individual recruiters. The BLS' summary of this occupation specifically notes, "Some organizations contract recruitment and placement work to outside firms, such as those in the employment services industry or consulting firms in the professional, scientific, and technical industry."[2] The BLS' data shows that the top industry for employment in this occupation is *Employment Services* (the BLS category for the recruitment process outsourcing (RPO) industry), while the second is *Professional, scientific, and technical services*, which is the setting of our field experiment.

Figure 1: **Growth in recruiters in employment services, 2002-2019**

EmploymentGrowthComparison.pdf goes here

Notes: This figure compares growth for recruiters in employment services versus overall US employment using data from the OES. We concentrate our attention on the "human resource specialists" occupation (SOC code 13-1071), whose primary task is to "recruit, screen, interview, or place individuals within an organization." Because recruiters can be either in-house or outsourced, we examine only workers in the employment services industry (NAICS code 56-1300).

[2] https://www.bls.gov/OOH/business-and-financial/human-resources-specialists.htm

Prevalence of Outsourced Hiring. Because comprehensive data on firm hiring practices is lacking (Oyer et al., 2011), the percentage of positions filled through an outsourced recruiter is unknown. However, several industry sources suggest the prevalence is very high. In 2017, RPO was a five billion dollar per year global industry,[3] and it is projected to more than double by 2023. The RPO industry serves a variety of industries, and eight large RPO companies reported filling just under one million positions in 2018, or about 36% of the total jobs created in the US that same year.[4] For example, the state of New York hired an RPO firm to review over 75,000 applications and hire over 7,000 contact tracers over an eight-month period during the COVID-19 pandemic.[5] Although we lack data on the number of vacancies that were filled by outsourced recruiters, we can use employment statistics to document growth in this occupation.

Figure 1 visualizes annual growth in the employment services industry using BLS' Occupational Employment Statistics data. The figure illustrates rapid growth in outsourced recruiting since 2000. While US employment grew by 15 percent from 2002 to 2019, the number of recruiters in employment services more than doubled. Appendix **??** presents US Economic Census data about RPO occupations and industries. We find similar patterns for the number of establishments, total revenue, and annual payroll.

Why Outsource? Outsourced recruiting is popular with employers for several reasons. First, it may be a byproduct of secular growth in demand for employee screening (Autor, 2001; Cappelli et al., 1997). A major proposed theory for this growth is technological change that increases the returns to selectivity in hiring (Acemoglu, 1999, 2002; Levy and Murnane, 1996), thus increasing the demand for screening, and eventually warranting specialization in a new industry. Second, hiring increasingly requires expertise in compliance and information technology: firms must manage databases of applicants, advertise job

[3] `https://www.workforce.com/2018/01/25/sector-report-tech-gaining-foothold-rpo-space/`
[4] `https://www.workforce.com/2019/01/24/recruitment-process-outsourcing-providers-think/`
[5] `https://info.leveluphcs.com/nys-contact-tracing-video`

openings, screen applicants, and comply with the recordkeeping requirements of labor law. Indeed, recruiting intermediaries pitch clients by referencing the hassles of HR compliance ("You didn't start your business to spend time on HR compliance"),[6] which businesses do not regard as their core focus. Finally, outsourced recruiting is popular with firms whose hiring is seasonal, for whom a permanent staff is less useful. Together, these factors have created a rich third-party marketplace for contract recruiting.

Job Description. While the details of recruiters' work varies across settings[7], there are a few common themes that informed the details of our experimental design. First, screening and interviewing are the primary responsibilities of outsourced recruiters. Research by Korn Ferry and *HRO Today Magazine* reports that 91% of RPO clients purchase screening services, and 64% purchase interviewing services.[8] *Staffing Industry Analysts* estimates that over 90% of RPO buyers purchase screening services or interviewing services.[9] Rather than just manage HR infrastructure, outsourced recruiters play an active role in the job matching process.

Second, recruiters are compensated using both a flat rate (typically hourly) and a performance bonus. According to the BLS, median recruiter pay is $29.77/hour.[10] According to the National Compensation Survey (NCS), 43% of human resource specialists (#13-1071) receive performance pay as of Q1 2020 (Makridis and Gittleman, 2020).[11] Recruiters' bonuses are typically tied to their ability to hire: a survey reported by the Society of Human Resources Management found that of recruiters who receive performance-related pay, 60%

[6]This is an advertising slogan for Bambee, http://bambee.com.

[7]For example, a large survey of companies' hiring strategies includes 18 broad approaches (not mutually exclusive); 14 of these were used by over 10% of respondents. Only one hiring strategy (employee referrals) was used by 85% of respondents. See https://www.shrm.org/ResourcesAndTools/business-solutions/Documents/Talent-Acquisition-Report-All-Industries-All-FTEs.pdf for more information.

[8]https://staging.kornferry.com/media/sidebar_downloads/Measuring-Up-A-new-research-report-about-RPO-metrics.pdf

[9]Staffing Industry Analysts, RPO Market Developments, December 2017

[10]https://www.bls.gov/ooh/business-and-financial/human-resources-specialists.htm

[11]We thank the authors for private correspondence to help locate this figure.

are "primarily measured on the number of hires or placements made."[12] This encourages recruiters to care about hiring yield and monitor mutual interest from both job candidates and employers.

Finally, outsourced recruiters typically have one of two business models: The first, known as the "relational" model, features firms aiming to provide the recruiting arm of a company for several years. These recruiters make investments in customizing and integrating deeply with the client. The second approach, called the "transactional" model, features less customization as well as less commitment from client and vendor. Although exact measurements are scarce, the consensus in this industry is that the "transactional" part is much larger, for both vendor firms as well as subcontracted individual recruiters. Although some providers of recruiting services enjoy repeat business from clients, this mostly happens without long-term contracts. As a result, outsourced recruiters must manage reputations to cultivate future business.

3 Experimental Design

The subjects in our experiment are professional recruiters whose jobs we detail above. To find subjects, we aimed to 1) identify recruiters who are typical of those hired by companies, and 2) engage them in natural ways for this industry in order to measure realistic field behavior. To achieve this, we identified and contacted professional recruiters following the procedures in Appendix **??**. Our main criteria were prior recruiting experience and a U.S.-based location.[13]

We hired 54 external recruiters to review 16 applications each, or 864 job applications in

[12] https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/rewarding-recruiters-for-performance.aspx

[13] Some of our empirical research questions required awareness of US educational institutions, companies, and locations. However, the larger RPO industry sometimes sends recruiting materials overseas for examination by low-wage workers.

total. Table 1 contains full descriptive statistics on the recruiters. 83% of screeners were female, almost sixty percent identified as white, and twenty two percent were black. 100% had prior recruiting experience. The recruiters received an average hourly rate of $37.48. This is comparable to national representative data about recruiters from the BLS.[14] Agan and Starr (2018) use a near-identical subject pool and find that 71% of subjects have over three years of experience in hiring HR roles.

Table 1: **Recruiter Characteristics**

	Mean	SD	Lower 95% CI	Upper 95% CI
Female	0.83	0.05	0.73	0.94
White	0.57	0.07	0.44	0.71
Black	0.22	0.06	0.11	0.34
Prior Recruiting Experience	1.00	0.00	1.00	1.00
Hourly Rate	37.48	2.79	31.88	43.07
Total Hours Spent on 16 Resumes	1.70	0.08	1.55	1.86
Observations	54			

Notes: This table displays descriptive statistics of recruiter characteristics.

3.1 Task

The primary task of the subjects was evaluating a group of 16 job applications for a software engineering position. In this section, we describe how recruiters were asked to evaluate the candidates. Our design was informed by the occupational details in Section 2 about recruiting, as well as our extensive informal interviews with recruiters. Appendix **??** contains all task files, including the job description, recruiter instructions, and a sample feedback form. Recruiters in our experiment were asked three evaluation questions about each candidate, to include optional notes or comments or explanations, and to make a recommendation for a callback (or not). Below, we describe the recruiters' task in more detail and connect their work to our research questions.

[14] https://www.bls.gov/ooh/business-and-financial/human-resources-specialists.htm

Payment. Recruiters were paid hourly based on the posted rate on their profile. We offered recruiters a bonus in addition to their hourly rate. This bonus mimicked the institutional setting described in Section 2, rewarded truthful reporting, and helped align the interests of the employer and the recruiter (see Appendix **??** for more details). In the main text of our communications with recruiters, we described the goals above as the basis of the bonus (in simple, non-technical language), which was likely sufficient for many of our recruiters. We offered additional details in a FAQ.

Callback Decisions. We asked each recruiter to make a Yes/No decision about contacting each of the sixteen candidates. Subjects were told that our employer could potentially hire multiple candidates from the applicant pool, which is common in high-tech labor settings featuring shortages of qualified workers.

Prediction. In addition to the callback measure, we asked recruiters to anticipate the reaction to a callback. A recruiter may decline to give a callback to a given candidate because the recruiter thinks the employer will not hire them, or because the candidate may be unlikely to accept an offer. For this reason, we asked recruiters to share their beliefs about whether each candidate would (i) agree to be interviewed if an interview offer were extended, (ii) pass the interview (and receive an offer) if they were interviewed, and (iii) accept the offer if it were extended. We asked recruiters to report these probabilities on a 0-100 probability scale.[15]

We interpret these probabilities as reflecting the recruiter's beliefs about the candidate's and manager's behavior in the hiring process. We take *P(Accept Interview)* and *P(Accept Job Offer | Pass Interview)* as our measures of candidate behavior, and *P(Pass Interview | Accept Interview)* as our measure of employer behavior. The latter lets us understand the types of candidates that recruiters believe that employers seek, while the former two probabilities allow us to investigate the types of employers that recruiters believe that candidates seek.

[15] Recruiters were welcome to approximate using a round number, and over 96% responses corresponded to a multiple of 0.05 (when expressed as probabilities).

We specifically asked for the recruiter's beliefs *conditional* on making it to the previous round of the hiring process.[16] In our discussion and tables, we refer to these probabilities using the shorthand of *P(Pass)*, *P(Accepts Interview)*, etc. However, all probabilities should be read conditionally — e.g., *P(Pass)* means *P(Pass Interview | Employer Offers Interview)* – and we use the abbreviations for readability. Table 2 reports descriptive statistics on recruiters' feedback.

Table 2: **Descriptive Statistics: Recruiter Feedback**

	Mean	SD	Lower 95% CI	Upper 95% CI
Interviewed	0.66	0.02	0.62	0.69
P(Accept Interview)	0.75	0.01	0.74	0.76
P(Pass Interview)	0.70	0.01	0.69	0.72
P(Accept Job Offer)	0.69	0.01	0.68	0.70
P(Hired)	0.39	0.01	0.37	0.40
Explained Choice	0.61	0.02	0.57	0.64
Observations	864			

Notes: This table displays descriptive statistics of recruiter feedback. The variable names above use shorthand of *P(Pass)*, *P(Accepts)*, etc. However, all probabilities should be read conditionally (e.g., *P(Pass)* means *P(Accept Interview | Employer Offers Interview)*) as described in Section 3.1, and we use the abbreviations for parsimony.

3.2 Theoretical Interpretation

The data collected above are predictions of behavior. We offer two important caveats to the interpretation of this data. First, the data speaks to recruiters' *beliefs* about workers/employers behavior. These beliefs could be inaccurate (for example, if recruiters believe that managers discriminate on the basis of gender, but managers do not). In Section 5.2, we assess the accuracy of these beliefs more directly using a survey of participants in this labor market. Even insofar as the predictions are inaccurate, they may still be an important factor in how workers are matched to employers within outsourced recruiting.

[16] We test if subjects understood this conditioning in Appendix **??**.

Second, behaviors by market participants could arise either for taste-based or statistical reasons. For example: A manager might have a low probability of passing a male candidate because he (or she) dislikes men (i.e., taste-based) or because they are gender-neutral but believe that on average men are associated with (say) too much aggression (i.e., statistical). It is also possible that discrimination on either side of the market comes from different sources (i.e., taste-based for candidates and statistical for employers). The current form of our experiment cannot distinguish between these two mechanisms that drive discrimination. However, this is a question that a future version of a two-sided audit study can shed light upon.

Relationship with Hiring Outcomes. Critics of audit studies have noted that we do not know (from the audit study alone) if the disparities in the initial contact phase (callbacks) lead to inequalities at later phases (Heckman, 1998; Heckman and Siegelman, 1993), though some researchers have used nationally representative data to simulate how employer callback discrimination affects wages (Lanning, 2013).

Like other audit studies, we do not collect data about final hiring outcomes in this study. However, our paper does have an avenue for assessing how callback outcomes might translate into differences in hiring. We collect three conditional probabilities that include a) the probability of receiving an offer conditional on being interviewed, and b) the probability of accepting an offer conditional on receiving one. Given their conditional nature, the product of these probabilities equals the probability that the candidate will be hired, conditional on being offered an interview. We call this p_{hire} and use the product in some of our analysis to offer insights into how callback decisions relate to differences in hiring outcomes.

3.3 Experimental Manipulation

We now discuss our randomized experimental treatments. Because our experiment uses outsourced recruiters, our two-sided design allows us to randomize both candidate and employer characteristics simultaneously.

Candidate Side. For each job application, we assigned the candidates a gender (male or female), race (white or black), education (undergraduate degree from elite or non-elite university), and prior employer (large or small firm). We chose candidate first names and last names to suggest a gender and race, and listed colleges and employers directly on the job application. Appendix **??** displays an example job application, and Appendix **??** lists the set of names, universities and employers. All applicants graduated with undergraduate degrees in computer science and related coursework.

Our candidate manipulations were meant to induce differences across candidates about the likelihood of receiving and accepting offers as well as the callback decision. They also embody candidate characteristics about which prior research documents discrimination. For example, a variety of studies have found that female and black candidates are less likely to receive callbacks than male and white candidates, respectively (see, for example Bertrand and Mullainathan, 2004). This could arise due to discriminatory behavior on the part of employers, or because of behavior on behalf of job applicants (for example, female candidates being less likely to accept job offers). Prior work has also found that credentials from elite universities increase callback rates (Gaddis, 2015). These degrees may increase desirability to employers by imparting valuable skills and networks, but these candidates may be less likely to accept job offers. Finally, our large company intervention is intended to capture the large-firm wage premium (Abowd et al., 1999; Song et al., 2019). For example, large company applicants may be desirable to employers if they were exposed to high productivity business practices, but may also have lower probabilities of accepting offers than similar candidates from small companies. Overall, the candidate manipulations were intended to create meaningful differences in candidates' likelihood of receiving and accepting offers and to focus on candidate traits featuring multiple, competing mechanisms for sorting in prior work.

Employer Side. On the demand side, we manipulate the characteristics of the firm that are sent to the recruiter. The instructional materials mentioned that the decision-to-interview bonus depended on interviews conducted by a hiring manager whose biographical information was disclosed. We randomly assigned each recruiter to one of nine different hiring managers. We randomly manipulated each hiring manager's gender (male or female), race (white or black) and prior education (elite or non-elite) and included a 9th demand-side treatment where gender, race, and education were blinded.[17]

Like our candidate manipulations, our manager manipulations were intended to affect the likelihood that job offers would be accepted or extended. Several prior papers have documented patterns of discrimination correlated with manager characteristics, although this could arise for several plausible reasons. A growing literature studies worker or candidate discrimination against managers (Stoll et al., 2004; Giuliano et al., 2009; Chakraborty et al., 2018; Doerrenberg et al., 2020; Ayalew et al., 2019; Abel, 2019; Abraham and Burbano, 2019). In this case, candidates may exhibit lower likelihoods of accepting job interviews and offers for managers they discriminate against. The manager manipulations may also influence a candidate's likelihood to pass the interview. Prior research suggests that certain managers have been pre-sorted into better organizations or more powerful positions (Brooks et al., 2014; Gornall and Strebulaev, 2019; Grossman et al., 2019), so that the manager manipulations influence the likelihood of extending job offers. For example, if VCs prefer funding white men (Brooks et al., 2014), firms with white male managers may have greater financial resources and ability to hire. Overall, the manager manipulations aimed to create differences in the probability of extending and accepting job offers.

[17] Our assumption is that hiring managers' characteristics are not perfectly known at the time of application, and that an offer of an interview with the hiring manager contains some new information about the hiring manager's characteristics.

Details of Simultaneous Randomization. For all 16 candidate types, we created three distinct instances for a total of 48 different candidates.[18] We compiled these into three "packets" of 16 job applications. In each packet, at least one instance of all 16 types appeared. All three packets were then matched with all nine hiring managers. This resulted in 27 unique sets of recruiter materials. Each unique set of materials was then evaluated twice by two separate recruiters, requiring hiring 54 recruiters (27×2).[19]

Balance. Because we randomize both sides of the market, we check for randomization balance in both candidate and manager characteristics. Appendix **??** shows that our random assignments are well-balanced across candidate and hiring manager manipulations (partly by construction), while Appendix **??** elaborates on the balance requirements of our design.

3.4 Specifications

To analyze our experiment, we use the five regression equations below (OLS). All equations are estimated with robust standard errors clustered at the screener level (Abadie et al., 2017).

Belief Correlations. Our first set of analyses are about whether participants on the same side of the market agree on the other side. We compute the correlation of *P(Accept Job | Employer Offers Job)* for every pair of candidates (and the same for *P(Accept Interview | Employer Offers Interview)*). If two candidates are chosen at random, how correlated are their behaviors about the same employers?

In addition, we place this question into a regression framework. The regression asks,

[18] For example, there were three white, female candidates from an elite university. Each of the three had a different white, female-sounding name and a different elite university.

[19] In Appendix **??**, we measure levels of cross-validation in recruiter assessments. Overall, the levels of cross-validation are relatively low but positive and statistically significant, but are relatively low. We discuss reasons for this in Appendix **??**.

"How well can we predict candidate c's likelihood of accepting a job from employer h as a function of the set of other candidates' ($K \neq c$) likelihood of accepting the job from k?" The regression we run is:

$$Y_{c,h,s} = \beta_0 + \beta_1 \underbrace{\left[\frac{1}{N}\sum_{k \neq c, s' \neq s} Y_{k,h,s'}\right]}_{\substack{\text{Average of other} \\ \text{candidates' } Y \text{ for} \\ \text{the same manager } h.}} + \epsilon_{c,h,s} \tag{1}$$

where c indexes candidates, h indexes hiring managers, and s indexes screeners. A high β_1 means that candidates largely agree about managers, and a β_1 close to zero means that candidates' probabilities are uncorrelated. We also run these analyses in the reverse direction, showing how correlated managers' views of workers are using $P(Pass)$.

Lastly, we estimate how much demand is reciprocated across the two sides of the market. If a candidate is likely to accept an interview or offer from a manager, is the manager likely to give an offer to the candidate? We measure this in three ways. First, we measure the simple Spearman correlation between the recruiter's beliefs about the probability that candidate i would accept a job from manager j if extended, and the probability that candidate i would pass an interview by manager j if an interview were held. We then place this analysis into a regression, by predicting $P(\text{Accepts Job Offer})_{i,j}$ from $P(\text{Passes Interview})_{i,j}$. Finally, we report the probability that a manager j ranks a candidate i above the median of $P(Passes)$, given that the candidate i ranked j above their median in $P(AcceptsJobOffer)$.

Impact of Candidate Characteristics. To study the impact of candidate characteristics, we estimate the regression:

$$\begin{aligned} Y_{c,h,s} = &\beta_0 + \beta_1 * Female_c + \beta_2 * Black_c + \beta_3 * EliteUniversity_c + \beta_4 * LargeCompany_c + \\ &\alpha * S_s + \gamma * HM_h + \epsilon_{c,h,s} \end{aligned} \tag{2}$$

$Female_c$, $Black_c$, $EliteUniversity_c$, and $LargeCompany_c$ are binary indicators of candidate c's characteristics, S_s is a vector of screener controls, and HM_h is a vector of hiring manager fixed effects. $Y_{c,h,s}$ measures an outcome Y (callback, p_{hire}, or one of the three underlying probabilities) for candidate c assigned to hiring manager h and screener s. β_1, β_2, β_3, and β_4 capture the effects of our supply-side treatment arms on screener beliefs about labor demand and labor supply.

Impact of Employer Characteristics. To study how manager characteristics impact screener beliefs, we estimate the following regression:

$$\begin{aligned} Y_{c,h,s} =& \beta_0 + \beta_1 * Female_h + \beta_2 * Black_h + \beta_3 * EliteUniversity_h \\ & + \beta_4 * Blinded_h + \alpha * S_s + \delta * C_c + \epsilon_{c,h,s} \end{aligned} \tag{3}$$

where S_s is a vector of screener controls and C_c is a vector of candidate fixed effects. This equation is similar to Equation 2, but binary indicators correspond to hiring manager characteristics rather than candidate characteristics and the fixed effects are now at the candidate level. In Equation 3, $Female_h$, $Black_h$, $EliteUniversity_h$, and $Blinded_h$ measure whether hiring manager h is female, black, from an elite university, or blinded, respectively, so β_1, β_2, β_3, and β_4 estimate the effects of our employer treatment arms.

Match Specific Effects. We are also interested in examining whether there are match-specific qualities that drive recruiter callback decisions. To do so, we estimate the following model:

$$Y_{c,h,s} = \beta_0 + \lambda * M_{c,h} + \alpha * S_s + \gamma * HM_h + \delta * C_c + \epsilon_{c,h,s} \tag{4}$$

where $M_{c,h}$ is a vector of match fixed effects for each possible candidate-manager pair.[20] The regression also includes screener controls plus fixed effects for candidates and for man-

[20] Given that we have 16 unique candidate profiles and 9 unique hiring manager profiles, there are a total of 144 fixed effects.

agers. The match fixed effects capture differences driven by matching specific candidates to specific managers (above each side's fixed effect). We estimate this equation for our four main dependent variables and run an F-test to see if the match fixed effects jointly predict outcomes.

Synthesis into Callback Choices. Our final set of specifications is about how beliefs about worker and manager behavior are combined into a callback decision. We estimate the following equation:

$$\begin{aligned} Y_{c,h,s} =& \beta_0 + \beta_1 * ProbabilityAcceptInterview_{c,h,s} + \beta_2 * ProbabilityPassInterview_{c,h,s} + \\ & \beta_3 * ProbabilityAcceptOffer_{c,h,s} + \alpha * S_s + \epsilon_{c,h,s} \end{aligned} \tag{5}$$

$Y_{c,h,s}$ measures the callback choice Y for candidate c assigned to hiring manager h and screener s. S_s is a vector of screener-level controls, and $\epsilon_{c,h,s}$ is the error term. $ProbabilityAcceptInterview_{c,h,s}$, $ProbabilityPassInterview_{c,h,s}$ and $ProbabilityAcceptOffer_{c,h,s}$ are probabilities measured through the procedure in Section 3.1.

Multiple Comparisons: To assuage multiple comparisons concerns (List et al., 2019), we apply the free step-down resampling methodology of Westfall and Young (1993) to control the probability of a Type 1 error.[21] Our tables containing multiple hypothesis tests display both conventional standard errors as well as adjusted p-values.

[21] We use the bootstrapped version of the Westfall and Young (1993) adjustment, which Jones et al. (2019) use to correct for multiple comparisons.

4 Results

4.1 Belief Correlations

We begin by studying the correlations between recruiters' beliefs about candidates and about employers. In Table 3, we study how much recruiters believe that candidates (or managers) will behave similarly to others on their own side of the market. For candidates, we study their willingness to accept offers from the same set of employers. For employers, we study their willingness to extend offers to the same set of workers. Our results show

Table 3: **Same-Sided Belief Correlations**

	Candidates	**Employers**
Measure of Demand for Other Side	P(Accepts Job)	P(Extends Offer)
Mean Pairwise Preference Correlation (Same Side)	*+0.48*	*+0.14*
Minimum Pairwise Correlation	-0.13	-0.57
25th Percentile	+0.32	-0.05
Median Pairwise Correlation	+0.49	+0.17
75th Percentile	+0.64	+0.34
Maximum Pairwise Correlation	+0.93	+0.77
Regression Coefficient	0.89**	0.59***
	(0.26)	(0.11)

Notes: This table shows correlations between the behaviors among the various participants in our paper. We assesses whether two candidates have the same behavior by measuring how correlated their willingness to accept (or reject) the same jobs are (when offered). In the final row, we show a regression version of this analysis using Equation 1. Section 3.4 contains additional details of this analysis. The variable names above use shorthand of *P(Pass)*, *P(Accepts)*, etc.; full variable definitions are in Section 3.1.

that candidates' predicted behaviors largely overlap, but that employers are expected to diverge idiosyncratically. The correlation between two randomly-chosen candidates' behavior is 0.48; by contrast, this value is much lower for managers (0.14). In the final row of Table 3, we show that any given candidate's willingness to accept a job can be easily predicted from other candidates' willingness to accept a job using our regression setup. Predicting managers' behavior from their peers is far less informative.

Table 4 examines cross-sided correlations to study whether recruiters believe that candidate-employer interest tends to be reciprocated. We find very weak levels of correlation and reciprocity. Our measures of demand are correlated at only $\rho = 0.12$ across the market. If a candidate ranks an employer in her top half, only 54% of employers reciprocate (ranking the candidate in their top half)— just slightly more than random.[22]

Table 4: **Cross-Sided Belief Correlations**

		Estimate	SE
Correlation:	$P(\text{Accepts Job Offer})_{i,j}$ with $P(\text{Extends Job Offer})_{i,j}$	+0.12***	
Coefficient:	$P(\text{Accepts Job Offer})_{i,j} = \beta_0 + \beta_1 P(\text{Extends Job Offer})_{i,j} + \epsilon$	+0.13***	(0.07)
Probability:	Employer j Ranks Candidate i above median, given i ranked j above median	+0.54***	(0.01)

Notes: If a candidate likes an employers, does the employer like the candidate back on average? In row one, we measure the simple Spearman correlation between the recruiter's beliefs about the probability candidate i would accept a job from j if extended, and the probability that candidate i would pass an interview by employer j if an interview were held. In row 2, we place this question into a regression framework. In the final row, we measure the probability that an employer j ranks a candidate i above the median, given that the candidate i ranked j above their median.

4.2 How do candidate characteristics affect screener beliefs?

Table 5 shows how candidate characteristics affect recruiter beliefs and callback behavior using Equation 2. The results indicate that screeners believe female applicants are more likely to accept interview offers and job offers. Placing greater weight on candidate expectations would therefore improve the callback rate for women. By contrast, the point estimates on black, elite university, and large company candidates are statistically indistinguishable from zero and are estimated precisely enough to mostly rule out large effect sizes (or effects as large as the female result).[23]

While expectations about candidate behavior help female applicants, expectations about

[22]The results in this section use P(Accepts Job Offer) as the measure of candidate behavior; we get similar results when using P(Accepts Interview).

[23]The one exception is the effects of attending an elite university (a negative effect). Our estimates of this coefficient are slightly less precise, and in some specifications we cannot rule out an effect as large as the one for female (but in the opposite direction).

Table 5: **Effects of Candidate Characteristics**

	Supply side		Demand side	Overall	
	P(Accepts Interview)	P(Accepts Offer | Passes Interview)	P(Passes Interview | Accepts Interview)	P(Hired)	Interviewed
Female Job Applicant	0.03***	0.04***	0.01	0.04***	0.04
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
Black Job Applicant	0.01	0.01	0.03**	0.02**	0.04
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)
Elite University Job Applicant	-0.01	-0.01	0.05***	0.01	0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)
Large Company Job Applicant	0.00	-0.02	0.05***	0.02	0.09***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
R^2	0.33	0.35	0.16	0.38	0.08
Observations	864	864	864	864	864
Fixed effects	Manager	Manager	Manager	Manager	Manager
Controls	Screener	Screener	Screener	Screener	Screener
Control mean	0.74	0.70	0.65	0.35	0.50
F-test	0.06	0.02	0.00	0.00	0.00
P-values:					
Female	0.04	0.01	0.87	0.01	0.72
Black	0.84	0.87	0.12	0.22	0.85
Elite university	0.87	0.87	0.01	0.87	0.01
Large company	0.87	0.72	0.00	0.72	0.05

Notes: This table displays the results of Equation 2 on predictions about the supply-side (columns 1 and 2), predictions about the demand-side (column 3), and overall hiring beliefs and behavior (columns 4 and 5). The regression controls for screener characteristics and includes robust standard errors clustered at the screener level. The bottom panel displays p-values adjusted for multiple comparisons (4 treatments $\times$ 5 outcomes) using the free step-down procedure of Westfall and Young (1993). The probability outcomes should be read conditionally as described in Section 3.1.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

employer behavior help elite university and large company applicants.[24] On gender, screeners believe that hiring managers have approximately similar behavior with male and female candidates, and our confidence intervals rule out effects as large as those above.

In our final two columns, we show how candidate characteristics affect callback decisions. Screeners believe that female candidates are more likely to result in a hire (if called back), and this effect is driven by beliefs about *candidate* behavior. We also find that black applicants are believed to be more likely to result in a hire (an effect driven by beliefs about hiring manager behavior), although this result does not fully survive our multiple hypothesis $p-$value corrections. By contrast, our subjects do *not* believe that

[24]We also find a small effect on black applicants, although these don't survive our multiple-hypothesis adjustments.

elite university and large company callbacks are more likely to produce a hire: although screeners believe managers would hire such applicants, these applicants are *not* more likely to accept interviews and final offers.

The final column in Table 5 shows that recruiters' beliefs about hiring probabilities are not fully incorporated into interview decisions. While screeners believe female and black applicants have higher probabilities of leading to a successful hire (compared to male and white applicants), screeners do not extend more interviews to these candidates. Instead, screeners are more likely to extend interview offers to applicants who attended elite universities and come from large companies. These effect sizes represent 28 and 18 percent increases in callback rates, respectively, relative to the control mean of 0.500. They are also statistically significant after our p-value adjustments.

4.3 How do hiring manager characteristics affect screener beliefs?

Table 6 shows how hiring manager characteristics impact recruiters' beliefs and callback behavior using Equation 3. Our estimates in this section are generally less precise because standard errors are (necessarily) clustered at the screener level, and each screener reviews 16 candidates. Our multiple hypothesis adjustments also reduce the significance of our tests. Nonetheless, we do have suggestive evidence about the effects of manager characteristics.

Our strongest results are about recruiter beliefs about candidate discrimination against the manager (measured by candidates' willingness to accept interviews and job offers). Blinded, black and female managers face lower probabilities of candidate acceptances, leading to a lower probability of a hire (even conditional on a callback). We can compare these to our previous results about the effects of candidates' race and gender. The effects of being a black or female manager on candidate labor supply are clearly larger than the effects of being a black or female candidate on employer willingness to hire (both statistically and in magnitude).

Table 6: **Effects of Hiring Manager Characteristics**

	Supply side		Demand side	Overall	
	P(Accepts Interview)	P(Accepts Offer | Passes Interview)	P(Passes Interview | Accepts Interview)	P(Hired)	Interviewed
Female Hiring Manager	-0.06*	-0.05*	-0.00	-0.07*	-0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)
Black Hiring Manager	-0.06*	-0.09***	0.00	-0.08**	0.01
	(0.03)	(0.03)	(0.03)	(0.04)	(0.07)
Elite University Hiring Manager	0.03	0.02	0.03	0.04	0.00
	(0.03)	(0.02)	(0.02)	(0.03)	(0.06)
Blinded Hiring Manager	-0.12**	-0.15***	-0.01	-0.17***	-0.10
	(0.05)	(0.04)	(0.04)	(0.05)	(0.08)
R^2	0.39	0.38	0.25	0.42	0.18
Observations	864	864	864	864	864
Fixed effects	Candidate	Candidate	Candidate	Candidate	Candidate
Controls	Screener	Screener	Screener	Screener	Screener
Control mean	0.77	0.75	0.65	0.40	0.68
F-test	0.03	0.00	0.71	0.00	0.78
P-values:					
Female	0.97	0.97	0.99	0.96	0.99
Black	0.97	0.80	0.99	0.94	0.99
Elite school	0.99	0.99	0.98	0.97	0.99
Blinded	0.89	0.66	0.99	0.76	0.99

Notes: This table displays the results of Equation 3 on predictions about the supply-side (columns 1 and 2), predictions about the demand-side (column 3), and overall hiring beliefs and behavior (columns 4 and 5). The regression controls for screener characteristics and includes robust standard errors clustered at the screener level. Finally, the bottom panel displays p-values adjusted for multiple comparisons (4 treatments $\times$ 5 outcomes) using the free step-down procedure of Westfall and Young (1993). The probability outcomes should be read conditionally as described in Section 3.1.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

4.4 Match-specific Factors

We now test for the presence of match-specific factors, or combinations of characteristics on the supply and demand sides that could change beliefs or callback rates. This would occur, for example, if a recruiter might have higher expectations and callback rates when a white male boss interviews a white male candidate (versus other types of managers and candidates). We study this using Equation 4, which includes fixed effects for all combinations of candidate $\times$ manager characteristics.

Our first result is to show that these factors exist and are detectable. In Appendix **??**, we display the results of the joint test that all fixed effects are zero (plus a count of the individual fixed effects that individually return $p < 0.05$). The results provide strong evidence that recruiters expect match-specific behavior. Across each of our four dependent variables in Table **??**, the joint-test of match fixed effects returns $p < 0.05$. These joint tests indicate that there are certain worker-manager combinations with higher outcomes above and beyond the average outcome for the focal worker and manager.

Our a priori hypothesis was that homophily could explain these match-specific patterns.[25] In Appendix Section **??**, we investigate this question but do not find strong evidence of homophily. In Table 7, we re-estimate Equation 4 for callback rates on subsets of the data covering all candidate (or manager) characteristics and present the p-value of the joint test of no match-specific effects. Appendix **??** displays the corresponding table for all key outcomes.

Table 7: **Match-specific Effects on Interviews (by characteristics)**

Match–specific analysis, Interview				
	Candidate		Manager	
Type	F-stat	p-value	F-stat	p-value
Men	44.3	< 0.01	46.3	< 0.01
Women	79.3	< 0.01	11.9	< 0.01
White	21.1	< 0.01	18.3	< 0.01
Black	114.7	< 0.01	30.3	< 0.01
Elite	26.8	< 0.01	14.4	< 0.01
Non-elite	22	< 0.01	71.7	< 0.01
Large company	32.7	< 0.01		
Small company	8363.1	< 0.01		
Blinded			0.39	0.69

Notes: This table displays the results of a regression of interviews on match-type fixed effects, screener controls, and candidate and hiring manager fixed effects using equation 4. We subset the regression for each candidate/manager type, with candidate types on the left of the table and hiring manager types on the right. The table displays the p-value and F-statistic from a joint test that all match type fixed effects (within the given type) are equal. In Appendix **??**, we display the corresponding table for all key outcomes.

Because this analysis was not pre-registered, we interpret these results as exploratory. The results show that we have stronger evidence for match specific effects for female and

[25] For example, beyond the potential benefit of being a male candidate or hiring manager (on average), male candidates may be particularly advantageous when evaluated by other men (see McPherson et al. 2001 for an excellent review of this literature).

black candidates. Stated differently, these candidates' outcomes are more variable (i.e., dependent on who the interviewer is) than for male and white candidates, respectively. In Appendix **??** we display the distribution of fixed effects by race and gender (collapsing the education and prior experience manipulations for ease of interpretability). The results indicate that callback rates for white male candidates vary the least with the identity of their hiring manager. Callback rates for white women, black men, and black women are more variable with the identity of their hiring manager: black men and white women have the highest callback rates when assigned to a black male manager, while black women have the lowest callback rates when assigned to a white male manager.

4.5 How are beliefs about candidate and hiring manager behavior synthesized into decisions?

We finally explore the role that expectations about candidate and hiring manager behavior play in callback decisions in Table 8 using Equation 5. Across a broad variety of specifications, we find several patterns. First, our results show that beliefs about candidate behavior play a significant role in who is selected for an interview. This contrasts with the typical interpretation of an audit study as reflecting only employers' behavior. Throughout all of our specifications, coefficients on our candidate behavior measures are statistically and economically significant. Second, we find that recruiters place a greater weight on candidate behavior *early* in the hiring process than later. Beliefs about the candidate's willingness to be interviewed are especially influential. Beliefs about the likelihood of a candidate accepting a job offer are also influential, but less than beliefs about accepting interviews. Overall, these results indicate that hiring managers are placing weight on their expectations of candidate behavior.

Finally, we find that recruiters place *much* greater weight on hiring managers' behavior than on candidates. In theory, if the recruiter was maximizing their expected task payment, then two candidates with the same p_{hire} should be equally attractive regardless of the underlying probabilities. We do find that the candidate measures receive some positive weight. However, if we compare coefficients from candidates versus employers, we see that our measures of employer behavior receive over three times the weight of candidate ones in most of our specifications.

Table 8: **Callback Decisions and Predictions**

	Interviewed	Interviewed	Interviewed	Interviewed
P(Accepts Interview)	0.53***			0.38*
	(0.15)			(0.22)
P(Passes Interview)	1.64***	1.65***	1.65***	1.48***
	(0.11)	(0.11)	(0.11)	(0.23)
P(Accepts Offer)	0.09	0.41***		-0.08
	(0.15)	(0.15)		(0.13)
Avg P(Applicants Accept)			0.60***	
			(0.17)	
R^2	0.45	0.43	0.44	0.45
Observations	864	864	864	864
P(Hire) decile control	No	No	No	Yes
P-values:				
Pass interview = Accept interview	< 0.01			< 0.01
Pass interview = Accept offer	< 0.01	< 0.01	< 0.01	< 0.01

Notes: This table examines the relationship between call-back decisions and supply- and demand-side behavior through our specification in Equation 5. All regressions control for screener characteristics and include robust standard errors clustered at the screener level. The bottom panel displays p-values testing whether demand-side behavior receive the same weight as supply-side ones. Columns 1 and 4 test two hypotheses each, so these columns display p-values adjusted for multiple comparisons using the free step-down procedure of Westfall and Young (1993). The probability outcomes should be read conditionally as described in Section 3.1. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

One possibility is that reputational incentives may compel recruiters to impress employers by delivering employer-approved candidates (at the expense of yield management). If this were true, then recruiters with a more established reputation may be more willing to engage in yield management. In Appendix **??**, we investigate this hypothesis. Across several measures, our results show that more established recruiters place more weight on candidate acceptance behavior suggesting that callback decisions may in part reflect reputation-seeking behavior by recruiters.

5 Discussion

The results in this paper open up three important points related to external validity, the accuracy of recruiter beliefs, and representation in recruiting.

5.1 External validity

In this section, we briefly consider the external validity of our results using the selection, attrition, naturalness, and scalability (SANS) conditions described in List (2020).

Selection. The subject pool for our experiment is recruiters from an online labor market, which are broadly representative of the population of recruiters (Agan et al., 2021). Moreover, the materials given to recruiters, including the resumes and job postings, were based on actual job applications and firm openings in the technology industry. Using the technology sector impacts the external validity of our results. We find that in this sector, recruiters determine callbacks by placing around 2/3rds of the weight on manager behavior and 1/3rd of the weight on candidate behavior. Given the sector regularly features labor shortages of talented software engineers, we might expect to see recruiters place even less weight on candidate behavior in settings where there is less competition for workers.

Attrition. We consider attrition in both the sample of recruiters and in the data collected for each resume. To meet our sample size goals from Section 3 (54 recruiters such that each unique set of material could be evaluated by twice), we invited 83 recruiters using the sequential re-randomization procedure outlined in Appendix **??**. One-third of which did not complete the task. In Appendix Section **??**, we show that attrition is not correlated with treatment assignment nor recruiter characteristics. Meanwhile, no candidate outcomes (for example, the interview decision) are missing for recruiters who sent work back.

Naturalness. We designed our experiment to mimic the type of work that recruiters would encounter in their jobs. The resumes and job descriptions were based on actual workers and firms in the industry, and the financial incentives mirrored those that recruiters face. We further elaborate on these design choices in Sections 2 and 3.

Scalability. Given this is not a programmatic study making recommendations to policymakers, we do not consider scaling. Given our paper focuses on establishing causality and illustrating the mechanisms behind employee-employer matching, we intend for our paper to serve as a wave 1 study using the typology in List (2020). Although our evidence comes from a particular industry (software engineering), other researchers may adapt our conceptual framework and research design. More broadly, the two-sided audit design introduced in this paper can serve as the basis for future work on employee-employer matching in other settings and for other research questions.

5.2 Are Recruiter Beliefs Accurate?

Recruiters in our study anticipate managers and candidates' future behavior. Are recruiters' beliefs about these choices accurate, and to what extent do recruiters differ in their callback decisions versus managers? To answer these questions, we collected survey data of job candidates and hiring managers. Although the candidates and hiring managers seen by recruiters were fictitious, we found survey subjects with similar characteristics and inquired about their behavior directly. Our goal was to collect data directly from workers and hiring managers about their likely outcomes in the scenario of our experiment – without the intermediation of recruiters. These survey responses permit a test of recruiter accuracy. Appendix **??** contains full details of these surveys.[26] Because this data was

[26] Because of limitations of the Prolific survey sample, we were not able to gather data from black engineers and managers. According to most industry statistics, black representation in the software industry is low. According to the BLS, the software engineering workforce is approximately 5% black (https://www.bls.gov/cps/cpsaat11.htm) As a result, we could not find a large sample of black engineers and engineering managers to survey. However, the group we do study contains the people most likely to be making choices

collected through surveys about hypothetical situations, we interpret these results as suggestive.

Study Design. Our source for these subjects was Prolific.co, a survey company that maintains a survey panel of software engineers and their managers. On the employer side, we presented approximately 250 software engineering managers with a job description and company like the one in our main experiment. We then asked the managers to assess a series of eight candidates that paralleled the candidates in our main recruiter experiment.

On the candidate side, we also recruited about 250 software engineering workers to perform a similar task from the perspective of the job-seekers. Each candidate in our survey reviewed nine hiring managers (including a blinded one) that were also parallel to those in our experiment. They reported if they were likely to accept an interview and/or job offer if one was extended by the company with this person as their manager.

We measure the accuracy of recruiter beliefs by how biased their forecasts were.[27] To quantify this, we combine our survey data with the experimental data into a single dataset. On the supply side of the market, each observation consists of an evaluation of hiring managers. Some are by recruiters anticipating how candidates would respond (from the main experiment) and some are from the candidates themselves (through the survey). Appendix Table **??** examines how the evaluations of hiring managers change depending on whether recruiters or job candidates perform the evaluation, and for which type of hiring managers. Appendix Tables **??** and **??** do the same analysis for the demand side of the market, comparing evaluations of job candidates by recruiters versus hiring managers.

Forecast Accuracy. Overall, our results suggest that recruiters were relatively unbiased in forecasting the behavior of job candidates and managers, with the exception of gender.

in this industry (either as managers or as job candidates). We do capture how these subjects react to the possibility of a black manager or job candidate, which is a critical question for increasing representation and part of the motivation of our study.

[27] In the spirit of Bohren et al. (2023), we measure the bias of the forecasts and not the variance, although variance is also a component of some methods for studying forecast accuracy.

For most of our estimates, we cannot reject a null hypothesis of zero difference between recruiter forecasts and subjects' reports. Our standard errors are precise enough that our 95% confidence intervals rule out large effects in either direction. Even when we can reject zero differences, we can rule out large differences. In some cases, we find that recruiters forecast overall higher levels for all types (i.e., a level effect) compared to survey-takers, although these level differences are also relatively small. We focus on interactions, or situations where recruiters report systematically different forecasts for particular types of candidates.

On the supply side (Table **??**), recruiters are relatively accurate in candidate perceptions of the manager's race, education, and the blinded condition. However, recruiters believe that candidates are 12 percentage points less likely to accept positions from companies with a female hiring manager (than from companies with male hiring managers). The candidates themselves report being slightly more likely ($\approx$1pp) to accept offers from companies with female hiring managers. The candidates in our survey also believe that elite-university hiring managers have a lower $P(PassInterview)$, whereas the recruiters believe they are more likely to pass candidates. For all of the outcomes we collect from candidates, we can reject the joint hypothesis that the recruiters' coefficients were the same as the candidates'. However, the magnitude of differences appears to be small.

On the demand side (Tables **??** and **??**), we also similarly find relatively small differences. Recruiters appear to be slightly overoptimistic about the prospects of black candidates. Managers themselves report no differences in their propensity to extend offers to candidates with black names (versus white ones). By contrast, recruiters view these candidates as more likely to pass the interview and receive an offer, but only by 3 percentage points. We also find that recruiters were more optimistic than actual hiring managers about the potential for women to accept offers. When we conduct a joint test for each of the three probability measures of whether recruiters differ in their assessments compared to managers, we fail to reject the null, with p-values ranging from 0.26 to 0.15. Thus, while recruiters were inaccurate in their assessments on the likelihood of accepting offers for female job candidates and on passing interviews for black candidates, our overall results suggest a moderately high degree of accuracy in the three probability measures.

Do Recruiters Decide Differently? While recruiters' forecasts were relatively accurate, there were larger differences about who the recruiters recommended to interview (compared the managers). Table **??** indicates that recruiters place a much higher weight on going to an elite university when making callback choices. Managers were 5 percentage points more likely to interview elite university versus non-elite university candidates, but recruiters were 21 percentage points more likely to interview these candidates. We can reject the joint hypothesis that the recruiters' coefficients were the same as the managers for the interview choices ($p = 0.01$).

In Table **??**, we examine how the probabilities are synthesized into interview choices by recruiters (versus managers themselves). Our analysis here is similar to Table 8 in the main experiment (discussed in Section 4.5), but we now measure the differences between recruiters and managers. Our results suggest that recruiters place a higher emphasis on finding candidates who will pass interviews, and that managers themselves place a higher weight on yield. These results suggest that recruiters shift interviews towards candidates who do well on measures of passing interviews (in our study, these are elite university job applicants), while forgoing job candidates who do well on measures of accepting job interviews and offers.

Our survey results have three implications for how recruiters influence employee-employer matching. First, recruiters appear moderately well-versed in understanding job candidate and manager behavior. Outside of gender, recruiters' beliefs are relatively accurate about how managers and candidates might respond to each other.

Regarding gender, these inaccuracies suggest several ways that recruiters impact labor market sorting. By introducing inaccurate beliefs, recruiters could create (or prevent) matches that might not happen if job candidates and employers matched without a recruiting intermediary. The gender inaccuracies also show why treating the two sides of the market distinctly is an important feature of our design: Inaccuracies impact women differently, depending on the side of the market. Recruiters' inaccuracy appears to favor females on the candidate side, but penalizes them on the manager side. On average, the inaccuracy does not necessarily help or hurt women uniformly, but has differential effects depending on their role. A promising avenue for future research is to test interventions that influence recruiter accuracy about characteristics such as gender.

Finally, we find that recruiters shift the types of and amount of candidates who are interviewed, despite having relatively accurate beliefs. We show how these differences can arise not necessarily from differences in information (e.g., inaccuracy), but from different costs and objectives between the recruiter and the manager. Our results suggest that differences are correlated with how important passing each step of the hiring process is. Recruiters place more weight on finding candidates they believe the employer will pass, and managers place more weight on mutual candidate-manager interest. Job candidates with higher chances of passing interviews appear to benefit from delegated recruiting, even if they do not necessarily share mutual interest with the firm. We find some suggestive evidence (in Appendix Section **??**) that one potential reason is the recruiters' need to maintain a reputation or relationship with employers.

5.3 Recruiter Demographics and Representativeness

Taken together, recruiter beliefs contain several instances of inaccuracy but most deviations are relatively small (around 1-3 percentage points). Because this data was collected through survey vignettes involving hypothetical situations — rather than from real randomized hiring – we interpret these results as suggestive. However, they may be related to a more general feature of outsourced recruiting in practice: Recruiters are demographically very different from the candidates and employers they serve.

Outsourced recruiting is often used to lower firms' HR costs. To achieve cost savings, these firms delegate recruiting choices from high-wage, high skill workers to lower-skilled, lower wage ones. The backgrounds of these recruiters are very different from those in their client industry.[28] Such differences may in theory introduce biases and distortions in the hiring process.

It is unlikely, however, that non-representation drives the instances of inaccurate beliefs that we see in our sample given that the majority of our recruiters were females. However, the extent to which better representation in the recruiter pool leads to more accurate assessments of candidate and manager behavior is an open topic for future work. Understanding the beliefs and behavior of recruiters, including interventions to correct their beliefs, presents a promising direction to further our understanding of employee-employer matching, and our paper aims to take a small step in this direction.

6 Conclusion

The use of third-party specialists in recruiting introduces new theoretical and empirical questions regarding the hiring process. We use this setting to study how beliefs about candidate and manager behavior are integrated for callback decisions. Hiring requires that both the worker and manager agree to a match, but prior work has struggled to decompose callback decisions into their candidate- and firm-specific parts. To do so, we run a novel two-sided audit study where we hire professional recruiters, assign them a job screening

[28]As an example, according to the BLS, the highest industry of employment for recruiters (aside from the RPO industry itself) is the *professional, scientific, and technical services*. This industry features a much more male workforce and higher wages. By contrast, the BLS's occupational data suggest that human resource work is mid-skill work requiring a bachelor's degree, but no related work experience or prior on-the-job training (see https://www.bls.gov/ooh/business-and-financial/human-resources-specialists.htm). In 2018 Human Resource workers across all industries were 69.7% female, 10.5% black, while the median hourly wage was $29.01 (see https://www.bls.gov/ooh/business-and-financial/human-resources-specialists.htm).

task, and manipulate the identity of workers in the candidate pool and the identity of the hiring managers responsible for conducting the interview. We then test recruiters' beliefs about both sides of the labor market (for example, how likely they believe a candidate is to accept an offer from a given manager), how these beliefs are influenced by candidate or hiring manager manipulations (for example, whether screeners believe women are more likely than men to accept job offers), and how these beliefs are integrated into callback decisions.

We find evidence that both candidate and hiring manager discrimination exists in the hiring process, but stronger evidence for candidate discrimination against managers. Screeners believe that candidates are less likely to accept job offers from black and female hiring managers. We find robust evidence that match-specific factors affect recruiters' beliefs and choices, causing the same candidate's outcome to vary widely depending on the employer. Because employer behavior is more idiosyncratic, match-specific variability is introduced by their behavior (more than candidates'). Black and female candidates face particularly high uncertainty, as employers' views of them vary widely.

Finally, we find that recruiters place about $\frac{2}{3}$rds weight on employer behavior (e.g., the more idiosyncratic and horizontal, demographically neutral side). Candidates' behavior (the more vertical, demographically sensitive side) receive about $\frac{1}{3}$rd weight. Our paper finds suggestive evidence that reputational incentives may compel recruiters to impress employers by catering to their wishes. Instead, employers often care about yield, and may not prefer this form of catering. An avenue for future research is studying why recruiters weigh employer behavior so strongly despite the incentives to manage yield.

In sum, the rise of outsourced recruiting will continue to have important implications for how employees are matched to employers in the modern economy. Our paper takes a small step in documenting these patterns by examining recruiter beliefs regarding employee and employer behavior, and how recruiters integrate these beliefs in determining callback decisions. Our hope is to inspire future work to gain a fuller understanding of how recruiters shape employee-employer matching and labor market sorting.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *NBER Working Paper 24003*, 2017.

Abel, Martin, "Do Workers Discriminate against Female Bosses?," 2019.

Abowd, John M, Francis Kramarz, and David N Margolis, "High wage workers and high wage firms," *Econometrica*, 1999, *67* (2), 251–333.

Abraham, Mabel and Vanessa C. Burbano, "The Importance of Gender Congruence in Corporate Social Responsibility: Field Experimental Evidence of Applicant Interest," *Working Paper*, 2019.

Acemoglu, Daron, "Changes in unemployment and wage inequality: An alternative theory and some evidence," *American economic review*, 1999, *89* (5), 1259–1278.

__ , "Technical change, inequality, and the labor market," *Journal of economic literature*, 2002, *40* (1), 7–72.

Agan, Amanda and Sonja Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 2018, *133* (1), 191–235.

Agan, Amanda Y, Bo Cowgill, and Laura K Gee, "Salary History and Employer Demand: Evidence from a Two-Sided Audit," Technical Report, National Bureau of Economic Research 2021.

Autor, David H, "Why do temporary help firms provide free general skills training?," *The Quarterly Journal of Economics*, 2001, *116* (4), 1409–1448.

Ayalew, Shibiru, Shanthi Manian, and Ketki Sheth, "Discrimination from below: Experimental evidence from Ethiopia," 2019.

Bertrand, Marianne and Esther Duflo, "Field experiments on discrimination," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 309–393.

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

Black, Ines, Sharique Hasan, and Rembrand Koning, "Hunting for Talent: Firm-driven Labor Market Search in the United State," 2022.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope, "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 2023, pp. 1–45.

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E Murray, "Investors prefer entrepreneurial ventures pitched by attractive men," *Proceedings of the National Academy of Sciences*, 2014, *111* (12), 4427–4431.

Cappelli, Peter, "Your approach to hiring is all wrong," *Harvard Business Review*, 2019, *97* (3), 48–58.

__ **, Steffanie L Wilk et al.**, *Understanding selection processes: organization determinants and performance outcomes*, Center for Economic Studies, US Department of Commerce, Bureau of the Census, 1997.

Carrillo-Tudela, Carlos, Bart Hobijn, Patryk Perkowski, and Ludo Visschers, "Majority of Hires Never Report Looking for a Job," *FRBSF Economic Letter, Available online at https://www.frbsf.org/economic-research/publications/economic-letter/2015/march/labor-market-turnover-new-hire-recruitment/*, 2015.

Chakraborty, Priyanka, Danila Serra et al., "Gender differences in top leadership roles: Does aversion to worker backlash matter?," Technical Report 2018.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan, "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, May 2016, *106* (5), 124–27.

Charles, Kerwin Kofi and Jonathan Guryan, "Prejudice and wages: an empirical assessment of Becker's The Economics of Discrimination," *Journal of political economy*, 2008, *116* (5), 773–809.

Cowgill, Bo, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening," *Working paper; Available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584916*, 2020.

Doerrenberg, Philipp, Denvil Duncan, and Danyang Li, "The (in) visible hand: do workers discriminate against employers?," 2020.

Gaddis, S Michael, "Discrimination in the credential society: An audit study of race and college selectivity in the labor market," *Social Forces*, 2015, *93* (4), 1451–1479.

__ , *Audit studies: Behind the scenes with theory, method, and nuance*, Vol. 14, Springer, 2018.

Giuliano, Laura, David I Levine, and Jonathan Leonard, "Manager race and the race of new hires," *Journal of Labor Economics*, 2009, *27* (4), 589–631.

Gornall, Will and Ilya A Strebulaev, "Gender, race, and entrepreneurship: A randomized field experiment on venture capitalists and angels," *Available at SSRN 3301982*, 2019.

Grossman, Philip J, Catherine Eckel, Mana Komai, and Wei Zhan, "It pays to be a man: Rewards for leaders in a coordination game," *Journal of Economic Behavior & Organization*, 2019, *161*, 197–215.

Heckman, James J, "Detecting discrimination," *Journal of economic perspectives*, 1998, *12* (2), 101–116.

— **and Peter Siegelman**, "The Urban Institute audit studies: Their methods and findings," *Clear and convincing evidence: Measurement of discrimination in America*, 1993, pp. 187–258.

Hunkenschroer, Anna Lena and Christoph Luetge, "Ethics of Ai-Enabled Recruiting and Selection: A Review and Research Agenda," *Journal of Business Ethics*, 2022, *178* (4), 977–1007.

Jones, Damon, David Molitor, and Julian Reif, "What do workplace wellness programs do? Evidence from the Illinois Workplace Wellness Study," *Quarterly Journal of Economics*, 2019, *134* (4), 1973–2042.

Kessler, Judd B., Corinne Low, and Colin D. Sullivan, "Incentivized Resume Rating: Eliciting Employer Preferences without Deception," *American Economic Review*, November 2019, *109* (11), 3713–44.

Kim, Danny and Mike Pergler, "Startup Hiring through Firm-Driven Search: Evidence from VFA," *Working paper*, 2022.

Kline, Patrick and Christopher Walters, "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 2021, *89* (2), 765–792.

Kohlhepp, Jacob and Stepan Aleksenko, "Delegated Recruitment and Hiring Distortions," *Available at SSRN 3905019*, 2021.

Kuncel, Nathan R., Deniz S. Ones, and David M. Klieger, "In Hiring, Algorithms Beat Instinct," *Harvard Business Review*, 2014, *May Print Issue.*

Lanning, Jonathan A, "Opportunities denied, wages diminished: Using search theory to translate audit-pair study findings into wage differentials," *The BE Journal of Economic Analysis & Policy*, 2013, *13* (2), 921–958.

Levy, Frank and Richard J Murnane, "With what skills are computers a complement?," *The American Economic Review*, 1996, *86* (2), 258–262.

Li, Danielle, Lindsey Raymond, and Peter Bergman, "Hiring as Exploration," *Working paper, available online at https://danielle-li.github.io/assets/docs/HiringAsExploration.pdf*, 2021.

List, John A, "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial," Working Paper 27535, National Bureau of Economic Research July 2020.

— **, Azeem M Shaikh, and Yang Xu**, "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 2019, *22* (4), 773–793.

Makridis, Christos and Maury Gittleman, "On the Cyclicality of Real Wages and Employment: New Evidence and Stylized Facts from Performance Pay and Fixed Wage Jobs," *Available at SSRN 3017034*, 2020.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 2001, *27*, 415–444.

Neumark, David, Roy J. Bank, and Kyle D. Van Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics*, 1996, *111* (3), 915–941.

Oyer, Paul, Scott Schaefer et al., "Personnel Economics: Hiring and Incentives," *Handbook of Labor Economics*, 2011, *4*, 1769–1823.

Perkowski, Patryk, "Gender representation and the adoption of hiring algorithms: Evidence from MBA students and executives," *Working paper*, 2023.

Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter, "Firming up inequality," *The Quarterly journal of economics*, 2019, *134* (1), 1–50.

Stoll, Michael A, Steven Raphael, and Harry J Holzer, "Black job applicants and the hiring officer's race," *ILR Review*, 2004, *57* (2), 267–287.

Westfall, Peter H and S. Stanley Young, *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*, Vol. 279, Hoboken, NJ: John Wiley & Sons, 1993.