Cowgill, Bo; Perkowski, Patryk

Working Paper

Delegation in Hiring: Evidence from a Two-Sided Audit

IZA Discussion Papers, No. 17004

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Cowgill, Bo; Perkowski, Patryk (2024) : Delegation in Hiring: Evidence from a Two-Sided Audit, IZA Discussion Papers, No. 17004, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/299932

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 2/3rds weight on employer's expected behavior and 1/3rd 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."

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.

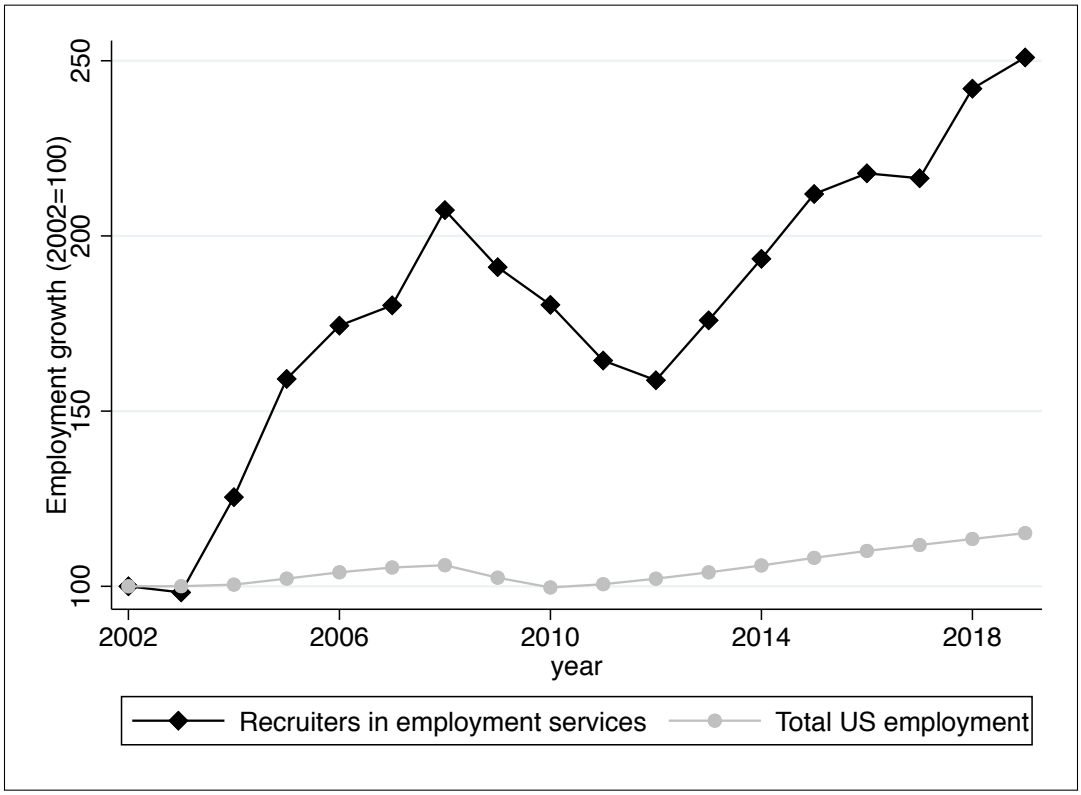
2 Institutional Setting

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.

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

[2] https://www.bls.gov/OOH/business-and-financial/human-resources-specialists.htm#tab-3
[3] https://www.workforce.com/2018/01/25/sector-report-tech-gaining-foothold-rpo-space/

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



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).

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

[4] https://www.workforce.com/2019/01/24/recruitment-process-outsourcing-providers-think/
[5] https://info.leveluphcs.com/nys-contact-tracing-video

number of recruiters in employment services more than doubled. Appendix A.1 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 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

[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

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% 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.

[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.

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

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 A.2. 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 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 et al. (2021) 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.

[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.

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

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 A.3 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.

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 A.4 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.

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.

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

[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).

[16] We test if subjects understood this conditioning in Appendix B.1.

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.

important factor in how workers are matched to employers within outsourced recruiting.

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 A.3.3 displays an example job application, and Appendix A.5 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,

[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.

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.

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 A.6 shows that our random assignments are well-balanced across candidate and hiring manager manipulations (partly by construction), while Appendix A.7 elaborates on the balance requirements of our design.

[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 B.2, 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 B.2.

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, "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 managers. 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.

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

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.

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.

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

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

[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).

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]

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

While expectations about candidate behavior help female applicants, expectations about 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.

[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).

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

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 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 B.3.1, 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 B.5, 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 B.3.2, 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 B.3.3 displays 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 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 B.3.4 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

[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).

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 B.3.3, we display the corresponding table for all key outcomes.

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 B.4, 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 A.7. One-third of which did not complete the task. In Appendix Section A.8, 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 C contains full details of these surveys.[26] Because this data was 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

[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 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.

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 C.1 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 C.2 and C.3 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. 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 C.1), 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 C.3 and C.2), 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 C.4 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 C.5, 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 B.4) 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

[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#tab-1). 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).

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 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," 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 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.

Appendix: For Online Publication Only

A Setting and Experimental Protocol Details

A.1 Growth in Outsourced Recruiting

Table A.1 uses government data sources to document the growth in outsourced recruiting. The table illustrates rapid growth in outsourced recruiting since 2000.

Table A.1: **Growth in Outsourced Recruiting**

Panel A: BLS' Occupational Employment Statistics data

	% change, 2002 to 2019		
	Employment	Mean hourly wage (nominal)	Mean annual wage (nominal)
Recruiters in employment services	+151%	+52%	+52%
Overall US economy	+15%	+50%	+50%

Panel B: Economic Census

	% change, 2002 to 2017			
	Employment	Establishments	Revenue	Annual payroll
	(1)	(2)	(3)	(4)
Employment placement agencies	+102%	+44%	+ 225%	+224%
Overall US economy	+24%	+17%	+83%	+90%

Notes: This panel examines growth in outsourced recruiting. Panel A uses data on human resource specialists (occupation code 13-1071) in the employment services industry (NAICS code 561300). We concentrate our attention on the "human resource specialists" occupation, 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). The employment services industry includes: (i) employment placement agencies (56-1310), (ii) temporary help services (56-1320), and (iii) professional employer organizations (56-1330). Ideally, we would concentrate our attention on employment placement agencies, but the OES data does not let us disaggregate to this industry level. We use this more narrow classification for the Economic Census data.

Panel B uses data from the Economic Census for employment placement agencies and executive search services (NAICS code 561310) for select industries that appear in both the 2002 and 2017 Economic Census data. The US Economic Census collects data on employment, payroll, and revenue for U.S. businesses every five years. Although the Economic Census does not break down its data by occupation, it contains more detailed industry classifications compared to the OES data. This allows us to focus our attention on "employment placement agencies" (NAICS code 56-1310, "establishments primarily engaged in listing employment vacancies and in referring or placing applicants for employment.")
The wage, payroll, and revenue figures reflect nominal increases.

A.2 Recruitment of Recruiter Subjects

We created a private job posting on a recruiting platform that was visible only to recruiters we invited.[1] We then used the platform's search function to generate a list of all possible eligible candidates. Candidates were eligible if they appeared in search results for specific keywords or categories,[2] were asking for an hourly rate of $100 or less,[3] and were located inside the U.S.[4]

We then sent a random sample of eligible recruiters a private message including a description of our task and an offer to hire the recruiter for two hours at the hourly rate listed on their profile. For screeners who accepted our offer, we undertook several verification measures.[5] We removed all subjects who did not report and/or document prior human resources experience in hiring. Before sending our application materials, we also asked each subject to answer three questions about the task to confirm their understanding of the instructions.[6]

A.3 Experimental Materials

A.3.1 Job Posting

Title:

Brief Feedback on 16 Job Applications

Description:

We are hiring a junior software engineer at our company. We have sixteen applications, and

[1] We did not make the post publicly available so that our research team could control the selection of recruiters into the task.

[2] We specifically used keywords "recruiting," "sourcing," "staffing," "human resources," "talent acquisition" (and minor variations).

[3] According to the BLS data on human resources, this includes all except the top $< 0.01\%$ of recruiters.

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

[5] We asked that subjects send us a resume or a publicly-facing LinkedIn profile so that we could learn more about their experience. Most sent a publicly-facing LinkedIn profile. We used this to confirm the subject's location and search for evidence of prior experience. In addition, the platform reported an estimated location based on the subject's IP address that we used to validate the location.

[6] In addition, we asked each recruiter to sign a non-disclosure agreement. This is a common practice in real-world recruitment outsourcing aimed at protecting employer and candidate privacy. All screeners signed the NDA, although some thought it was unnecessary because it was covered by the platform's terms-of-service.

would like some brief feedback on them (four short questions). We ask that you NOT contact the candidate, and instead send your feedback after reading the one-page job applications.
Each question can be answered either with a yes/no, or number between 0 and 100. There are only four such questions. No essay questions. You can send your answer back in a spreadsheet we'll provide you.
You should be either based in the US, or familiar with US institutions and culture (such as US schools and employers) so that you can better understand the candidates' qualifications.
We'll pay you hourly. We think you can easily finish the job within two hours, so we'll set a max of two billable hours. You can do the work anytime in the week after we hire you.
To accept this offer, please send us your LinkedIn profile or resume. We'll send you a packet of instructions. We'll verify that you read the instructions, send you the job applications then await your feedback.

A.3.2 Recruiter Instructions

Our instructions were approximately three pages with an additional six pages of FAQ. A sample set of instructions for hiring manager Katie Schmidt (University of Oklahoma) is below.

Screener Instructions

Thank you for your help screening candidates. This document contains instructions for providing your input. Please read this information carefully and fully before you begin.

1 About our Hiring Needs

We are interested in finding candidates that would succeed in a full-stack software engineering position at a mid-sized software start-up company. Later in this document we will tell you more about the company – for now, let's focus on the hiring needs and process.

Qualified candidates should have a working understanding of hardware systems, infrastructure, creating and manipulating databases, writing back-end code in one or more languages (e.g., Ruby, Java, Python, C#), and writing front-end code in one or more languages (e.g., HTML, Javascript). Other responsibilities include project management and technical documentation.

The compensation for the job openings is $103K annually and includes benefits, stock and a performance-based annual bonus. *Compensation is non-negotiable.*

We could potentially hire multiple people from the candidates you evaluate – as many candidates as would be a good fit. We are seeking great candidates who will succeed with the company. We do *not* have a limit for how many can (or must) be hired. You should let us know about any of the candidates worth pursuing.

2 Compensation for you, our recruiter, for this task

For your assistance with our hiring needs, you will be paid the base rate in your contract, plus a bonus. You will receive the base for submitting complete feedback about our sixteen candidates. We ask for answers to only four simple questions for each candidate, all of which can be answered with a yes/no or a number between 0 and 100.

You will also be paid a bonus in two parts. The first bonus depends on whether we hire the candidates you recommend. After we receive your suggestions, **Ms. Katie Schmidt**, an engineering manager at our company, will interview all of the candidates you recommend in person.

Katie will not be available to answer additional questions. **However for your reference, additional information about Katie Schmidt's background is below**.

> **Katie Schmidt** *is a Senior Engineering Manager in our company's San Jose office. In her current role, she focuses on the mission-critical jobs of finding, building and scaling large-scale systems, operations and shared services. She has expertise in deploying scalable consumer technology products that combine cloud-based computing resources with user-interface design. Prior to her current position, Katie was a senior member of infrastructure engineering at* **eBay**, *where she led the development of* **eBay***'s operational strategy and information architecture. Katie graduated from* **University of Oklahoma** *in* **Norman, OK** *with both a BS in Computer Science (2000) and an MBA (2007).*

Although we could hire several people from these candidates, **we do not want to waste Katie's time on candidates who might not be a good match**. In order to encourage you to exercise judge-

ment, your bonus will also include a light penalty for recommending candidates who don't pass the interview or don't take our offer. The lowest your bonus can be is zero – you cannot earn a "negative bonus," even if there are many "light penalties" for unproductive interviews.

We would like you to be candid in your responses. Your assessments are confidential, and we will never send your assessments to others involved in our process so that everyone can develop independent views. This includes Katie, who will not know the details of your feedback about the candidates (so that she can assess them independently through the interview).

For the second part of your bonus, we ask you to predict whether each candidate will i) accept an interview invite if one is offered, ii) pass the interview if an interview takes place, and iii) accept an offer if one has been extended.

We recognize there is some randomness and unpredictability in any hiring process, so we ask that you report a *probability* of these outcomes. If you are having trouble estimating a probability, think about it like this: If the outcome is equally likely to happen as not happen, then you would report 50% (like an even coin toss).

You will be paid more for greater accuracy in these probabilities. For example: Suppose a candidate were only 50% likely to pass the interview, but you claim 75% on your spreadsheet. If the candidate passes, you'd be paid more. However you'd also be penalized more if the candidate fails.

Our penalty for overstating (or understating) is enough to outweigh the potential gains. The best strategy is to simply report your honest estimate of the chances of each outcome.

In the appendix, you can see the exact formula we use to reward you for accuracy. This is a conventional method for rewarding accurate predictions with money.

Please provide predictions for all candidates, even those you do not suggest interviewing. Katie may interview some candidates you did *not* recommend if someone else suggests interviewing them. If Katie interviews a candidate because of someone else's suggestion, you will still be rewarded for accurately predicting what happens.

Please note: We will *not* use these probabilities to influence our hiring decisions – **we will use your yes/no recommendations only**. We will look at these probabilities *only* after our process is finished. Your predictions help us evaluate how well our group of recruiters understands our company's job market. We value accurate predictions and want you to forecast well.

So that you take the forecasts seriously, we'll pay you extra if your predictions are accurate. There are 16 candidates, and three prediction questions each. Given this, the bonuses for being accurate can add up very quickly.

3 Your Feedback about the Candidates

In your packet, you will see a file called "FeedbackForms-Katie-Schmidt.csv." This is a spreadsheet containing the names of all candidates, plus four columns. Please fill out the columns below for all candidates. Save the spreadsheet and send it back to us.

Required Feedback:

1) **Should we interview this candidate?**

- Write "Yes" or "No" in the column marked "Interview?"

2) **Which candidates would accept an invitation to interview for this position?** Suppose we were to invite each candidate to an in-person interview with Katie. Who would meet Katie for an interview?

- For each candidate, enter a number 0-100 representing the probability the candidate would meet Katie.
- Write these probabilities in the column marked "Probability of Accepting Invitation to Interview."
- Answer this question for all candidates, including candidates you do not recommend interviewing.

3) **Who would perform well in the interview?** Suppose each candidate were invited to an interview with Katie and accepted the interview invitation. Using the interview, Katie assesses how each one would perform on the job and extends an offer to those she thinks would perform well. Who would perform well in these interviews and receive an offer?

- For each candidate, enter a number 0-100 representing the probability that Katie would decide to extend the candidate an offer after the interview.
- Write these probabilities in the column marked "Probability of Interview Success."
- Answer this question for all candidates, including candidates you do not recommend interviewing.

4) **Who would accept our offer (if one were extended)?** Now suppose we interviewed the candidates and each passed and received an offer.

- For each candidate, enter a number 0-100 representing the probability each candidate would accept Katie's offer.
- Write these probabilities in the column marked "Probability of Accepting Offer If Extended."
- Answer this question for all candidates, including candidates you do not recommend interviewing.

Optional Feedback:

4) **Comments**: Please leave any notes, observations or comments in this column, such as why you made the assessments you did.

4 Additional Information

Please do not contact any of these candidates. We are asking you only to evaluate them and send us your candid assessments privately. We will contact the candidates you identify and arrange an interview. In the candidates' applications, we have blacked out their contact information.

Please do not consult any information outside of the packets we send you. For example, do not look up the candidates on Google or LinkedIn. Our philosophy is to make *interview* decisions based only on the information on the job application.

5 The Remainder of this Document

- **Ms. Katie Schmidt's background**. Katie is the engineering manager who will be interviewing candidates and making decisions about which candidates to give offers.
- More details about the performance bonus (including the exact formula and examples).
- Additional information about the company.
- Additional information about the job opening for a full-stack software engineer.

Katie Schmidt (Our Hiring Manager for this Position)

Katie Schmidt is a Senior Engineering Manager in our company's San Jose office. In her current role, she focuses on the mission-critical jobs of finding, building and scaling large-scale systems, operations and shared services. She has expertise in deploying scalable consumer technology products that combine cloud-based computing resources with user-interface design. Prior to her current position, Katie was a senior member of infrastructure engineering at **eBay**, where she led the development of **eBay**'s operational strategy and information architecture. Katie graduated from **University of Oklahoma** in **Norman, OK** with both a BS in Computer Science (2000) and an MBA (2007).

A Exact Formula for Calculating Your Bonus

Your bonus will be the **sum of your decision-to-interview bonus** (for question one) plus your **prediction accuracy bonus** (for questions two, three and four).

1) Decision-to-Interview Bonus. This bonus is equal to:

($10 $\times$ the number of hires) - ($4 $\times$ the number candidates interviewed but not hired)

2) Prediction accuracy bonus. This gives you more money for submitting accurate probabilities. The formula below pays you more for the most honest, accurate predictions. In the bonus structure below, the penalty for overstating (or understating) the chances of an outcome outweigh the potential gains. To make the most possible money, you should report your honest opinion about what will happen.

Note that there are 16 candidates, and three prediction questions each, so the bonuses for being accurate can add up very quickly. Given this, it's possible to be paid more in bonuses for prediction accuracy than your hourly charges.

Exact formulas: For the formulas below, p is the probability you enter between 0 and 100.

- For candidates who are invited to an interview and participate in the interview, you will receive a bonus equal to $1-((p/100)-1)^2$. For candidates who are invited and decline, you will receive a bonus equal to $1-(p/100)^2$. p is the probability you entered for the candidate accepting our invitation if we extended one.
- For candidates who are interviewed and pass, you will receive a bonus equal to $1-((p/100)-1)^2$. For candidates who are interviewed and fail, you will receive a bonus equal to $1-(p/100)^2$. p is the probability you entered for the candidate passing if they were interviewed.
- For each candidate who is extended an offer and accepts, you will receive a bonus equal to $1-((p/100)-1)^2$. For each candidate who is extended an offer and rejects, you will receive a bonus equal to $1-(p/100)^2$. p is the probability you entered for the candidate accepting an offer if extended.

Examples for prediction accuracy bonus:

- Suppose a candidate were 80% likely to accept an interview invitation, and you write down 80%. Using the formula above, you'll be paid $0.96 for this candidate if they accept. If the candidate does not accept, you're paid $0.36. If the candidate is 80% likely to pass and you wrote 80%, then you'll be paid $0.84 on average.
- Note that there are 16 candidates, and three prediction questions each, so the bonuses for being accurate can add up very quickly. Given this, it's possible to be paid more in bonuses for prediction accuracy than your hourly charges.
- Now suppose candidate were 80% likely to accept, but you instead wrote down 50%. Using the formula above, you'll be paid $0.75 for this candidate if they accept. If the candidate does not accept, you're paid $0.75. If the candidate is 80% likely to pass and you wrote 50%, then you'll be paid $0.75 on average. Notice you could have made 12% more by simply reporting the true estimate of 80%.
- Again, suppose the candidate were 80% likely to accept, but you instead wrote down 95%. Using the formula above, you'll be paid $0.99 for this candidate if they accept. If the candidate does not accept, you're paid $0.09. If the candidate is 80% likely to pass and you wrote

95%, then you'll be paid $0.81 on average. Notice you could have made 4% more by simply reporting the true estimate of 80%.

B Additional information about our company

Our company strives to shape the future of technology by seizing the opportunities of tomorrow and creating value for our customers, employees, and investors. Throughout our business, customers come first, and we are fiercely dedicated to creating lifelong customer partnerships and working with them to recognize their needs and provide solutions that will help drive success.

We believe in what individuals make possible. Our mission is to provide the resources and opportunities to enable people and organizations to achieve more by capturing the best of what employees, customers, and partners have to offer. Our core competency is software development. The expertise we provide spans many disciplines, providing clients with software solutions that create unprecedented value and dramatically drive results. We hold ourselves to high standards of accountability for ensuring that our customers have the right technology to help them grow and stay ahead of their competitors. Our products allow customers to become faster, smarter, and more innovative.

We cater to a global community of clients and aspire to be an organization that reflects the diverse audience that our technology and products serve. We believe that in addition to hiring the most qualified talent, a diversity of ideas, perspectives, and cultures lead to the creation of better products and services in addition to a stronger community.

We celebrate, support, and thrive on diversity for the benefit of our employees, our products, and our community. We do not discriminate based upon race, color, religion, gender, gender identity, age, national origin, sexual orientation, status as a protected veteran, status as an individual with a disability, or other applicable legally protected characteristics. We are proud to be an Equal Employment Opportunity and Affirmative Action employer.

C Additional information about our job opening for a full-stack software engineer

Our team's software engineers are at the forefront of altering how users connect, explore, and interact with information and one another by developing the next-generation technologies. Our products require handling information at massive scale. We seek engineers who are innovative and bring fresh ideas from all disciplines, including (but not limited to) natural language processing, artificial intelligence, information retrieval, large-scale system design, networking and data storage, distributed computing, UI design, and mobile.

The role as a software engineer will be to work on specific projects critical to our users' needs, with the opportunities to change projects and teams as you and our rapidly-evolving businesses grow. We require our engineers to be multifaceted, display successful leadership abilities, and be enthusiastic to tackle new and challenging problems as we continue to advance our technologies.

We hire talented individuals with an extensive set of technical skills who are eager to take on some of technology's greatest challenges. Throughout our team, engineers consistently work on large-scale applications, massive scalability and storage solutions, and new platforms for developers.

Responsibilities:

- Design, develop, test, deploy, maintain, and improve software.
- Manage individual project priorities, deliverables, and deadlines.
- Collaborate with other specialists in development teams.
- Analyze and improve efficiency, scalability, and stability of various system resources.

Minimum qualifications:

- BS degree in Computer Science or related technical field.
- Software development experience in one or more general purpose programming languages.
- Experience working with two or more from the following: web application development, Unix/Linux environments, mobile application development, distributed and parallel systems, machine learning, information retrieval, natural language processing, networking, developing large software systems, and/or security software development.
- Working proficiency and communication skills in verbal and written English.

Preferred qualifications:

- Master's, PhD degree, further education or relevant experience in engineering, computer science or related technical field.
- Experience with one or more general purpose programming languages including but not limited to: Java, C/C++, C#, Objective C, Python, JavaScript, or Go.
- Interest and ability to learn other coding languages as needed.

A.3.3 Sample Job Application

Each recruiter received a ZIP file containing 16 job applications as PDFs. Below is an example of a job application from Cindy Olson.

Cindy Olson

Details for Software Engineering Job Application

Candidate Information

Candidate Id: 1526328980A7 **Mailing Address**: [redacted] **City/State**: San Francisco, CA

ZIP: [redacted] **Phone**: ([redacted]) [redacted]-[redacted] **Email**: [redacted] **URL**: http://[redacted]

Are you legally authorized to work in the US? ⊠ **Are you over the age of 18?**: ⊠

Are you willing to relocate for this position? ⊠ **Will you now (or in the future) require visa sponsorship?** ☐

Preferred methods of contact: Phone? ⊠ Email? ⊠

Education

College/University: Harvard University **Location**: Cambridge, MA **Dates**: 2011 - 2015 **Graduated?** ⊠

Level: BS (Bachelor of Science) **Subject/Major**: Computer Science

Related Coursework: Coursework Includes: Data Structures and Algorithms, Network-Based Applications, Java, Artificial Intelligence, Cloud Computing.

Employment History

Job Title: Software Engineer **Employer**: Airware **Location**: San Francisco, CA **Dates**: 10/2016 - Present

Duties, Responsibilities and Accomplishments: Implement designs, including experimentation and multiple iterations. Provide user requirements analysis, design and programming support for enhancement of Web application accessed by 5 million users worldwide. Fueled additional revenue stream through responsive customer support, generating $18K in new license sales within first few weeks of new product release.

Job Title: Developer **Employer**: MedicAnimal **Location**: London, UK **Dates**: 05/2015 - 10/2016

Duties, Responsibilities and Accomplishments: Participated in application modification and development of new applications to meet business needs. Provided full life-cycle project expertise. Project work focused on business applications and e-business solutions. Responsibilities included application integration and development using .NET including C#, ASP.Net, WinForms, MS Exchange, and Microsoft Sharepoint Portal Server.

Job Title: Programming Intern **Employer**: Datascape **Location**: Atlanta, GA **Dates**: Summer 2014

Duties, Responsibilities and Accomplishments: Provided Customer and System support for Transaction Service Provider using Terminal (dial-up), Web Browser-based applications, and Microsoft SQL Server Databases. Monitored software systems and customer support provided to AT&T, Verizon, Bluegrass, SouthernLinc, and US Cellular agents.

Additional Information, Experiences, Skills and/or Capabilities

Skills, Accomplishments or Experience: Java, JavaScript, .Net, XML, J2EE, HTML, TCP/IP, REST, SOAP, SOA, Visual Studio .Net, Eclipse, SQL

1

A.3.4 Sample Feedback Form

The below tables were given to the recruiters in Microsoft Excel format. All columns were empty except for the first three (which identified each candidate). The candidates were placed in random order in the feedback form.

Row	Candidate Unique Identifier	Candidate Name	Interview? (Y/N)	Probability Would Accept Interview Invite (0-100)	Probability Would Pass Interview (0-100)	Probability Would Accept Offer If Extended (0-100)	Comments/ Notes
1	1526328980A7	Cindy Olson					
2	1526328980A4	DeShawn Washington					
3	1526328980A8	Billy Snyder					
...	...	...	...	...	...	...	...

A.4 Recruiter Bonus Details

As discussed in the main text, we gave recruiters incentives for accurate probabilities through a bonus payment. In the main text of our instructions, we described our bonus philosophy non-technically, stating that the bonus rewarded truthful, non-exaggerated evaluations. Our communication then provided subjects with an FAQ entry containing a bonus formula based on these reports and some examples.

The formula in the FAQ implemented the well-known quadratic scoring rule (QSR) originally studied in (Brier, 1950) (the "Brier Scoring Rule") and subsequently used in many experimental economics papers (Murphy and Winkler, 1970; Camerer, 1995). Huck and Weizsäcker (2002) compare beliefs elicited via a QSR with beliefs elicited via a Becker-DeGroot-Marshak pricing rule, and conclude that the QSR procedure yields more accurate beliefs.

The Brier Scoring Rule is a proper (i.e. incentive-compatible) scoring rule. In the case of accepting the interview invitation, screeners received a bonus of $1 - ((p/100) - 1)^2$ for applicants who participated in the interview, and a bonus of $1 - (p/100)^2$ for applicants who were invited to interview but did not. Candidates received similar bonuses for the other two probabilities.

A.4.1 Payment of Bonuses

Recruiters were paid bonuses. However, since there were no actual candidates nor firms, the bonus amounts were based on simulated outcomes using data from similar real-world firms. We used data from other, similar firms to simulate likely outcomes to use as the basis for the bonuses. For example, if our simulation suggested that a candidate's $p_1 = 0.3$, $p_2 = 0.5$ and $p_3 = 0.4$, we would draw three random numbers (all zero or one) with $P(1)$ equal to the probabilities above. Based on these binary variables and the recruiter's submissions, we would apply the bonus formulas described above.

Using this approach, each screener gets 4 bonuses per candidate (three predictions and one decision) $\times$ 16 candidates, or 64 total bonuses. Note that some of these 64 could in theory be negative. We added all 64 up to arrive at a single number and sent the total amount to the recruiter. More granular breakdowns of this bonus (for example, by candidate or by the reason for the bonus) were not provided nor requested. The average bonus was approximately \$15, or around 20% of the total wage bill. No recruiter's total bonus was negative or zero. Bonuses were paid after waiting at least 30 to 45 days after

we received feedback from a recruiter, which we believed was a reasonable time period for all candidates to make it through a hiring process.

A.5 Candidates' Names, Universities and Employers

For our candidates' names, we compiled a list of first and last names associated with gender and race. To compile this list, we utilized data from a variety of sources to identify distinctively white, black, male and female names. Our data sources included the US Census Bureau, the Social Security Baby Names data (for the years of our fictional subjects' births), as well as lists compiled by other academics using administrative data. We sought to identify names that were both distinctive but also common so that our recruiter subjects would recognize them. We defined this as occurring in more than 100K individuals with the associated characteristics in the Census for the approximate ages of our subjects. We also manually edited our list in order to avoid names of celebrities arising in recent years as well as popular fictional characters that we felt might distract from the work. Since our project began, Gaddis (2017) developed new methods to help researchers choose names.

First names for white male candidates were Billy, Jimmy, Joey, and Jonny, while white female candidates included Ashly, Brittany, Cindy, and Tabatha. Last names for white candidates were Hansen, Hoffman, O'Brien, Olson, Schmidt, Schneider, Schultz, and Snyder. First names for black male candidates were Darnell, DeAndre, DeShawn, and Marquis, while black female applicants included Aaliyah, Ebony, Imani, and Precious. Last names for black applicants were Alston, Banks, Battle, Booker, Jackson, Jefferson, Mosley, and Washington. Hiring manager names included Tyler Schmidt (white male), Katie Schmidt (white female), Tyrone Jackson (black male), and Shanice Jackson (black female).

Elite universities included Stanford, Harvard, and Yale, while the non-elite group consisted of public universities like University of Tennessee, Alabama, and Florida State. Large employers comprised mostly of top tech firms such as Google, Microsoft, and Facebook, while the group of small employers contains start-ups like LendUp, Gusto, Greenhouse, and MedicAnimal.

A.6 Balance

Because we randomize both sides of the market, we check for balance in both candidate and manager characteristics. The supply-side test examines whether (for example) female recruiters are disproportionately assigned to screen job applications of white candidates. In our design, each recruiter was sent a packet containing all 16 candidate types.[7] As a result, balance was automatically achieved on these dimensions.

On the employer side, balance tests would examine whether (for example) male recruiters are disproportionately assigned to work for particular manager types. Avoiding this kind of imbalance may be particularly useful in two-sided designs. Because of the limits on sample size (constrained by the price of recruiting time), small imbalances could severely reduce statistical efficiency. To avoid demand-side imbalance, we implemented a stratified randomization procedure to guarantee covariate similarity on key recruiter characteristics.[8] We detail this procedure in Appendix A.7.

Table A.2 shows the corresponding tables for both candidate and hiring manager characteristics. The table notes, as we broadly find balance on both sides of the experiment. The tables notes of Table A.2 contain additional discussion.

A.7 Randomization Procedure

Our randomization procedure was designed to address covariate balance across screeners of different types. Randomization was sequential and proceeded in batches. Recruiters were randomly assigned to packets. Before sending out the experimental materials for recruiters' feedback, we performed a covariate balance check (described below). If our covariate balance checked passed, we would send the experimental materials to the recruiters. If the balance checks failed, we re-randomized the current batch (previous batches had already been sent to recruiters who'd already begun work and could not be re-randomized).

Our balance test checked for equality of the average of eight covariates across treatment arms. The covariates were: 1) Race (dummy variables for white and black), 2) gender (dummy variable for male), 3) the recruiters' advertised hourly rate, 4) total hours billed, 5) average feedback score, 6) number of simultaneous (concurrent) assignments, 7) the

[7] Black, white $\times$ male, female $\times$ elite university, not $\times$ large company, not = 2^4 =16.

[8] The recruiter characteristics balanced were the recruiter's hourly rate, gender, race, prior and current outcomes on the platform, location and the keywords used to find the subject.

Table A.2: **Balance Tests: Candidate and Manager Characteristics**

Panel A: Candidate Characteristics Balance

	Female screener	White screener	Black screener	Hourly rate
Male candidate	5.1e-19	-1.5e-18	1.8e-18	-3.3e-17
	(.03)	(.03)	(.03)	(1.40)
White candidate	5.1e-19	-1.3e-18	1.5e-18	-9.9e-17
	(.03)	(.03)	(.03)	(1.40)
Elite university candidate	-5.1e-19	5.1e-19	2.6e-19	3.3e-17
	(.03)	(.03)	(.03)	(1.40)
Large company candidate	-5.1e-19	1.5e-18	-7.7e-19	2.1e-16
	(.03)	(.03)	(.03)	(1.40)
R^2	0	0	5.7e-15	1.0e-15
Observations	864	864	864	864
F-test p-value	1	1	1	1

Panel B: Hiring Manager Characteristics Balance

	Female screener	White screener	Black screener	Hourly rate
Male hiring manager	-1.7e-17	-.042	.042	-7.6
	(.11)	(.15)	(.11)	(6.0)
White hiring manager	-.17	-.042	-.21*	-2.7
	(.11)	(.15)	(.11)	(6.0)
Elite university hiring manager	2.5e-17	.12	-.12	5
	(.11)	(.15)	(.11)	(6.0)
Blinded hiring manager	-.08	-.25	.17	-22***
	(.18)	(.25)	(.25)	(6.8)
R^2	.044	.047	.13	.13
Observations	54	54	54	54
F-test p-value	.58	.61	.15	.012

Notes: This table displays balance tests on both candidate and hiring manager characteristics. The table displays the results of a regression of screener attributes on either candidate attributes (Panel A) or hiring manager attributes (Panel B), with robust standard errors clustered at the screener level. The table also displays the p-value from an F-test of all candidate or hiring manager characteristics jointly predicting screener characteristics.
If randomization was done successfully, we would expect to see that candidate and hiring manager treatment assignments are uncorrelated with screener characteristics. This is indeed what we see. As discussed in Section 3.3, our design ensures complete covariate balance on the candidate side, and this is what the table reflects.
On the demand side, one of the conditions exhibits some imbalance: screeners assigned to the blind hiring manager condition are likely to report smaller hourly rates. However, given we are conducting a total of 20 tests on the hiring manager side, we would expect one test to return a p-value of $p < 0.05$ by random chance. We control for screener attributes in all the regressions displayed in the paper.

search keywords that lead us to the recruiter's profile, and 8) location (a dummy variable for Californians, the most populous group in the data). Race and gender were manually coded, and all other variables appeared in the recruiter's profile.

We tested for equality of these means across all treatment groups through a separate per-variable test of equality across all eight treatment arms described in Section 3. For assignments where these tests' p-values were less than 0.2, we re-randomized. The sequential balance checks were "cumulative." The tests above included observations for all prior assignments including the current batch. However, the current batch was the only batch that could be potentially adjusted if re-randomization were necessary.

A.8 Attrition

In Table A.3 below, we conduct an attrition analysis to see whether attrition was correlated with any of the treatment arms. We regress a binary indicator for attrition (that equals one if the recruiter left the study, and zero otherwise) on binary indicators for the hiring manager treatment assignments. We do not need to test for the candidate treatment assignments given that these were conducted within each recruiter. The results indicate that we do not observe differential attrition by treatment arm. None of the manipulations are significant predictors of whether a recruiter exited the study or not. Meanwhile, Table A.4 conducts the same analysis for recruiter characteristics, and shows that they do not predict attrition.

Table A.3: **Attrition analysis with treatment assignment**

	Dropped Out	Dropped Out
Male Hiring Manager	.03	.06
	(.11)	(.12)
White Hiring Manager	.03	.04
	(.11)	(.13)
Elite University Hiring Manager	.02	.01
	(.11)	(.12)
Hiring Manager Blinded	-.00	.03
	(.19)	(.20)
Balance Vars	N	Y
R^2	.01	.04
Observations	83	83

Notes: This table conducts an attrition analysis by regressing a binary indicator for attrition on the hiring manager treatment assignments. The second column includes controls for recruiter characteristics.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table A.4: **Attrition analysis with recruiter characteristics**

	Dropped Out	Dropped Out
Is white	-.02	-.02
	(.11)	(.12)
Is male	-.02	-.04
	(.15)	(.16)
Number of current assignments	-.02	-.02
	(.03)	(.04)
Average feedback score	-.03	-.03
	(.03)	(.03)
Total actual hours billed	.00	.00
	(.00)	(.00)
Hourly rate amount	-.00	-.00
	(.00)	(.00)
HM assignment	N	Y
R^2	.03	.035
Observations	83	83

Notes: This table conducts an attrition analysis by regressing a binary indicator for attrition on recruiter characteristics. The second column includes controls for hiring manager assignment.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B Additional Empirical Analysis

B.1 Potential Subject Misunderstanding

We also test the robustness of our findings to potential misunderstandings of our instructions or subject mistakes.

First, we examine robustness with respect to three ways that screeners may have misunderstood the instructions. First, the screeners may have misunderstood the bonus formula. Screeners may have thought the bonus does not penalize screeners for forwarding candidates who do not accept the interview, causing them to only consider $P(PassInterview|AcceptInterview)$ and $P(AcceptJobOffer|PassInterview)$ when allocating interviews. Column 2 of Table 8 in the main-text examines how these two probabilities only, and not $P(AcceptInterview)$, influence which candidates receive interviews. The results confirm our main results that screeners place much greater weight on hiring managers' expected behavior.

A second way is through the reporting of probabilities. Although our instructions clearly stressed that screeners should report conditional probabilities, they may have mistakenly reported non-conditional ones. If screeners reported non-conditional probabilities, we would expect that $p_1 \geq p_2 \geq p_3$ for all candidates: the chance of reaching each state should lower (or stay the same) as the stages of hiring progress. Nonetheless, we see the opposite: 100 percent of recruiters report probabilities at later stages that are larger than earlier stages for at least one candidate.

In other words, they report that (for example), the probability of getting an offer is higher than the probability of getting an interview. We interpret this as evidence all subjects know they are being asked about conditional probabilities.

Our third test examines whether our results are robust to using the rank order of predictions about subjects as dependent variables (rather than the levels of predictions). We estimate Equations 2 and 3 but use the ranking of each probability by screener as the outcome measure. For each probability measure and screener, we sort candidates by the probability and assign a rank to each one from 1 (highest probability within this pool of candidates) to 16 (lowest probability). We then examine how candidate and hiring manager attributes impact these rankings.

Table B.1 displays the results for candidate characteristics while Table B.2 displays them for hiring manager characteristics. The results line up with Tables 5 and 6 in the main text.

Screeners believe female candidates are more likely to accept interviews and offers (the negative coefficient corresponds to a lower rank (closer to one) so these candidates have a higher probability) candidates from elite universities and large companies are more likely to pass interviews. As in the main text, none of the hiring manager manipulations survive multiple comparisons adjustments.

Table B.1: **Impact of candidate characteristics on probability rank**

	Supply side		Demand side	Overall	
	Rank of P(Accept Interview)	Rank of P(Accept Offer)	Rank of P(Pass Interview)	Rank of P(Hire)	Interviewed
Female Job Applicant	-1.08***	-1.31***	-0.22	-1.28***	0.04
	(0.37)	(0.35)	(0.31)	(0.37)	(0.03)
Black Job Applicant	-0.27	-0.18	-0.94**	-0.68*	0.04
	(0.40)	(0.32)	(0.40)	(0.35)	(0.04)
Elite University Job Applicant	0.78*	0.37	-1.20***	-0.13	0.14***
	(0.40)	(0.33)	(0.39)	(0.38)	(0.04)
Large Company Job Applicant	-0.33	0.32	-1.45***	-0.63*	0.09***
	(0.39)	(0.42)	(0.30)	(0.37)	(0.03)
R^2	0.06	0.05	0.07	0.04	0.08
Observations	864.00	864.00	864.00	864.00	864.00
Fixed effects	Manager	Manager	Manager	Manager	Manager
Controls	Screener	Screener	Screener	Screener	Screener
Control mean	6.70	6.22	8.69	8.89	0.50
F-test	0.06	0.01	0.00	0.00	0.00
WY P-value: Female	0.05	0.01	0.93	0.02	0.70
WY P-value: Black	0.93	0.93	0.19	0.39	0.87
WY P-value: Elite university	0.39	0.87	0.04	0.93	0.02
WY P-value: Large company	0.93	0.93	0.00	0.53	0.06

Notes: This table displays the results of Equation 2 on supply-side behavior (columns 1 and 2), demand-side behavior (column 3), and overall hiring beliefs and behavior (columns 4 and 5). Columns 1–4 use the rank of the probability as the outcome. The regression controls for screener characteristics and includes robust standard errors clustered at the screener level. The table also displays the p-value from an F-test of all candidate characteristics jointly predicting each dependent variable, plus the mean for the control group (where all binary indicators are equal to zero).
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.2 Cross-validation of Recruiters' Beliefs

Given that each set of materials was evaluated by two recruiters, we can test for cross validation in their assessments. We do so in Table B.3 below. We assemble all subjects into pairs of recruiters who evaluate the same packets, and then predict the first evaluation as a function of the second. For ease of interpretation, we standardize both evaluations. We cluster standard errors by the pair.

Table B.3 shows that two out of three correlations are statistically significant and positive, although the magnitudes of the coefficients are all small. Overall, the levels of cross-validation are positive and statistically significant, but relatively low. We can reject the null

Table B.2: **Impact of hiring manager characteristics on probability rank**

	Supply side		Demand side	Overall	
	Rank of P(Accept Interview)	Rank of P(Accept Offer)	Rank of P(Pass Interview)	Rank of P(Hire)	Interviewed
Female Hiring Manager	0.82*	0.24	-0.26	-0.08	-0.04
	(0.41)	(0.21)	(0.20)	(0.14)	(0.06)
Black Hiring Manager	-0.68	-0.09	-0.48**	-0.11	0.01
	(0.46)	(0.23)	(0.19)	(0.13)	(0.07)
Elite University Hiring Manager	-0.17	0.05	0.03	-0.16	0.00
	(0.41)	(0.20)	(0.17)	(0.13)	(0.06)
Blinded Hiring Manager	-0.18	0.08	-0.56	-0.20	-0.10
	(0.61)	(0.48)	(0.35)	(0.23)	(0.08)
R^2	0.15	0.12	0.18	0.13	0.18
Observations	864.00	864.00	864.00	864.00	864.00
Fixed effects	Candidate	Candidate	Candidate	Candidate	Candidate
Controls	Screener	Screener	Screener	Screener	Screener
Control mean	7.12	7.43	7.55	8.45	0.68
F-test	0.16	0.77	0.07	0.33	0.78
WY P-value: Female	0.99	1.00	1.00	1.00	1.00
WY P-value: Black	1.00	0.96	1.00	1.00	1.00
WY P-value: Elite University	1.00	1.00	1.00	1.00	1.00
WY P-value: Blinded	1.00	1.00	1.00	1.00	1.00

Notes: This table displays the results of Equation 3 on supply-side behavior (columns 1 and 2), demand-side behavior (column 3), and overall hiring beliefs and behavior (columns 4 and 5). Columns 1–4 use the rank of the probability as the outcome. The regression controls for screener characteristics includes robust standard errors clustered at the screener level. The table also displays the p-value from an F-test of all hiring manager characteristics jointly predicting each dependent variable, plus the mean for the control group (where all binary indicators are equal to zero).
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

hypothesis that they are completely uncorrelated, however the correlations are relatively weak.

Table B.3: **Cross Validation of Recruiter Evaluations**

	(1) P(Accepts Interview) Eval #1	(2) P(Accepts Offer) Eval #1	(3) P(Passes Interview) Eval #1	(4) Interview Eval #1
Second Evaluator	.52	.17***	.11*	.082*
	(.39)	(.05)	(.06)	(.05)
Observations	864	864	864	864
R^2	.01	.03	.01	.01

Notes: In this table, we assemble all subjects into pairs of recruiters who evaluate the same packets. We then predict the first evaluation as a function of the second. For ease of interpretation, we standardize both evaluations. We cluster standard errors by the pair.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

As a basis of comparison, we examine levels of cross validation in candidates' own stated

preferences in our accuracy survey (Section 5.2, page 29 and Appendix C). As with the table above, we assemble all survey respondents (hiring managers) subjects into pairs who evaluate the same candidates. We then remove all pairs except for pairs that are observably identical on the three variables we collect (gender, race and education). We then predict the first evaluation as a function of the second. Table B.4 shows our results. We find that observably similar hiring managers also have low levels of correlation when evaluating the same candidates.

Table B.4: **Cross Validation of Evaluations of Candidates (Accuracy Survey)**

	(1) P(Accepts Interview) Eval #1	(2) P(Accepts Offer) Eval #1	(3) P(Passes Interview) Eval #1
Second Evaluator	.009	.029**	.019
	(.014)	(.014)	(.012)
Observations	35,392	35,392	35,392
R^2	.00009	.00082	.00036

Notes: In this table, we used our data about candidates' own stated preferences in our accuracy survey (Section 5.2, page 29). We assemble all survey respondents (hiring managers) subjects into pairs who evaluate the same candidates. We then remove all pairs except for pairs that are observably identical on the three variables we collect (gender, race and education). We then predict the first evaluation as a function of the second. We find that observably similar hiring managers also have low levels of correlation when evaluating the same candidates.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.3 Match-specific analysis

B.3.1 Match analysis

In this subsection, we conduct a match analysis for all key outcomes. Table B.5 displays the results of the joint test plus a count of the individual fixed effects that return $p < 0.05$ obtained from Equation 4. The results in Table B.5 provide strong support for the existence of match-specific preferences in our setting. Across all four dependent variables, 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 worker and manager.

Table B.5: **Match-specific quality**

Outcome	Interview	P(Accept Interview)	P(Pass Interview)	P(Accept Offer)
R^2	0.77	0.97	0.96	0.96
Joint test, F-stat	11.57	12.13	15.72	11.16
Joint test, p-value	< 0.01	< 0.01	< 0.01	< 0.01
Fixed effects with $p < 0.05$	44	9	12	37

Notes: This table displays the results of a regression of our key outcomes on match-type fixed effects, screener controls, and candidate and hiring manager fixed effects. The table displays the R^2 from the model, the p-value and F-statistic from a joint test that all match type fixed effects are equal, as well as the number of specific pair fixed effects with $p < 0.05$.

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.

B.3.2 Homophily

Prior to running the experiment, we hypothesized that homophily (McPherson et al., 2001) would explain why certain worker-manager matches were more likely to receive interviews. Homophily could drive our results if, for example, screeners believed that male candidates preferred male hiring managers, and male hiring managers preferred male candidates. We test for homophily in worker-manager pairs in three ways. However, all show limited support for homophily driving the candidate-hiring manager matches we observe.

First, we test whether screeners believe there is homophily across broad demographic characteristics. To do so, we run the following equation:

$$\begin{aligned} Y_{c,h,s} =& \beta_0 + \beta_1 * SameGender_{c,h} + \beta_2 * SameRace_{c,h} + \beta_3 * SameEducationalStatus_{c,h} + \\ & \alpha * S_s + \gamma * HM_h + \delta * C_c + \epsilon_{c,h,s} \end{aligned} \tag{6}$$

This equation includes three binary variables that measure whether both candidate and manager are matched on gender, race, and educational background, plus both candidate and hiring manager fixed effects (C_c and HM_h, respectively), and screener controls. The comparison group for these indicators is the mixed-attribute pair. For example, β_1 returns the effect of a worker being matched to a same-gender manager versus a different gender one. As in Equations 2 and 3, standard errors are clustered at the screener level. We display these results in Table B.6, with a joint p-value testing whether the three homophily

measures are jointly equal to zero, and whether the sum of all three equals zero. The results, however, show limited support for homophily across broad demographic characteristics– same gender/race/education status pairs are just as likely to be interviewed as different gender/race/pairs.

Table B.6: **Homophily across broad demographic characteristics**

	Interviewed	P(Accepts Interview)	P(Passes Interview)	P(Accepts Offer)
Same gender	0.01	0.00	0.00	0.01
	(0.03)	(0.01)	(0.01)	(0.01)
Same race	0.03	0.01	0.00	-0.00
	(0.03)	(0.01)	(0.01)	(0.01)
Same education status	0.00	0.01	-0.01	0.01
	(0.04)	(0.01)	(0.01)	(0.01)
R^2	0.19	0.40	0.26	0.40
Observations	864.00	864.00	864.00	864.00
Sum p-value	0.53	0.25	0.60	0.22
Joint p-value	0.87	0.70	0.74	0.44

Notes: This table displays the results of Equation 6 on supply-side behavior (columns 1 and 2), demand-side behavior (column 3), and overall hiring beliefs and behavior (columns 4 and 5). The regression includes candidate and hiring manager fixed effects, screener controls, and robust standard errors clustered at the screener level. The table also displays the p-value from an F-test of all same-characteristic pairs jointly predicting each dependent variable, as well as from a test of whether the sum of all three coefficients equals zero.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

The results in Table B.6 pool across the demographic categories and return the average effect across the gender/race/education categories in our experiment. However, it is possible that homophily exists only within certain groups within these broad categories (for example, there is only homophily for men, but not women). To test for this, we run the following equation:

$$\begin{aligned} Y_{c,h,s} =& \beta_0 + \beta_1 * BothMale_{c,h} + \beta_2 * BothFemale_{c,h} + \beta_3 * BothWhite_{c,h} + \beta_4 * BothBlack_{c,h} + \\ & \beta_5 * BothElite_{c,h} + \beta_6 * BothNonElite_{c,h} + \alpha * S_s + \gamma * HM_h + \delta * C_c + \epsilon_{c,h,s} \end{aligned} \tag{7}$$

This equation mimics Equation 6 but replaces each demographic categorical variable with its underlying attribute so that there are now six binary variables instead of three. We display these results in Table B.7 below. Table B.7 also does not provide support for homophily. At first glance, it may seem that there are is homophily by race– randomly assigning a white candidate with a white hiring manager increases the probability of a hire by 22 percentage points while assigning a black candidate with a black hiring

manager depresses this probability by 20 percentage points. However, decomposing this into demand- and supply-side behavior reveals that this result is not due to homophily. Although screeners believe that white and black candidates exhibit different behavior with same-race hiring managers, screeners do not believe that hiring managers exhibit such behavior since they are not more likely to make more or fewer offers to same-race candidates. Thus, what seems like homophily from the aggregate probability of hire measure is not once we decompose it into the screener's beliefs about the behavior of each side[9]

Table B.7: **Homophily across specific demographic characteristics**

	Supply side		Demand side	Overall	
	P(Accepts Interview)	P(Accepts Offer)	P(Passes Interview)	P(Hired)	Interviewed
Both male	-0.02	-0.04	-0.02	0.01	0.05
	(0.05)	(0.04)	(0.07)	(0.06)	(0.09)
Both female	0.02	0.06	0.03	0.02	-0.03
	(0.04)	(0.04)	(0.07)	(0.06)	(0.09)
Both elite	-0.03	-0.06	0.01	-0.03	0.04
	(0.05)	(0.05)	(0.04)	(0.06)	(0.08)
Both non-elite	0.04	0.08	-0.04	0.02	-0.03
	(0.05)	(0.05)	(0.04)	(0.07)	(0.09)
Both white	0.19***	0.24***	0.03	0.22***	0.05
	(0.06)	(0.06)	(0.06)	(0.07)	(0.12)
Both black	-0.17***	-0.24***	-0.03	-0.20***	0.00
	(0.06)	(0.06)	(0.06)	(0.07)	(0.11)
R^2	0.40	0.40	0.26	0.43	0.19
Observations	864	864	864	864	864
Fixed effects	C+HM	C+HM	C+HM	C+HM	C+HM
Controls	Screener	Screener	Screener	Screener	Screener
F-test	0.06	0.00	0.79	0.01	0.89
P-values:					
Both male	1.00	1.00	1.00	1.00	1.00
Both female	1.00	1.00	1.00	1.00	1.00
Both elite	1.00	1.00	1.00	1.00	1.00
Both non-elite	1.00	1.00	1.00	1.00	1.00
Both white	0.95	0.85	1.00	0.95	1.00
Both black	0.96	0.87	1.00	0.96	1.00

Notes: This table displays the results of Equation 7 on supply-side behavior (columns 1 and 2), demand-side behavior (column 3), and overall hiring beliefs and behavior (columns 4 and 5). The regression includes candidate and hiring manager fixed effects, screener controls, and robust standard errors clustered at the screener level. The table also displays the p-value from an F-test of all same-characteristic pairs jointly predicting each dependent variable. The bottom of the table displays p-values adjusted for multiple comparisons (6 treatments $\times$ 5 outcomes) using the free step-down procedure of Westfall and Young (1993). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

The results in Tables B.6 and B.7 examine homophily within a single demographic characteristic. An alternative possibility is that homophily exists on the extensive margin, in that matching on more characteristics increases the likelihood of an interview. To test for this possibility, we create a 0–3 homophily index by summing up $SameGender_{c,h}$,

[9]This reinforces the benefits of a two-sided audit– the design allows us to collect data on screener beliefs about the behavior of both sides to understand if the matching we observe is driven by homophily.

$SameRace_{c,h}$, and $SameEducationalStatus_{c,h}$. A candidate-hiring manager pair who match on gender, race, and educational status receive a score of three, while those match on none receive a score of zero. We then estimate the following equation:

$$Y_{c,h,s} = \beta_0 + \beta_1 * HomophilyIndex_{c,h} + \alpha * S_s + \gamma * HM_h + \delta * C_c + \lambda * M_{c,h} + \epsilon_{c,h,s} \quad (8)$$

This equation includes candidate and hiring manager fixed effects (C_c and HM_h, respectively), screener controls, and match-type fixed effects ($M_{c,h}$). β_1 captures the impact of being matched on one additional attribute, and is displayed in Table B.8. The results again indicate no support for homophily driving the matches we observe.

Table B.8: **Homophily index results**

	Interviewed	P(Accepts Interview)	P(Passes Interview)	P(Accepts Offer)
Homophily index, 0–4	0.17	0.04	0.02	-0.00
	(0.22)	(0.04)	(0.05)	(0.02)
R^2	0.32	0.48	0.35	0.47
Observations	864.00	864.00	864.00	864.00
p-value	0.45	0.32	0.67	0.94

Notes: This table displays the results of Equation 8 on supply-side behavior (columns 1 and 2), demand-side behavior (column 3), and overall hiring beliefs and behavior (columns 4 and 5). The regression includes candidate and hiring manager fixed effects, match-type fixed effects, screener controls, and robust standard errors clustered at the screener level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.3.3 Match analysis by type

In Table B.9, we re-estimate Equation 4 for all key outcomes on subsets of the data covering all candidate (or manager) characteristics. For each subset, we present the p-value of joint test of no match-specific effects.

B.3.4 Distribution of fixed effects by race and gender (pooling across the education or experience manipulations)

In Figure B1, we display the distribution of match-type fixed effects from equation 4 by race and gender. We collapse the education and prior experience manipulations for ease of interpretability and display the average fixed effect by candidate's and hiring manager's race and gender across the education and prior experience manipulations.

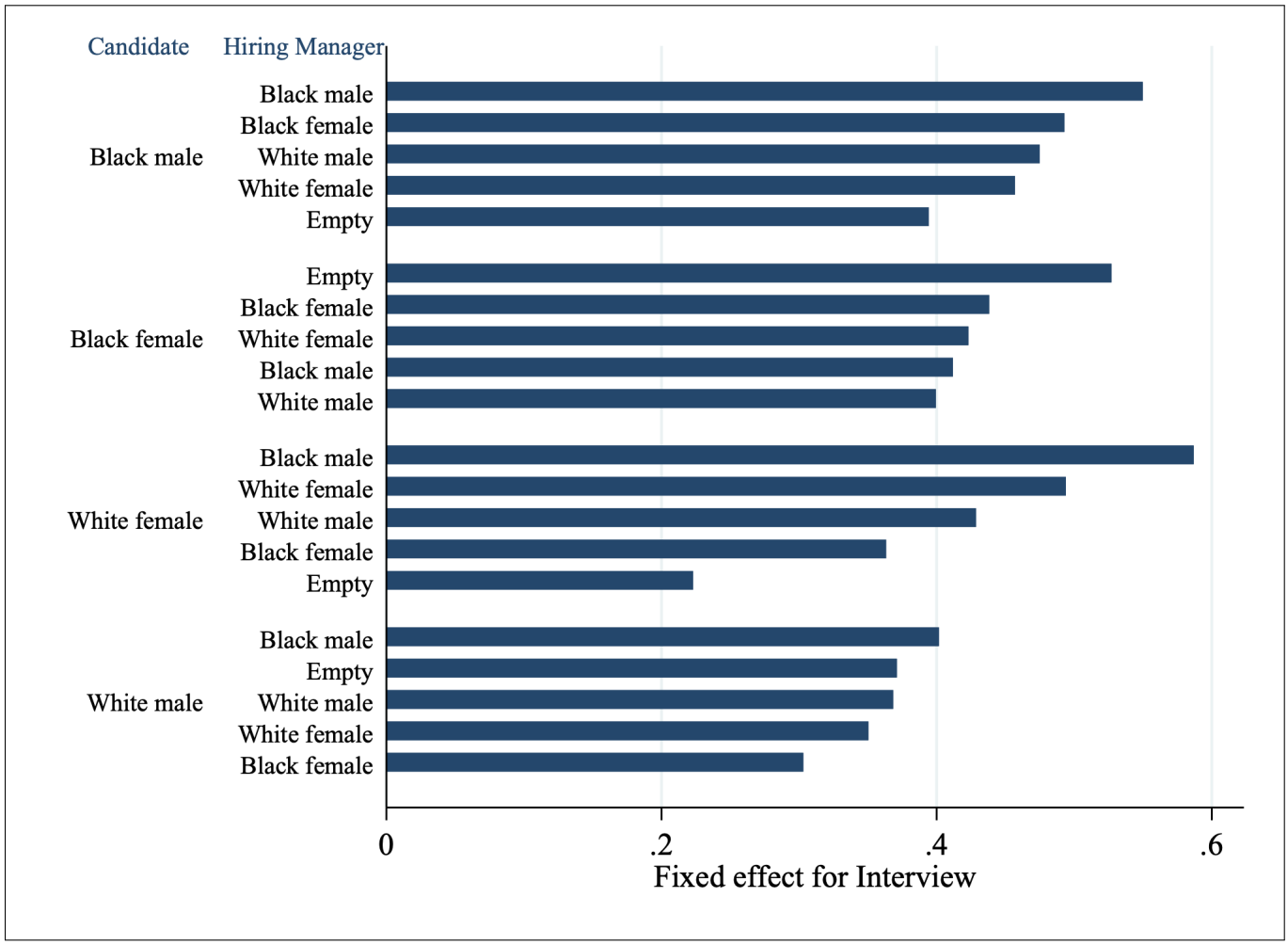
Table B.9: **Match-specific analysis, by type**

Candidate level								
	Interview		P(Accept Interview)		P(Pass Interview)		P(Accept Job Offer)	
Type	F-stat	p-value	F-stat	p-value	F-stat	p-value	F-stat	p-value
Men	44.3	< 0.01	12.3	< 0.01	68.3	< 0.01	10.5	< 0.01
Women	79.3	< 0.01	25.6	< 0.01	17.5	< 0.01	96.1	< 0.01
White	21.1	< 0.01	197	< 0.01	27.7	< 0.01	38.9	< 0.01
Black	114.7	< 0.01	55.3	< 0.01	72.5	< 0.01	63.3	< 0.01
Elite	26.8	< 0.01	31.5	< 0.01	7.9	< 0.01	78.1	< 0.01
Non-elite	22	< 0.01	27.7	< 0.01	41.2	< 0.01	27.9	< 0.01
Large company	32.7	< 0.01	23.1	< 0.01	28.7	< 0.01	67.5	< 0.01
Small company	8363.1	< 0.01	23.6	< 0.01	54.5	< 0.01	38.7	< 0.01

Hiring manager level								
	Interview		P(Accept Interview)		P(Pass Interview)		P(Accept Job Offer)	
Type	F-stat	p-value	F-stat	p-value	F-stat	p-value	F-stat	p-value
Men	46.3	< 0.01	75.1	< 0.01	82.5	< 0.01	351.5	< 0.01
Women	11.9	< 0.01	63.4	< 0.01	3.9	< 0.01	13.2	< 0.01
White	18.3	< 0.01	34.9	< 0.01	10.2	< 0.01	15.8	< 0.01
Black	30.3	< 0.01	232.6	< 0.01	5.5	< 0.01	419.8	< 0.01
Elite	14.4	< 0.01	14.9	< 0.01	298.5	< 0.01	28.1	< 0.01
Non-elite	71.7	< 0.01	164.6	< 0.01	52.6	< 0.01	30.5	< 0.01
Blinded	0.39	0.69	< 0.01	0.90	1.1	0.34	0.8	0.42

Notes: This table displays the results of a regression of outcomes 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.

Figure B1: **Distribution of match-type fixed effects by race and gender**



Notes: This figure displays the distribution of match-type fixed effects from equation 4 by race and gender using interview as the dependent variable. It plots the estimated fixed effect and corresponding confidence interval. Some coefficients are missing because they are dropped from the regression.

B.4 Screeners' Reputational Considerations

In the main text of the manuscript, we document that screeners put more weight on their beliefs of manager behavior versus candidate behavior. In this subsection, we present evidence that reputational considerations may explain the strong weighting of manager behavior. Recruiters of unknown quality on the platform (possibly who are early in their careers on the platform) have greater incentives to signal their quality to employers, and reputationally-secure "incumbents" face less. We now operationalize and test these

predictions. We obtained six variables directly from the recruiting platform to help identify reputationally-secure "incumbent" recruiters. *Yes* answers to any of the questions below indicate greater incumbency on the platform.

1. *Has Prior Reviews*: Does the recruiter's profile feature any reviews from prior employers?[10]
2. *Has Other Assignments*. Is the recruiter simultaneously working for another client?
3. *Has Prior Platform Income*. Has the recruiter earned any prior revenue on the platform?
4. *Has Certifications*. Has the recruiter earned one of the platform's skill certifications?
5. *Has Previously Billed Hours*. Has the recruiter logged any billable hour revenue? This form of contract is associated with longer-term, open-ended relationships with clients.
6. *Fully Billed*. Did the recruiter bill our employer account for the full amount? Some recruiters did not. We interpret this as an effort to forgo money for a better review.[11]

All measures are binary.[12] All variables except #6 are publicly displayed on the recruiters' profiles for clients to see and were measured before the subject was contacted.[13] We also measure effects on a composite index variable (the normalized sum of the six variables above).

Candidates with these features enjoy greater reputational security, and thus *lower* incentives to engage in strategic reweighting. In Table B.10, we address this hypothesis by interacting each of the above dummy variables with our three probabilities. The raw probabilities capture how much weight less-established recruiters place on each stage, while the interaction terms measure how much greater (or less) weight reputationally-incumbent recruiters place on each probability compared with less-established ("non-incumbent") recruiters in our sample.

We broadly find that the established/incumbent recruiters indeed place *less* weight on

[10] As in other platforms (Nosko and Tadelis, 2015; Filippas et al., 2018), most feedback is positive if it is left at all.

[11] Technically, the "fully billed" variable is an outcome of our experiment, and thus different from our other measures of reputational sensitivity that were collected at the baseline. We believe this variable measures a latent recruiter characteristic that was present when we hired the recruiter and unaffected by our task. We find no effect of our interventions on the "fully billed" outcomes.

[12] Several of our measures capture an amount which we converted into a binary variable for observations above zero. We considered examining heterogeneity using the amount instead of the binary. However, for many of our variables, the largest plurality of subjects was at zero.

[13] These measures may appear to capture the same underlying dimension (prior experience on the platform), however, the average correlation between any two variables is only 0.42.

managers' behavior, and *greater* weight on beliefs about candidates' behavior (compared to their non-incumbent counterparts). They also place a greater weight on candidate behavior at the end of the hiring process (job acceptances) and less weight at the beginning (interview participation) than their non-incumbent counterparts. These differences *diminish* (but do not entirely eliminate) the overall trends we found in Table 8. This is consistent with the idea that reputational concerns cause screeners to heavily weigh the behavior of employers. More-established recruiters still engage in strategic reweighing (i.e., they still may have reputational concerns), but less than non-incumbents do.

Table B.10: **Recruiter Incumbency Characteristics and Supply/Demand Considerations:**

Panel A: Logistic Regressions

	Interview	Interview	Interview	Interview	Interview	Interview	Interview
P(Accept Interview), Log	1.8***	2***	2.3**	2.3**	1.8	2.3***	1.7
	(.65)	(.73)	(.93)	(.91)	(1.2)	(.83)	(1.1)
P(Pass Interview), Log	6.1***	6.4***	6.8***	6.6***	6.3***	6.5***	8***
	(.57)	(.7)	(.8)	(.8)	(.83)	(.71)	(1.1)
P(Accept Job), Log	-.96**	-1.4***	-1.7***	-1.8***	-1.6**	-1.7***	-1.5*
	(.44)	(.46)	(.63)	(.62)	(.68)	(.57)	(.83)
P(Accept Interview), Log × Characteristic	-.71	-2	-1.5	-1.4	-.81	-1.5	-.35
	(.61)	(1.4)	(1.3)	(1.3)	(1.5)	(1.4)	(1.3)
P(Pass Interview), Log × Characteristic	-1*	-1.3	-2.1*	-1.6	-.92	-1.5	-3**
	(.58)	(1.2)	(1.2)	(1.1)	(1.2)	(1.2)	(1.3)
P(Accept Job), Log × Characteristic	1.4***	3.1***	2.3***	2.6***	2.3**	2.5***	2.1**
	(.36)	(.8)	(.91)	(.9)	(.95)	(.87)	(1.1)
R^2	.33	.33	.32	.32	.32	.32	.33
Observations	864	864	864	864	864	864	864
Recruiter Characteristic	Index(σ)	Has Reviews	Has Other Assign-ments	Has Prior Revenue	Has Certif-ications	Previously Billed Hours	Fully-billed

Panel B: OLS

	Interview	Interview	Interview	Interview	Interview	Interview	Interview
P(Accept Interview)	.36**	.47**	.43*	.49**	.25	.45**	.34
	(.17)	(.19)	(.23)	(.24)	(.28)	(.2)	(.25)
P(Accept Pass Interview)	1.5***	1.6***	1.6***	1.6***	1.7***	1.6***	1.7***
	(.08)	(.11)	(.1)	(.11)	(.14)	(.1)	(.15)
P(Accept Job)	-.10	-.28**	-.23	-.3*	-.19	-.25*	-.32**
	(.14)	(.11)	(.14)	(.16)	(.16)	(.14)	(.15)
P(Accept Interview)× Characteristic	-.10	-.51	-.18	-.28	.09	-.25	.06
	(.19)	(.38)	(.37)	(.36)	(.37)	(.39)	(.35)
P(Accept Pass Interview)× Characteristic	-.19**	-.3*	-.31*	-.32*	-.3	-.29*	-.39**
	(.08)	(.18)	(.17)	(.16)	(.18)	(.17)	(.19)
P(Accept Job)× Characteristic	.28*	.76**	.37	.48	.26	.45	.44
	(.16)	(.33)	(.33)	(.32)	(.29)	(.34)	(.28)
R^2	.36	.36	.36	.36	.35	.36	.37
Observations	864	864	864	864	864	864	864
Recruiter Characteristic	Index(σ)	Has Reviews	Has Other Assign-ments	Has Prior Revenue	Has Certif-ications	Previously Billed Hours	Fully-billed

Notes: This table examines the relationship between call-back decisions and supply- and demand-side behavior. In this table, we examine heterogeneity by the type of recruiter. The types of heterogeneity we examine are in the final rows of the tables above. They were chosen because of their relation to career concerns (we explain these choices in Section 4.5). Each regression contains estimates of Equation 5 with the characteristic (binary in columns 2–7, and a normalized index of all six in column 1) variable for that column added, as well as three interactions with that variable (the probabilities). Panel A displays the results of a logistic regression on the log of various hiring probabilities while Panel B displays the results of an OLS regression on various hiring probabilities. The regression use robust standard errors clustered at the screener level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

A potential explanation for these results is that these recruiters have learned to integrate candidate behavior through job or life experience, and not because their reputations are secure. The difference between age, experience and reputational security is a classic problem that comes up in many papers about reputation; many papers directly model reputation-seeking behavior as a function of age or experience (Holmström, 1999; Tadelis, 2002).

Our setting has several attractive features for addressing this question empirically. To place our data in context, note that all of our measures of reputational security are *specific to the platform we used to engage our subjects.* For example, one measure is how many prior reviews the recruiter received on the platform we used to hire them. As a result, there are recruiters who are relatively experienced (and/or old) who are *not* reputational incumbents on this platform, depending on how and when they engaged with the platform.

Our analysis below uses data from our recruiters' resumes. We collected these in order to verify prior human resources experience when we screened recruiters for participation in our experiment. In particular, we code their age (based on their year of college entry and/or high school graduation) and their years of experience working specifically in hiring. This allows us to separately analyze age, recruiter experience and reputational incumbency.

We code each variable into a binary representing above or below the median. In our sample, age and recruiting experience are weakly correlated ($\rho = 0.14$).[14] We also find that both age and recruiting experience are relatively *uncorrelated* with our measures of reputational incumbency on the platform.[15]

Given the above, it is unsurprising that our results are robust to including controls for age and recruiting experience. In Table B.11, we rerun Table B.10 including controls for age and experience (and interactions with the three probabilities). Although the precision of our results is lower in some cases, we find the same overall pattern and several precisely estimated results. These generally show that more reputationally secure workers weigh their beliefs about employer behavior less (and beliefs about candidate behavior more).

[14]This is for several likely reasons. First, recruiting does not require highly specialized skills. As a result, many recruiters have spent their careers between recruiting and other business generalist roles.

[15]For our six measures of incumbency, prior experience was correlated between -0.05 and +0.06. For age, our measures were correlated between -0.23 and -0.07. Note that age is mostly *negatively* correlated with our measures of reputational incumbency. This is natural because the recruiting platform we used is online only. Younger, digital-native workers may have utilized it first (while older workers continued to find recruiters through offline means).

Table B.11: **Recruiter Incumbency Characteristics and Supply/Demand Considerations, Controlling for Age and Experience:**

	Interview	Interview	Interview	Interview	Interview	Interview	Interview
P(Accept Interview)× Characteristic	-.07	-.46	-.07	-.21	.11	-.14	.10
	(.18)	(.36)	(.35)	(.35)	(.36)	(.37)	(.33)
P(Accept Pass Interview)× Characteristic	-.19**	-.31*	-.28*	-.33**	-.28	-.28*	-.36*
	(.08)	(.17)	(.16)	(.15)	(.18)	(.16)	(.18)
P(Accept Job)× Characteristic	.27*	.74**	.37	.44	.22	.4	.44
	(.15)	(.32)	(.32)	(.31)	(.29)	(.33)	(.27)
Observations	864	864	864	864	864	864	864
Recruiter Characteristic	Index(σ)	Has Reviews	Has Other Assignments	Has Prior Revenue	Has Certifications	Previously Billed Hours	Fully-billed

Notes: This table examines the relationship between call-back decisions and supply- and demand-side behavior. In this table, we examine heterogeneity by the type of recruiter. The types of heterogeneity we examine are in the final rows of the tables above. They were chosen because of their relation to career concerns (we explain these choices in Section 4.5). Each regression contains estimates of Equation 5 with the characteristic (binary in columns 2–7, and a normalized index of all six in column 1) variable for that column added, as well as three interactions with that variable (the probabilities). To flexibly control for age and experience, we place these dummy variables, we place each of these variables (described above), plus interactions with the three probabilities, into our regressions. We then use the post-double-selection lasso methodology (Belloni et al., 2011, 2012, 2014, 2016; Ahrens et al., 2018) for a principled approach to adding high-dimensional controls. All regressions use robust standard errors clustered at the screener level.

C Accuracy of Recruiters' Beliefs

Our paper examines how third-party recruiters integrate their beliefs regarding candidate and manager behavior into callback decisions. In this section, we explore the accuracy of these recruiter beliefs, and whether recruiter callback decisions mirror callback decisions by managers. We are particularly interested in understanding (i) whether recruiters have unbiased beliefs regarding candidate and manager behavior, and (ii) whether recruiters' choices about callbacks differ from those of managers. To answer these questions, we ran two online survey experiments. Section C.1 describes our participants, C.2 outlines the design of our experiments, section C.3 overviews our specifications, and section C.4 discusses our results.

Summary of Results. Overall, our results suggest that recruiters were relatively unbiased regarding the behavior of job candidates and managers, with the exception of gender. Our estimates examine the probability that a subject (such as a manager or candidate) would take a particular action (such as accept or reject an offer) – and we compare the difference between the recruiter's probabilistic forecast against the subject's self-reported likelihood. For most of our estimates, we cannot reject a null hypothesis of zero difference between recruiter predictions and subjects' reports. Our standard errors are precise enough that our 95% confidence intervals rule out differences of $\approx$ 5-10 percentage points (and larger) in both directions. Even when we can reject zero differences, we can rule out differences of larger than 5-10 percentage points for several estimates. We do see several statistically significant differences about gender between recruiter forecasts and subject reports. However, even these are mostly small in economic significance.

We separately assess how recruiters make callback choices similarly (or differently) than managers. This is not necessarily a forecast accuracy question, because a callback choice could incorporate other considerations besides predictions. However, this analysis could shed light on whether recruiters make different decisions than managers themselves. We find several instances suggesting that callback decisions by recruiters are different. Recruiters placed a higher weight on candidates likely to pass interviews, and a lower weight on candidates likely to accept interviews and offers. When managers report their own decisions, they appear to place more consideration on the candidates' probability of accepting (or lack thereof).

C.1 Survey Subjects

Our accuracy exercise occurred on Prolific.co, a survey company that composes survey panels for academic research. We used Prolific.co because the company maintains a participant pool of software engineers and software engineering managers. Researchers such as ourselves can use this survey pool to ask questions of the participants of this industry (both rank-and-file software engineers as well as managers). While the candidates and hiring managers in our main experiment were fictitious, we used Prolific to find survey subjects with similar characteristics. We then inquired about the behavior of these subjects directly. The subjects provided us with self-reported data from the participants to compare (and contrast) with recruiters' beliefs about them (as collected in our main experiment). To identify software engineers, we filtered for subjects with "computer programming" job-related skills. For managers, we filtered for subjects with "computer programming" job-related skills and with management experience. Prolific pre-screened subjects for these attributes for participation in its survey panel.

C.2 Survey Design

We ran two separate surveys (one about each side of the market).

Survey of Managers about Candidates. In the first survey, we presented approximately 250 software engineering managers with a job description and company like the one in our main experiment. Next, we asked participants to assess a series of job candidates.

Each hiring manager in our survey reviewed eight candidates. These candidates in the survey parallelled candidates from our main experiment on recruiters. For each candidate, we showed the same names as in our main survey (indicative of race and gender), as well as the university where they received their BA in Computer Science. We also showed a randomized prior employer (either a large company or small one, as in our main experiment).

We showed all combinations of race, gender and education (eight candidates), and randomized the prior employer. This generated eight candidates.[16] For each job candidate, the manager reported $P(AcceptInterview)$, $P(PassInterview|AcceptInterview)$ and $P(AcceptOffer|PassInterview)$. We also asked the managers to suggest a callback decision

[16]Our main experiment on recruiters contained 16 candidate types because we showed all combinations of (gender, race, education, prior employer). We simplified this in our accuracy experiment in order to avoid subject fatigue.

for each job candidate.

Survey of Job Candidates about Potential Managers. In the second survey, we recruited about 250 software engineering workers (rather than managers). These workers performed a similar task, but from the perspective of the job-seekers. Each candidate in the survey reviewed nine hiring managers (including the blinded condition) that were similarly parallel to hiring managers in our main experiment. For each manager, we again showed their name using the same names as in our main survey (which were indicative of race and gender), as well as the university where they received their BA in Computer Science and MBA. We also showed a randomized prior employer (either a large company or small one, as in our main experiment). This produced eight types of managers, plus the blinded condition for a total of nine. For each manager, the participant reported *P(Accept Interview)*, *P(Pass Interview | Accept Interview)* and *P(Accept Offer | Pass Interview)*.

C.3 Specifications

These surveys present an opportunity to assess the accuracy of recruiter beliefs and compare the callback behavior of recruiters versus managers. To do so, we append the datasets from Prolific with the dataset from the main experiment. We then run three sets of analyses.

Predictions about Candidates. First, in order to understand whether the recruiters in our sample had accurate (unbiased) beliefs about manager evaluations of job candidates, we estimate the following models:

$$\begin{aligned} Y_{c,e} =& \beta_0 + \beta_1 * Female_c + \beta_2 * Black_c + \beta_3 * EliteUniversity_c + \beta_4 * Female_c * Recruiter_e \\ &+ \beta_5 * Black_c * Recruiter_e + \beta_6 * EliteUniversity_c * Recruiter_e + \gamma * E_e + \epsilon_{c,e} \end{aligned} \tag{9}$$

where c indexes job candidates and e indicates evaluators. As in equation 2, $Female_c$, $Black_c$, and $EliteUniversity_c$ are binary indicators of candidate c's characteristics. E_e is a vector of evaluator fixed effects. $Y_{c,e}$ measures an outcome Y (callback or one of the three underlying probabilities) for candidate c assigned to evaluator e. Coefficients β_1, β_2, and β_3 capture the effects of our candidate manipulations on the beliefs of hiring managers.

One limitation of the equation 9 is that the fixed effects absorb the level effects of

recruiters. In order to express this, we also estimate a second equation:

$$\begin{aligned} Y_{c,e} =& \beta_0 + \beta_1 * Female_c + \beta_2 * Black_c + \beta_3 * EliteUniversity_c + \beta_4 * Recruiter_e \\ &+ \beta_5 * Female_c * Recruiter_e + \beta_6 * Black_c * Recruiter_e \\ &+ \beta_7 * EliteUniversity_c * Recruiter_e + \epsilon_{c,e} \end{aligned} \tag{10}$$

The notation is the same as equation 9, but $Recruiter_e$ is a binary indicator that equals one if the evaluator is the recruiter, and zero if the evaluator is the manager in the Prolific sample. For both specifications, the coefficients of interest are β_4, β_5, and β_6: these indicators capture the extent to which recruiter evaluations differ from manager evaluations. We also estimate a joint F-test under the null hypothesis that recruiter beliefs are the same as manager beliefs (i.e., $\beta_4 = \beta_5 = \beta_6 = 0$).

Predictions about Hiring Managers. In order to understand whether the recruiters in our sample had accurate (unbiased) beliefs about candidate evaluations of managers, we estimate:

$$\begin{aligned} Y_{h,e} =& \beta_0 + \beta_1 * Female_h + \beta_2 * Black_h + \beta_3 * EliteUniversity_h + \beta_4 * Blind_h \\ &+ \beta_5 * Female_h * Recruiter_e + \beta_6 * Black_h * Recruiter_e + \beta_7 * EliteUniversity_h * Recruiter_e \\ &+ \beta_8 * Blind_h * Recruiter_e + \gamma * E_e + \epsilon_{h,e} \end{aligned} \tag{11}$$

where h indexes hiring managers and e indicates evaluators. Like in equation 3, $Female_h$, $Black_h$, $EliteUniversity_h$, and $Blind_h$ measure whether hiring manager h is female, black, from an elite university, or blinded, respectively. $Recruiter_e$ is a binary indicator that equals one if the evaluator is the recruiter, and zero if the evaluator is the job candidate in the Prolific sample. $Y_{h,e}$ measures an outcome Y (callback or one of the three underlying probabilities) for hiring manager h assigned to evaluator e.

Because each screener in our main experiment only saw one hiring manager, we cannot include evaluator fixed effects as we do in Equation 9. Instead we include E_e, a vector of evaluator characteristics (gender, race, education, firm background, and whether the evaluator was a recruiter or manager).[17] Although the evaluator controls absorb some variation, they are less powerful than evaluator fixed effects. As a result, our estimates in this section will be less precise than Equation 9.

[17] For survey subjects, we collected these variables in the survey. For recruiters, we collected it from their profiles. Where this information was missing, we included a "missing" dummy variable. Our results are robust to including these controls in the regression or not.

The coefficients of interest are β_5, β_6, β_7, and β_8: these indicators capture the extent to which recruiter evaluations differ from candidate evaluations. We also estimate a joint F-test under the null hypothesis that recruiter beliefs are the same as candidate beliefs (i.e., $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$).

Callback Decisions. Finally, we are interested in testing whether the recruiters in our sample put the same weight on the three probability measures as hiring managers would, when deciding who to interview. In order to estimate this, we run the following regressions:

$$\begin{aligned} Y_{c,e} =& \beta_0 + \beta_1 * ProbabilityAcceptInterview_{c,e} + \beta_2 * ProbabilityPassInterview_{c,e} \\ &+ \beta_3 * ProbabilityAcceptOffer_{c,e} + \beta_4 * ProbabilityAcceptInterview_{c,e} * Recruiter_e \\ &+ \beta_5 * ProbabilityPassInterview_{c,e} * Recruiter_e + \\ &\beta_6 * ProbabilityAcceptOffer_{c,e} * Recruiter_e + \gamma * E_e + \epsilon_{c,e} \end{aligned} \tag{12}$$

where c indexes job candidates and e indicates evaluators. As in equation 5, $ProbabilityAcceptInterview_{c,h,s}$, $ProbabilityPassInterview_{c,h,s}$ and $ProbabilityAcceptOffer_{c,h,s}$ are probabilities measured through the procedure in Section 3.1. E_e is a vector of evaluator fixed effects, and $Y_{c,e}$ measures the callback choice Y for candidate c assigned to evaluator e. In order to report the levels, we also run the following regression without fixed effects:

$$\begin{aligned} Y_{c,e} =& \beta_0 + \beta_1 * ProbabilityAcceptInterview_{c,e} + \beta_2 * ProbabilityPassInterview_{c,e} \\ &+ \beta_3 * ProbabilityAcceptOffer_{c,e} + \beta_4 * Recruiter_e \\ &+ \beta_5 * ProbabilityAcceptInterview_{c,e} * Recruiter_e \\ &+ \beta_6 * ProbabilityPassInterview_{c,e} * Recruiter_e \\ &+ \beta_7 * ProbabilityAcceptOffer_{c,e} * Recruiter_e + \epsilon_{c,e} \end{aligned} \tag{13}$$

$Recruiter_e$ is a binary indicator that equals one if the evaluator is the recruiter, and zero if the evaluator is the manager in the Prolific sample, Equation 13 includes a binary indicator for recruiters to estimate the difference in levels of callbacks, while equation 12 has evaluator fixed effects that absorb the level difference.

The coefficients of interest are β_4, β_5, and β_6: these indicators capture the extent to which recruiter weights on the probability measure differ from the manager weights in determining callback decisions. We also estimate a joint F-test under the null hypothesis that recruiter weighting of the probabilities is the same as the managers (i.e., $\beta_4 = \beta_5 = \beta_6 = 0$).

C.4 Results

Accuracy of Candidate Evaluations of Managers. We begin by examining whether the recruiters in our sample had accurate (unbiased) beliefs about candidate evaluations of managers. In Table C.1, we estimate equation 11 for the three probability measures. Our results suggest that overall, recruiters had unbiased beliefs about candidate reactions to manager characteristics on all dimensions except gender.

Table C.1: **Accuracy of Candidate Evaluations of Hiring Managers**

	(1) P(Accepts Interview)	(2) P(Passes Interview)	(3) P(Accepts Offer)
Female Hiring Manager	.01***	.01**	.01***
	(.005)	(.004)	(.005)
Black Hiring Manager	-.01	-.003	-.01
	(.01)	(.01)	(.01)
Elite University Hiring Manager	.02***	-.02***	.01***
	(.01)	(.005)	(.01)
Blinded Hiring Manager	-.02*	-.04***	.01
	(.01)	(.01)	(.01)
Recruiter Eval	.02	.12*	.01
	(.06)	(.07)	(.05)
Recruiter Eval× Female Hiring Manager	-.07*	-.03	-.11***
	(.04)	(.03)	(.03)
Recruiter Eval× Black Hiring Manager	-.05	.01	-.04
	(.04)	(.03)	(.03)
Recruiter Eval× Elite University Hiring Manager	.02	.07**	.02
	(.03)	(.03)	(.03)
Recruiter Eval× Blinded Hiring Manager	.06	.05	.03
	(.06)	(.05)	(.04)
P(Recruiter Coeffs All Zero)	0.06	0.12	0.00
Observations	2776	2776	2776
R^2	0.03	0.05	0.04

Notes: This table examines recruiter accuracy of candidate evaluations of hiring managers. It regresses the probability measures on the hiring manager manipulations, a binary indicator for the evaluator, and interactions between the two, using equation 11. All regressions include robust standard errors clustered at the evaluator level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

We begin by analyzing *P(Accept Interviews)* and *P(Accept Offer)* in Columns 1 and 3. For these two outcomes, our survey subjects speak for themselves about their own probability of acceptance under different circumstances. On these outcomes, recruiters appear to

believe (incorrectly) that candidates would avoid accepting offers and interviews with female leaders. The recruiters in our sample believed that job candidates were less likely to accept interviews and offers from female hiring managers (see Table 6). However, job candidates in our validation exercise were 1.4 and 1.3 percentage points more likely to accept interviews and job offers from female hiring managers, respectively. For the other characteristics of hiring managers, we cannot reject a null hypothesis of zero differences between recruiter forecasts and candidate responses. Our standard errors are generally small enough to rule out large differences.

We also study *P(Pass Interview)*. In this survey, we asked candidates to assess the probability that they would pass the interview. Unlike their reports about their likelihood of accepting, this variable is not about their own behavior (it is about the behavior of managers). As such, these survey responses are less connected to ground-truth. However, these responses can be used to measure the level of disagreement between the recruiters and the candidates about the likelihood of passing interviews. The results in Table C.1 suggest that recruiters and job candidates were generally similar on the interaction terms. We do observe a level effect, indicating that recruiters are more optimistic than job candidates about their chances of passing interviews. On the interaction terms, we also find differences regarding hiring managers from elite universities. Job candidates reported being 1.6 percentage points less likely to pass interviews with hiring managers from elite versus non-elite universities, while recruiters reported no differences in likelihood of passing interviews with managers from elite versus non-elite universities.

Accuracy of Manager Evaluations of Candidates. We now examine the other side of the market: whether the recruiters in our sample had accurate beliefs about manager evaluations of job candidates. We begin by using our specification without fixed effects so that we can level effects. Table C.2 estimates equation 10.

Our results here suggest that recruiters had also held relatively accurate beliefs regarding manager evaluations of candidates. Of the three probability measure, the respondents in this survey (managers) are in the best position to evaluate column #2 (the probability they would pass this candidate). Here we see level effects of around +8 percentage points. However, our coefficients on the interaction terms in this column are also mostly small and cannot be rejected from zero, and the standard errors on these coefficients rule out large effects. One exception is for black candidates. Managers themselves report differences for candidates with black names. Recruiters view these candidates as more likely to pass the interview, but only by 3 percentage points. In Table C.3, we add fixed effects (equation 9). This eliminates the level effects, but the interaction terms are almost identical.

Table C.2: **Accuracy of Hiring Manager Evaluations of Candidates (No Evaluator Fixed Effects)**

	(1) P(Accepts Interview)	(2) P(Passes Interview)	(3) P(Accepts Offer)
Female Job Applicant	-.002	.01*	.002
	(.004)	(.005)	(.004)
Black Job Applicant	-.004	-.01	.003
	(.01)	(.01)	(.01)
Elite University Job Applicant	-.01**	.05***	-.02***
	(.01)	(.01)	(.01)
Recruiter Eval	.00	-.08***	-.04
	(.03)	(.03)	(.03)
Recruiter Eval× Female Applicant	.03**	.02	.03*
	(.01)	(.02)	(.02)
Recruiter Eval× Black Applicant	.01	.03*	.01
	(.02)	(.02)	(.01)
Recruiter Eval× Elite University Applicant	.01	.02	.02
	(.02)	(.02)	(.02)
P(Recruiter Coeffs All Zero)	0.15	0.21	0.26
Observations	2328	2328	2328
R^2	0.01	0.03	0.01

Notes: This table examines recruiter accuracy of manager evaluations of candidates. It regresses the probability measures on the candidate manipulations, a binary indicator for the evaluator, and interactions between the two, using equation 10. All regressions include robust standard errors clustered at the evaluator level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

We also ask the managers in our survey what they believe about likely candidate behavior. Although the managers' views could themselves be inaccurate, this would show evidence of disagreement between recruiters and their clients. On these outcomes (Columns 1 and 3), we find no evidence of level effects. (both in Tables C.2 and C.3). However, we do see gender differences. Managers believe female candidates are approximately equally likely to accept as men. Recruiters report greater optimism about women accepting. The joint F-test returns $p > 0.10$ for all three probability measures.

Differences about Whom to Interview. The results in this section thus far examine the probability that a subject (such as a manager or candidate) would take a particular action (such as accept or reject an offer) – and we compare the difference between the recruiter's probabilistic forecast against the subject's self-reported likelihood.

Table C.3: **Accuracy of Hiring Manager Evaluations of Candidates**

	(1) P(Accepts Interview)	(2) P(Passes Interview)	(3) P(Accepts Offer)
Female Job Applicant	-.002	.01*	.002
	(.004)	(.005)	(.004)
Black Job Applicant	-.004	-.01	.003
	(.01)	(.01)	(.01)
Elite University Job Applicant	-.01**	.05***	-.02***
	(.01)	(.01)	(.01)
Recruiter Eval× Female Applicant	.03**	.02	.03*
	(.01)	(.02)	(.02)
Recruiter Eval× Black Applicant	.01	.03*	.01
	(.02)	(.02)	(.01)
Recruiter Eval× Elite University Applicant	.01	.02	.02
	(.02)	(.02)	(.02)
P(Recruiter Coeffs All Zero)	0.15	0.21	0.26
Evaluator FEs	Y	Y	Y
Observations	2328	2328	2328
R^2	0.57	0.54	0.59

Notes: This table examines recruiter accuracy of manager evaluations of candidates. It regresses the probability measures on the candidate manipulations, evaluator fixed effects, and interactions between the two, using equation 9. All regressions include robust standard errors clustered at the evaluator level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

We now study how recruiters make callback choices similarly (or differently) than managers. This is not necessarily a forecast accuracy question (i.e., a callback choice could incorporate other considerations besides predictions). However, this analysis can shed light on whether recruiters make different decisions than managers themselves. We begin by considering Table C.4: In our survey, we asked managers to make an interviewing choice about each of the eight candidates. From this, we can see whether recruiters and their clients (the managers themselves) would make different decisions (and about whom).

We use this outcome to estimate equations 9 and 10 (column 1 containing level effects column 2 with fixed effects). Both suggest that recruiters and managers differ about who is chosen for callbacks. In Column 1, we find a large level effect: Recruiters are much more selective than managers. We also see that managers are more likely to interview elite versus non-elite university applicants, but recruiters are even more likely to give them a callback (both with and without fixed effects). The joint F-test p-value is $p = 0.01$ in both

columns.

Table C.4: **Callback Decisions: Recruiters versus Surveyed Hiring Managers**

	(1) Interview	(2) Interview
Female Job Applicant	.001	.001
	(.01)	(.01)
Black Job Applicant	-.03*	-.02*
	(.02)	(.02)
Elite University Job Applicant	.05***	.05***
	(.02)	(.02)
Recruiter Eval	-.33***	
	(.06)	
Recruiter Eval× Female Applicant	.05	.05
	(.04)	(.04)
Recruiter Eval× Black Applicant	.04	.04
	(.05)	(.05)
Recruiter Eval× Elite University Applicant	.16***	.16***
	(.06)	(.06)
P(Recruiter Coeffs All Zero)	0.01	0.01
Evaluator FEs		Y
Observations	2328	2328
R^2	0.06	0.30

Notes: This table examines differences in callbacks from recruiters versus hiring managers by regressing the interview outcome on the candidate manipulations, indicators for the evaluator, and interactions between the two. Column 1 displays the results of equation 13 using a binary indicator for whether the evaluator is a recruiter, so we can estimate level differences. Column 2 displays the results of equation 12 using evaluator fixed effects. All regressions include robust standard errors clustered at the evaluator level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Why do recruiters and managers recommend different candidates? Table C.5 estimates equations 12 and 13. The joint F-test returns $p < 0.01$ (both with fixed effects and level effects), which indicates that recruiters put different weights on the three probability measures than managers do. In particular, our results suggest that compared to managers, recruiters choices about interviews are much more correlated with who they think will pass interviews (*P(Pass Interview | Accept Interview)*, and less correlated with who they think is interested in the firm (*P(Accept Interview)* and *P(Accept Offer | Pass Interview)*. Thus, recruiters shift the types of candidates that receive callbacks. 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.

Table C.5: **Callback Decisions: Recruiters versus Surveyed Hiring Managers**

	(1) Interview	(2) Interview	(3) Interview	(4) Interview
P(Accepts Interview)	.16	.18		
	(.11)	(.11)		
P(Passes Interview)	.88***	1.20***	.90***	1.20***
	(.10)	(.10)	(.10)	(.10)
P(Accepts Offer)	-.21**	.26**	-.11	.36***
	(.10)	(.13)	(.08)	(.11)
Recruiter Eval	-.89***		-.69***	
	(.25)		(.20)	
Recruiter Eval× P(Accepts Interview)	.44	.54**		
	(.29)	(.26)		
Recruiter Eval× P(Passes Interview)	.59***	.69***	.58***	.69***
	(.15)	(.15)	(.16)	(.15)
Recruiter Eval× P(Accepts Offer)	-.05	-.36**	.16	-.12
	(.17)	(.15)	(.29)	(.26)
P(Recruiter Coeffs All Zero)	0.00	0.00	0.00	0.00
Evaluator FEs		Y		Y
Observations	2328	2328	2328	2328
R^2	0.28	0.55	0.27	0.54

Notes: This table examines differences in callbacks from recruiters versus hiring managers by regressing the interview outcome on the three probability measures, indicators for the evaluator, and interactions between the two. Columns 1 and 3 display the results of equation 13 using a binary indicator for whether the evaluator is a recruiter, so we can estimate level differences. Columns 2 and 4 display the results of equation 12 using evaluator fixed effects. All regressions include robust standard errors clustered at the evaluator level.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

References

Ahrens, Achim, Christian B. Hansen, and Mark E Schaffer, "PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference," Statistical Software Components, Boston College Department of Economics February 2018.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 2012, *80* (6), 2369–2429.

_ , **Victor Chernozhukov, and Christian Hansen**, "Inference for high-dimensional sparse econometric models," *arXiv preprint arXiv:1201.0220*, 2011.

_ , _ , **and** _ , "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 2014, *28* (2), 29–50.

_ , _ , _ , **and Damian Kozbur**, "Inference in high-dimensional panel models with an application to gun control," *Journal of Business & Economic Statistics*, 2016, *34* (4), 590–605.

Brier, Glenn W, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, 1950, *78* (1), 1–3.

Camerer, Colin, "Individual decision making," *Handbook of experimental economics*, 1995.

Filippas, Apostolos, John Joseph Horton, and Joseph Golden, "Reputation inflation," in "Proceedings of the 2018 ACM Conference on Economics and Computation" 2018, pp. 483–484.

Gaddis, S Michael, "How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies," *Sociological Science*, 2017, *4*, 469–489.

Holmström, Bengt, "Managerial incentive problems: A dynamic perspective," *The review of Economic studies*, 1999, *66* (1), 169–182.

Huck, Steffen and Georg Weizsäcker, "Do players correctly estimate what others do?: Evidence of conservatism in beliefs," *Journal of Economic Behavior & Organization*, 2002, *47* (1), 71–85.

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.

Murphy, Allan H and Robert L Winkler, "Scoring rules in probability assessment and evaluation," 1970.

Nosko, Chris and Steven Tadelis, "The limits of reputation in platform markets: An empirical analysis and field experiment," Technical Report, National Bureau of Economic Research 2015.

Tadelis, Steven, "The market for reputations as an incentive mechanism," *Journal of political Economy*, 2002, *110* (4), 854–882.

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.