NBER WORKING PAPER SERIES

INCENTIVIZED RESUME RATING: ELICITING EMPLOYER PREFERENCES WITHOUT DECEPTION

Judd B. Kessler
Corinne Low
Colin Sullivan

Working Paper 25800
http://www.nber.org/papers/w25800

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019, Revised May 2019

We thank the participants of the NBER Summer Institute Labor Studies, the Berkeley Psychology and Economics Seminar, the Stanford Institute of Theoretical Economics Experimental Economics Session, Advances with Field Experiments at the University of Chicago, the Columbia-NYU-Wharton Student Workshop in Experimental Economics Techniques, and the Wharton Applied Economics Workshop for helpful comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentivized Resume Rating: Eliciting Employer Preferences without Deception
Judd B. Kessler, Corinne Low, and Colin Sullivan
NBER Working Paper No. 25800
May 2019, Revised May 2019
JEL No. C90,J24,J71

ABSTRACT

We introduce a new experimental paradigm to evaluate employer preferences, called Incentivized Resume Rating (IRR). Employers evaluate resumes they know to be hypothetical in order to be matched with real job seekers, preserving incentives while avoiding the deception necessary in audit studies. We deploy IRR with employers recruiting college seniors from a prestigious school, randomizing human capital characteristics and demographics of hypothetical candidates. We measure both employer preferences for candidates and employer beliefs about the likelihood candidates will accept job offers, avoiding a typical confound in audit studies. We discuss the costs, benefits, and future applications of this new methodology.

Judd B. Kessler
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
judd.kessler@wharton.upenn.edu

Corinne Low
1461 Steinberg Hall Dietrich Hall
3620 Locust Walk
Philadelphia, PA 19104
corlow@wharton.upenn.edu

Colin Sullivan
3620 Locust Walk
Philadelphia, PA 19104
colins@wharton.upenn.edu

An online appendix is available at http://www.nber.org/data-appendix/w25800

1 Introduction

How labor markets reward education, work experience, and other forms of human capital is of fundamental interest in labor economics and the economics of education (e.g., Autor and Houseman [2010], Pallais [2014]). Similarly, the role of discrimination in labor markets is a key concern for both policy makers and economists (e.g., Altonji and Blank [1999], Lang and Lehmann [2012]). Correspondence audit studies, including resume audit studies, have become powerful tools to answer questions in both domains.[1] These studies have generated a rich set of findings on discrimination in employment (e.g., Bertrand and Mullainathan [2004]), real estate and housing (e.g., Hanson and Hawley [2011], Ewens et al. [2014]), retail (e.g., Pope and Sydnor [2011], Zussman [2013]), and other settings (see Bertrand and Duflo [2016]). More recently, resume audit studies have been used to investigate how employers respond to other characteristics of job candidates, including unemployment spells [Kroft et al., 2013, Eriksson and Rooth, 2014, Nunley et al., 2017], for-profit college credentials [Darolia et al., 2015, Deming et al., 2016], college selectivity [Gaddis, 2015], and military service [Kleykamp, 2009].

Despite the strengths of this workhorse methodology, however, resume audit studies are subject to two major concerns. First, they use deception, generally considered problematic within economics [Ortmann and Hertwig, 2002, Hamermesh, 2012]. Employers in resume audit studies waste time evaluating fake resumes and pursuing non-existent candidates. If fake resumes systematically differ from real resumes, employers could become wary of certain types of resumes sent out by researchers, harming both the validity of future research and real job seekers whose resumes are similar to those sent by researchers. These concerns about deception become more pronounced as the method becomes more popular.[2] To our knowledge, audit and correspondence audit studies are the only experiments within economics for which deception has been permitted, presumably because of the importance of the underlying research questions and the absence of a method to answer them without deception.

A second concern arising from resume audit studies is their use of "callback rates" (i.e., the rates at which employers call back fake candidates) as the outcome measure that proxies for employer

[1] Resume audit studies send otherwise identical resumes, with only minor differences associated with a treatment (e.g., different names associated with different races), to prospective employers and measure the rate at which candidates are called back by those employers (henceforth the "callback rate"). These studies were brought into the mainstream of economics literature by Bertrand and Mullainathan [2004]. By comparing callback rates across groups (e.g., those with white names to those with minority names), researchers can identify the existence of discrimination. Resume audit studies were designed to improve upon traditional audit studies of the labor market, which involved sending matched pairs of candidates (e.g., otherwise similar study confederates of different races) to apply for the same job and measure whether the callback rate differed by race. These traditional audit studies were challenged on empirical grounds for not being double-blind [Turner et al., 1991] and for an inability to match candidate characteristics beyond race perfectly [Heckman and Siegelman, 1992, Heckman, 1998].

[2] Baert [2018] notes 90 resume audit studies focused on discrimination against protected classes in labor markets alone between 2005 and 2016. Many studies are run in the same venues (e.g., specific online job boards), making it more likely that employers will learn to be skeptical of certain types of resumes. These harms might be particularly relevant if employers become aware of the existence of such research. For example, employers may know about resume audit studies since they can be used as legal evidence of discrimination [Neumark, 2012].

interest in candidates. Since recruiting candidates is costly, firms may be reluctant to pursue candidates who will be unlikely to accept a position if offered. Callback rates may therefore conflate an employer's interest in a candidate with the employer's expectation that the candidate would accept a job if offered one.[3] This confound might contribute to counterintuitive results in the resume audit literature. For example, resume audit studies typically find higher callback rates for unemployed than employed candidates [Kroft et al., 2013, Nunley et al., 2017, 2014, Farber et al., 2018], results that seem much more sensible when considering this potential role of job acceptance. In addition, callback rates can only identify preferences at one point in the quality distribution (i.e., at the threshold at which employers decide to call back candidates). While empirically relevant, results at this callback threshold may not be generalizable [Heckman, 1998, Neumark, 2012]. To better understand the underlying structure of employer preferences, we may also care about how employers respond to candidate characteristics at other points in the distribution of candidate quality.

In this paper, we introduce a new experimental paradigm, called Incentivized Resume Rating (IRR), which avoids these concerns. Instead of sending fake resumes to employers, IRR invites employers to evaluate resumes known to be hypothetical—avoiding deception—and provides incentives by matching employers with real job seekers based on employers' evaluations of the hypothetical resumes. Rather than relying on binary callback decisions, IRR can elicit much richer information about employer preferences; any information that can be used to improve the quality of the match between employers preferences and real job seekers can be elicited from employers in an incentivized way. In addition, IRR gives researchers the ability to elicit a single employer's preferences over multiple resumes, to randomize many candidate characteristics simultaneously, and to collect supplemental data about the employers reviewing resumes and their firms. Finally, IRR allows researchers to study employers who would not respond to unsolicited resumes.

We deploy IRR in partnership with the University of Pennsylvania (Penn) Career Services office to study the preferences of employers hiring graduating seniors through on-campus recruiting. This market has been unexplored by the resume audit literature since firms in this market hire through their relationships with schools rather than by responding to cold resumes. Our implementation of IRR asked employers to rate hypothetical candidates on two dimensions: (1) how interested they would be in hiring the candidate and (2) the likelihood that the candidate would accept a job offer if given one. In particular, employers were asked to report their interest in hiring a candidate on a 10-point Likert scale under the assumption that the candidate would accept the job if offered—mitigating concerns about a confound related to the likelihood of accepting the job. Employers were additionally asked the likelihood the candidate would accept a job offer on a 10-point Likert scale. Both responses were used to match employers with real Penn graduating seniors.

[3]Researchers who use audit studies aim to mitigate such concerns through the content of their resumes (e.g., Bertrand and Mullainathan [2004] notes that the authors attempted to construct high-quality resumes that did not lead candidates to be "overqualified," page 995).

We find that employers value higher grade point averages as well as the quality and quantity of summer internship experiences. Employers place extra value on prestigious and substantive internships but do not appear to value summer jobs that Penn students typically take for a paycheck, rather than to develop human capital for a future career, such as barista, server, or cashier. This result suggests a potential benefit on the post-graduate job market for students who can afford to take unpaid or low-pay internships during the summer rather than needing to work for an hourly wage.

Our granular measure of hiring interest allows us to consider how employer preferences for candidate characteristics respond to changes in overall candidate quality. Most of the preferences we identify maintain sign and significance across the distribution of candidate quality, but we find that responses to major and work experience are most pronounced towards the middle of the quality distribution and smaller in the tails.

While we do not find that employers are more or less interested in female and minority candidates on average, we find some evidence of discrimination against white women and minority men among employers looking to hire candidates with Science, Engineering, and Math majors.[4]

The employers in our study report having a positive preference for diversity in hiring.[5] In addition, employers report that white female candidates are less likely to accept job offers than their white male counterparts, suggesting a novel channel for discrimination.

Of course, the IRR method also comes with some drawbacks. First, while we attempt to directly identify employer interest in a candidate, our Likert-scale measure is not a step in the hiring process and thus—in our implementation of IRR—we cannot draw a direct link between our Likert-scale measure and hiring outcomes. However, we imagine future IRR studies could make advances on this front (e.g., by asking employers to guarantee interviews to matched candidates). Second, because the incentives in our study are similar but not identical to those in the hiring process, we cannot be sure that employers evaluate our hypothetical resumes with the same rigor or using the same criteria as they would real resumes. Again, we hope future work might validate that the time and attention spent on resumes in the IRR paradigm is similar to resumes evaluated as part of standard recruiting processes.

Our implementation of IRR was the first of its kind and thus left room for improvement on a few fronts. For example, as discussed in detail in Section 4, we attempted to replicate our study at the University of Pittsburgh to evaluate preferences of employers more like those traditionally targeted by resume audit studies. We underestimated how much Pitt employers needed candidates

[4]We find suggestive evidence that discrimination in hiring interest is due to implicit bias by observing how discrimination changes as employers evaluate multiple resumes. In addition, consistent with results from the resume audit literature finding lower returns to quality for minority candidates (see Bertrand and Mullainathan [2004]), we also find that—relative to white males—other candidates receive a lower return to work experience at prestigious internships.

[5]In a survey employers complete after evaluating resumes in our study, over 90% of employers report that both "seeking to increase gender diversity / representation of women" and "seeking to increase racial diversity" factor into their hiring decisions, and 82% of employers rate both of these factors at 5 or above on a Likert scale from 1 = "Do not consider at all" to 10 = "This is among the most important things I consider."

with specific majors and backgrounds, however, and a large fraction of resumes that were shown to Pitt employers were immediately disqualified based on major. This mistake resulted in highly attenuated estimates. Future implementations of IRR should more carefully tailor the variables for their hypothetical resumes to the needs of the employers being studied. We emphasize other lessons from our implementation in Section 5.

Despite the limitations of IRR, our results highlight that the method can be used to elicit employer preferences and suggest that it can also be used to detect discrimination. Consequently, we hope IRR provides a path forward for those interested in studying labor markets without using deception. The rest of the paper proceeds as follows: Section 2 describes in detail how we implement our IRR study; Section 3 reports on the results from Penn and compares them to extant literature; Section 4 describes our attempted replication at Pitt; and Section 5 concludes.

2 Study Design

In this section, we describe our implementation of IRR, which combines the incentives and ecological validity of the field with the control of the laboratory. In Section 2.1, we outline how we recruit employers who are in the market to hire elite college graduates. In Section 2.2, we describe how we provide employers with incentives for reporting preferences without introducing deception. In Section 2.3, we detail how we created the hypothetical resumes and describe the extensive variation in candidate characteristics that we included in the experiment, including grade point average and major (see 2.3.1), previous work experience (see 2.3.2), skills (see 2.3.3), and race and gender (see 2.3.4). In Section 2.4, we highlight the two questions that we asked subjects about each hypothetical resume, which allowed us to get a granular measure of interest in a candidate without a confound from the likelihood that the candidate would accept a job if offered.

2.1 Employers and Recruitment

IRR allows researchers to recruit employers in the market for candidates from particular institutions and those who do not screen unsolicited resumes and thus may be hard—or impossible—to study in audit or resume audit studies. To leverage this benefit of the experimental paradigm, we partnered with the University of Pennsylvania (Penn) Career Services office to identify employers recruiting highly skilled generalists from the Penn graduating class.

Penn Career Services sent invitation emails (see Appendix Figure A.1 for recruitment email) in two waves during the 2016-2017 academic year to employers who historically recruited Penn seniors (e.g., firms that recruited on campus, regularly attended career fairs, or otherwise hired students). The first wave was around the time of on-campus recruiting in the fall of 2016. The second wave was around the time of career-fair recruiting in the spring of 2017. In both waves, the recruitment email invited employers to use "a new tool that can help you to identify potential job candidates." While the recruitment email and the information that employers received before rating resumes (see

Appendix Figure A.3 for instructions) noted that anonymized data from employer responses would be used for research purposes, this was framed as secondary. The recruitment process and survey tool itself both emphasized that employers were using new recruitment software. For this reason, we note that our study has the ecological validity of a field experiment.[6] As was outlined in the recruitment email (and described in detail in Section 2.2), each employer's one and only incentive for participating in the study is to receive 10 resumes of job seekers that match the preferences they report through rating the hypothetical resumes.

2.2 Incentives

The main innovation of IRR is its method for incentivized preference elicitation, a variant of a method pioneered by Low [2017] in a different context. In its most general form, the method asks subjects to evaluate candidate profiles, which are known to be hypothetical, with the understanding that more accurate evaluations will maximize the value of their participation incentive. In our implementation of IRR, each employer evaluates 40 hypothetical candidate resumes and their participation incentive is a packet of 10 resumes of real job seekers from a large pool of Penn seniors. For each employer, we select the 10 real job seekers based on the employer's evaluations.[7] Consequently, the participation incentive in our study becomes more valuable as employers' evaluations of candidates better reflect their true preferences for candidates.[8]

A key design decision to help ensure subjects in our study truthfully and accurately report their preferences is that we provide no additional incentive (i.e., beyond the resumes of the 10 real job seekers) for participating in the study, which took a median of 29.8 minutes to complete. Limiting the incentive to the resumes of 10 job seekers makes us confident that participants value the incentive, since they have no other reason to participate in the study. Since subjects value the incentive, and since the incentive becomes more valuable as preferences are reported more accurately, subjects have good reason to report their preferences accurately.

2.3 Resume Creation and Variation

Our implementation of IRR asked each employer to evaluate 40 unique, hypothetical resumes, and it varied multiple candidate characteristics simultaneously and independently across resumes,

[6] Indeed, the only thing that differentiates our study from a "natural field experiment" as defined by Harrison and List [2004] is that subjects know that academic research is ostensibly taking place, even though it is framed as secondary relative to the incentives in the experiment.

[7] The recruitment email (see Appendix Figure A.1) stated: "the tool uses a newly developed machine-learning algorithm to identify candidates who would be a particularly good fit for your job based on your evaluations." We did not use race or gender preferences when suggesting matches from the candidate pool. The process by which we identify job seekers based on employer evaluations is described in detail in Appendix A.3.

[8] In Low [2017], heterosexual male subjects evaluated online dating profiles of hypothetical women with an incentive of receiving advice from an expert dating coach on how to adjust their own online dating profiles to attract the types of women that they reported preferring. While this type of non-monetary incentive is new to the labor economics literature, it has features in common with incentives in laboratory experiments, in which subjects make choices (e.g., over monetary payoffs, risk, time, etc.) and the utility they receive from those choices is higher as their choices more accurately reflect their preferences.

allowing us to estimate employer preferences over a rich space of baseline candidate characteristics.[9] Each of the 40 resumes was dynamically populated when a subject began the survey tool. As shown in Table 1 and described below, we randomly varied a set of candidate characteristics related to education; a set of candidate characteristics related to work, leadership, and skills; and the candidate's race and gender.

We made a number of additional design decisions to increase the realism of the hypothetical resumes and to otherwise improve the quality of employer responses. First, we built the hypothetical resumes using components (i.e., work experiences, leadership experiences, and skills) from real resumes of seniors at Penn. Second, we asked the employers to choose the type of candidates that they were interested in hiring, based on major (see Appendix Figure A.4). In particular, they could choose either "Business (Wharton), Social Sciences, and Humanities" (henceforth "Humanities & Social Sciences") or "Science, Engineering, Computer Science, and Math" (henceforth "STEM"). They were then shown hypothetical resumes from the set of majors they selected. As described below, this choice affects a wide range of candidate characteristics; majors, internship experiences, and skills on the hypothetical resumes varied across these two major groups. Third, to enhance realism, and to make the evaluation of the resumes less tedious, we used 10 different resume templates, which we populated with the candidate characteristics and component pieces described below, to generate the 40 hypothetical resumes (see Appendix Figure A.5 for a sample resume). We based these templates on real student resume formats (see Appendix Figure A.6 for examples).[10] Fourth, we gave employers short breaks within the study by showing them a progress screen after each block of 10 resumes they evaluated. As described in Section 3.4 and Appendix B.4, we use the change in attention induced by these breaks to construct tests of implicit bias.

2.3.1 Education Information

In the education section of the resume, we independently randomized each candidate's grade point average (GPA) and major. GPA is drawn from a uniform distribution between 2.90 and 4.00, shown to two decimal places and never omitted from the resume. Majors are chosen from a list of Penn majors, with higher probability put on more common majors. Each major was associated with a degree (BA or BS) and with the name of the group or school granting the degree within Penn (e.g., "College of Arts and Sciences"). Appendix Table A.3 shows the list of majors by major category, school, and the probability that the major was used in a resume.

[9] In a traditional resume audit study, researchers are limited in the number of resumes and the covariance of candidate characteristics that they can show to any particular employer. Sending too many fake resumes to the same firm, or sending resumes with unusual combinations of components, might raise suspicion. For example, Bertrand and Mullainathan [2004] send only four resumes to each firm and create only two quality levels (i.e., a high quality resume and a low quality resume, in which various candidate characteristics vary together).

[10] We blurred the text in place of a phone number and email address for all resumes, since we were not interested in inducing variation in those candidate characteristics.

Table 1: Randomization of Resume Components

Resume Component	Description	*Analysis Variable*
Personal Information		
First & last name	Drawn from list of 50 possible names given selected race and gender (names in Tables A.1 & A.2) Race drawn randomly from U.S. distribution (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian) Gender drawn randomly (50% male, 50% female)	*Female, White* (32.85%) *Male, Non-White* (17.15%) *Female, Non-White* (17.15%) *Not a White Male* (67.15%)
Education Information		
GPA	Drawn $Unif[2.90, 4.00]$ to second decimal place	*GPA*
Major	Drawn from a list of majors at Penn (Table A.3)	*Major* (weights in Table A.3)
Degree type	BA, BS fixed to randomly drawn major	*Wharton* (40%)
School within university	Fixed to randomly drawn major	*School of Engineering and Applied Science* (70%)
Graduation date	Fixed to upcoming spring (i.e., May 2017)	
Work Experience		
First job	Drawn from curated list of top internships and regular internships	*Top Internship* (20/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate's junior year (i.e., 2016)	
Second job	Left blank or drawn from curated list of regular internships and work-for-money jobs (Table A.5)	*Second Internship* (13/40) *Work for Money* (13/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate's sophomore year (i.e., 2015)	
Leadership Experience		
First & second leadership	Drawn from curated list	
Title and activity	Fixed to randomly drawn leadership	
Location	Fixed to Philadelphia, PA	
Description	Bullet points fixed to randomly drawn leadership	
Dates	Start and end years randomized within college career, with more recent experience coming first	
Skills		
Skills list	Drawn from curated list, with two skills drawn from {Ruby, Python, PHP, Perl} and two skills drawn from {SAS, R, Stata, Matlab} shuffled and added to skills list with probability 25%.	*Technical Skills* (25%)

Resume components are listed in the order that they appear on hypothetical resumes. Italicized variables in the right column are variables that were randomized to test how employers responded to these characteristics. Degree, first job, second job, and skills were drawn from different lists for Humanities & Social Sciences resumes and STEM resumes (except for work-for-money jobs). Name, GPA, work-for-money jobs, and leadership experience were drawn from the same lists for both resume types. Weights of characteristics are shown as fractions when they are fixed across subjects (e.g., each subject saw exactly 20/40 resumes with a *Top Internship*) and percentages when they represent a draw from a probability distribution (e.g., each resume a subject saw had a 32.85% chance of being assigned a white female name).

2.3.2 Work Experience

We included realistic work experience components on the resumes. To generate the components, we scraped more than 700 real resumes of Penn students. We then followed a process described in Appendix A.2.5 to select and lightly sanitize work experience components so that they could be randomly assigned to different resumes without generating conflicts or inconsistencies (e.g., we eliminated references to particular majors or to gender or race). Each work experience component included the associated details from the real resume from which the component was drawn, including an employer, position title, location, and a few descriptive bullet points.

Our goal in randomly assigning these work experience components was to introduce variation along two dimensions: *quantity* of work experience and *quality* of work experience. To randomly assign quantity of work experience, we varied whether the candidate only had an internship in the summer before senior year, or also had a job or internship in the summer before junior year. Thus, candidates with more experience had two jobs on their resume (before junior and senior years), while others had only one (before senior year).

To introduce random variation in *quality* of work experience, we selected work experience components from three categories: (1) "top internships," which were internships with prestigious firms as defined by being a firm that successfully hires many Penn graduates; (2) "work-for-money" jobs, which were paid jobs that—at least for Penn students—are unlikely to develop human capital for a future career (e.g., barista, cashier, waiter, etc.); and (3) "regular" internships, which comprised all other work experiences.[11]

The first level of quality randomization was to assign each hypothetical resume to have either a top internship or a regular internship in the first job slot (before senior year). This allows us to detect the impact of having a higher quality internship.[12]

The second level of quality randomization was in the kind of job a resume had in the second job slot (before junior year), if any. Many students may have an economic need to earn money during the summer and thus may be unable to take an unpaid or low-pay internship. To evaluate whether employers respond differentially to work-for-money jobs, which students typically take for pay, and internships, resumes were assigned to have either no second job, a work-for-money job, or a standard internship, each with (roughly) one-third probability (see Table 1). This variation

[11] See Appendix Table A.4 for a list of top internship employers and Table A.5 for a list of work-for-money job titles. As described in Appendix A.2.5, different internships (and top internships) were used for each major type but the same work-for-money jobs were used for both major types. The logic of varying internships by major type was based on the intuition that internships could be interchangeable within each group of majors (e.g., internships from the Humanities & Social Sciences resumes would not be unusual to see on any other resume from that major group) but were unlikely to be interchangeable across major groups (e.g., internships from Humanities & Social Sciences resumes would be unusual to see on STEM resumes and vice versa). We used the same set of work-for-money jobs for both major types, since these jobs were not linked to a candidate's field of study.

[12] Since the work experience component was comprised of employer, title, location, and description, a higher quality work experience necessarily reflects all features of this bundle; we did not independently randomize the elements of work experience.

allows us to measure the value of having a work-for-money job and to test how it compares to the value of a standard internship.

2.3.3 Leadership Experience and Skills

Each resume included two leadership experiences as in typical student resumes. A leadership experience component includes an activity, title, date range, and a few bullet points with a description of the experience (Philadelphia, PA was given as the location of all leadership experiences). Participation dates were randomly selected ranges of years from within the four years preceding the graduation date. For additional details, see Appendix A.2.5.

With skills, by contrast, we added a layer of intentional variation to measure how employers value technical skills. First, each resume was randomly assigned a list of skills drawn from real resumes. We stripped from these lists any reference to Ruby, Python, PHP, Perl, SAS, R, Stata, and Matlab. With 25% probability, we appended to this list four technical skills: two randomly drawn advanced programming languages from {Ruby, Python, PHP, Perl} and two randomly drawn statistical programs from {SAS, R, Stata, Matlab}.

2.3.4 Names Indicating Gender and Race

We randomly varied gender and race by assigning each hypothetical resume a name that would be indicative of gender (male or female) and race (Asian, Black, Hispanic, or White).[13] To do this randomization, we needed to first generate a list of names that would clearly indicate both gender and race for each of the groups. We used birth records and Census data to generate first and last names that would be highly indicative of race and gender, and combined names within race.[14] The full lists of names are given in Appendix Tables A.1 and A.2 (see Appendix A.2.3 for additional details).

For realism, we randomly selected races at rates approximating the distribution in the US population (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian). While a more uniform variation in race would have increased statistical power to detect race-based discrimination, such an approach would have risked signaling to subjects our intent to study racial preferences. In our analysis, we pool non-white names to explore potential discrimination of minority candidates.

[13] For ease of exposition, we will refer to race / ethnicity as "race" throughout the paper.

[14] For first names, we used a dataset of all births in the state of Massachusetts between 1989-1996 and New York City between 1990-1996 (the approximate birth range of job seekers in our study). Following Fryer and Levitt [2004], we generated an index for each name of how distinctively the name was associated with a particular race and gender. From these, we generated lists of 50 names by selecting the most indicative names and removing names that were strongly indicative of religion (such as Moshe) or gender ambiguous in the broad sample, even if unambiguous within an ethnic group (such as Courtney, which is a popular name among both black men and white women). We used a similar approach to generating racially indicative last names, assuming last names were not informative of gender. We used last name data from the 2000 Census tying last names to race. We implemented the same measure of race specificity and required that the last name make up at least 0.1% of that race's population, to ensure that the last names were sufficiently common.

2.4 Rating Candidates on Two Dimensions

As noted in the Introduction, audit and resume audit studies generally report results on callback, which has two limitations. First, callback only identifies preferences for candidates at one point in the quality distribution (i.e., at the callback threshold), so results may not generalize to other environments or to other candidate characteristics. Second, while callback is often treated as a measure of an employer's interest in a candidate, there is a potential confound to this interpretation. Since continuing to interview a candidate, or offering the candidate a job that is ultimately rejected, can be costly to an employer (e.g., it may require time and energy and crowd out making other offers), an employer's callback decision will optimally depend on both the employer's interest in a candidate and the employer's belief about whether the candidate will accept the job if offered. If the likelihood that a candidate accepts a job when offered is decreasing in the candidate's quality (e.g., if higher quality candidates have better outside options), employers' actual effort spent pursuing candidates may be non-monotonic in candidate quality. Consequently, concerns about a candidate's likelihood of accepting a job may be a confound in interpreting callback as a measure of interest in a candidate.[15]

An advantage of the IRR methodology is that researchers can ask employers to provide richer, more granular information than a binary measure of callback. We leveraged this advantage to ask two questions, each on a Likert scale from 1 to 10. In particular, for each resume we asked employers to answer the following two questions (see an example at the bottom of Appendix Figure A.5):

1. "How interested would you be in hiring **[Name]**?"
   (1 = "Not interested"; 10 = "Very interested")

2. "How likely do you think **[Name]** would be to accept a job with your organization?"
   (1 = "Not likely"; 10 = "Very likely")

In the instructions (see Appendix Figure A.3), employers were specifically told that responses to both questions would be used to generate their matches. In addition, they were told to focus only on their interest in hiring a candidate when answering the first question (i.e., they were instructed to assume the candidate would accept an offer if given one). We denote responses to this question "hiring interest." They were told to focus only on the likelihood a candidate would accept a job offer when answering the second question (i.e., they were instructed to assume they candidate had been given an offer and to assess the likelihood they would accept it). We denote responses to this question a candidate's "likelihood of acceptance." We asked the first question to assess how resume characteristics affect hiring interest. We asked the second question both to encourage employers

[15] Audit and resume audit studies focusing on discrimination do not need to interpret callback as a measure of an employer's interest in a candidate to demonstrate discrimination (any difference in callback rates is evidence of discrimination).

to focus only on hiring interest when answering the first question and to explore employers' beliefs about the likelihood that a candidate would accept a job if offered.

The 10-point scale has two advantages. First, it provides additional statistical power, allowing us to observe employer preferences toward characteristics of inframarginal resumes, rather than identifying preferences only for resumes crossing a binary callback threshold in a resume audit setting. Second, it allows us to explore how employer preferences vary across the distribution of hiring interest, an issue we explore in depth in Section 3.3.

3 Results

3.1 Data and Empirical Approach

We recruited 72 employers through our partnership with the University of Pennsylvania Career Services office in Fall 2016 (46 subjects, 1840 resume observations) and Spring 2017 (26 subjects, 1040 resume observations).[16]

As described in Section 2, each employer rated 40 unique, hypothetical resumes with randomly assigned candidate characteristics. For each resume, employers rated hiring interest and likelihood of acceptance, each on a 10-point Likert scale. Our analysis focuses initially on hiring interest, turning to how employers evaluate likelihood of acceptance in Section 3.5. Our main specifications are ordinary least squares (OLS) regressions. These specifications make a linearity assumption with respect to the Likert-scale ratings data. Namely, they assume that, on average, employers treat equally-sized increases in Likert-scale ratings equivalently (e.g., an increase in hiring interest from 1 to 2 is equivalent to an increase from 9 to 10). In some specifications, we include subject fixed effects, which account for the possibility that employers have different mean ratings of resumes (e.g., allowing some employers to be more generous than others with their ratings across all resumes), while preserving the linearity assumption. To complement this analysis, we also run ordered probit regression specifications, which relax this assumption and only require that employers, on average, consider higher Likert-scale ratings more favorably than lower ratings.

In Section 3.2, we examine how human capital characteristics (e.g., GPA, major, work experience, and skills) affect hiring interest. These results report on the mean of preferences across

[16]The recruiters who participated in our study as subjects were primarily female (59%) and primarily white (79%) and Asian (15%). They reported a wide range of recruiting experience, including some who had been in a position with responsibilities associated with job candidates for one year or less (28%); between two and five years (46%); and six or more years (25%). Almost all (96%) of the participants had college degrees, and many (30%) had graduate degrees including an MA, MBA, JD, or Doctorate. They were approximately as likely to work at a large firm with over 1000 employees (35%) as a small firm with fewer than 100 employees (39%). These small firms include hedge fund, private equity, consulting, and wealth management companies that are attractive employment opportunities for Penn undergraduates. Large firms include prestigious Fortune 500 consumer brands, as well as large consulting and technology firms. The most common industries in the sample are finance (32%); the technology sector or computer science (18%); and consulting (16%). The sample had a smaller number of sales/marketing firms (9%) and non-profit or public interest organizations (9%). The vast majority (86%) of participating firms had at least one open position on the East Coast, though a significant number also indicated recruiting for the West Coast (32%), Midwest (18%), South (16%), or an international location (10%).

the distribution; we show how our results vary across the distribution of hiring interest in Section 3.3. In Section 3.4, we discuss how employers' ratings of hiring interest respond to demographic characteristics of our candidates. In Section 3.5, we investigate the likelihood of acceptance ratings and identify a potential new channel for discrimination. In Section 3.6, we compare our results to prior literature.

3.2 Effect of Human Capital on Hiring Interest

Employers in our study are interested in hiring graduates of the University of Pennsylvania for full-time employment, and many recruit at other Ivy League schools and other top colleges and universities. This labor market has been unexplored by resume audit studies, in part because the positions employers aim to fill through on-campus recruiting at Penn are highly unlikely to be filled through online job boards or by screening unsolicited resumes. In this section, we evaluate how randomized candidate characteristics—described in Section 2.3 and Table 1—affect employers' ratings of hiring interest.

We denote an employer i's rating of a resume j on the 1–10 Likert scale as V_{ij} and estimate variations of the following regression specification (1). This regression allows us to investigate the average response to candidate characteristics across employers in our study.

$$\begin{aligned} V_{ij} = {} & \beta_0 + \beta_1 \, GPA + \beta_2 \, Top \ Internship + \beta_3 \, Second \ Internship + \beta_4 \, Work \ for \ Money + \\ & \beta_5 \, Technical \ Skills + \beta_6 \, Female, \ White + \beta_7 \, Male, \ Non\text{-}White + \\ & \beta_8 \, Female, \ Non\text{-}White + \mu_j + \gamma_j + \omega_j + \alpha_i + \varepsilon_{ij} \end{aligned} \tag{1}$$

In this regression, *GPA* is a linear measure of grade point average. *Top Internship* is a dummy for having a top internship, *Second Internship* is a dummy for having an internship in the summer before junior year, and *Work for Money* is a dummy for having a work-for-money job in the summer before junior year. *Technical Skills* is a dummy for having a list of skills that included a set of four randomly assigned technical skills. Demographic variables *Female, White*; *Male, Non-White*; and *Female, Non-White* are dummies equal to 1 if the name of the candidate indicated the given race and gender.[17] μ_j are dummies for each major. Table 1 provides more information about these dummies and all the variables in this regression. In some specifications, we include additional controls. γ_j are dummies for each of the leadership experience components. ω_j are dummies for the number of resumes the employer has evaluated as part of the survey tool. Since leadership experiences are independently randomized and orthogonal to other resume characteristics of interest, and since resume characteristics are randomly drawn for each of the 40 resumes, our results should be robust

[17] Coefficient estimates on these variables report comparisons to white males, which is the excluded group. While we do not discuss demographic results in this section, we include controls for this randomized resume component in our regressions and discuss the results in Section 3.4 and Appendix B.4.

to the inclusion or exclusion of these dummies. Finally, α_i are employer (i.e., subject) fixed effects that account for different average ratings across employers.

Table 2 shows regression results where V_{ij} is *Hiring Interest*, which takes values from 1 to 10. The first three columns report OLS regressions with slightly different specifications. The first column includes all candidate characteristics we varied to estimate their impact on ratings. The second column adds leadership dummies γ and resume order dummies ω. The third column also adds subject fixed effects α. As expected, results are robust to the addition of these controls. The fourth column, labeled *GPA-Scaled OLS*, rescales all coefficients from the third column by the coefficient on GPA (2.196) so that the coefficients on other variables can be interpreted in GPA points. These regressions show that employers respond strongly to candidate characteristics related to human capital.

GPA is an important driver of hiring interest. An increase in GPA of one point (e.g., from a 3.0 to a 4.0) increases ratings on the Likert scale by 2.1–2.2 points. The standard deviation of quality ratings is 2.6, suggesting that a point improvement in GPA moves hiring interest ratings by about 0.8 of a standard deviation.

As described in Section 2.3.2, we created *ex ante* variation in both the quality and quantity of candidate work experience. Both affect employer interest. The quality of a candidate's work experience in the summer before senior year has a large impact on hiring interest ratings. The coefficient on *Top Internship* ranges from 0.9–1.0 Likert-scale points, which is roughly a third of a standard deviation of ratings. As shown in the fourth column of Table 2, a top internship is equivalent to a 0.41 improvement in GPA.

Employers value a second work experience on the candidate's resume, but only if that experience is an internship and not if it is a work-for-money job. In particular, the coefficient on *Second Internship*, which reflects the effect of adding a second "regular" internship to a resume that otherwise has no work experience listed for the summer before junior year, is 0.4–0.5 Likert-scale points—equivalent to 0.21 GPA points. While listing an internship before junior year is valuable, listing a work-for-money job that summer does not appear to increase hiring interest ratings. The coefficient on *Work for Money* is small and not statistically different from zero in our data. While it is directionally positive, we can reject that work-for-money jobs and regular internships are valued equally ($p < 0.05$ for all tests comparing the *Second Internship* and *Work for Money* coefficients). This preference of employers may create a disadvantage for students who cannot afford to accept (typically) unpaid internships the summer before their junior year.[18]

We see no effect on hiring interest from increased *Technical Skills*, suggesting that employers on average do not value the technical skills we randomly added to candidate resumes or that listing technical skills does not credibly signal sufficient mastery to affect hiring interest (e.g., employers may consider skills listed on a resume to be cheap talk).

[18]These results are consistent with a penalty for working-class candidates. In a resume audit study of law firms, Rivera and Tilcsik [2016] found that resume indicators of lower social class (such as receiving a scholarship for first generation college students) led to lower callback rates.

Table 2: Hiring Interest

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.125	2.190	2.196	1.000	0.891
	(0.145)	(0.150)	(0.129)	(.)	(0.063)
Top Internship	0.902	0.900	0.897	0.409	0.378
	(0.095)	(0.099)	(0.081)	(0.043)	(0.040)
Second Internship	0.465	0.490	0.466	0.212	0.206
	(0.112)	(0.118)	(0.095)	(0.045)	(0.047)
Work for Money	0.116	0.157	0.154	0.070	0.052
	(0.110)	(0.113)	(0.091)	(0.042)	(0.046)
Technical Skills	0.046	0.053	-0.071	-0.032	0.012
	(0.104)	(0.108)	(0.090)	(0.041)	(0.043)
Female, White	-0.152	-0.215	-0.161	-0.073	-0.061
	(0.114)	(0.118)	(0.096)	(0.044)	(0.048)
Male, Non-White	-0.172	-0.177	-0.169	-0.077	-0.075
	(0.136)	(0.142)	(0.115)	(0.053)	(0.058)
Female, Non-White	-0.009	-0.022	0.028	0.013	-0.014
	(0.137)	(0.144)	(0.120)	(0.055)	(0.057)
Observations	2880	2880	2880	2880	2880
R^2	0.129	0.181	0.483		
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

Ordered probit cutpoints: 1.91, 2.28, 2.64, 2.93, 3.26, 3.60, 4.05, 4.51, and 5.03.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. The p-values of tests of joint significance of major fixed effects are indicated (F-test for OLS, likelihood ratio test for ordered probit).

Table 2 also reports the p-value of a test of whether the coefficients on the major dummies are jointly different from zero. Results suggest that the randomly assigned major significantly affects hiring interest. While we do not have the statistical power to test for the effect of each major, we can explore how employers respond to candidates being from more prestigious schools at the University of Pennsylvania. In particular, 40% of the Humanities & Social Sciences resumes are assigned a BS in Economics from Wharton and the rest have a BA major from the College of Arts and Sciences. In addition, 70% of the STEM resumes are assigned a BS from the School of Engineering and Applied Science and the rest have a BA major from the College of Arts and Sciences. As shown in Appendix Table B.2, in both cases, we find that being from the more prestigious school—and thus receiving a BS rather than a BA—is associated with an increase in hiring interest ratings of about 0.4 Likert-scale points or 0.18 GPA points.[19]

We can loosen the assumption that employers treated the intervals on the Likert scale linearly by treating *Hiring Interest* as an ordered categorical variable. The fifth column of Table 2 gives the results of an ordered probit specification with the same variables as the first column (i.e., omitting the leadership dummies and subject fixed effects). This specification is more flexible than OLS, allowing the discrete steps between Likert-scale points to vary in size. The coefficients reflect the effect of each characteristic on a latent variable over the Likert-scale space, and cutpoints are estimated to determine the distance between categories. Results are similar in direction and statistical significance to the OLS specifications described above.[20]

As discussed in Section 2, we made many design decisions to enhance realism. However, one might be concerned that our independent cross-randomization of various resume components might lead to unrealistic resumes and influence the results we find. We provide two robustness checks in the appendix to address this concern. First, our design and analysis treat each work experience as independent, but, in practice, candidates may have related jobs over a series of summers that create a work experience "narrative." In Appendix B.1 and Appendix Table B.1, we describe how we construct a measure of work experience narrative, we test its importance, and find that while employers respond positively to work experience narrative ($p = 0.054$) our main results are robust to its inclusion. Second, the GPA distribution we used for constructing the hypothetical resumes did not perfectly match the distribution of job seekers in our labor market. In Appendix B.2, we re-weight our data to match the GPA distribution in the candidate pool of real Penn job seekers and show that our results are robust to this re-weighting. These exercises provide some assurance that our results are not an artifact of how we construct hypothetical resumes.

[19]Note that since the application processes for these different schools within Penn are different, including the admissions standards, this finding also speaks to the impact of institutional prestige, in addition to field of study (see, e.g., Kirkeboen et al. [2016]).

[20]The ordered probit cutpoints (2.14, 2.5, 2.85, 3.15, 3.46, 3.8, 4.25, 4.71, and 5.21) are approximately equally spaced, suggesting that subjects treated the Likert scale approximately linearly. Note that we only run the ordered probit specification with the major dummies and without leadership dummies or subject fixed effects. Adding too many dummies to an ordered probit can lead to unreliable estimates when the number of observations per cluster is small [Greene, 2004].

3.3 Effects Across the Distribution of Hiring Interest

The regression specifications described in Section 3.2 identify the average effect of candidate characteristics on employers' hiring interest. As pointed out by Neumark [2012], however, these average preferences may differ in magnitude—and even direction—from differences in callback rates, which derive from whether a characteristic pushes a candidate above a specific quality threshold (i.e., the callback threshold). For example, in the low callback rate environments that are typical of resume audit studies, differences in callback rates will be determined by how employers respond to a candidate characteristic in the right tail of their distribution of preferences.[21] To make this concern concrete, Appendix B.3 provides a simple graphical illustration in which the average preference for a characteristic differs from the preference in the tail of the distribution. In practice, we may care about preferences in any part of the distribution for policy. For example, preferences at the callback threshold may be relevant for hiring outcomes, but those thresholds may change with a hiring expansion or contraction.

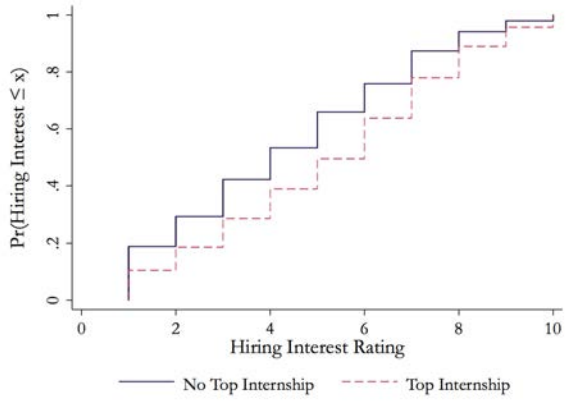
An advantage of the IRR methodology, however, is that it can deliver a granular measure of hiring interest to explore whether employers' preferences for characteristics do indeed differ in the tails of the hiring interest distribution. We employ two basic tools to explore preferences across the distribution of hiring interest: (1) the empirical cumulative distribution function (CDF) of hiring interest ratings and (2) a "counterfactual callback threshold" exercise. In the latter exercise, we impose a counterfactual callback threshold at each possible hiring interest rating (i.e., supposing that employers called back all candidates that they rated at or above that rating level) and, for each possible rating level, report the OLS coefficient an audit study researcher would find for the difference in callback rates.

While the theoretical concerns raised by Neumark [2012] may be relevant in other settings, the average results we find in Section 3.2 are all consistent across the distribution of hiring interest, including in the tails (except for a preference for Wharton students, which we discuss below). The top half of Figure 1 shows that *Top Internship* is positive and statistically significant at all levels of selectivity. Panel (a) reports the empirical CDF of hiring interest ratings for candidates with and without a top internship. Panel (b) shows the difference in callback rates that would arise for *Top Internship* at each counterfactual callback threshold. The estimated difference in callback rates is positive and significant everywhere, although it is much larger in the midrange of the quality distribution than at either of the tails.[22] The bottom half of Figure 1 shows that results across

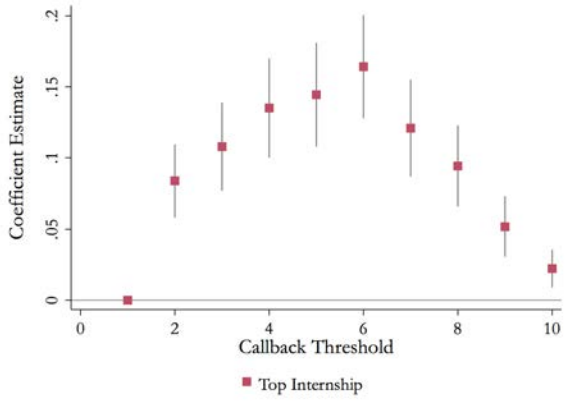
[21] A variant of this critique was initially brought up by Heckman and Siegelman [1992] and Heckman [1998] for in-person audit studies, where auditors may be imperfectly matched, and was extended to correspondence audit studies by Neumark [2012] and Neumark et al. [2015]. A key feature of the critique is that certain candidate characteristics might affect higher moments of the distribution of employer preferences so that how employers respond to a characteristic on average may be different than how an employer responds to a characteristic in the tail of their preference distribution.

[22] This shape is partially a mechanical feature of low callback rate environments: if a threshold is set high enough that only 5% of candidates with a desirable characteristic are being called back, the difference in callback rates can be no more than 5 percentage points. At lower thresholds (e.g., where 50% of candidates with desirable characteristics

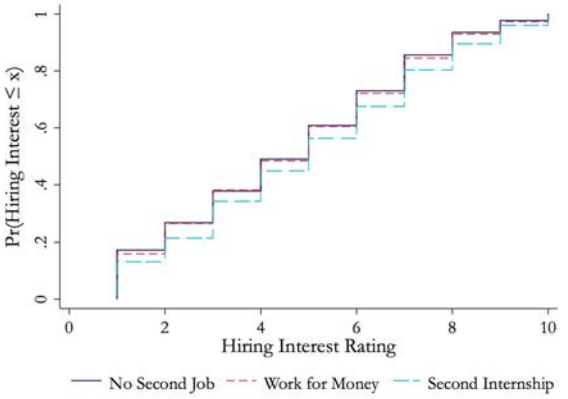
Figure 1: Value of Quality of Experience Over Selectivity Distribution



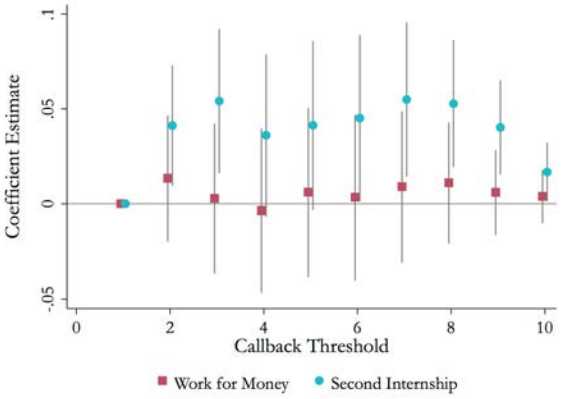
(a) Empirical CDF for Top Internship



(b) Linear Probability Model for Top Internship



(c) Empirical CDF for Second Job Type



(d) Linear Probability Model for Second Job Type

Empirical CDF of *Hiring Interest* (Panels 1a & 1c) and difference in counterfactual callback rates (Panels 1b & 1d) for *Top Internship*, in the top row, and *Second Internship* and *Work for Money*, in the bottom row. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

the distribution for *Second Internship* and *Work for Money* are also consistent with the average results from Section 3.2. *Second Internship* is positive everywhere and almost always statistically significant. *Work for Money* consistently has no impact on employer preferences throughout the distribution of hiring interest.

As noted above, our counterfactual callback threshold exercise suggests that a well-powered audit study would likely find differences in callback rates for most of the characteristics that we estimate as statistically significant on average in Section 3.2, regardless of employers' callback threshold. This result is reassuring both for the validity of our results and in considering the generalizability of results from the resume audit literature. However, even in our data, we observe a case where a well-powered audit study would be unlikely to find a result, even though we find one on average. Appendix Figure B.1 mirrors Figure 1 but focuses on having a Wharton degree among employers seeking Humanities & Social Sciences candidates. Employers respond to Wharton in the middle of the distribution of hiring interest, but preferences seem to converge in the right tail (i.e., at hiring interest ratings of 9 or 10), suggesting that the best students from the College of Arts and Sciences are not evaluated differently than the best students from Wharton.

3.4 Demographic Discrimination

In this section, we examine how hiring interest ratings respond to the race and gender of candidates. As described in Section 2 and shown in Table 1, we use our variation in names to create the variables: *Female, White*; *Male, Non-White*; and *Female, Non-White*. As shown in Table 2, the coefficients on the demographic variables are not significantly different from zero, suggesting no evidence of discrimination on average in our data.[23] This null result contrasts somewhat with existing literature—both resume audit studies (e.g., Bertrand and Mullainathan [2004]) and laboratory experiments (e.g., Bohnet et al. [2015]) generally find evidence of discrimination in hiring. Our differential results may not be surprising given that our employer pool is different than those usually targeted through resume audit studies, with most reporting positive tastes for diversity.

While we see no evidence of discrimination on average, a large literature addressing diversity in the sciences (e.g., Carrell et al. [2010], Goldin [2014]) suggests we might be particularly likely to see discrimination among employers seeking STEM candidates. In Table 3, we estimate the regression in Equation (1) separately by major type. Results in Columns 5-10 show that employers looking for STEM candidates display a large, statistically significant preference for white male candidates over white females and non-white males. The coefficients on *Female, White* and *Male, Non-White* suggest that these candidates suffer a penalty of 0.5 Likert-scale points—or about 0.27 GPA points—that is robust across our specifications. These effects are at least marginally significant even after multiplying our p-values by two to correct for the fact that we are

are called back), differences in callback rates can be much larger. In Appendix B.3, we discuss how this feature of difference in callback rates could lead to misleading comparisons across experiments with very different callback rates.

[23] In Appendix Table B.6, we show that this effect does not differ by the gender and race of the employer rating the resume.

Table 3: Hiring Interest by Major Type

	Dependent Variable: Hiring Interest									
	Humanities & Social Sciences					STEM				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.208	2.304	2.296	1.000	0.933	1.932	1.885	1.882	1.000	0.802
	(0.173)	(0.179)	(0.153)	(.)	(0.074)	(0.267)	(0.309)	(0.242)	(.)	(0.112)
Top Internship	1.075	1.043	1.033	0.450	0.452	0.398	0.559	0.545	0.289	0.175
	(0.108)	(0.116)	(0.095)	(0.050)	(0.046)	(0.191)	(0.216)	(0.173)	(0.100)	(0.078)
Second Internship	0.540	0.516	0.513	0.224	0.240	0.242	0.307	0.311	0.165	0.111
	(0.132)	(0.143)	(0.114)	(0.051)	(0.056)	(0.208)	(0.246)	(0.189)	(0.103)	(0.088)
Work for Money	0.087	0.107	0.116	0.050	0.037	0.151	0.275	0.337	0.179	0.076
	(0.129)	(0.134)	(0.110)	(0.048)	(0.055)	(0.212)	(0.254)	(0.187)	(0.102)	(0.088)
Technical Skills	0.063	0.084	-0.050	-0.022	0.013	-0.028	-0.113	-0.180	-0.096	-0.001
	(0.122)	(0.130)	(0.106)	(0.046)	(0.052)	(0.197)	(0.228)	(0.186)	(0.100)	(0.083)
Female, White	-0.047	-0.117	-0.054	-0.024	-0.015	-0.419	-0.612	-0.545	-0.290	-0.171
	(0.134)	(0.142)	(0.117)	(0.051)	(0.057)	(0.215)	(0.249)	(0.208)	(0.115)	(0.089)
Male, Non-White	-0.029	-0.010	-0.026	-0.011	-0.007	-0.567	-0.617	-0.507	-0.270	-0.265
	(0.158)	(0.169)	(0.137)	(0.059)	(0.066)	(0.271)	(0.318)	(0.257)	(0.136)	(0.111)
Female, Non-White	0.085	0.101	0.091	0.040	0.024	-0.329	-0.260	-0.046	-0.025	-0.142
	(0.160)	(0.171)	(0.137)	(0.060)	(0.068)	(0.264)	(0.301)	(0.261)	(0.138)	(0.111)
Observations	2040	2040	2040	2040	2040	840	840	840	840	840
R^2	0.128	0.196	0.500			0.119	0.323	0.593		
p-value for test of joint significance of Majors	0.021	0.027	0.007	0.007	0.030	< 0.001	0.035	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No	No	No	Yes	Yes	No

Ordered probit cutpoints (Column 5): 2.25, 2.58, 2.96, 3.26, 3.60, 3.94, 4.41, 4.86, 5.41.

Ordered probit cutpoints (Column 10): 1.44, 1.90, 2.22, 2.51, 2.80, 3.14, 3.56, 4.05, 4.48.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 and Column 8 divided by the Column 3 and Column 8 coefficients on GPA, with standard errors calculated by delta method. The p-values of tests of joint significance of major fixed effects are indicated (F-test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

analyzing our results within two subgroups (uncorrected p-values are: $p = 0.009$ for *Female, White*; $p = 0.049$ for *Male, Non-White*). Results in Columns 1-5 show no evidence of discrimination in hiring interest among Humanities & Social Sciences employers.

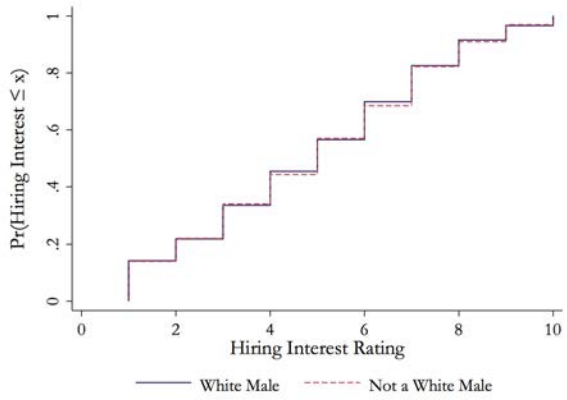
As in Section 3.3, we can examine these results across the hiring interest rating distribution. Figure 2 shows the CDF of hiring interest ratings and the difference in counterfactual callback rates. For ease of interpretation and for statistical power, we pool female and minority candidates and compare them to white male candidates in these figures and in some analyses that follow. The top row shows these comparisons for employers interested in Humanities & Social Sciences candidates and the bottom row shows these comparisons for employers interested in STEM candidates. Among employers interested in Humanities & Social Sciences candidates, the CDFs of *Hiring Interest* ratings are nearly identical. Among employers interested in STEM candidates, however, the CDF for white male candidates first order stochastically dominates the CDF for candidates who are not white males. At the point of the largest counterfactual callback gap, employers interested in STEM candidates would display callback rates that were 10 percentage points lower for candidates who were not white males than for their white male counterparts.

One might be surprised that we find any evidence of discrimination, given that employers may have (correctly) believed we would not use demographic tastes in generating their matches and given that employers may have attempted to override any discriminatory preferences to be more socially acceptable. One possibility for why we nevertheless find discrimination is the role of implicit bias [Greenwald et al., 1998, Nosek et al., 2007], which Bertrand et al. [2005] has suggested is an important channel for discrimination in resume audit studies. In Appendix B.4, we explore the role of implicit bias in driving our results.[24] In particular, we leverage a feature of implicit bias—that it is more likely to arise when decision makers are fatigued [Wigboldus et al., 2004, Govorun and Payne, 2006, Sherman et al., 2004]—to test whether our data are consistent with employers displaying an implicit racial or gender bias. As shown in Appendix Table B.7, employers spend less time evaluating resumes both in the latter half of the study and in the latter half of each set of 10 resumes (after each set of 10 resumes, we introduced a short break for subjects), suggesting evidence of fatigue. Discrimination is statistically significantly larger in the latter half of each block of 10 resumes, providing suggestive evidence that implicit bias plays a role in our findings, although discrimination is not larger in the latter half of the study.

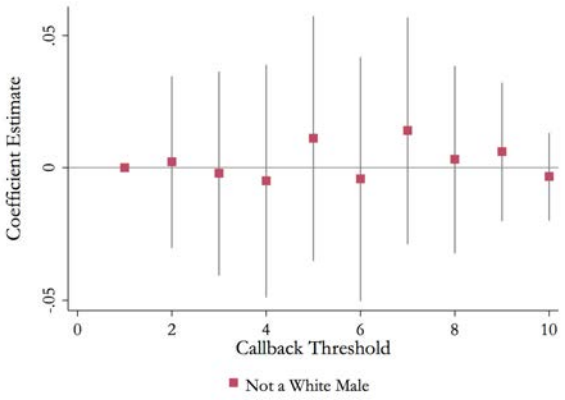
Race and gender could also subconsciously affect how employers view other resume components. We test for negative interactions between race and gender and desirable candidate characteristics, which have been found in the resume audit literature (e.g., minority status has been shown to lower returns to resume quality [Bertrand and Mullainathan, 2004]). Appendix Table B.8 interacts *Top*

[24] Explicit bias might include an explicit taste for white male candidates or an explicit belief they are more prepared than female or minority candidates for success at their firm, even conditional on their resumes. Implicit bias [Greenwald et al., 1998, Nosek et al., 2007], on the other hand, may be present even among employers who are not explicitly considering race (or among employers who are considering race but attempting to suppress any explicit bias they might have).

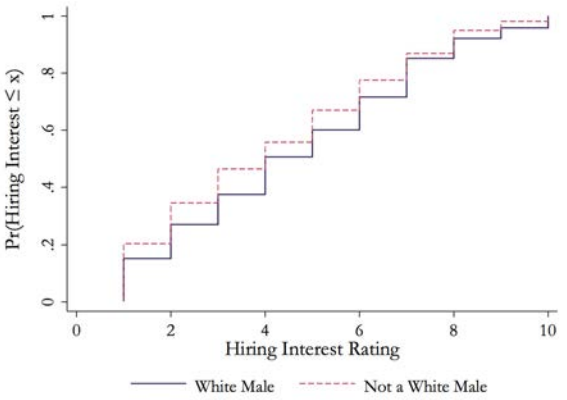
Figure 2: Demographics by Major Type Over Selectivity Distribution



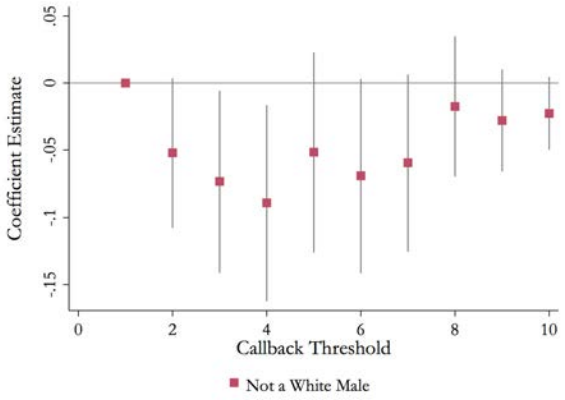
(a) Empirical CDF: Not a White Male, Humanities & Social Sciences



(b) Linear Probability Model: Not a White Male, Humanities & Social Sciences



(c) Empirical CDF: Not a White Male, STEM



(d) Linear Probability Model: Not a White Male, STEM

Empirical CDF of *Hiring Interest* (Panels 2a & 2c) and difference in counterfactual callback rates (Panels 2b & 2d) for *White Male* and *Not a White Male*. Employers interested in Humanities & Social Sciences candidates are shown in the top row and employers interested in STEM candidates are shown in the bottom row. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

Internship, our binary variable most predictive of hiring interest, with our demographic variables. These interactions are all directionally negative, and the coefficient *Top Internship* × *Female, White* is negative and significant, suggesting a lower return to a prestigious internships for white females. One possible mechanism for this effect is that employers believe that other employers exhibit positive preferences for diversity, and so having a prestigious internship is a less strong signal of quality if one is from an under-represented group. This aligns with the findings shown in Appendix Figure B.6, which shows that the negative interaction between *Top Internship* and demographics appears for candidates with relatively low ratings and is a fairly precisely estimated zero when candidates receive relatively high ratings.

3.5 Candidate Likelihood of Acceptance

In resume audit studies, traits that suggest high candidate quality do not always increase employer callback. For example, several studies have found that employers call back employed candidates at lower rates than unemployed candidates [Kroft et al., 2013, Nunley et al., 2017, 2014, Farber et al., 2018], but that longer periods of unemployment are unappealing to employers. This seeming contradiction is consistent with the hypothesis that employers are concerned about the possibility of wasting resources pursuing a candidate who will ultimately reject a job offer. In other words, hiring interest is not the only factor determining callback decisions. This concern has been acknowledged in the resume audit literature, for example when Bertrand and Mullainathan [2004, p. 992] notes, "In creating the higher-quality resumes, we deliberately make small changes in credentials so as to minimize the risk of overqualification."

As described in Section 2.4, for each resume we asked employers "How likely do you think **[Name]** would be to accept a job with your organization?" Asking this question helps ensure that our measure of hiring interest is unconfounded with concerns that a candidate would accept a position when offered. However, the question also allows us to study this second factor, which also affects callback decisions.

Table 4 replicates the regression specifications from Table 2, estimating Equation (1) when V_{ij} is *Likelihood of Acceptance*, which takes values from 1 to 10. Employers in our sample view high quality candidates as *more likely* to accept a job with their firm than low quality candidates. This suggests that employers in our sample believe candidate fit at their firm outweighs the possibility that high quality candidates will be pursued by many other firms. In Appendix B.5, we further consider the role of horizontal fit and vertical quality and find that—holding hiring interest in a candidate constant—reported likelihood of acceptance falls as evidence of vertical quality (e.g., GPA) increases. This result highlights that there is independent information in the likelihood of acceptance measure.

Table 4 shows that employers report female and minority candidates are less likely to accept a position with their firm, by 0.2 points on the 1–10 Likert scale (or about one tenth of a standard deviation). This effect is robust to the inclusion of a variety of controls, and it persists when we

hold hiring interest constant in Appendix Table B.9. Table 5 splits the sample and shows that while the direction of these effects is consistent among both groups of employers, the negative effects are particularly large among employers recruiting STEM candidates.

If minority and female applicants are perceived as less likely to accept an offer, this could induce lower callback rates for these candidates. Our results therefore suggest a new channel for discrimination observed in the labor market, which is worth exploring. Perhaps due to the prevalence of diversity initiatives, employers expect that desirable minority and female candidates will receive many offers from competing firms and thus will be less likely to accept any given offer. Alternatively, employers may see female and minority candidates as less likely to fit in the culture of the firm, making these candidates less likely to accept an offer. This result has implications for how we understand the labor market and how we interpret the discrimination observed in resume audit studies.[25]

[25]In particular, while audit studies can demonstrate that groups are not being treated equally, differential callback rates need not imply a lack of employer interest. The impact of candidate characteristics on likelihood of acceptance is a case of omitted variable bias, but one that is not solved by experimental randomization, since the randomized trait endows the candidate with hiring interest and likelihood of acceptance simultaneously.

Table 4: Likelihood of Acceptance

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.605	0.631	0.734	0.263
	(0.144)	(0.150)	(0.120)	(0.060)
Top Internship	0.683	0.677	0.664	0.285
	(0.094)	(0.098)	(0.076)	(0.040)
Second Internship	0.418	0.403	0.394	0.179
	(0.112)	(0.119)	(0.091)	(0.047)
Work for Money	0.197	0.192	0.204	0.088
	(0.111)	(0.116)	(0.090)	(0.047)
Technical Skills	-0.051	-0.059	-0.103	-0.025
	(0.104)	(0.108)	(0.086)	(0.044)
Female, White	-0.231	-0.294	-0.258	-0.093
	(0.114)	(0.118)	(0.094)	(0.048)
Male, Non-White	-0.125	-0.170	-0.117	-0.060
	(0.137)	(0.142)	(0.110)	(0.057)
Female, Non-White	-0.221	-0.236	-0.162	-0.103
	(0.135)	(0.142)	(0.112)	(0.057)
Observations	2880	2880	2880	2880
R^2	0.070	0.124	0.492	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: -0.26, 0.13, 0.49, 0.75, 1.12, 1.49, 1.94, 2.46, and 2.83.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-values of tests of joint significance of major fixed effects are indicated (F-test for OLS, likelihood ratio test for ordered probit).

Table 5: Likelihood of Acceptance by Major Type

	Dependent Variable: Likelihood of Acceptance							
	Humanities & Social Sciences				STEM			
	OLS	OLS	OLS	Ordered Probit	OLS	OLS	OLS	Ordered Probit
GPA	0.581	0.610	0.694	0.251	0.688	0.724	0.813	0.314
	(0.176)	(0.186)	(0.142)	(0.072)	(0.251)	(0.287)	(0.237)	(0.110)
Top Internship	0.786	0.773	0.754	0.316	0.391	0.548	0.527	0.190
	(0.111)	(0.118)	(0.089)	(0.046)	(0.178)	(0.199)	(0.171)	(0.078)
Second Internship	0.481	0.422	0.424	0.201	0.254	0.324	0.301	0.119
	(0.136)	(0.148)	(0.109)	(0.055)	(0.198)	(0.230)	(0.187)	(0.088)
Work for Money	0.206	0.173	0.187	0.084	0.155	0.346	0.350	0.092
	(0.135)	(0.144)	(0.108)	(0.055)	(0.194)	(0.239)	(0.186)	(0.088)
Technical Skills	-0.094	-0.103	-0.106	-0.046	0.050	0.000	-0.116	0.032
	(0.125)	(0.134)	(0.104)	(0.052)	(0.190)	(0.217)	(0.179)	(0.083)
Female, White	-0.175	-0.211	-0.170	-0.062	-0.365	-0.572	-0.577	-0.177
	(0.139)	(0.148)	(0.116)	(0.056)	(0.198)	(0.236)	(0.194)	(0.089)
Male, Non-White	-0.069	-0.076	-0.046	-0.030	-0.269	-0.360	-0.289	-0.147
	(0.161)	(0.172)	(0.130)	(0.066)	(0.259)	(0.302)	(0.246)	(0.110)
Female, Non-White	-0.244	-0.212	-0.163	-0.107	-0.200	-0.108	-0.010	-0.105
	(0.162)	(0.175)	(0.130)	(0.068)	(0.243)	(0.278)	(0.245)	(0.110)
Observations	2040	2040	2040	2040	840	840	840	840
R^2	0.040	0.107	0.516		0.090	0.295	0.540	
p-value for test of joint significance of Majors	0.798	0.939	0.785	0.598	< 0.001	0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No	No	Yes	Yes	No
Subject FEs	No	No	Yes	No	No	No	Yes	No

Ordered probit cutpoints (Column 4): -0.23, 0.14, 0.50, 0.75, 1.11, 1.48, 1.93, 2.42, 2.75.

Ordered probit cutpoints (Column 8): -0.23, 0.20, 0.55, 0.83, 1.25, 1.64, 2.08, 2.71, 3.57.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1). *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. The p-values of tests of joint significance of major fixed effects are indicated (F-test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

3.6 Comparing our Demographic Results to Previous Literature

3.6.1 Qualitative comparison

Our results can be compared to those from other studies of employer preferences, with two caveats. First, our measure of the firms' interest in hiring a candidate may not be directly comparable to findings derived from callback rates, which likely combine both hiring interest and likelihood of acceptance into a single binary outcome. Second, our subject population is made up of firms that would be unlikely to respond to cold resumes and thus may have different preferences than the typical firms audited in prior literature.

Resume audit studies have consistently shown lower callback rates for minorities. We see no evidence of lower ratings for minorities on average, but we do see lower ratings of minority male candidates by STEM employers. Results on gender in the resume audit literature have been mixed. In summarizing results from 11 studies conducted between 2005 and 2016, [Baert, 2018] finds four studies with higher callback rates for women, two with lower callback rates, and five studies with no significant difference. None of these studies found discrimination against women in a U.S. setting. This may be due to resume audit studies targeting female-dominated occupations, such as clerical or administrative work. Riach and Rich [2006], which specifically targets male-dominated occupations, shows lower callback rates for women. Outside the labor market, Bohren et al. [2018] and Milkman et al. [2012] found evidence of discrimination against women using audit-type methodologies. We find that firms recruiting STEM candidates give lower ratings to white women, demonstrating the importance of being able to reach new subject pools with IRR. We also find that white women receive a lower return to prestigious internships. This result matches a type of discrimination—lower return to quality—seen in Bertrand and Mullainathan [2004], but we find it for gender rather than race.

We also find that employers believe white women are less likely to accept positions if offered, which could account for discrimination found in the resume audit literature. For example, Quadlin [2018] finds that women with very high GPAs are called back at lower rates than women with lower GPAs, which could potentially arise from a belief these high quality women will be recruited by other firms, rather than from a lack of hiring interest.

3.6.2 Quantitative comparison using GPA as a numeraire

In addition to making qualitative comparisons, we can conduct some back-of-the-envelope calculations to compare the magnitude of our demographic effects to those in previous studies, including Bertrand and Mullainathan [2004]. We conduct these comparisons by taking advantage of the ability—in our study and others—to use GPA as a numeraire.

In studies that randomize GPA, we can divide the observed effect due to race or gender by the effect due to GPA to compare with our GPA-scaled estimates. For example, exploiting the random variation in GPA and gender from Quadlin [2018], we calculate that being female leads

to a decrease in callback equivalent to 0.23 GPA points.[26] Our results (shown in Tables 2 and 3) suggest that being a white female, as compared to a white male, is equivalent to a decrease of 0.073 GPA points overall and 0.290 GPA points among employers recruiting for STEM.

When a study does not vary GPA, we can benchmark the effect of demographic differences on callback to the effect of GPA on counterfactual callback in our study. For example, in Bertrand and Mullainathan [2004], 8% of all resumes receive callbacks, and having a black name decreases callback by 3.2 percentage points. 7.95% of resumes in our study receive a 9 or a 10 rating, suggesting that receiving a 9 or higher is a similar level of selectivity as in Bertrand and Mullainathan [2004]. A linear probability model in our data suggests that each 0.1 GPA point increases counterfactual callback at this threshold by 1.13 percentage points. Thus, the Bertrand and Mullainathan [2004] race effect is equivalent to an increase of 0.28 GPA points in our study.[27] This effect can be compared to our estimate that being a minority male, as compared to a white male, is equivalent to a decrease of 0.077 GPA points overall and 0.270 GPA points among employers recruiting for STEM.

4 Pitt Replication: Results and Lessons

In order to explore whether preferences differed between employers at Penn (an elite, Ivy League school) and other institutions where recruiters might more closely resemble the employers of typical resume audit studies, we reached out to several Pennsylvania schools in hopes of running an IRR replication. We partnered with the University of Pittsburgh (Pitt) Office of Career Development and Placement Assistance to run two experimental rounds during their spring recruiting cycle.[28] Ideally, the comparison between Penn and Pitt would have given us additional insight into the extent to which Penn employers differed from employers traditionally targeted by audit studies.

[26] Quadlin [2018] reports callback rate in four GPA bins. The paper finds callback is lower in the highest GPA bin than the second highest bin, which may be due to concerns about likelihood of acceptance. Looking at the second and third highest bins (avoiding the non-monotonic bin), we see that an increase in GPA from the range [2.84, 3.20] to [3.21, 3.59]—an average increase of 0.38 GPA points—results in a callback rate increase of 3.5 percentage points. Dividing 0.38 by 3.5 suggests that each 0.11 GPA points generates 1 percentage point difference in callback rates. Quadlin [2018] also finds a callback difference of 2.1 percentage points between male (14.0%) and female (11.9%) candidates. Thus, applicant gender has about the same effect as a 0.23 change in GPA.

[27] Bertrand and Mullainathan [2004] also varies quality, but through changing multiple characteristics at once. Using the same method, these changes, which alter callback by 2.29 percentage points, are equivalent to a change of 0.20 GPA points, providing a benchmark for their quality measure in our GPA points.

[28] Unlike at Penn, there is no major fall recruiting season with elite firms at Pitt. We recruited employers in the spring semester only, first in 2017 and again in 2018. The Pitt recruitment email was similar to that used at Penn (Figure A.1), and originated from the Pitt Office of Career Development and Placement Assistance. For the first wave at Pitt we offered webinars, as described in Appendix A.1, but since attendance at these sessions was low, we did not offer them in the second wave. We collected resume components to populate the tool at Pitt from real resumes of graduating Pitt seniors. Rather than collect resumes from clubs, resume books, and campus job postings as we did at Penn, we used the candidate pool of job-seeking seniors both to populate the tool and to suggest matches for employers. This significantly eased the burden of collecting and scraping resumes. At Pitt, majors were linked to either the "Dietrich School of Arts and Sciences" or the "Swanson School of Engineering". Table C.1 lists the majors, associated school, major category, and the probability that the major was drawn. We collected top internships at Pitt by identifying the firms hiring the most Pitt graduates, as at Penn. Top internships at Pitt tended to be less prestigious than the top internships at Penn.

Instead, we learned that we were insufficiently attuned to how recruiting differences between Penn and Pitt employer populations should influence IRR implementation. Specifically, we observed significant attenuation over nearly all candidate characteristics in the Pitt data. Table 6 shows fully controlled OLS regressions highlighting that our effects at Pitt (shown in the second column) are directionally consistent with those at Penn (shown in the first column for reference), but much smaller in size. For example, the coefficient on GPA is one-tenth the size in the Pitt data. We find similar attenuation on nearly all characteristics at Pitt for both *Hiring Interest* and *Likelihood of Acceptance*, in the pooled sample and separated by major type. We find no evidence of Pitt employers responding to candidate demographics. (Appendix C provides details for our experimental implementation at Pitt and Tables C.2, C.3, and C.4 display the full results.)

We suspect the cause of the attenuation at Pitt was our failure to appropriately tailor resumes to meet the needs of Pitt employers who were seeking candidates with specialized skills or backgrounds. A large share of the resumes at Pitt (33.8%) received the lowest possible *Hiring Interest* rating, more than double the share at Penn (15.5.%). Feedback from Pitt employers suggested that they were also less happy with their matches: many respondents complained that the matches lacked a particular skill or major requirement for their open positions.[29] In addition, the importance of a major requirement was reflected on the post-survey data in which 33.7% of Pitt employers indicated that candidate major was among the most important considerations during recruitment, compared to only 15.3% at Penn.

After observing these issues in the first wave of Pitt data collection, we added a new checklist question to the post-tool survey in the second wave: "I would consider candidates for this position with any of the following majors...." This question allowed us both to restrict the match pool for each employer, improving match quality, and to directly assess the extent to which our failure to tailor resumes was attenuating our estimates of candidate characteristics. Table 6 shows that when splitting the data from the second wave based on whether a candidate was in a target major, the effect of GPA is much larger in the target major sample (shown in the fourth column), and that employers do not respond strongly to any of the variables when considering candidates with majors that are not *Target Majors*.

The differential responses depending on whether resumes come from *Target Majors* highlights the importance of tailoring candidate resumes to employers when deploying the IRR methodology. We advertised the survey tool at both Pitt and Penn as being particularly valuable for hiring skilled generalists, and we were ill equipped to measure preferences of employers looking for candidates with very particular qualifications.

This was a limitation in our implementation at Pitt rather than in the IRR methodology itself. That is, one could design an IRR study specifically for employers interested in hiring registered nurses, or employers interested in hiring mobile software developers, or employers interested in

[29] As one example, a firm wrote to us in an email: "We are a Civil Engineering firm, specifically focused on hiring students out of Civil and/or Environmental Engineering programs... there are 0 students in the group of real resumes that you sent over that are Civil Engineering students."

Table 6: Hiring Interest at Penn and Pitt

	Dependent Variable: Hiring Interest			
	Penn	Pitt	Pitt, Wave 2 Non-Target Major	Pitt, Wave 2 Target Major
GPA	2.196	0.265	-0.196	0.938
	(0.129)	(0.113)	(0.240)	(0.268)
Top Internship	0.897	0.222	0.020	0.098
	(0.081)	(0.074)	(0.142)	(0.205)
Second Internship	0.466	0.212	0.095	0.509
	(0.095)	(0.085)	(0.165)	(0.220)
Work for Money	0.154	0.153	0.144	0.378
	(0.091)	(0.081)	(0.164)	(0.210)
Technical Skills	-0.071	0.107	0.125	-0.035
	(0.090)	(0.077)	(0.149)	(0.211)
Female, White	-0.161	0.028	-0.015	-0.151
	(0.096)	(0.084)	(0.180)	(0.212)
Male, Non-White	-0.169	-0.040	0.002	-0.331
	(0.115)	(0.098)	(0.185)	(0.251)
Female, Non-White	0.028	-0.000	0.182	-0.332
	(0.120)	(0.100)	(0.197)	(0.256)
Observations	2880	3440	642	798
R^2	0.483	0.586	0.793	0.596
p-value for test of joint significance of Majors	< 0.001	< 0.001	0.120	0.850
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes	Yes
Order FEs	Yes	Yes	Yes	Yes
Subject FEs	Yes	Yes	Yes	Yes

Table shows OLS regressions of hiring interest from Equation (1). Sample differs in each column as indicated by the column header. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in all specifications. R^2 is indicated for each OLS regression. The p-value of an F-test of joint significance of major fixed effects is indicated for all models.

hiring electrical engineers. Our failure at Pitt was in showing all of these employers resumes with the same underlying components. We recommend that researchers using IRR either target employers that specifically recruit high quality generalists, or construct resumes with appropriate variation within the employers' target areas. For example, if we ran our IRR study again at Pitt, we would ask the *Target Majors* question first and then only generate hypothetical resumes from those majors.

5 Conclusion

This paper introduces a novel methodology, called Incentivized Resume Rating (IRR), to measure employer preferences. The method has employers rate candidate profiles they know to be hypothetical and provides incentives by matching employers to real job seekers based on their reported preferences.

We deploy IRR to study employer preferences for candidates graduating from an Ivy League university. We find that employers highly value both more prestigious work experience the summer before senior year and additional work experience the summer before junior year. We use our rating data to demonstrate that preferences for these characteristics are relatively stable throughout the distribution of candidate quality.

We find no evidence that employers are less interested in female or minority candidates on average, but we find evidence of discrimination among employers recruiting STEM candidates. Moreover, employers report that white female candidates are less likely to accept job offers than their white male counterparts, a novel channel for discrimination. We also find evidence of lower returns to prestigious internships for women and minorities. One possible story that can explain these results is that employers believe other firms have a positive preference for diversity, even though they do not display this preference themselves. It may thus be fruitful to examine in future research whether employers have distorted beliefs about aggregate preferences for diversity, which could harm female and minority candidates in the job market.

Here, we further discuss the benefits and costs of the IRR methodology, highlight lessons learned from our implementation—which point to improvements in the method—and discuss directions for future research.

A key advantage of the IRR methodology is that it avoids the use of deception. Economics experiments generally aspire to be deception-free, but the lack of an incentivized method to elicit employer preferences without deception has made correspondence audits a default tool of choice for labor economists. Audit studies have also expanded into areas where the continued use of deception may be more fraught, since such deception has the potential to alter the preferences or beliefs of subjects.[30] We hope that further development of the IRR method will provide a useful

[30] As prominent examples, researchers have recently audited college professors, requesting in-person meetings [Milkman et al., 2012, 2015], and politicians, requesting information [Butler and Broockman, 2011, Distelhorst and Hou, 2017]. Professors are likely to learn about audit studies *ex post* and may take the existence of such studies as an

alternative for researchers, reducing the need for deceptive field experiments. This would both limit any potential harms of deception—such as to applicants whose profiles may resemble researcher-generated ones—as well as provide a positive externality for researchers whose work requires an audit design (by reducing potential contamination of the subject pool).

A second advantage of the IRR method is that it elicits richer preference information than binary callback decisions.[31] In our implementation, we elicit granular measures of employers' hiring interest and of employers' beliefs about the likelihood of job acceptance. We also see the potential for improvements in preference elicitation by better mapping these metrics into hiring decisions, by collecting additional information from employers, and by raising the stakes, which we discuss below.

The IRR method has other advantages. IRR can access subject populations that are inaccessible with audit or resume audit methods. IRR allows researchers to gather rich data from a single subject—each employer in our implementation rates 40 resumes—which is helpful for power and makes it feasible to identify preferences for characteristics within individual subjects. IRR allows researchers to randomize many candidate characteristics independently and simultaneously, which can be used to explore how employers respond to interactions of candidate characteristics. Finally, IRR allows researchers to collect supplemental data about research subjects, which can be correlated with subject-level preference measures and allows researchers to better understand their pool of employers.

A final advantage of IRR is that it may provide direct benefits to subjects and other participants in the labor market being studied; this advantage stands in stark contrast to using subject time without consent, as is necessary in audit studies. We solicited subject feedback at numerous points throughout the study and heard very few concerns.[32] Instead, many employers reported positive feedback. Positive feedback also came by way of the career services offices at Penn and Pitt, which were in more-direct contact with our employer subjects. Both offices continued the experiment for a second wave of recruitment and expressed interest in making the experiment a permanent feature of their recruiting processes. In our meetings, the career services offices reported seeing value in IRR to improve their matching process and to learn how employers valued student characteristics (e.g., informing the advice they could give to students about pursuing summer work

excuse to ignore emails from students in the future. Audits of politicians' responses to correspondence from putative constituents might distort politicians' beliefs about the priorities of the populations they serve, especially when researchers seek a politician-level audit measure, which requires sending many fake requests to the same politician.

[31] Bertrand and Duflo [2016] argues that the literature has generally not evolved past measuring differences in callback means between groups, and that it has been less successful in illuminating mechanisms driving these differences. That said, there have been some exceptions, like Bartoš et al. [2016], which uses emails containing links to learn more about candidates to show that less attention is allocated to candidates who are discriminated against. Another exception is Bohren et al. [2018], which uses evaluations of answers posted on an online Q&A forum—which are not conflated with concerns about likelihood of acceptance—to test a dynamic model of mistaken discriminatory beliefs.

[32] First, we solicited feedback in an open comments field of the survey itself. Second, we invited participants to contact us with questions or requests for additional matches when we sent the 10 resumes. Third, we ran a follow-up survey in which we asked about hiring outcomes for the recommended matches (unfortunately, we offered no incentive to complete the follow-up survey and so its participation was low).

and leadership experience and how to write their resumes). While we did not solicit feedback from student participants in the study, we received hundreds of resumes from students at each school, suggesting that they valued the prospect of having their resumes sent to employers.[33]

Naturally, IRR also has some limitations. Because the IRR method informs subjects that responses will be used in research, it may lead to experimenter demand effects (see, e.g., de Quidt et al. [2018]). We believe the impact of any experimenter demand effects is likely small, as employers appeared to view our survey tool as a way to identify promising candidates, rather than as being connected to research (see discussion in Section 2). For this reason, though, as well as others highlighted in Section 3.4, IRR may be less well equipped to identify explicit bias than implicit bias. More broadly, we cannot guarantee that employers treat our hypothetical resumes as they would real job candidates. As discussed in the Introduction, however, future work could help validate employer attention in IRR studies.[34] In addition, because the two outcome measures in our study are hypothetical objects rather than stages of the hiring process, in our implementation of IRR we cannot draw a direct link between our findings and hiring outcomes. Below, we discuss how this might be improved in future IRR implementations.

A final cost of running an IRR study is that it requires finding an appropriate subject pool and candidate matching pool, which may not be available to all researchers. It also requires an investment in constructing the hypothetical resumes (e.g., scraping and sanitizing resume components) and developing the process to match employer preferences to candidates. Fortunately, the time and resources we devoted to developing the survey tool software can be leveraged by other researchers.

Future research using IRR can certainly improve upon our implementation. First, as discussed at length in Section 4, our failed attempt to replicate at Pitt highlights that future researchers must take care to effectively tailor the content of resumes to match the hiring needs of their subjects. Second, we suggest developing a way to translate Likert-scale responses to the callback decisions typical in correspondence audit studies. One idea is to ask employers to additionally answer, potentially for a subset of resumes, a question of the form: "Would you invite **[Candidate Name]** for an interview?" By having the Likert-scale responses and this measure, researchers could identify what combination of the hiring interest and likelihood of acceptance responses translates into a typical callback decision (and, potentially, how the weight placed on each component varies by firm). Researchers could also explore the origin and accuracy of employer beliefs about likelihood of acceptance by asking job candidates about their willingness to work at participating firms. Third, researchers could increase the stakes of IRR incentives (e.g., by asking employer subjects

[33]Student involvement only required uploading a resume and completing a short preference survey. We did not notify students when they were matched with a firm, in order to give the firms freedom to choose which students to contact. Thus, most students were unaware of whether or not they were recommended to a firm. We recommended 207 unique student resumes over the course of the study, highlighting the value to students.

[34]The time employers spent evaluating resumes in our study at Penn had a median of 18 seconds and a mean that was substantially higher (and varies based on how outliers are handled). These measures are comparable to estimates of time spent screening real resumes (which include estimates of 7.4 seconds per resume [Dishman, 2018] and a mean of 45 seconds per resume [Culwell-Block and Sellers, 1994]).

to guarantee interviews to a subset of the recommended candidates) and gather more information on resulting interviews and hiring outcomes (e.g., by building or leveraging an existing platform to measure employer and candidate interactions).[35]

While we used IRR to measure the preferences of employers in a particular labor market, the underlying incentive structure of the IRR method is much more general, and we see the possibility of it being applied outside of the resume rating context. At the heart of IRR is a method to elicit preference information from experimental subjects by having them evaluate hypothetical objects and offering them an incentive that increases in value as preference reports become more accurate. Our implementation of IRR achieves this by eliciting continuous Likert-scale measures of hypothetical resumes, using machine learning to estimate the extent to which employers care about various candidate characteristics, and providing employers with resumes of real candidates that they are estimated to like best. Researchers could take a similar strategy to explore preferences of professors over prospective students, landlords over tenants, customers over products, individuals over dating profiles, and more, providing a powerful alternative to deceptive studies.

[35] An additional benefit of collecting data on interviews and hiring is that it would allow researchers to measure the impact of IRR matches on hiring, in addition to validating the quality of the matches (e.g., researchers could identify 12 potential matches and randomize which 10 are sent to employers, identifying the effect of sending a resume to employers on interview and hiring outcomes). If employers do respond to the matches, one could imagine using IRR as an intervention in labor markets to help mitigate discrimination in hiring, since IRR matches can be made while ignoring race and gender.

References

Joseph G Altonji and Rebecca M Blank. Race and gender in the labor market. *Handbook of Labor Economics*, 3:3143–3259, 1999.

David H Autor and Susan N Houseman. Do temporary-help jobs improve labor market outcomes for low-skilled workers? Evidence from "Work First". *American Economic Journal: Applied Economics*, pages 96–128, 2010.

Stijn Baert. Chapter 3: Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. In Michael S. Gaddis, editor, *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, chapter 3, pages 63–77. Springer, 2018.

Vojtěch Bartoš, Michal Bauer, Julie Chytilová, and Filip Matějka. Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–75, June 2016. doi: 10.1257/aer.20140571.

Marianne Bertrand and Esther Duflo. Field experiments on discrimination. NBER Working Papers 22014, National Bureau of Economic Research, Inc, Feb 2016.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? *The American Economic Review*, 94(4):991–1013, 2004.

Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan. Implicit discrimination. *American Economic Review*, 95(2):94–98, 2005.

Iris Bohnet, Alexandra Van Geen, and Max Bazerman. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5):1225–1234, 2015.

J. Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American Economic Review (Forthcoming)*, 2018.

Daniel M. Butler and David E. Broockman. Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*, 55 (3):463–477, 2011.

Scott E. Carrell, Marianne E. Page, and James E. West. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144, 2010.

Beverly Culwell-Block and Jean Anna Sellers. Resume content and format - do the authorities agree?, Dec 1994. URL https://www.questia.com/library/journal/1G1-16572126/resume-content-and-format-do-the-authorities-agree.

Rajeev Darolia, Cory Koedel, Paco Martorell, Katie Wilson, and Francisco Perez-Arce. Do employers prefer workers who attend for-profit colleges? Evidence from a field experiment. *Journal of Policy Analysis and Management*, 34(4):881–903, 2015.

Jonathan de Quidt, Johannes Haushofer, and Christopher Roth. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, November 2018. doi: 10.1257/aer.20171330. URL http://www.aeaweb.org/articles?id=10.1257/aer.20171330.

David J. Deming, Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz. The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review*, 106(3):778–806, March 2016.

Lydia Dishman. Your resume only gets 7.4 seconds to make an impression–here's how to stand out, Nov 2018. URL https://www.fastcompany.com/90263970/your-resume-only-gets-7-4-seconds-to-make-an-impression-heres-how-to-stand-out.

Greg Distelhorst and Yue Hou. Constituency service under nondemocratic rule: Evidence from China. *The Journal of Politics*, 79(3):1024–1040, 2017. doi: 10.1086/690948.

Stefan Eriksson and Dan-Olof Rooth. Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. *American Economic Review*, 104(3):1014–39, March 2014. doi: 10.1257/aer.104.3.1014.

Michael Ewens, Bryan Tomlin, and Liang Choon Wang. Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, 96(1):119–134, 2014.

Henry S Farber, Chris M Herbst, Dan Silverman, and Till von Wachter. Whom do employers want? The role of recent employment and unemployment status and age. Working Paper 24605, National Bureau of Economic Research, May 2018.

Roland G. Fryer and Steven D. Levitt. The causes and consequences of distinctively black names. *Quarterly Journal of Economics*, 119:767–805, 2004.

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

Claudia Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104 (4):1091–1119, 04 2014.

Olesya Govorun and B Keith Payne. Ego–depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, 24(2):111–136, 2006.

William Greene. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, 7(1):98–119, 2004.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Daniel Hamermesh. Are fake resumes ethical for academic research? *Freakonomics Blog*, 2012.

Andrew Hanson and Zackary Hawley. Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics*, 70(2-3): 99–114, 2011.

Glenn W. Harrison and John A. List. Field experiments. *Journal of Economic Literature*, 42(4): 1009–1055, December 2004. doi: 10.1257/0022051043004577.

James J. Heckman. Detecting discrimination. *Journal of Economic Perspectives*, 12(2):101–116, 1998.

James J. Heckman and Peter Siegelman. The Urban Institute audit studies: their methods and findings. In *Clear and convincing evidence: measurement of discrimination in America*, pages 187–258. Lanhan, MD: Urban Institute Press, 1992.

Lars J Kirkeboen, Edwin Leuven, and Magne Mogstad. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111, 2016.

Meredith Kleykamp. A great place to start?: The effect of prior military service on hiring. *Armed Forces & Society*, 35(2):266–285, 2009. doi: 10.1177/0095327X07308631.

Kory Kroft, Fabian Lange, and Matthew J. Notowidigdo. Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics*, 128(3): 1123–1167, 2013. doi: 10.1093/qje/qjt015.

Kevin Lang and Jee-Yeon K Lehmann. Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, 50(4):959–1006, 2012.

Corinne Low. A "reproductive capital" model of marriage market matching. *Manuscript, Wharton School of Business*, 2017.

Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. Temporal distance and discrimination: an audit study in academia. *Psychological Science*, 23(7):710–717, 2012.

Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, 100(6), 2015.

David Neumark. Detecting discrimination in audit and correspondence studies. *Journal of Human Resources*, 47(4):1128–1157, 2012.

David Neumark, Ian Burn, and Patrick Button. Is it harder for older workers to find jobs? New and improved evidence from a field experiment. Technical report, National Bureau of Economic Research, 2015.

Brian A Nosek, Frederick L Smyth, Jeffrey J Hansen, Thierry Devos, Nicole M Lindner, Kate A Ranganath, Colin Tucker Smith, Kristina R Olson, Dolly Chugh, Anthony G Greenwald, et al. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1):36–88, 2007.

John M Nunley, Adam Pugh, Nicholas Romero, and R Alan Seals. Unemployment, underemployment, and employment opportunities: Results from a correspondence audit of the labor market for college graduates. *Auburn University Department of Economics Working Paper Series*, 4, 2014.

John M. Nunley, Adam Pugh, Nicholas Romero, and R. Alan Seals. The effects of unemployment and underemployment on employment opportunities: Results from a correspondence audit of the labor market for college graduates. *ILR Review*, 70(3):642–669, 2017.

Andreas Ortmann and Ralph Hertwig. The costs of deception: Evidence from psychology. *Experimental Economics*, 5(2):111–131, 2002.

Amanda Pallais. Inefficient hiring in entry-level labor markets. *American Economic Review*, 104 (11):3565–99, 2014.

Devin G Pope and Justin R Sydnor. What's in a picture? Evidence of discrimination from prosper.com. *Journal of Human resources*, 46(1):53–92, 2011.

Natasha Quadlin. The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review*, 83(2):331–360, 2018. doi: 10.1177/0003122418762291. URL https://doi.org/10.1177/0003122418762291.

Peter A Riach and Judith Rich. An experimental investigation of sexual discrimination in hiring in the English labor market. *Advances in Economic Analysis & Policy*, 5(2), 2006.

Lauren A. Rivera and András Tilcsik. Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review*, 81(6):1097–1131, 2016. doi: 10.1177/0003122416668154. URL https://doi.org/10.1177/0003122416668154.

Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

Jeffrey W Sherman, Frederica R Conrey, and Carla J Groom. Encoding flexibility revisited: Evidence for enhanced encoding of stereotype-inconsistent information under cognitive load. *Social Cognition*, 22(2):214–232, 2004.

Margery Turner, Michael Fix, and Raymond J. Struyk. Opportunities denied, opportunities, diminished: racial discrimination in hiring. *Washington, DC: Urban Institute Press*, 1991.

Daniël HJ Wigboldus, Jeffrey W Sherman, Heather L Franzese, and Ad van Knippenberg. Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition*, 22(3):292–309, 2004.

Asaf Zussman. Ethnic discrimination: Lessons from the Israeli online market for used cars. *The Economic Journal*, 123(572):F433–F468, 2013.