

Incentivized Resume Rating: Eliciting Employer Preferences without Deception[†]

By JUDD B. KESSLER, CORINNE LOW, AND COLIN D. SULLIVAN*

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 that candidates will accept job offers, avoiding a typical confound in audit studies. We discuss the costs, benefits, and future applications of this new methodology. (JEL D83, I26, J23, J24, M51)

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 policymakers 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, Tomlin, and Wang 2014), retail (e.g., Pope

* Kessler: Wharton School, University of Pennsylvania, 3733 Spruce Street, Vance Hall, Philadelphia, PA 19104 (email: judd.kessler@wharton.upenn.edu); Low: Wharton School, University of Pennsylvania, 3733 Spruce Street, Vance Hall, Philadelphia, PA 19104 (email: corlow@wharton.upenn.edu); Sullivan: Stanford University, 579 Serra Mall, Landau Economics Building, Stanford, CA 94305 (email: [cgsulliv@stanford.edu](mailto:cdsulliv@stanford.edu)). Stefano DellaVigna was the coeditor for this article. 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.

[†] Go to <https://doi.org/10.1257/aer.20181714> to visit the article page for additional materials and author disclosure statements.

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, Fix, and Struyk 1991) and for an inability to match candidate characteristics beyond race perfectly (Heckman and Siegelman 1992, Heckman 1998).

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, Lange, and Notowidigdo 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, 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 nonexistent 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 is currently 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 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, Lange, and Notowidigdo 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

²Baert (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).

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

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: (i) how interested they would be in hiring the candidate and (ii) 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 that 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.

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 toward 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 Employers in our study report having a

⁴We 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.

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 III, 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 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 IV.

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 I describes in detail how we implement our IRR study; Section II reports on the results from Penn and compares them to extant literature; Section III describes our attempted replication at Pitt; and Section IV concludes.

I. 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 IA, we outline how we recruit employers who are in the market to hire elite college graduates. In Section IB, we describe how we provide employers with incentives for reporting preferences without introducing deception. In Section IC, 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, previous work experience, skills, and race and gender.

⁵In a survey that employers complete after evaluating resumes in our study, over 90 percent 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 percent 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.”

In Section ID, 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.

A. *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 online 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 online 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 IB), 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.

B. *Incentives*

The main innovation of IRR is its method for incentivized preference elicitation, a variant of a method pioneered by Low (2019) 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, selected based on the employer’s evaluations.⁷ Consequently, the

⁶Indeed, 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.

⁷The recruitment email (see online 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

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.

C. 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, 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 students at Penn. Second, we asked the employers to choose the type of candidates that they were interested in hiring, based on major (see online Appendix Figure A.4). In particular, they could choose either "Business (Wharton), Social Sciences, and Humanities" (henceforth, "Humanities and 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 ten different resume templates, which we populated with the candidate characteristics and component pieces described below, to generate the 40 hypothetical resumes (see online

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 online Appendix A.3.

⁸In Low (2019), 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 nonmonetary 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.

⁹In 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).

TABLE 1—RANDOMIZATION OF RESUME COMPONENTS

Resume component	Description	Analysis variable
<i>Personal information</i>		
First and last name	Drawn from list of 50 possible names given selected race and gender (names in Tables A.1 and A.2) Race drawn randomly from US 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%)
<i>Education information</i>		
GPA	Drawn <i>Unif</i> [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)	
<i>Work experience</i>		
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)	
<i>Leadership experience</i>		
First and 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	
<i>Skills</i>		
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%)

Notes: Resume components are listed in the order that they appear on hypothetical resumes. Variables in the right-hand column were randomized to test how employers responded to these characteristics. Degree, first job, second job, and skills were drawn from different lists for Humanities and 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 percent chance of being assigned a white female name).

Appendix Figure A.5 for a sample resume). We based these templates on real student resume formats (see online 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 ten resumes they evaluated. As described in Section IID and online Appendix B.4, we use the change in attention induced by these breaks to construct tests of implicit bias.

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.

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”). Online Appendix Table A.3 shows the list of majors by major category, school, and the probability that the major was used in a resume.

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 online 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 and to gender). 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 had an internship only 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: (i) “top internships,” which were internships with prestigious firms as defined by being a firm that successfully hires many Penn graduates; (ii) “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 (iii) “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

¹¹ See online 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 online 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 and 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 and 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.

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 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.

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 online 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 percent 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}.

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 online Appendix Tables A.1 and A.2 (see online Appendix A.2.3 for additional details).

For realism, we randomly selected races at rates approximating the distribution in the US population (65.7 percent white, 16.8 percent Hispanic, 12.6 percent black, 4.9 percent Asian). While a more uniform variation in race would have increased statistical power to detect race-based discrimination, such an approach would have

¹³For ease of exposition, we will refer to race/ethnicity as “race” throughout the paper.

¹⁴For 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 percent of that race’s population, to ensure that the last names were sufficiently common.

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.

D. 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 candidates of higher or lower quality. 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 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 nonmonotonic 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 online Appendix Figure A.5):

- (i) "How interested would you be in hiring [Name]?"
(1 = "Not interested"; 10 = "Very interested")
- (ii) "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 online 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 the candidate had been given an offer and to assess the likelihood the candidate would accept it). We denote responses to this question a candidate's "likelihood of

¹⁵Note that audit and resume audit studies focusing on discrimination may not need to interpret callback as a measure of an employer's interest in a candidate as any difference in callback rates is evidence of unequal treatment.

acceptance.” We asked the first question to assess how resume characteristics affect hiring interest. We asked the second question both to encourage employers 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 infra-marginal 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 IIC.

II. Results

A. Data and Empirical Approach

We recruited 72 employers through our partnership with the University of Pennsylvania Career Services office in Fall 2016 (46 subjects, 1,840 resume observations) and Spring 2017 (26 subjects, 1,040 resume observations).¹⁶

As described in Section I, 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 IIE. 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.

¹⁶The recruiters who participated in our study as subjects were primarily female (59 percent) and primarily white (79 percent) and Asian (15 percent). 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 percent); between two and five years (46 percent); and six or more years (25 percent). Almost all (96 percent) of the participants had college degrees, and many (30 percent) had graduate degrees including an MA, MBA, JD, or doctorate. They were approximately as likely to work at a large firm with over 1,000 employees (35 percent) as a small firm with fewer than 100 employees (39 percent). 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 percent); the technology sector or computer science (18 percent); and consulting (16 percent). The sample had a smaller number of sales/marketing firms (9 percent) and nonprofit or public interest organizations (9 percent). The vast majority (86 percent) 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 percent), Midwest (18 percent), South (16 percent), or an international location (10 percent).

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

B. 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 IC 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. Formally,

$$(1) \quad V_{ij} = \beta_0 + \beta_1 GPA + \beta_2 Top\ Internship + \beta_3 Second\ Internship \\ + \beta_4 Work\ for\ Money + \beta_5 Technical\ Skills + \beta_6 Female,\ White \\ + \beta_7 Male,\ Non-White + \beta_8 Female,\ Non-White \\ + \mu_j + \gamma_j + \omega_j + \alpha_i + \varepsilon_{ij}.$$

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 Variables 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. Variables γ_j are dummies for each of the leadership experience components. Variables ω_j are dummies for the number of

¹⁷Coefficient 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 IID and online Appendix B.4.

resumes the employer has evaluated so far 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 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 standard deviations.

As described in Section IC, 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 one-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

¹⁸These 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: <i>Hiring Interest</i>				
	OLS	OLS	OLS	GPA-scaled OLS	Ordered probit
<i>GPA</i>	2.125 (0.145)	2.190 (0.150)	2.196 (0.129)	1.000 (·)	0.891 (0.063)
<i>Top Internship</i>	0.902 (0.095)	0.900 (0.099)	0.897 (0.081)	0.409 (0.043)	0.378 (0.040)
<i>Second Internship</i>	0.465 (0.112)	0.490 (0.118)	0.466 (0.095)	0.212 (0.045)	0.206 (0.047)
<i>Work for Money</i>	0.116 (0.110)	0.157 (0.113)	0.154 (0.091)	0.070 (0.042)	0.052 (0.046)
<i>Technical Skills</i>	0.046 (0.104)	0.053 (0.108)	-0.071 (0.090)	-0.032 (0.041)	0.012 (0.043)
<i>Female, White</i>	-0.152 (0.114)	-0.215 (0.118)	-0.161 (0.096)	-0.073 (0.044)	-0.061 (0.048)
<i>Male, Non-White</i>	-0.172 (0.136)	-0.177 (0.142)	-0.169 (0.115)	-0.077 (0.053)	-0.075 (0.058)
<i>Female, Non-White</i>	-0.009 (0.137)	-0.022 (0.144)	0.028 (0.120)	0.013 (0.055)	-0.014 (0.057)
Observations	2,880	2,880	2,880	2,880	2,880
R^2	0.129	0.181	0.483		
<i>p</i> -value for test of joint significance of majors	<0.001	<0.001	<0.001	<0.001	<0.001
Major fixed effects	Yes	Yes	Yes	Yes	Yes
Leadership fixed effects	No	Yes	Yes	Yes	No
Order fixed effects	No	Yes	Yes	Yes	No
Subject fixed effects	No	No	Yes	Yes	No

Notes: 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 IIC and in online 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).

mastery to affect hiring interest (e.g., employers may consider skills listed on a resume to be cheap talk).

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 percent of the humanities and 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 percent 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 online 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 I, 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 online 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 online Appendix B.1 and online Appendix Table B.1, we describe how we construct a measure of work experience narrative and test its importance. We 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 online Appendix B.2, we reweight 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 reweighting. These exercises provide some assurance that our results are not an artifact of how we construct hypothetical resumes.

C. *Effects across the Distribution of Hiring Interest*

The regression specifications described in Section IIB 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

¹⁹Note 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, Leuven, and Mogstad 2016).

²⁰The 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).

of preferences.²¹ To make this concern concrete, online 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: (i) the empirical cumulative distribution function (CDF) of hiring interest ratings and (ii) 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.

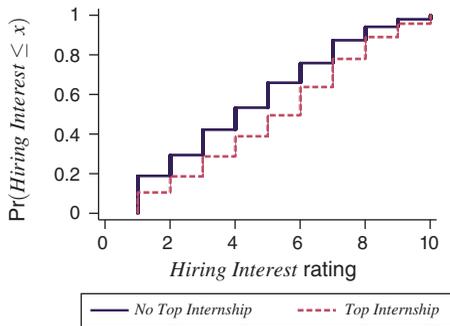
While the theoretical concerns raised by Neumark (2012) may be relevant in other settings, the average results we find in Section IIB are all consistent across the distribution of hiring interest, including in the tails (except for a preference for Wharton students, which we discuss below). The top half 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 the distribution for *Second Internship* and *Work for Money* are also consistent with the average results from Section IIB. *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 IIB, 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

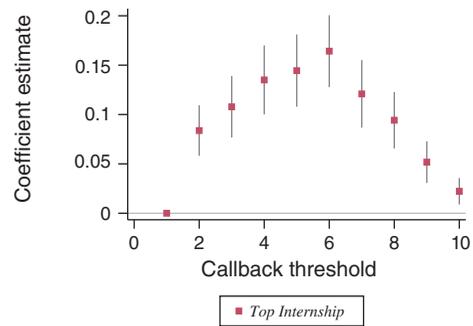
²¹ A variant of this critique was initially brought up by Heckman and Siegelman (1992) and Heckman (1998) for in-person audit studies, where auditors may be imperfectly matched, and was extended to correspondence audit studies by Neumark (2012) and Neumark, Burn, and Button (2015). A key feature of the 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 percent of candidates with a desirable characteristic are being called back, the difference in callback rates from removing that characteristic can be no more than 5 percentage points. At lower thresholds (e.g., where 50 percent of candidates with desirable characteristics are called back), differences in callback rates can be much larger. In online 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.

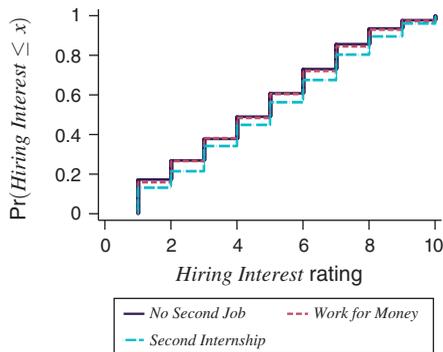
Panel A. Empirical CDF for *Top Internship*



Panel B. Linear probability model for *Top Internship*



Panel C. Empirical CDF for second job type



Panel D. Linear probability model for second job type

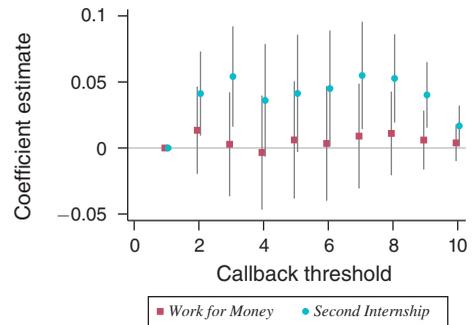


FIGURE 1. VALUE OF QUALITY OF EXPERIENCE OVER SELECTIVITY DISTRIBUTION

Notes: Empirical CDF of *Hiring Interest* (panels A and C) and difference in counterfactual callback rates (panels B and D) 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. Ninety-five percent confidence intervals in panels B and D are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

a well-powered audit study would be unlikely to find a result, even though we find one on average. Online Appendix Figure B.1 mirrors Figure 1 but focuses on having a Wharton degree among employers seeking humanities and 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.

D. Demographic Discrimination

In this section, we examine how hiring interest ratings respond to the race and gender of candidates. As described in Section I 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, van Geen, and Bazerman 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 employers in our sample 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, Page, and West 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 panel B 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 2 to correct for the fact that we are 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 panel A show no evidence of discrimination in hiring interest among humanities and social sciences employers.

As in Section IIC, 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 and social sciences candidates and the bottom row shows these comparisons for employers interested in STEM candidates. Among employers interested in humanities and 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, McGhee, and Schwartz 1998; Nosek et al. 2007), which Bertrand, Chugh, and Mullainathan (2005) has suggested is an important channel for discrimination in

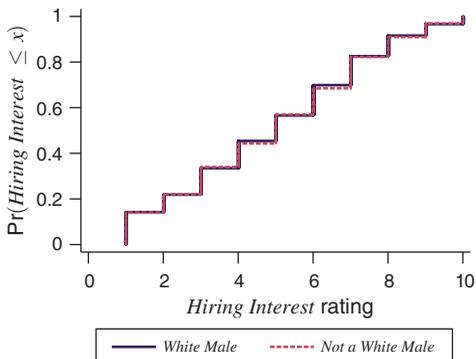
²³In online Appendix Table B.6, we show that this effect does not differ by the gender and race of the employer rating the resume.

TABLE 3—HIRING INTEREST BY MAJOR TYPE

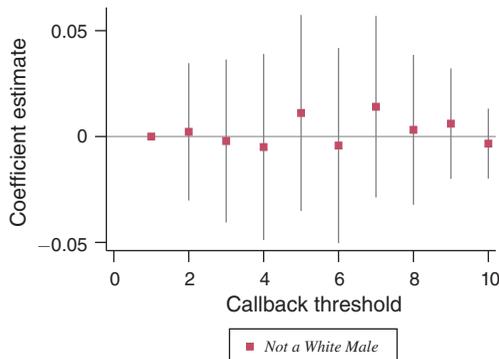
	Dependent variable: <i>Hiring Interest</i>				
	OLS	OLS	OLS	GPA-scaled OLS	Ordered probit
<i>Panel A. Humanities and Social Sciences</i>					
<i>GPA</i>	2.208 (0.173)	2.304 (0.179)	2.296 (0.153)	1.000 (·)	0.933 (0.074)
<i>Top Internship</i>	1.075 (0.108)	1.043 (0.116)	1.033 (0.095)	0.450 (0.050)	0.452 (0.046)
<i>Second Internship</i>	0.540 (0.132)	0.516 (0.143)	0.513 (0.114)	0.224 (0.051)	0.240 (0.056)
<i>Work for Money</i>	0.087 (0.129)	0.107 (0.134)	0.116 (0.110)	0.050 (0.048)	0.037 (0.055)
<i>Technical Skills</i>	0.063 (0.122)	0.084 (0.130)	-0.050 (0.106)	-0.022 (0.046)	0.013 (0.052)
<i>Female, White</i>	-0.047 (0.134)	-0.117 (0.142)	-0.054 (0.117)	-0.024 (0.051)	-0.015 (0.057)
<i>Male, Non-White</i>	-0.029 (0.158)	-0.010 (0.169)	-0.026 (0.137)	-0.011 (0.059)	-0.007 (0.066)
<i>Female, Non-White</i>	0.085 (0.160)	0.101 (0.171)	0.091 (0.137)	0.040 (0.060)	0.024 (0.068)
Observations	2,040	2,040	2,040	2,040	2,040
R^2	0.128	0.196	0.500		
p -value for test of joint significance of majors	0.021	0.027	0.007	0.007	0.030
Major fixed effects	Yes	Yes	Yes	Yes	Yes
Leadership fixed effects	No	Yes	Yes	Yes	No
Order fixed effects	No	Yes	Yes	Yes	No
Subject fixed effects	No	No	Yes	Yes	No
<i>Panel B. STEM</i>					
<i>GPA</i>	1.932 (0.267)	1.885 (0.309)	1.882 (0.242)	1.000 (·)	0.802 (0.112)
<i>Top Internship</i>	0.398 (0.191)	0.559 (0.216)	0.545 (0.173)	0.289 (0.100)	0.175 (0.078)
<i>Second Internship</i>	0.242 (0.208)	0.307 (0.246)	0.311 (0.189)	0.165 (0.103)	0.111 (0.088)
<i>Work for Money</i>	0.151 (0.212)	0.275 (0.254)	0.337 (0.187)	0.179 (0.102)	0.076 (0.088)
<i>Technical Skills</i>	-0.028 (0.197)	-0.113 (0.228)	-0.180 (0.186)	-0.096 (0.100)	-0.001 (0.083)
<i>Female, White</i>	-0.419 (0.215)	-0.612 (0.249)	-0.545 (0.208)	-0.290 (0.115)	-0.171 (0.089)
<i>Male, Non-White</i>	-0.567 (0.271)	-0.617 (0.318)	-0.507 (0.257)	-0.270 (0.136)	-0.265 (0.111)
<i>Female, Non-White</i>	-0.329 (0.264)	-0.260 (0.301)	-0.046 (0.261)	-0.025 (0.138)	-0.142 (0.111)
Observations	840	840	840	840	840
R^2	0.119	0.323	0.593		
p -value for test of joint significance of majors	<0.001	0.035	<0.001	<0.001	<0.001
Major fixed effects	Yes	Yes	Yes	Yes	Yes
Leadership fixed effects	No	Yes	Yes	Yes	No
Order fixed effects	No	Yes	Yes	Yes	No
Subject fixed effects	No	No	Yes	Yes	No

Notes: Panel A: Ordered probit cutpoints: 2.25, 2.58, 2.96, 3.26, 3.60, 3.94, 4.41, 4.86, 5.41. Panel B: Ordered probit cutpoints: 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 IC and in online 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 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) after a Bonferroni correction for analyzing two subgroups.

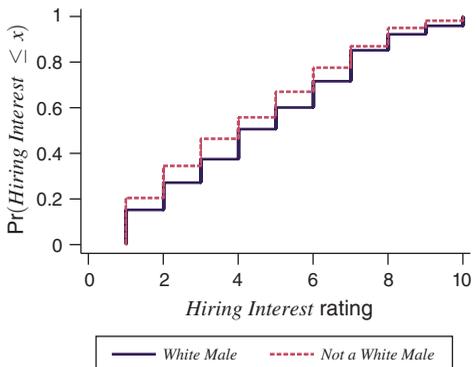
Panel A. Empirical CDF for *Not a White Male*, Humanities and Social Sciences



Panel B. Linear probability model for *Not a White Male*, Humanities and Social Sciences



Panel C. Empirical CDF for *Not a White Male*, STEM



Panel D. Linear probability model for *Not a White Male*, STEM

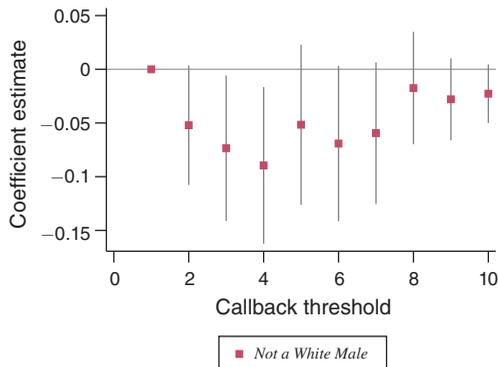


FIGURE 2. DEMOGRAPHICS BY MAJOR TYPE OVER SELECTIVITY DISTRIBUTION

Notes: Empirical CDF of *Hiring Interest* (panels A and C) and difference in counterfactual callback rates (panels B and D) for *White Male* and *Not a White Male*. Employers interested in Humanities and 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. Ninety-five percent confidence intervals in panels B and D are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

resume audit studies. In online 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, Conrey, and Groom 2004), to test whether our data are consistent with employers displaying an implicit racial or gender bias. As shown in online Appendix Table B.7, employers spend less time evaluating resumes

²⁴Explicit 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 the firm, even conditional on their resumes. Implicit bias (Greenwald, McGhee, and Schwartz 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).

both in the latter half of the study and in the latter half of each set of ten resumes (after each set of ten 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 ten 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)). Online Appendix Table B.8 interacts *Top Internship*, our binary variable most predictive of hiring interest, with our demographic variables. These interactions are all directionally negative, and the coefficient *Top Internship* \times *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 online 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.

E. 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, Lange, and Notowidigdo 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 ID, 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 online Appendix B.5, we further

TABLE 4—LIKELIHOOD OF ACCEPTANCE

	Dependent variable: <i>Likelihood of Acceptance</i>			
	OLS	OLS	OLS	Ordered probit
<i>GPA</i>	0.605 (0.144)	0.631 (0.150)	0.734 (0.120)	0.263 (0.060)
<i>Top Internship</i>	0.683 (0.094)	0.677 (0.098)	0.664 (0.076)	0.285 (0.040)
<i>Second Internship</i>	0.418 (0.112)	0.403 (0.119)	0.394 (0.091)	0.179 (0.047)
<i>Work for Money</i>	0.197 (0.111)	0.192 (0.116)	0.204 (0.090)	0.088 (0.047)
<i>Technical Skills</i>	-0.051 (0.104)	-0.059 (0.108)	-0.103 (0.086)	-0.025 (0.044)
<i>Female, White</i>	-0.231 (0.114)	-0.294 (0.118)	-0.258 (0.094)	-0.093 (0.048)
<i>Male, Non-White</i>	-0.125 (0.137)	-0.170 (0.142)	-0.117 (0.110)	-0.060 (0.057)
<i>Female, Non-White</i>	-0.221 (0.135)	-0.236 (0.142)	-0.162 (0.112)	-0.103 (0.057)
Observations	2,880	2,880	2,880	2,880
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 fixed effects	Yes	Yes	Yes	Yes
Leadership fixed effects	No	Yes	Yes	No
Order fixed effects	No	Yes	Yes	No
Subject fixed effects	No	No	Yes	No

Notes: 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 IC and in online 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).

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 online 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,

TABLE 5—LIKELIHOOD OF ACCEPTANCE BY MAJOR TYPE

	Dependent variable: <i>Likelihood of Acceptance</i>							
	Humanities and Social Sciences				STEM			
	OLS	OLS	OLS	Ordered probit	OLS	OLS	OLS	Ordered probit
<i>GPA</i>	0.581 (0.176)	0.610 (0.186)	0.694 (0.142)	0.251 (0.072)	0.688 (0.251)	0.724 (0.287)	0.813 (0.237)	0.314 (0.110)
<i>Top Internship</i>	0.786 (0.111)	0.773 (0.118)	0.754 (0.089)	0.316 (0.046)	0.391 (0.178)	0.548 (0.199)	0.527 (0.171)	0.190 (0.078)
<i>Second Internship</i>	0.481 (0.136)	0.422 (0.148)	0.424 (0.109)	0.201 (0.055)	0.254 (0.198)	0.324 (0.230)	0.301 (0.187)	0.119 (0.088)
<i>Work for Money</i>	0.206 (0.135)	0.173 (0.144)	0.187 (0.108)	0.084 (0.055)	0.155 (0.194)	0.346 (0.239)	0.350 (0.186)	0.092 (0.088)
<i>Technical Skills</i>	-0.094 (0.125)	-0.103 (0.134)	-0.106 (0.104)	-0.046 (0.052)	0.050 (0.190)	0.000 (0.217)	-0.116 (0.179)	0.032 (0.083)
<i>Female, White</i>	-0.175 (0.139)	-0.211 (0.148)	-0.170 (0.116)	-0.062 (0.056)	-0.365 (0.198)	-0.572 (0.236)	-0.577 (0.194)	-0.177 (0.089)
<i>Male, Non-White</i>	-0.069 (0.161)	-0.076 (0.172)	-0.046 (0.130)	-0.030 (0.066)	-0.269 (0.259)	-0.360 (0.302)	-0.289 (0.246)	-0.147 (0.110)
<i>Female, Non-White</i>	-0.244 (0.162)	-0.212 (0.175)	-0.163 (0.130)	-0.107 (0.068)	-0.200 (0.243)	-0.108 (0.278)	-0.010 (0.245)	-0.105 (0.110)
Observations	2,040	2,040	2,040	2,040	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 fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership fixed effects	No	Yes	Yes	No	No	Yes	Yes	No
Order fixed effects	No	Yes	Yes	No	No	Yes	Yes	No
Subject fixed effects	No	No	Yes	No	No	No	Yes	No

Notes: 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 IC and in online 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.

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

F. Comparing our Demographic Results to Previous Literature

Qualitative Comparison.—Our results can be compared to those from other studies of employer preferences, with two caveats. First, our measure of the firms'

²⁵In 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.

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 US 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, Imas, and Rosenberg (2019) and Milkman, Akinola, and Chugh (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.

Quantitative Comparison Using GPA as a Numéraire.—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 numéraire.

In studies that randomize GPA, we can divide the observed effect of race or gender by the effect of GPA to compare with our GPA-scaled estimates. For example, using the variation in callback by 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

²⁶Quadlin (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 nonmonotonic 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 a 1 percentage point difference in callback rates. Quadlin (2018) also finds a callback difference of 2.1 percentage points between male (14.0 percent) and female (11.9 percent) candidates. Thus, applicant gender has about the same effect as a 0.23 change in GPA.

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 percent of all resumes receive callbacks, and having a black name decreases callback by 3.2 percentage points; 7.95 percent 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 points 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.

III. 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 in 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.

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

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 online 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. Online Appendix 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.

TABLE 6—HIRING INTEREST AT PENN AND PITT

	Dependent variable: <i>Hiring Interest</i>			
	Penn	Pitt	Pitt, wave 2 non-target major	Pitt, wave 2 target major
<i>GPA</i>	2.196 (0.129)	0.265 (0.113)	-0.196 (0.240)	0.938 (0.268)
<i>Top Internship</i>	0.897 (0.081)	0.222 (0.074)	0.020 (0.142)	0.098 (0.205)
<i>Second Internship</i>	0.466 (0.095)	0.212 (0.085)	0.095 (0.165)	0.509 (0.220)
<i>Work for Money</i>	0.154 (0.091)	0.153 (0.081)	0.144 (0.164)	0.378 (0.210)
<i>Technical Skills</i>	-0.071 (0.090)	0.107 (0.077)	0.125 (0.149)	-0.035 (0.211)
<i>Female, White</i>	-0.161 (0.096)	0.028 (0.084)	-0.015 (0.180)	-0.151 (0.212)
<i>Male, Non-White</i>	-0.169 (0.115)	-0.040 (0.098)	0.002 (0.185)	-0.331 (0.251)
<i>Female, Non-White</i>	0.028 (0.120)	-0.000 (0.100)	0.182 (0.197)	-0.332 (0.256)
Observations	2,880	3,440	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 fixed effects	Yes	Yes	Yes	Yes
Leadership fixed effects	Yes	Yes	Yes	Yes
Order fixed effects	Yes	Yes	Yes	Yes
Subject fixed effects	Yes	Yes	Yes	Yes

Notes: 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 IC and in online 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.

type. We find no evidence of Pitt employers responding to candidate demographics. (Online 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 percent) received the lowest possible *Hiring Interest* rating, more than double the share at Penn (15.5 percent). 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 percent of Pitt employers indicated that candidate major was among the

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.”

most important considerations during recruitment, compared to only 15.3 percent 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 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.

IV. 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 that 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 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.

³⁰ As prominent examples, researchers have recently audited college professors, requesting in-person meetings (Milkman, Akinola, and Chugh 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 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, Imas, and Rosenberg (2019), 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.

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 frequent 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 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, Haushofer, and Roth 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 I). For this reason, though, as well as others highlighted in Section IID, 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 III, our failed attempt to replicate at Pitt highlights

³²First, 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 ten 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 participation was low).

³³Student 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 they were recommended to a firm. We recommended 207 unique student resumes over the course of the study at Penn, highlighting the value to students.

³⁴The 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)).

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 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 granular 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.

REFERENCES

- Altonji, Joseph G., and Rebecca M. Blank.** 1999. “Race and Gender in the Labor Market.” In *Handbook of Labor Economics*, Vol. 3C, edited by Orley C. Ashenfelter and David Card, 3143–259. Amsterdam: North-Holland.
- Autor, David H., and Susan N. Houseman.** 2010. “Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from “Work First.”” *American Economic Journal: Applied Economics* 2 (3): 96–128.
- Baert, Stijn.** 2018. “Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments since 2005.” In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by Michael S. Gaddis, 63–77. Berlin: Springer.
- Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka.** 2016. “Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition.” *American Economic Review* 106 (6): 1437–75.
- Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan.** 2005. “Implicit Discrimination.” *American Economic Review* 95 (2): 94–98.

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.

- Bertrand, Marianne, and Esther Dufo.** 2016. "Field Experiments on Discrimination." NBER Working Paper 22014.
- Bertrand, Marianne, and Sendhil Mullainathan.** 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.
- Bohnet, Iris, Alexandra van Geen, and Max Bazerman.** 2015. "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation." *Management Science* 62 (5): 1225–34.
- Bohren, Aislinn, Alex Oleg Imas, and Michael Rosenberg.** 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109 (10): 3395–3436.
- Butler, Daniel M., and David E. Broockman.** 2011. "Do Politicians Racially Discriminate against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55 (3): 463–77.
- Carrell, Scott E., Marianne E. Page, and James E. West.** 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *Quarterly Journal of Economics* 125 (3): 1101–44.
- Culwell-Block, Beverly, and Jean Anna Sellers.** 1994. "Resume Content and Format: Do the Authorities Agree?" *Bulletin of the Association for Business Communication* 57 (4). <https://www.questia.com/library/journal/1G1-16572126/resume-content-and-format-do-the-authorities-agree>.
- Darolia, Rajeev, Cory Koedel, Paco Martorell, Katie Wilson, and Francisco Perez-Arce.** 2015. "Do Employers Prefer Workers Who Attend For-Profit Colleges? Evidence from a Field Experiment." *Journal of Policy Analysis and Management* 34 (4): 881–903.
- Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz.** 2016. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study." *American Economic Review* 106 (3): 778–806.
- de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth.** 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–302.
- Dishman, Lydia.** 2018. "Your Resume Only Gets 7.4 Seconds to Make an Impression—Here's How to Stand Out." *Fast Company*, November 7, 2018. <https://www.fastcompany.com/90263970/your-resume-only-gets-7-4-seconds-to-make-an-impression-heres-how-to-stand-out>.
- Distelhorst, Greg, and Yue Hou.** 2017. "Constituency Service under Nondemocratic Rule: Evidence from China." *Journal of Politics* 79 (3): 1024–40.
- Eriksson, Stefan, and Dan-Olof Rooth.** 2014. "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment." *American Economic Review* 104 (3): 1014–39.
- Ewens, Michael, Bryan Tomlin, and Liang Choon Wang.** 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *Review of Economics and Statistics* 96 (1): 119–34.
- Farber, Henry S., Chris M. Herbst, Dan Silverman, and Till von Wachter.** 2018. "Whom Do Employers Want? The Role of Recent Employment and Unemployment Status and Age." NBER Working Paper 24605.
- Fryer, Roland G., and Steven D. Levitt.** 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119 (3): 767–805.
- Gaddis, S. Michael.** 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93 (4): 1451–79.
- Goldin, Claudia.** 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4): 1091–119.
- Govorun, Olesya, and B. Keith Payne.** 2006. "Ego–Depletion and Prejudice: Separating Automatic and Controlled Components." *Social Cognition* 24 (2): 111–36.
- Greene, William.** 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7 (1): 98–119.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz.** 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464–80.
- Hamermesh, Daniel.** 2012. "Are Fake Resumes Ethical for Academic Research?" *Freakonomics Blog*, March 15, 2012. <http://freakonomics.com/2012/03/15/are-fake-resumes-ethical-for-academic-research/>.
- Hanson, Andrew, and Zackary Hawley.** 2011. "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.
- Harrison, Glenn W., and John A. List.** 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55.
- Heckman, James J.** 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12 (2): 101–16.

- Heckman, James J., and Peter Siegelman.** 1992. "The Urban Institute Audit Studies: Their Methods and Findings." In *Clear and Convincing Evidence: Measurement of Discrimination in America*, 187–258. Lanhan, MD: Urban Institute Press.
- Kessler, Judd B., Corinne Low, and Colin D. Sullivan.** 2019. "Incentivized Resume Rating: Eliciting Employer Preferences without Deception: Dataset." *American Economic Review*. <https://doi.org/10.1257/aer.20181714>.
- Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad.** 2016. "Field of Study, Earnings, and Self-Selection." *Quarterly Journal of Economics* 131 (3): 1057–111.
- Kleykamp, Meredith.** 2009. "A Great Place to Start? The Effect of Prior Military Service on Hiring." *Armed Forces & Society* 35 (2): 266–85.
- Kroft, Kory, Fabian Lange, and Matthew J. Notowidigdo.** 2013. "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment." *Quarterly Journal of Economics* 128 (3): 1123–67.
- Lang, Kevin, and Jee-Yeon K. Lehmann.** 2012. "Racial Discrimination in the Labor Market: Theory and Empirics." *Journal of Economic Literature* 50 (4): 959–1006.
- Low, Corinne.** 2019. "A 'Reproductive Capital' Model of Marriage Market Matching." Unpublished.
- Milkman, Katherine L., Modupe Akinola, and Dolly Chugh.** 2012. "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science* 23 (7): 710–17.
- Milkman, Katherine L., Modupe Akinola, and Dolly Chugh.** 2015. "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): 1678–1712.
- Neumark, David.** 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47 (4): 1128–57.
- Neumark, David, Ian Burn, and Patrick Button.** 2019. "Is It Harder for Older Workers to Find Jobs?" *Journal of Political Economy* 127 (2): 922–70.
- Nosek, Brian A., Frederick L. Smyth, Jeffrey J. Hansen, Thierry Devos, Nicole M. Lindner, Kate A. Ranganath, Colin Tucker Smith, et al.** 2007. "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes." *European Review of Social Psychology* 18 (1): 36–88.
- Nunley, John M., Adam Pugh, Nicholas Romero, and Richard Alan Seals.** 2014. "Unemployment, Underemployment, and Employment Opportunities: Results from a Correspondence Audit." Auburn University Department of Economics Working Paper 4.
- Nunley, John M., Adam Pugh, Nicholas Romero, and Richard Alan Seals.** 2017. "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): 642–69.
- Ortmann, Andreas, and Ralph Hertwig.** 2002. "The Costs of Deception: Evidence from Psychology." *Experimental Economics* 5 (2): 111–31.
- Pallais, Amanda.** 2014. "Inefficient Hiring in Entry-Level Labor Markets." *American Economic Review* 104 (11): 3565–99.
- Pope, Devin G., and Justin R. Sydnor.** 2011. "What's in a Picture? Evidence of Discrimination from prosper.com." *Journal of Human Resources* 46 (1): 53–92.
- Quadlin, Natasha.** 2018. "The Mark of a Woman's Record: Gender and Academic Performance in Hiring." *American Sociological Review* 83 (2): 331–60.
- Riach, Peter A., and Judith Rich.** 2006. "An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market." *BE Journal of Economic Analysis & Policy* 5 (2).
- Rivera, Lauren A., and András Tilcsik.** 2016. "Class Advantage, Commitment Penalty: The Gendered Effect of Social Class Signals in an Elite Labor Market." *American Sociological Review* 81 (6): 1097–131.
- Sherman, Jeffrey W., Frederica R. Conrey, and Carla J. Groom.** 2004. "Encoding Flexibility Revisited: Evidence for Enhanced Encoding of Stereotype-Inconsistent Information under Cognitive Load." *Social Cognition* 22 (2): 214–32.
- Turner, Margery, Michael E. Fix, and Raymond J. Struyk.** 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, DC: Urban Institute Press.
- Wigboldus, Daniël H. J., Jeffrey W. Sherman, Heather L. Franzese, and Ad van Knippenberg.** 2004. "Capacity and Comprehension: Spontaneous Stereotyping under Cognitive Load." *Social Cognition* 22 (3): 292–309.
- Zussman, Asaf.** 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal* 123 (572): F433–68.