Incentivized Resume Rating: Eliciting Employer Preferences without Deception

Judd B. Kessler, Corinne Low, and Colin D. Sullivan*

November 6, 2018

Abstract

We introduce a new experimental paradigm to evaluate employer preferences for job candidates, called Incentivized Resume Rating (IRR). Employers evaluate resumes they know to be hypothetical in order to be matched with real job seekers, allowing researchers to randomly assign resume characteristics while preserving incentives and avoiding deception. We deploy IRR to investigate preferences of employers recruiting prestigious college graduates and find evidence of discrimination. Employers recruiting for science and engineering positions are less interested in hiring females and minorities. Moreover, employers believe females and minorities are less likely to accept jobs when offered, suggesting a novel channel for discrimination.

*The Wharton School, University of Pennsylvania, 3620 Locust Walk, Steinberg Hall-Dietrich Hall, Philadelphia, PA 19104 (email: judd.kessler@wharton.upenn.edu, corlow@wharton.upenn.edu, colins@wharton.upenn.edu). 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.

1 Introduction

How labor markets reward education, work experience, and other forms of human capital is of fundamental interest in labor economics and the economics of education (e.g., Autor and Houseman [2010], Pallais [2014]). Similarly, the role of discrimination in labor markets is a key concern for both policy makers and economists (e.g., Altonji and Blank [1999], Lang and Lehmann [2012]). Correspondence audit studies, including resume audit studies, have become powerful tools to answer questions in both domains.[1] These studies have generated a rich set of findings on discrimination in settings as diverse as employment (e.g., Bertrand and Mullainathan [2004]), real estate and housing (e.g., Hanson and Hawley [2011], Ewens et al. [2014]), and retail (e.g., Pope and Sydnor [2011], Zussman [2013]).[2] More recently, resume audit studies have been used to investigate how employers respond to other characteristics of job candidates, including unemployment spells [Kroft et al., 2013, Eriksson and Rooth, 2014, Nunley et al., 2017], for-profit college credentials [Darolia et al., 2015, Deming et al., 2016], college selectivity [Gaddis, 2015], and military service [Kleykamp, 2009].

Despite the strengths of this workhorse methodology, resume audit studies also face some limitations. First, resume audit studies require deception. While there is generally a prohibition on deception within experimental economics [Ortmann and Hertwig, 2002], audit and resume audit studies are granted an exception, presumably because of the importance of the questions they tackle combined with a lack of alternative options to answer them. Second, audit studies generally report callback results, which (a) may conflate an employer's interest in a candidate with the employer's expectation of a candidate's interest in the employer and (b) can only

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

[2]See Bertrand and Duflo [2016] for a thorough summary of field experiments on discrimination in myriad domains.

identify preferences at one point in the quality distribution (i.e., at the callback threshold), which may not be generalizable [Heckman, 1998, Neumark, 2012].[3] In this paper, we pioneer a new experimental paradigm that addresses these challenges and allows us to study employers in markets not accessible to resume audit studies.

Our paradigm is called Incentivized Resume Rating (IRR). IRR invites employers to evaluate resumes that they know to be hypothetical—avoiding deception—and provides incentives by matching employers with real job seekers based on these evaluations. As we demonstrate through our implementation of IRR, researchers using the paradigm can: (a) recruit employers who would not respond to unsolicited resumes, (b) elicit a single employer's preferences over multiple resumes, (c) randomize many candidate characteristics simultaneously, (d) collect a granular measure of employer interest in each candidate rather than a binary callback decision, (e) ask employers to provide a separate measure of likelihood of offer acceptance to mitigate against it confounding the rating of interest in a candidate, and (f) collect supplemental data about the specific individuals reviewing resumes and their firms.

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, a market largely unexplored by the resume audit literature.[4] We find that employers value higher grade point averages as well as the quality and quantity of summer internship experience. 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 (i.e., 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 rather than needing to work for an hourly wage.

[3] The former concern might contribute to counterintuitive results in the resume audit literature, such as higher callback rates for unemployed than employed candidates [Kroft et al., 2013, Nunley et al., 2017, 2014, Farber et al., 2018]. Researchers who use audit studies aim to mitigate such concerns (e.g., Bertrand and Mullainathan [2004] attempted to construct high-quality resumes that did not lead candidates to be "overqualified," page 995). While the latter concern might be mitigated by the callback threshold being empirically relevant—we may particularly care about detecting preferences at the callback margin—such preferences may not generalize to other relevant margins (e.g., when firms are expanding and hiring more candidates).

[4] Presumably, the resume audit literature has left this market unexplored because firms hiring students coming out of prestigious colleges and universities typically do not recruit candidates by responding to cold resumes.

Studying this population of employers also allows us to investigate whether race and gender discrimination is active in this market—a market in which many employers claim to have a preference for diversity.[5] We do not find evidence that employers are less interested in female and minority candidates on average, but we do find discrimination among employers looking to hire candidates with STEM majors.[6] We find suggestive evidence that discrimination in hiring interest is due to implicit bias by observing how discrimination changes as employers evaluate multiple resumes.[7]

By asking employers to evaluate their interest in candidates on a 10-point Likert scale (i.e., rather than asking for binary callback decisions), we are able to identify how employer preferences for specific candidate characteristics differ as candidate quality changes. Our results suggest that most of the preferences we identify maintain sign and significance across the distribution of quality. However, we find that the effects of major and of work experience are most pronounced towards the middle of the quality distribution and smaller in the tails.[8]

To mitigate against employers' ratings of interest in a candidate being confounded with employers' beliefs about the likelihood that a candidate would accept a job offer, we asked employers to rate their interest in each candidate assuming the candidate would accept a job if offered, and we separately asked employers for their evaluations of each candidate's likelihood of accepting a job. The likelihood of offer acceptance data provide additional insights. Employers report believing that minority and female candidates are less likely to accept job offers than their White male counterparts, which suggests a novel channel for discrimination. Due to the cost of recruiting, firms may be reluctant to pursue female and minority candidates who they do not believe will accept a position if offered.

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

[6] STEM is a common acronym for "Science, Technology, Engineering, and Math."

[7] In addition, consistent with results from the resume audit literature finding lower returns to quality for minority candidates (see Bertrand and Mullainathan [2004]), we find that candidates who are not White males receive a lower return to work experience at prestigious internships.

[8] For example, employers display a strong preference for candidates who attend Wharton (the business school at Penn) over candidates from the College of Arts and Sciences on average but such a preference would be difficult or impossible to detect in the low callback rate environments typical of resume audit studies.

In addition to presenting results on employer preferences, this paper describes in detail how to implement an IRR study, which we hope will be of value to future researchers. IRR provides a path forward for those interested in studying discrimination and other employer preferences without using deception. While correspondence audit studies have been tremendously useful, their proliferation has raised concerns about whether the use of unwitting subjects' time (see Hamermesh [2012] for a critique) and other possible harms of deception are outweighed by the social benefits of the research.[9] Audit studies have typically been insulated against a principal critique of deception in the lab—that it pollutes the subject pool for future research—because of the notion that the experimenter's manipulation is small relative to the market as a whole. However, this argument becomes less valid as the number of audit studies grows, particularly now that many studies are run in the same venues (e.g., certain online job boards). Thus, the continued expansion of resume audit studies raises the concern that employers could become wary of certain types of resumes sent out by researchers and that this could not only harm the validity of future research but also real job seekers whose resumes are similar to those sent by researchers.[10] Given these concerns, we believe IRR provides significant value as a deception-free alternative to measure employer preferences.

The remainder of this paper proceeds as follows. Section 2 describes the IRR paradigm and our specific experimental design. Section 3 reports on the results about the preferences of employers recruiting at the University of Pennsylvania.

[9]As evidence of their proliferation, Baert [2018] notes 90 resume audit studies focused on discrimination against protected classes in labor markets alone between 2005 and 2016. The method has also proliferated into other domains, such as requesting in-person meetings from professors [Milkman et al., 2012, 2015]. 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 difference. That said, there have been some exceptions, like Bartoš et al. [2016], which uses a novel design where employers and landlords can click a link in an email to acquire more information on the candidate to show that less attention is allocated to candidates one discriminates against.

[10]These harms might be particularly relevant when the populations being audited are likely to become aware of the research *ex post*. For example, college professors (audited in Milkman et al. [2012, 2015]) might take such studies as an excuse to ignore future student emails. The popularity of the method makes it likely that employers are aware of it as well, especially as audit studies can be used as legal evidence of discrimination [Neumark, 2012]. Other expansions of the audit methodology may be particularly concerning, such as auditing politicians' responses to putative constituents (see Butler and Broockman [2011], Distelhorst and Hou [2017]), which might distort politicians' beliefs about the priorities of the populations they serve. This may be particularly problematic when researchers seek a politician-level audit measure, which requires sending many fake requests to the same politician.

Section 4 discusses a replication study we conducted at the University of Pittsburgh. Section 5 discusses the costs and benefits of IRR, highlights the implications of our results, and concludes.

2 Study Design

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

2.1 Employers and Recruitment

IRR allows researchers to recruit employers in the market for candidates from particular institutions and those who do not screen unsolicited resumes and thus may be hard—or impossible—to study in audit or resume audit studies. To leverage this benefit of the experimental paradigm, we partnered with the University of Pennsylvania (Penn) Career Services office to identify employers recruiting highly skilled generalists from the Penn graduating class.[11]

Penn Career Services sent invitation emails (see Appendix Figure A.1 for recruitment email) in two waves during the 2016-2017 academic year to employers who historically recruited Penn seniors (e.g., firms that recruited on campus, regularly attended career fairs, or otherwise hired students). The first wave was around

[11] We initially partnered with Penn Career Services since all the authors were affiliated with the University of Pennsylvania. As described in detail in Section 4, we replicated our experiment at University of Pittsburgh (Pitt). We identified Pitt by reaching out to the career services departments of other Pennsylvania schools that differed from Penn on a number of dimensions. Pitt's career services department was the most willing to participate.

the time of on-campus recruiting in the fall of 2016. The second wave was around the time of career-fair recruiting in the spring of 2017. In both waves, the recruitment email invited employers to use "a new tool that can help you to identify potential job candidates." While the recruitment email and the information that employers received before rating resumes (see Appendix Figure A.3 for instructions) noted that anonymized data from employer responses would be used for research purposes, this was framed as secondary. The recruitment process and survey tool itself both emphasized that employers were using new recruitment software. For this reason, we note that our study has the ecological validity of a field experiment.[12] As was outlined in the recruitment email (and described in detail in Section 2.2), each employer's one and only incentive for participating in the study is to receive 10 resumes of job seekers that match the preferences they report in the survey tool.[13]

2.2 Incentives

The main innovation of IRR is its method for incentivized preference elicitation, a variant of a method pioneered by Low [2017] in a different context. In its most general form, the method asks subjects to evaluate candidate profiles, which are known to be hypothetical, with the understanding that more accurate evaluations will maximize the value of their participation incentive. In our implementation of IRR, each employer evaluates 40 hypothetical candidate resumes and their participation incentive is a packet of 10 resumes of real job seekers from a large pool of Penn seniors. For each employer, we select the 10 real job seekers based on the employer's evaluations.[14] Consequently, the participation incentive in our study be-

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

[13]We timed recruitment so that employers would receive these 10 resumes around the time they were on campus in order to facilitate meeting the job seekers. In addition, we offered webinars for employers who were interested in learning about the survey screening experience before they participated. Employers could anonymously join a call where they viewed a slideshow about the survey software and could submit questions via chat box. Attendance at these webinars was low.

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

comes more valuable as employers' evaluations of candidates better reflect their true preferences for candidates.[15]

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

2.3 Resume Creation and Variation

Our implementation of IRR asked each employer to evaluate 40 unique, hypothetical resumes, and varied multiple candidate characteristics simultaneously and independently across resumes. This allows us to estimate employer preferences for each characteristic, identified over a rich space of baseline candidate characteristics.[16] In particular, when an employer began the survey tool, each of the 40 resumes was dynamically populated with independently drawn, randomly selected candidate characteristics. 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.

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

[16]In a traditional audit or resume audit study, researchers are limited in the number of candidates or resumes that they can show to any particular employer—sending too many fake resumes to the same firm might trigger ethics concerns or cause employers to become suspicious. They are additionally limited in how candidate characteristics can covary across resumes—a resume with a very low GPA but very prestigious internships might appear unusual or suspicious. For example, in their seminal resume audit study paper, Bertrand and Mullainathan [2004] only send four resumes to each firm and only create two quality levels of resumes (i.e., a high quality resume and a low quality resume, in which various candidate characteristics vary together).

First, we built the hypothetical resumes using components (i.e., work experiences, leadership experiences, and skills) from real resumes of seniors at Penn. Second, we asked the employers to choose the type of candidates that they were interested in hiring, based on major (see Appendix Figure A.4). In particular, they could choose either "Business (Wharton), Social Sciences, and Humanities" (henceforth "Humanities & Social Sciences") or "Science, Engineering, Computer Science, and Math" (henceforth "STEM"). They were then shown hypothetical resumes focused on 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.[17] Third, to enhance realism, and to make the evaluation of the resumes less tedious, we used 10 different resume templates, which we populated with the candidate characteristics and component pieces described below, to generate the 40 hypothetical resumes (see Appendix Figure A.5 for a sample resume). We based these templates on real student resume formats (see Appendix Figure A.6 for examples).[18] Fourth, we gave employers short breaks within the study by showing them a progress screen after each block of 10 resumes they evaluated. In Section 3.4, we use the change in attention induced by these breaks to construct tests of implicit bias.

2.3.1 Education Information

In the education section of the resume, we independently randomized each candidate's grade point average (GPA) and major. GPA is drawn from a uniform distribution between 2.90 and 4.00, shown to two decimal places and never omitted from the resume.[19] 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").[20]

[17] The resumes of the 10 real job seekers the employers were sent after completion of the study were also drawn from a set of majors corresponding to the employer's choice.

[18] 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.

[19] Some students omit GPA from their resumes, presumably to avoid reporting a low GPA. We chose not to test how employers respond to the omission of the GPA for this experiment, but it would be an easy modification to our experimental design for future researchers.

[20] Table A.3 shows the list of majors by major category, school, and the probability that the major was used in a resume.

Table 1: Randomization of Resume Components

Resume Component	Description	*Analysis Variable*
Personal Information		
First & last name	Drawn from list of 50 possible names given selected race and gender (names in Tables A.1 & A.2)	*Not a White Male* (67.15%)
	Race drawn randomly from U.S. distribution (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian)	*Not White* (34.3%)
	Gender drawn randomly (50% male, 50% female)	*Female* (50%)
Education Information		
GPA	Drawn $Unif[2.90, 4.00]$ to second decimal place	*GPA*
Major	Drawn from a list of majors at Penn (Table A.3)	*Major* (weights in Table A.3)
Degree type	BA, BS fixed to randomly drawn major	*Wharton* (40%)
School within university	Fixed to randomly drawn major	*School of Engineering and Applied Science* (70%)
Graduation date	Fixed to upcoming spring (i.e., May 2017)	
Work Experience		
First job	Drawn from curated list of top internships and regular internships	*Top Internship* (20/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate's junior year (i.e., 2016)	
Second job	Left blank or drawn from curated list of regular internships and work-for-money jobs (Table A.5)	*Second Internship* (13/40) *Work for Money* (13/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate's sophomore year (i.e., 2015)	
Leadership Experience		
First & second leadership	Drawn from curated list	
Title and activity	Fixed to randomly drawn leadership	
Location	Fixed to Philadelphia, PA	
Description	Bullet points fixed to randomly drawn leadership	
Dates	Start and end years randomized within college career, with more recent experience coming first	
Skills		
Skills list	Drawn from curated list, with two skills drawn from {Ruby, Python, PHP, Perl} and two skills drawn from {SAS, R, Stata, Matlab} shuffled and added to skills list with probability 25%.	*Technical Skills* (25%)

Resume components are listed in the order that they appear on hypothetical resumes. Italicized variables in the right column are variables that were randomized to test how employers responded to these characteristics. Degree, first job, second job, and skills were drawn from the different lists for Humanities & Social Sciences resumes and STEM resumes. Name, GPA, and leadership 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 67.15% chance of having an assigned white name).

2.3.2 Work Experience

We included realistic work experience components on the resumes. To generate the components, we scraped more than 700 real resumes of Penn students. We then followed a process described in Appendix A.2.5 to select and lightly sanitize work experience components so that they could be randomly assigned to different resumes without generating conflicts or inconsistencies (e.g., we eliminated references to particular majors or to gender or race).[21]

Our goal in randomly assigning these work experience components was to introduce variation along two dimensions: *quantity* of work experience and *quality* of work experience. To randomly assign quantity of work experience, we varied whether the candidate only had an internship in the summer before senior year, or also had a job or internship in the summer before junior year. Thus, candidates with more experience had two jobs on their resume (before junior and senior years), while others had only one (before senior year).[22]

To introduce random variation in *quality* of work experience, we grouped work experience components into three categories: (1) "top internships," which were internships with prestigious firms as defined by being a firm that successfully hires many Penn graduates; (2) "work-for-money" jobs, which were paid jobs that—at least for Penn students—are unlikely to develop human capital for a future career (e.g., barista, cashier, waiter, etc.); and (3) "regular" internships, which comprised all other work experiences.[23]

---

[21] The work experience component included the associated details from the real resume from which the component was drawn. For example, a work experience component would include an employer, title, location, and a few bullet points with a description of the tasks completed while in that internship.

[22] The resume was constructed dynamically, so if only one job was listed on the resume, the work experience section of the resume appeared shorter (i.e., rather than introducing empty space).

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

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.[24]

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 have 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.

2.3.3 Leadership Experience and Skills

The procedure followed to select leadership experience components was similar to that for work experiences. A leadership experience component includes an activity, title, date, and a few bullet points with a description of the experience (Philadelphia, PA was given as the location of all leadership experiences).[25]

Leadership experiences were included so our resumes would more realistically resemble typical student resumes. We randomly assigned two leadership experiences, with randomly selected ranges of years from within the four years preceding the graduation date, to each resume.[26]

[24] Since the work experience component was comprised of employer, title, location, and bullets about the job—as in practice—a higher quality work experience necessarily reflects all features of this bundle. Consequently, while we could additionally estimate how employers respond to specific elements of the component (e.g., a candidate having an internship in a particular city or working for a Fortune 500 company), these elements were not independently randomized and so are correlated with other features of the work experience that employers likely also care about.

[25] We used the same leadership positions for both major types under the assumption that most extracurricular activities at Penn could plausibly include students from all majors; however, this required us to exclude the few leadership experiences that were too revealing of field of study (e.g., "American Institute of Chemical Engineers"). For additional details, see Appendix A.2.5.

[26] As with work experiences, the richness in detail of a leadership experience would theoretically allow us to estimate how employers respond to specific elements of the leadership experience (e.g., a candidate being a president of a club or the captain of a team), but, again, these elements were not independently randomized and are correlated with other features of a leadership experience that employers might care about.

With skills, by contrast, we added a layer of intentional variation. 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 75% probability, the hypothetical resume included one of these lists. With 25% probability, we appended to this list four technical skills: two randomly drawn advanced programming languages from {Ruby, Python, PHP, Perl} and two randomly drawn statistical programs from {SAS, R, Stata, Matlab}. This variation allows us to measure how employers value candidates with more technical skills listed on their resumes.

2.3.4 Names Indicating Gender and Race

We randomly varied gender and race by assigning each hypothetical resume a name that would be indicative of gender (male or female) and race / ethnicity (Asian, Black, Hispanic, or White).[27] In order to do this randomization, we needed to first generate a list of names that would clearly indicate both gender and race for each the these groups.

We created names by independently generating first names and last names. For first names, we used a dataset of all births in the state of Massachusetts between 1989-1996 and in 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 indicating 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 though they might be 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, under the assumption that last names should not be informative of gender. We used last name data from the 2000 Census tying last names to race. We implemented the same measure of race specificity and required that the last name make up at least 0.1% of that race's population, to ensure that the last names were sufficiently common. Finally, we

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

combined first and last names within race.[28] The full lists of names are given in Appendix Tables A.1 and A.2 (see Appendix A.2.3 for additional details).

For realism, we randomly selected races at rates approximating the distribution in the US population (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian). While a less subtle variation in race might increase statistical power to detect race-based discrimination, such an approach would risk signaling to subjects our intent to study racial preferences. As a result, we evaluate all discrimination on a pooled gender and race measure, looking at White males compared to all those who are not White males (minorities and women combined). In Appendix B.3.4, we examine these demographic groups separately to see if results are always directionally aligned between the two groups.

2.4 Rating Candidates on Two Dimensions

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

[28]This resulted in eight lists of first name-last name pairs for each of: Asian females, Asian males, Black females, Black males, Hispanic females, Hispanic males, White females, and White males. Each list contained 50 names.

sequently, 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.[29,30]

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.[31] In particular, for each resume we asked employers to answer the following two questions (see an example at the bottom of Appendix Figure A.5):

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

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

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

[29] A salient example of this potential confound for academic economists who have participated in junior recruiting (outside of the most desirable departments), is to imagine what an audit study of economics departments would find if it randomly added publications to the CVs of graduate students on the job market. Additional "top-five" publications could very likely decrease the chance that a student receives flyouts from departments that are concerned with whether the candidate is likely to be out of reach.

[30] Audit and resume audit studies focusing on discrimination do not need to interpret callback as a measure of an employer's interest in a candidate to demonstrate discrimination (any difference in callback rates is evidence of discrimination). Nevertheless, richer information might help identify the mechanism driving the discrimination and show whether its magnitude is likely to differ in other environments or as candidate characteristics vary.

[31] The 10-point scale has two advantages. First, it provides additional statistical power, allowing us to observe whether an employer values a characteristic even when it is added to an inframarginal resume that would not be pushed over a binary callback threshold in a resume audit setting. For example, if an employer were to call back anyone they rated an 8 or above on our Likert scale, we can still observe when a characteristic increases an employer's rating from a 6 to a 7 or from a 9 to a 10. Second, it allows us to explore how employer preferences vary across the distribution of hiring interest, an issue we explore in depth in Section 3.3.

would accept it). We denote responses to this question a candidate's "likelihood of acceptance." We asked the first question to assess how resume characteristics affect hiring interest. We asked the second question both to encourage employers 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.

3 Results

3.1 Data and Empirical Approach

We recruited 72 employers through our partnership with the University of Pennsylvania Career Services office in Fall 2016 (46 subjects, 1840 resume observations) and Spring 2017 (26 subjects, 1040 resume observations). The employers who participated in our study as subjects were primarily female (60%) and primarily White (79%) and Asian (15%). They were approximately as likely to work at a large firm with over 1000 employees (38%) as a small firm with less than 100 employees (40%).[32]

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

[32]These small firms are mostly hedge fund, private equity, consulting, or 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. Our sample also included a small number of nonprofit and public interest organizations.

that employers, on average, consider higher Likert-scale ratings more favorably than lower ratings.

In Section 3.2, we examine how human capital characteristics (e.g., GPA, major, work experience, and skills) affect hiring interest. These results report on the mean of preferences across the distribution; we show how our results vary across the distribution of hiring interest in Section 3.3.[33] In Section 3.4, we discuss how employers' ratings of hiring interest respond to demographic characteristics of our candidates. In Section 3.5, we investigate the likelihood of acceptance ratings and identify a potential new channel for discrimination.

3.2 Effect of Human Capital on Hiring Interest

Employers in our study are interested in hiring graduates of the University of Pennsylvania for full-time employment, and many recruit at other Ivy League schools and other top colleges and universities. This labor market has been unexplored by resume audit studies.[34] In this section, we evaluate how randomized candidate characteristics—described in Table 1 and Section 2.3—affect employers' ratings of hiring interest.

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

$$\begin{aligned} V_{ij} = \beta_0 + \beta_1\, \mathit{GPA} + \beta_2\, \mathit{Top\ Internship} + \beta_3\, \mathit{Second\ Internship} + \beta_4\, \mathit{Work\ for\ Money}\, + \\ \beta_5\, \mathit{Technical\ Skills} + \beta_6\, \mathit{Not\ a\ White\ Male} + \mu_j + \gamma_j + \omega_j + \alpha_i + \varepsilon_{ij} \end{aligned} \tag{1}$$

In this regression, *GPA* is a linear measure of grade point average. *Top Internship* is a dummy for having a top internship, *Second Internship* is a dummy for having an internship in the summer before junior year, and *Work for Money* is a dummy for having a work-for-money job in the summer before junior year. *Technical Skills* is a dummy for having a list of skills that included a set of four randomly

[33]Mean preferences may not be reflective of preferences at the tail of a distribution, as highlighted by the Neumark [2012] critique of correspondence audit studies.

[34]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.

assigned technical skills. *Not a White Male* is a dummy equal to 1 if the name of candidate was not indicative of a White male, a measure that allows us to combine discrimination based on race and discrimination based on gender.[35] μ_j are dummies for each major. Table 1 provides more information about these dummies and all the variables in this regression. In some specifications, we include additional controls. γ_j are dummies for each of the leadership experience components. ω_j are dummies for the number of resumes the employer has evaluated as part of the survey tool.[36] 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. These regression show that employers respond strongly to candidate characteristics related to human capital.

GPA is an important driver of hiring interest. The coefficient on *GPA* ranges from 2.1–2.2 Likert-scale points, suggesting 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 2.81, suggesting that a point improvement in GPA moves hiring interest ratings by about three quarters of a standard deviation.

Employers additionally value the quality and quantity of a candidate's work experience. The quality of a candidate's work experience in the summer before senior year has a large impact on hiring interest ratings. The coefficient on *Top Internship* ranges from 0.9–1.0 Likert-scale points, which is roughly a third of a standard deviation of ratings. Comparing regression coefficients, we see that a top internship is approximately equivalent to a half-point improvement in GPA, meaning that, on average, a candidate with a 3.5 GPA and a prestigious internship is viewed equivalently to a candidate with a 4.0 GPA and a regular internship.

[35] While we do not discuss results about *Not a White Male* in this section, we include a control for this randomized resume component in our regressions and discuss the results in Section 3.4 and Appendix B.3. In Appendix B.3.4, we analyze race and gender separately.

[36] 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.

Table 2: Human Capital Experience

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.124***	2.189***	2.195***	0.890***
	(0.145)	(0.150)	(0.129)	(0.0613)
Top Internship	0.905***	0.904***	0.902***	0.379***
	(0.0944)	(0.0988)	(0.0806)	(0.0396)
Second Internship	0.463***	0.487***	0.463***	0.206***
	(0.112)	(0.118)	(0.0947)	(0.0468)
Work for Money	0.112	0.152	0.149	0.0510
	(0.110)	(0.113)	(0.0913)	(0.0468)
Technical Skills	0.0499	0.0557	-0.0680	0.0133
	(0.104)	(0.108)	(0.0900)	(0.0441)
Not a White Male	-0.122	-0.157	-0.117	-0.0531
	(0.0987)	(0.103)	(0.0842)	(0.0415)
Observations	2880	2880	2880	2880
R^2	0.129	0.180	0.482	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 1.91, 2.28, 2.64, 2.94, 3.26, 3.6, 4.05, 4.51, and 5.03.

Table shows OLS and ordered probit regressions hiring interest from Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* and major are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS, likelihood ratio test for ordered probit).

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 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.[37] 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).

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

Table 2 also reports the p-value of a test of whether the coefficients on the major dummies are jointly different from zero. Results suggest that the randomly assigned major significantly affects hiring interest. While we do not have the statistical power to test for the effect of each major, we can explore how employers respond to candidates being from more prestigious schools at the University of Pennsylvania. In particular, 40% of the Humanities & Social Sciences resumes are assigned a BS in Economics from Wharton and the rest have a BA major from the College of Arts and Sciences. In addition, 70% of the STEM resumes are assigned a BS from the School of Engineering and Applied Science and the rest have a BA major from the College of Arts and Sciences. As shown in Appendix Table B.2, in both cases, we find that being from the more prestigious school—and thus receiving a BS rather than a BA—is associated with an increase in hiring interest ratings of about 0.4–0.5 Likert-scale points.[38]

[37]This effect is substantially smaller than the effect of *Top Internship* ($p < 0.05$ for all tests comparing the *Second Internship* and *Top Internship* coefficients). A second summer of internship experience is worth approximately half as much as having a higher-quality internship experience in the summer before senior year.

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

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 fourth 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.[39]

As discussed in Section 2, we made many design decisions to enhance realism. However, one might be concerned that our independent cross-randomization of various resume components might lead to unrealistic resumes and influence the results we find. We provide two robustness checks in the appendix to address this concern. First, we recognize that while our design and analysis treats each work experience as independent, in practice, candidates may have chosen to work in related jobs over a series of summers to create a compelling work experience "narrative." If employers value combinations of work experiences, narrative might be an omitted variable that could introduce bias (e.g., if our *Top Internships* are more likely to generate narratives than regular internships, we may misestimate its effect on hiring interest). In Appendix B.1, we describe how we test for the importance of work experience narrative, and show regression results in Appendix Table B.1. We find that employers do respond positively to work experience narrative ($p = 0.054$), but our main results are robust to the inclusion of this omitted variable.

Second, the GPA distribution we used for constructing the hypothetical resumes did not perfectly match the distribution of job seekers in our labor market. In Appendix B.2, we re-weight our data to match the GPA distribution in the candidate pool of real Penn job seekers and show that our results are robust to this

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

re-weighting.[40] These exercises provide some assurance that our results are not an artifact of how we construct hypothetical resumes.

3.3 Effects Across the Distribution of Hiring Interest

The regression specifications described in Section 3.2 identify the average effect of candidate characteristics on employers' hiring interest. As pointed out by Neumark [2012], however, these average preferences may differ in magnitude—and even direction—from differences in callback rates, which derive from whether a characteristic pushes a candidate above a specific quality threshold (i.e., the callback threshold). For example, in the low callback rate environments that are typical of resume audit studies, differences in callback rates will be determined by how employers respond to a candidate characteristic in the right tail of their distribution of preferences.[41] To make this concern concrete, Appendix B.4 provides a simple graphical illustration in which the average preference for a characteristic differs from the preference in the tail of the distribution.[42]

An advantage of the IRR methodology, however, is that it can deliver a granular measure of hiring interest to explore whether employers' preferences for characteristics do indeed differ in the tails of the hiring interest distribution. We employ two basic tools to explore preferences across the distribution of hiring interest: (1) the empirical cumulative distribution function (CDF) of hiring interest ratings and (2) a "counterfactual callback threshold" exercise. In the latter exercise, we impose a

[40] Matching the underlying distribution of characteristics in hypothetical resumes to the distribution of real candidates is also an issue for resume auditors who must contend with a limited number of underlying resumes (i.e., resumes that they alter to create treatment variation). Given uncertainty about the characteristics of candidates and the limited number of underlying resumes, resume auditors may not be able to perfectly match the distribution of characteristics of a target population. An additional advantage of the IRR methodology is that it involves collecting a large number of resumes from an applicant pool of real job seekers, which gives us information on the distribution of candidate characteristics that we can use to re-weight the data ex post.

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

[42] In a case where there is a conflict between average preferences and preferences in the tail, it is unclear which preference we care about for policy. For example, we may care about the preference at the relevant thresholds for callback or hiring, but we may also be interested in whether those preferences are robust to a hiring expansion or contraction where those thresholds change.

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.[43]

While the theoretical concerns raised by Neumark [2012] may be relevant in other settings, the direction of the average results we report in Section 3.2 is consistent across the distribution of hiring interest, including in the tails.[44] 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.[45] 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 3.2. *Second Internship* is positive everywhere and almost always statistically significant.[46] *Work for Money* consistently has no impact on employer preferences throughout the distribution of hiring interest.

As noted above, our counterfactual callback threshold exercise suggests that a well-powered audit study would likely find differences in callback rates for most of the characteristics that we estimate as statistically significant on average in Section 3.2, regardless of employers' callback threshold. This result is reassuring both for the validity of our results and in considering the generalizability of results from the resume audit literature.[47] However, even in our data, we observe a case where a

[43] We report the OLS coefficient since most resume audit studies report this measure. In Appendix B.4 we present alternative specifications (odds ratio and logit models) and discuss the implications of our results for measurement in audit studies.

[44] The one exception that we find, a preference for Wharton among employers looking to hire in the Humanities & Social Sciences, is discussed below.

[45] This shape is partially a mechanical feature of low callback rate environments: if a threshold is set high enough that only 5% of candidates with a desirable characteristic are being called back, the difference in callback rates can be no more than 5 percentage points. At lower thresholds (e.g., where 50% of candidates with desirable characteristics are called back), differences in callback rates can be much larger. In Appendix B.4, we discuss how this feature of difference in callback rates could lead to misleading comparisons across experiments with very different callback rates.

[46] Unlike *Top Internship*, *Second Internship* has a similar percentage point effect size throughout the distribution.

[47] Our results provide initial evidence that audit study results may be reflective of average preferences and may be robust to different callback thresholds. However, we may still be concerned

well-powered audit study would be unlikely to find a result, even though we find one on average. Appendix Figure B.1 mirrors Figure 1 but focuses on having a Wharton degree among employers seeking Humanities & Social Sciences candidates. Employers respond to Wharton in the middle of the distribution of hiring interest, but preferences seem to converge in the right tail (i.e., at hiring interest ratings of 9 or 10), suggesting that the best students from the College of Arts and Sciences are not evaluated differently than the best students from Wharton.

3.4 Demographic Discrimination

In this section, we examine how hiring interest ratings respond to the race and gender of candidates. As described in Section 2 and shown in Table 1, we use our variation in names to create a variable *Not a White Male* that indicates a candidate is either female or non-White. In Appendix B.3, we recreate all of the results described here breaking out the effects into *Female, White*, *Male, Non-White*, and *Female, Non-White* to see which groups our driving our results. As shown in Table 2, *Not a White Male*, while directionally negative, is not significantly different from zero, suggesting no evidence of discrimination on average in our data.[48] This null result contrasts somewhat with existing literature—both resume audit studies (e.g., Bertrand and Mullainathan [2004]) and laboratory experiments (e.g., Bohnet et al. [2015]) generally find evidence of discrimination in hiring. This may not be surprising given that our employer pool is different than those usually targeted through resume audit studies, with most reporting positive tastes for diversity.[49]

While we see little evidence of discrimination on average, a large literature addressing diversity in the sciences (e.g., Carrell et al. [2010], Goldin [2014]) suggests we might be particularly likely to see discrimination among employers seeking STEM candidates.[50] In Table 3, we estimate the regression in Equation (1) separately by major type. Since we are splitting the data, significance tests in Table 3 are presented with a Bonferroni correction in which we multiply p-values by 2. Results in Panel B show that employers looking for STEM candidates display a large, sta-

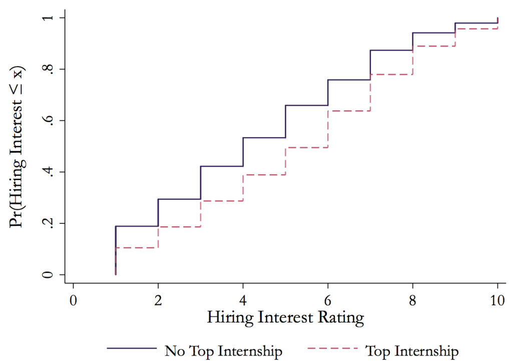
about inferences arising from comparing differences in callback rates at different levels of selectivity for reasons discussed in footnote 45 and Appendix B.4.

[48] In Appendix B.3.1, we show that this effect does not differ by the gender or race of the employer rating the resume.

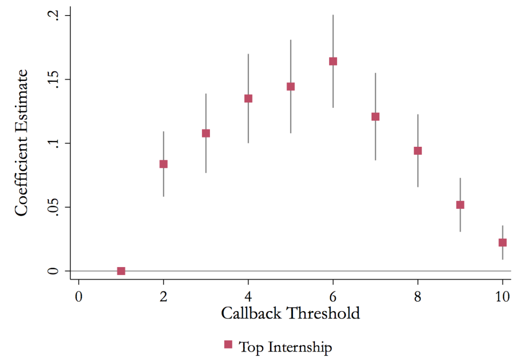
[49] See the discussion in footnote 5.

[50] As described in detail in Section 2.3, at the start of the study, each employer selected the group of majors (i.e., STEM or Humanities & Social Sciences) from which they were looking to hire.

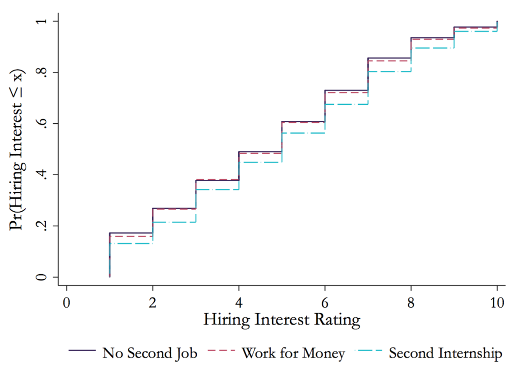
Figure 1: Value of Quality of Experience Over Selectivity Distribution



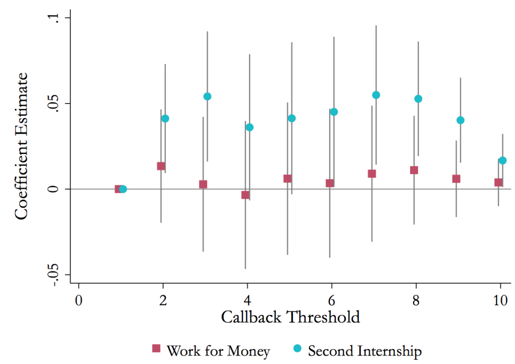
(a) Empirical CDF for Top Internship



(b) Linear Probability Model for Top Internship



(c) Empirical CDF for Second Job Type



(d) Linear Probability Model for Second Job Type

Empirical CDF of *Hiring Interest* (Panels 1a & 1c) and difference in counterfactual callback rates (Panels 1b & 1d) for *Top Internship* and *No Second Job*, *Second Internship*, and *Work for Money*. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

tistically significant preference for White male candidates. The coefficient on *Not a White Male* suggests that candidates who are not White males suffer a penalty of 0.4–0.5 Likert-scale points—a penalty of about 0.25 GPA points—that is robust across our specifications. Results in Panel A show that employers looking for Humanities & Social Sciences candidates do not show evidence of discrimination in hiring interest.[51] As shown in Appendix Table B.9) in Appendix B.3.4, this discrimination against candidates who are not White males is driven by discrimination against White women and minority men.

As in Section 3.3, we can examine these results across the hiring interest rating distribution. Figure 2 shows the CDF of hiring interest ratings and the difference in counterfactual callback rates, separately for employers interested in Humanities & Social Sciences candidates and STEM candidates. Among employers interested in Humanities & Social Sciences candidates, the CDFs of *Hiring Interest* ratings are nearly identical for White male candidates and candidates who are not White males. 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.

[51] Aside from discrimination and minor differences in the value of *Top Internship* and *Second Internship*, preferences for candidate characteristics appear to be similar across the different types of majors at Penn. That the coefficient estimates on *Top Internship* and *Second Internship* differ between STEM and Humanities & Social Sciences is not particularly surprising given that the specific jobs associated with those labels differ across the two major types, and we do not have an objective measure of the prestige of top STEM internships relative to those in Humanities & Social Sciences.

Table 3: Effects by Major Type (with Bonferroni-Corrected Significance Levels)

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
	Panel A: Humanities & Social Sciences			
GPA	2.210***	2.309***	2.300***	0.933***
	(0.172)	(0.179)	(0.153)	(0.0735)
Top Internship	1.080***	1.051***	1.039***	0.453***
	(0.108)	(0.116)	(0.0944)	(0.0461)
Second Internship	0.541***	0.518***	0.514***	0.241***
	(0.132)	(0.143)	(0.114)	(0.0555)
Work for Money	0.0856	0.102	0.114	0.0364
	(0.129)	(0.134)	(0.109)	(0.0554)
Technical Skills	0.0647	0.0851	-0.0492	0.0137
	(0.122)	(0.130)	(0.106)	(0.0522)
Not a White Male	-0.00876	-0.0334	-0.0110	-0.00304
	(0.115)	(0.122)	(0.0998)	(0.0487)
Observations	2040	2040	2040	2040
R^2	0.128	0.195	0.500	
p-value for test of joint significance of Majors	0.019	0.026	0.007	0.029
	Panel B: STEM			
GPA	1.925***	1.861***	1.852***	0.799***
	(0.267)	(0.309)	(0.243)	(0.112)
Top Internship	0.396*	0.548**	0.530***	0.175*
	(0.191)	(0.215)	(0.173)	(0.0784)
Second Internship	0.231	0.291	0.291	0.105
	(0.208)	(0.244)	(0.187)	(0.0878)
Work for Money	0.142	0.262	0.319	0.0720
	(0.212)	(0.252)	(0.185)	(0.0879)
Technical Skills	-0.0184	-0.103	-0.171	0.00445
	(0.196)	(0.227)	(0.186)	(0.0830)
Not a White Male	-0.433**	-0.516**	-0.399*	-0.186**
	(0.192)	(0.219)	(0.188)	(0.0798)
Observations	840	840	840	840
R^2	0.118	0.321	0.590	
p-value for test of joint significance of Majors	< 0.001	0.030	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints (A): 2.26, 2.58, 2.96, 3.27, 3.6, 3.94, 4.41, 4.87, 5.41.

Ordered probit cutpoints (B): 1.44, 1.89, 2.22, 2.5, 2.8, 3.13, 3.55, 4.05, 4.48.

Table shows OLS and ordered probit regressions of hiring interest from Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively, after Bonferroni correction for multiple hypothesis testing in which we multiplied p-values by 2. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS, likelihood ratio test for ordered probit).

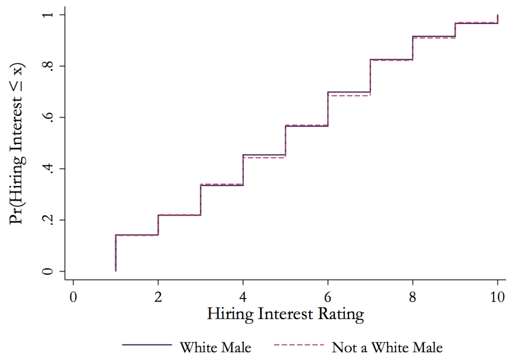
There are several stories consistent with lower hiring interest ratings for female and non-White candidates among employers recruiting candidates in STEM fields. To attempt to identify the drivers of discrimination, we first consider two types of discrimination that might be present among employers in our data: explicit bias and implicit bias. Explicit bias might include an explicit taste for White male candidates or an explicit belief they are more prepared than female or non-White candidates for success at their firm, even conditional on experience listed on a resume.[52] There are a few reasons why we might not expect to see explicit bias in our data. First, employers may attempt to consciously override their explicit biases when they believe expressing such biases may be socially unacceptable. Second, employers might (correctly) believe that our matching algorithm would ignore race and gender when producing matches and so might infer that the race and gender of the candidate could be safely ignored when evaluating candidates. Implicit bias [Greenwald et al., 1998, Nosek et al., 2007], on the other hand, may be present even among employers who are not explicitly considering race (or among employers who are considering race but attempting to suppress any explicit bias they might have). Bertrand et al. [2005] have suggested implicit bias as a channel for discrimination in resume audit studies, and so one might also expect any bias in our ratings to be implicit in nature.

We leverage features of implicit bias—that it is more likely to arise when decision makers are fatigued [Wigboldus et al., 2004, Govorun and Payne, 2006, Sherman et al., 2004]—to test whether our data are consistent with implicit bias. We evaluate whether the bias is larger—whether we estimate a more negative *Not a White Male* coefficient—when subjects are relatively more fatigued. In particular, we perform two sets of comparisons. First, we compare bias in the first and second half of the study under the hypothesis that employer might be more fatigued in the second half of the study. Second, we leverage the periodic breaks—after every 10 resumes that an employer completed—that we built into the survey tool.[53] Research suggests that such "micro breaks" can have relatively large effects on focus and attention [Rzeszotarski et al., 2013], and so we compare bias in the early half and latter half

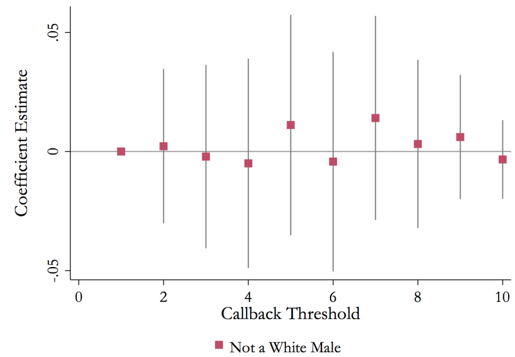
[52] For example, this may include an explicit belief that identical resume components signal different information for different types of candidates.

[53] As described in Section 2, after every 10 resumes an employer completed, the employer was shown a simple webpage with an affirmation that gave them a short break (e.g., after the first 10 resumes it read: "You have rated 10 of 40 resumes. Keep up the good work!").

Figure 2: Demographics by Major Type Over Selectivity Distribution

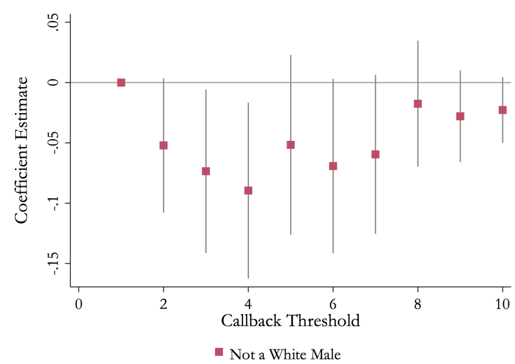


(a) Empirical CDF: Not a White Male, Humanities & Social Sciences



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

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



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

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

of each block of 10 resumes under the assumption that employers might be more fatigued in the latter half of each block of 10 resumes.

We report the results of both comparisons in Table 4. First, to validate our hypotheses that employers are more fatigued in the second half of the study and the latter half of each block of 10 resumes, we test whether the speed at which employers respond to a resume is faster during those parts of the study. The first and second columns of Table 4 show that, consistent with employers being fatigued—and thus having less mental energy to deliberate explicitly—subjects spend less time evaluating each resume in the second half of the study and in the latter half of each block of 10 resumes. The third and fourth columns of Table 4 show that the bias against candidates who are not White males is indeed statistically significantly larger in the latter half of each block of 10 resumes than the first half of each block. The third columns reports a statistically significant interaction on *Latter Half of Block* $\times$ *Not a White Male* of -0.385 Likert-scale points, equivalent to about 0.18 GPA points. The fourth columns reports, however, that the bias in the second half of the study is not statistically significantly larger than the bias in the first half.[54] These results provide suggestive, though not conclusive, evidence that the discrimination we detect may indeed be driven by implicit bias.

In addition, our cross-randomization of candidate names with other characteristics allows us to test an additional hypothesis from the resume audit literature in our data. Bertrand and Mullainathan [2004] finds that minority status decreases the return to resume quality, with smaller increases in callback rates for candidates with Black names between low and high quality resumes (compared to candidates with White names). We assign quality through cross-randomization of many individual characteristics (i.e., rather than updating multiple characteristics at once), so we perform this exercise by focusing on the binary variable that is the most predictive of hiring interest, *Top Internship*, and asking whether returns to having a top internship are lower for candidates who are not White males. Appendix Table B.7 reports on this exercise. The coefficient on *Top Internship* $\times$ *Not a White Male* is negative and at least marginally significant across all specifications, suggesting there is a lower return to a top internship for candidates who are not White males. One possible mechanism for this effect is that employers believe that other employers

[54] Note that Table 4 reports results among our full sample, where we do not find evidence of discrimination in the aggregate.

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.[55] This aligns with the effects over the distribution, shown in Appendix Figure B.3, where we see the penalty associated with not being a White male is larger and significantly different from zero at low thresholds, but a fairly precisely estimated zero at high thresholds. A candidate with sufficiently high observable quality may be inoculated against this particular form of bias. We discuss this result in Appendix B.3.2, and address our results in comparison with previous literature in Appendix B.3.3.

[55]To check whether this effect is absent for a marker of quality not evaluated by other employers, We also run this regression with a binary measure indicating high GPA. The interaction between *High GPA* and *Not a White Male* is negative in sign, but not statistically significant. The effect is smaller in magnitude than the interaction with *Top Employer*, although the two effects are not statistically different from one another.

Table 4: Implicit Bias

	Dependent Variable: Response Time		Dependent Variable: Hiring Interest	
Latter Half of Block	-3.518***		0.360***	
	(0.613)		(0.137)	
Second Half of Study		-4.668***		-0.142
		(0.598)		(0.138)
Not a White Male	-0.642	-0.648	0.0695	-0.107
	(0.666)	(0.665)	(0.115)	(0.118)
Latter Half of Block × Not a White Male			-0.385**	
			(0.165)	
Second Half of Study × Not a White Male				-0.0225
				(0.166)
GPA	2.791***	2.944***	2.187***	2.187***
	(0.961)	(0.949)	(0.128)	(0.128)
Top Internship	-0.799	-0.638	0.905***	0.904***
	(0.622)	(0.620)	(0.0802)	(0.0800)
Second Internship	2.163***	2.118***	0.471***	0.458***
	(0.752)	(0.750)	(0.0934)	(0.0934)
Work for Money	1.850**	1.813**	0.154*	0.140
	(0.741)	(0.740)	(0.0909)	(0.0910)
Technical Skills	0.881	0.892	-0.0668	-0.0780
	(0.715)	(0.713)	(0.0889)	(0.0890)
Observations	2880	2880	2880	2880
R^2	0.405	0.412	0.475	0.475
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes	Yes
Order FEs	No	No	No	No
Subject FEs	Yes	Yes	Yes	Yes

Standard errors in parentheses

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

Regressions of *Response Time* and *Hiring Interest* on resume characteristics and resume order variables. The first and second columns show *Response Time* regressions; the third and fourth columns show *Hiring Interest* regressions. *Response Time* is defined as the number of seconds before page submission, Winsorized at the 95^{th} percentile (77.9 seconds). Mean of *Response Time*: 23.6 seconds. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. *Latter Half of Block* is an indicator variable for resumes shown among the last five resumes within a 10-resume block. *Second Half of Study* is an indicator variable for resumes shown among the last 20 resumes viewed by a subject. Fixed effects for subjects, majors, and leadership experience 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.

3.5 Candidate Likelihood of Acceptance

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

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

Table 5 replicates the regression specifications from Table 2, estimating Equation (1) when V_{ij} is *Likelihood of Acceptance*, which takes values from 1 to 10. Table 5 shows two key features of likelihood of acceptance. First, we find that 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. Second, we find that firms believe that 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 represents a new channel for discrimination, as discussed below.

As noted above, the positive correlation between indicators of human capital—*GPA*, *Top Internship*, *Second Internship*, and *Work for Money*—and *Likelihood of Acceptance* suggests that part of the variation in likelihood of acceptance comes

Table 5: Likelihood of Acceptance

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.605***	0.632***	0.734***	0.263***
	(0.144)	(0.150)	(0.120)	(0.0601)
Top Internship	0.683***	0.678***	0.666***	0.284***
	(0.0942)	(0.0978)	(0.0763)	(0.0394)
Second Internship	0.419***	0.402***	0.393***	0.179***
	(0.112)	(0.119)	(0.0910)	(0.0467)
Work for Money	0.195*	0.190	0.200**	0.0878*
	(0.111)	(0.116)	(0.0895)	(0.0466)
Technical Skills	-0.0534	-0.0615	-0.105	-0.0258
	(0.104)	(0.108)	(0.0862)	(0.0440)
Not a White Male	-0.201**	-0.247**	-0.197**	-0.0867**
	(0.0981)	(0.102)	(0.0805)	(0.0413)
Observations	2880	2880	2880	2880
R^2	0.069	0.123	0.491	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: -.26, .13, .49, .74, 1.12, 1.49, 1.94, 2.46, and 2.83.

Table shows OLS and ordered probit regressions of likelihood of acceptance estimated from the specification in Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS, likelihood ratio test for ordered probit).

Table 6: Likelihood of Acceptance with Hiring Interest Controls

	Dependent Variable: Likelihood of Acceptance			
	OLS	Ordered Probit	OLS	Ordered Probit
GPA	-0.809***	-0.636***	-0.820***	-0.657***
	(0.0821)	(0.0641)	(0.0816)	(0.0646)
Top Internship	0.0318	-0.00194	0.0300	-0.00193
	(0.0535)	(0.0406)	(0.0535)	(0.0408)
Second Internship	0.0676	0.0526	0.0700	0.0507
	(0.0634)	(0.0477)	(0.0633)	(0.0480)
Work for Money	0.0953	0.0836*	0.0956	0.0879*
	(0.0610)	(0.0475)	(0.0609)	(0.0477)
Technical Skills	-0.0572	-0.0600	-0.0653	-0.0689
	(0.0596)	(0.0448)	(0.0595)	(0.0452)
Not a White Male	-0.115**	-0.0807*	-0.117**	-0.0819*
	(0.0544)	(0.0420)	(0.0543)	(0.0422)
Hiring Interest	0.703***	0.477***	FEs	FEs
	(0.0144)	(0.0104)		
Observations	2880	2880	2880	2880
R^2	0.766		0.767	
p-value for test of joint significance of Majors	0.028	< 0.001	0.034	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	No	Yes	No
Order FEs	Yes	No	Yes	No
Subject FEs	Yes	No	Yes	No

Cutpoints (Col 2): -1.83, -1.18, -0.55, -0.11, 0.49, 1.06, 1.71, 2.38, 2.80.

Cutpoints (Col 4): -2.00, -1.26, -0.59, -0.14, 0.45, 1.00, 1.62, 2.27, 2.69.

Table shows OLS and ordered probit regressions of likelihood of acceptance estimated from the specification in Equation (1), with additional controls for hiring interest. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS, likelihood ratio test for ordered probit).

from employers' beliefs that more qualified candidates will be more likely to accept a job with their firm. This pattern leads us to wonder how likelihood of acceptance responds to candidate characteristics after controlling for hiring interest.[56] We run regressions to answer this question in Table 6, focusing on the fully controlled OLS specification and the ordered probit specification. The first two columns of Table 6 include a linear control for *Hiring Interest*, which is estimated to be positive and significant (as expected given the results from Table 5). The latter two columns of Table 6 add flexible controls for hiring interest (i.e., dummies for each hiring interest rating). We find that after controlling for hiring interest, the relationship between *GPA* and *Likelihood of Acceptance* becomes negative and statistically significant under all specifications. Similarly, we find that the coefficients on other indicators of human capital (*Top Internship* and *Second Internship*) become statistically indistinguishable from zero. *Work for Money* is positive and marginally significant under some specifications. Under all specifications, however, after controlling for hiring interest, firms still believe candidates who are not White males are significantly less likely to accept job offers than their White male counterparts.

If minority and female applicants are perceived as less likely to accept an offer, this could induce lower callback rates for these candidates. This result suggests a new channel for discrimination observed in the labor market, and the causes of such a bias are worth exploring. Perhaps due to the proliferation 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.[57] Alternatively, employers may see female and minority candidates as less likely to fit in the culture of the firm, making these candidates less likely to accept an offer. This result has implications for how we understand the labor market and how we interpret the discrimination observed in resume audit studies.[58]

[56]The thought experiment is to imagine two candidates both of whom receive the same hiring interest rating from an employer and ask how their characteristics affect the employers' rating of their likelihood of acceptance.

[57]Our partners at the Penn Career Services Office interpreted the results in this light, based on anecdotal evidence and their previous conversations with employers.

[58]In particular, while audit studies can demonstrate that groups are not being treated equally, differential callback rates need not imply a lack of employer interest. In fact, if employers perceive likelihood of acceptance to vary by demographic group and by candidate quality, it could account for lower returns to quality for minority candidates (as identified by Bertrand and Mullainathan [2004]). This 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.

4 Pitt Replication: Results and Lessons

As discussed below in Section 5, one advantage of the IRR methodology is the relatively low marginal costs of running a similar study once the fixed costs to build the software for the survey tool have been invested. We leveraged this relatively low cost to repeat our experiment at a second school. In particular, we implemented IRR at the University of Pittsburgh ("Pitt") to explore the preferences of firms recruiting at a slightly less prestigious university.[59] One new finding from this replication study is that employers recruiting at Pitt care much more about horizontal differentiation (i.e., specific fit of the candidate for the job as indicated by, say, major) than employers recruiting at Penn. For example, on the post-survey, 33.7% of Pitt employers rated major a 10 out of 10 on its importance for candidate recruitment, whereas only 15.3% of Penn employers rated it that highly.

The two waves at Pitt in Spring 2017 (50 subjects, 2000 observations) and Spring 2018 (36 subjects, 1440 observations) provided a total of 86 subjects (3440 observations). Table 7 shows fully controlled OLS regressions and demonstrates 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 reasons we explore below.[60] There are also some interesting differences, such as that Pitt recruiters respond more to *Work for Money* relative to *Second Internship* than Penn employers. In fact, work-for-money jobs are not statistically different than second internships in the Pitt sample. This difference between Penn and Pitt is consistent with comments from the Pitt career services office that employers re-

[59]Our implementation at Pitt was very similar to our experimental waves at Penn. At Pitt, where the fall recruiting season plays a smaller role, 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 footnote 13, but since attendance at these sessions was low, we did not offer them in the second wave. We collected resume components to populate the tool at Pitt from real resumes of graduating Pitt seniors. Rather than collect resumes from clubs, resume books, and campus job postings as we did at Penn, we used the candidate pool of job-seeking seniors both to populate the tool and to suggest matches for employers. This significantly eased the burden of collecting and scraping resumes. At Pitt, majors were linked to either the "Dietrich School of Arts and Sciences" or the "Swanson School of Engineering". Table C.1 lists the majors, associated school, major category, and the probability that the major was drawn. We collected top internships at Pitt by identifying the firms hiring the most Pitt graduates, as at Penn. Top internships at Pitt tended to be less prestigious than the top internships at Penn.

[60]For example, the response to each GPA point is an order of magnitude smaller at Pitt than at Penn even though the standard deviation of hiring interest ratings and the real GPA distribution are similar at Pitt and Penn.

cruiting a Pitt are likely to particularly value the work ethic and customer service experience that students might acquire in work-for-money jobs.

We suspect that the smaller effect sizes observed at Pitt reflect attenuation due to the greater importance placed on horizontal differentiation.[61] We noticed this attention after the first wave at Pitt. To examine if the attenuation was due to horizontal match, in the second wave at Pitt, we added a question to the post-survey asking employers to indicate which academic majors they would consider to fill the position. We refer to these as *Target Majors*. Table 7 shows that when splitting 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.[62] The importance of careful tailoring is also a concern in resume audit studies, since poorly tailored resumes will generate a low callback rate and likely attenuate the effect of characteristics of interest.

5 Conclusion

This paper introduces a novel methodology, called Incentivized Resume Rating (IRR), to measure employer preferences. The method has employers rate candidate profiles they know to be hypothetical and provides incentives by matching employers to real job seekers based on their preferences. A key advantage of the IRR methodology is that it avoids the use of deception. As discussed in the Introduction, the expansion of resume audit studies—and other correspondence audit studies that

[61] Such attenuation would arise if employers were only interested in students from a particular major and thus did not respond to candidate GPA for candidates outside of their target majors.

[62] We advertised the survey tool at both Pitt and Penn as being particularly valuable for hiring skilled generalists, and we were ill equipped to find highly specialized candidates or those with very particular qualifications. This was a limitation in our application rather than in the methodology itself (e.g., one could design an IRR study specifically for employers interested in hiring registered nurses or candidates for another specialized position). We would recommend that future researchers using IRR either: target employers that specifically recruit high quality generalists or allow employers to indicate preferred candidate backgrounds and only show employers resumes that fall within their 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.

Table 7: Hiring Interest at Penn and Pitt

	Dependent Variable: Hiring Interest			
	Penn	Pitt	Pitt, Wave 2 Non-Target Major	Pitt, Wave 2 Target Major
GPA	2.195***	0.263**	-0.202	0.942***
	(0.129)	(0.113)	(0.238)	(0.267)
Top Internship	0.902***	0.222***	0.0202	0.0989
	(0.0806)	(0.0741)	(0.141)	(0.205)
Second Internship	0.463***	0.212**	0.0806	0.521**
	(0.0947)	(0.0844)	(0.164)	(0.219)
Work for Money	0.149	0.154*	0.130	0.397*
	(0.0913)	(0.0807)	(0.162)	(0.206)
Technical Skills	-0.0680	0.107	0.126	-0.0420
	(0.0900)	(0.0768)	(0.148)	(0.211)
Not a White Male	-0.117	0.00297	0.0311	-0.252
	(0.0842)	(0.0710)	(0.144)	(0.182)
Observations	2880	3440	642	798
R^2	0.482	0.586	0.793	0.595
p-value for test of joint significance of Majors	< 0.001	< 0.001	0.121	0.863
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes	Yes
Order FEs	Yes	Yes	Yes	Yes
Subject FEs	Yes	Yes	Yes	Yes

Table shows OLS regressions of hiring interest from Equation (1). Sample differs in each column as indicated by the column header. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in all specifications. R^2 is indicated for each OLS regression. The p-value of an F-test of joint significance of major fixed effects is indicated for all models.

also use deception—has raised additional concerns about the externalities associated with the use of deception. We see IRR as a valuable tool to explore employer preferences without expanding the use of deceptive research practice.

The IRR methodology has other key advantages that we leveraged to study employer preferences for candidates graduating from an Ivy League university. We find that these employers put a premium on evidence of educational success, including earning a high GPA and attending a selective school within Penn. Employers also value substantive work experience and experience at prestigious firms in particular. However, having held a work-for-money job does not increase average employer ratings of hiring interest. One interpretation of these results is that selective employers discriminate against working-class candidates, since these candidates may not have the resources to accept unpaid internships or internships that offer low pay. Another interpretation is that employers do not find value in work experience as a server or cashier (which may still result in a disadvantage for working-class candidates on the labor market). 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. We explore these effects across the range of hiring interest and show that the sign is consistent across the distribution.

We also highlight that a candidate's likelihood of acceptance might affect employer callback decisions. Employers in our sample report believing that female and minority candidates are less likely to accept job offers than their White male counterparts. This suggests a novel channel for discrimination. In a resume audit study, this mechanism would be indistinguishable from employers having a preference for White male candidates.

Running an IRR study involves initial set-up costs. We devoted significant time and resources to build the survey tool software, to construct the hypothetical resumes (e.g., to scrape and sanitize resume components), and to develop the process to match employer preferences to candidates. Fortunately, some of the investments we made can be leveraged by other researchers. And while there are some fixed costs of running IRR studies per research team, we found that the marginal cost of building the survey tool for a new setting—as we did in our replication at Pitt—is much more manageable.[63]

[63] Once the resume tool was up and running at Penn, the marginal cost for Pitt was mostly limited to constructing Pitt-specific hypothetical resumes.

These costs provide clear benefits. In addition to delivering rich data on employer preferences, IRR also provides value to participants. This value allowed us to recruit employers from selective firms who are unlikely to be accessible with audit or resume audit methods (e.g., because they do not accept unsolicited resumes). We thus believe the advantages are large enough—and the underlying methodology flexible enough—to justify using IRR to explore employer preferences for candidate characteristics in other settings and to fruitfully answer a variety of questions in labor economics.[64] Any researcher who has access to a subject pool of employers and a pool of real job seekers can run an incentivized resume rating study of their own.

[64] The general method of incentivizing subjects can also be used in other domains as shown by Low [2017].

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

FOR ONLINE PUBLICATION ONLY

Appendices

We provide three appendices. In Appendix A, we describe the design of our experiment in detail, including recruitment materials (A.1), survey tool construction (A.2), and the candidate matching process (A.3). In Appendix B, we present additional analyses and results, including human capital results (B.1), regressions weighted by GPA (B.2), a discussion of our discrimination results (B.3), and a discussion of preferences over the quality distribution (B.4). In Appendix C, we discuss additional details related to replicating our experiment at Pitt.

A Experimental Design Appendix

A.1 Recruitment Materials

University of Pennsylvania Career Services sent recruitment materials to both recruiting firms and graduating seniors to participate in the study. All materials marketed the study as an additional tool to connect students with firms, rather than a replacement for any usual recruiting efforts. The recruitment email for employers, shown in Figure A.1, was sent to a list of contacts maintained by Career Services and promised to use a "newly developed machine-learning algorithm to identify candidates who would be a particularly good fit for your job based on your evaluations." In our replication at the University of Pittsburgh, a similar email was sent from the Pitt Office of Career Development and Placement Assistance.

Penn Career Services recruited graduating seniors to participate as part of the candidate matching pool through their regular newsletter called the "Friday Flash." The relevant excerpt from this email newsletter is shown in Figure A.2.

A.2 Survey Tool Design

In this appendix, we describe the process of generating hypothetical resumes. This appendix should serve to provide additional details about the selection and randomization of resume components, and as a guide to researchers wishing to im-

Figure A.1: Employer Recruitment Email

From: upenn@csm.symplicity.com [mailto:upenn@csm.symplicity.com]
Sent: Tuesday, July 26, 2016 1:34 PM
To: [redacted]
Subject: Identify Top Penn Students for your Firm

Dear [redacted]

This year, Penn Career Services is participating in a pilot with two Wharton professors who are developing a new tool that can help you to identify potential job candidates from the University of Pennsylvania for post-graduate positions.

The tool is designed to identify top candidates for your open positions and provides you with those candidates' contact information and resumes so you can invite them to coffee chats, to info sessions, and to apply for a job at your organization. Since the tool uses data-driven methods to identify candidates, we see this as a useful complement to firms' existing methods for identifying promising candidates.

Completing the tool takes about 30 minutes and involves evaluating 40 hypothetical resumes. After evaluating these resumes, 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. The Wharton professors will also use a completely anonymized version of your data to perform research on broader trends in what firms value in hiring, and they will be glad to share these insights with your company once the research is complete. To be provided with potential candidates for a position, at least one person from your firm must complete the tool. If possible, having multiple individuals participate will help increase the accuracy of the algorithm's recommendations. Additionally, if you are hiring for different positions within your organization, we recommend at least one person from your organization take the tool for each open position so you get a list of candidates tailored for each job opening. Rising Penn seniors will be invited to participate in the trial by submitting their resumes beginning on August 22nd, and we plan to have candidate recommendations to you by early September.

To take the tool, please click the link here:

https://wharton.qualtrics.com/SE/?SID=SV_3I3ohtNPn2R8c97

If you you would like to discuss more about how the tool could be useful for your firm, or have any questions, please contact the Wharton researchers: Judd B. Kessler (judd.kessler@wharton.upenn.edu) and Corinne Low (corlow@wharton.upenn.edu).

Sincerely,

Barbara Hewitt, Senior Associate Director, Career Services

Email sent to firms recruiting at Penn originating from the Senior Associate Director of Career Services at the University of Pennsylvania. Subjects who followed the link in the email were taken to the instructions (Figure A.3).

Figure A.2: Email Announcement to Graduating Seniors

From: Career Services - Wharton Class of 2017 <CAREERSERVICES2017@LISTS.UPENN.EDU> On Behalf Of Ross, S. David
Sent: Friday, August 26, 2016 5:20 PM
To: CAREERSERVICES2017@LISTS.UPENN.EDU
Subject: Wharton Seniors: Penn Career Services Senior Friday Flash, August 26, 2016

Welcome back! I hope you had a wonderful and productive summer. This is the first issue of the senior Career Services Friday Flash for the year. Barbara Hewitt is the Senior Associate Director in the Career Services office working with Wharton undergraduate students and alumni - she will manage the Career Services listserv for Wharton seniors and will be sending you weekly Friday Flash e-mails to keep you updated on workshops, job postings, employer presentations, career resources and more. Barbara and I look forward to working with you this year as you begin (or continue!) to think about life after Penn. Please do come in to speak with either of us about your plans. Also, please note that On Campus Recruiting activities have started, so don't delay if you would like to participate!

[OTHER TEXT APPEARED HERE]

Announcements

An Opportunity To Reach More Employers

This year, Penn Career Services is working with two Wharton professors on a pilot that can help you get noticed by top employers in all fields. Wharton professors Judd B. Kessler and Corinne Low have developed a tool that analyzes employer preferences for job candidates and then uses machine learning to identify Penn seniors who may be a good fit for the employer's positions. Employers across a variety of industries (e.g. consulting, finance, technology, etc.) have already participated in the pilot by providing preferences for job candidates. Upload your resume now to be eligible to participate! Only candidates who upload their resume through this link can participate in the pilot. To upload your resume, click here: https://wharton.qualtrics.com/SE/?SID=SV_bryPbgBn4rEXD0h. If you have any questions about the pilot, please contact the Wharton professors running it: Judd B. Kessler (judd.kessler@wharton.upenn.edu) and Corinne Low (corlow@wharton.upenn.edu). (Note: this pilot will be run in parallel to all existing recruiting activities.)

Excerpt from email newsletter sent to the Career Services office mailing list. The email originated from the Senior Associate Director of Career Services at the University of Pennsylvania. Students following the link were taken to a survey page where they were asked to upload their resumes and to answer a brief questionnaire about their job search (page not shown).

plement our methodology. In Section A.2.1, we describe the structure of the IRR survey tool and participant experience. In Section A.2.2, we describe the structure of our hypothetical resumes. In Section A.2.3, we detail the randomization of candidate gender and race through names. Section A.2.4 details the randomization of educational background. Section A.2.5 describes the process we used to collect and scrape real resume components to randomize work experience, leadership experience, and skills.

A.2.1 Survey Tool Structure

We constructed the survey tool using Qualtrics software for respondents to access from a web browser. Upon opening the survey link, respondents must enter an email address on the instructions page (see Figure A.3) to continue. Respondents then select the type of candidates they will evaluate for their open position, either "Business (Wharton), Social Sciences, and Humanities" or "Science, Engineering, Computer Science, and Math." In addition, they may enter the position title they are looking to fill. The position title is not used in determining the content of the hypothetical candidate resumes. The major selection page is shown in Figure A.4. After this selection, the randomization software populates 40 resumes for the respondent to evaluate, drawing on different content by major type. The subject then evaluates 40 hypothetical resumes. After every 10 resumes, a break page encourages subjects to continue.

A.2.2 Resume Structure

We designed our resumes to combine realism with the requirements of experimental identification. We designed 10 resume templates to use as the basis for the 40 resumes in the tool. Each template presented the same information, in the same order, but with variations in page layout and font. Figures A.5 and A.6 show sample resume templates. All resumes contained five sections, in the following order: Personal Information (including name and blurred contact information); Education (GPA, major, school within university); Work Experience; Leadership Experience; and Skills.[65] While the real student resumes we encountered varied in content,

[65] These sections were not always labelled as such on candidate resumes. Personal Information was generally not identified, though each resume contained a name and blurred text in place of contact information. Skills were also marked as "Skills & Interests" and "Skill Summary".

Figure A.3: Survey Tool Instructions & Contact Information

Wharton
UNIVERSITY *of* PENNSYLVANIA

Wharton Professors Judd B. Kessler and Corinne Low are using a new survey tool to help understand what firms value in job candidates while matching firms with real candidates using machine learning.

The survey tool will ask you to evaluate resumes of hypothetical candidates. While the candidates shown are not real, your choices will be used to provide you with recommendations of *actual* Penn students who might be a good fit for your organization. The more carefully you complete the survey, the better we will be able to match you with these Penn students.

In this survey tool, you will be shown 40 resumes of hypothetical job candidates and asked to evaluate:
(a) how interested you would be in hiring the candidate; and
(b) how likely the candidate would be to accept a job at your organization

Importantly, when gauging your interest level in the candidate, imagine that the candidate is guaranteed to accept your job offer — think only about your perception of the candidate's quality.

When gauging how likely the candidate would be to accept a job, imagine that the candidate has been given a job — think only about whether you think the candidate would accept the job.

We will use both answers to recommend Penn students for you who we estimate to be strong candidates for your position(s).

After you finish evaluating the resumes, you will be asked to complete a few more questions to help us better understand your firm.

All of your data will be kept strictly confidential. After we send you resumes and contact information of the Penn students we recommend based on your responses, your data will be stripped of identifying information when used for research purposes.

If you have any questions, please contact the Wharton researchers who have built the tool: Judd B. Kessler (judd.kessler@wharton.upenn.edu) and Corinne Low (corlow@wharton.upenn.edu).

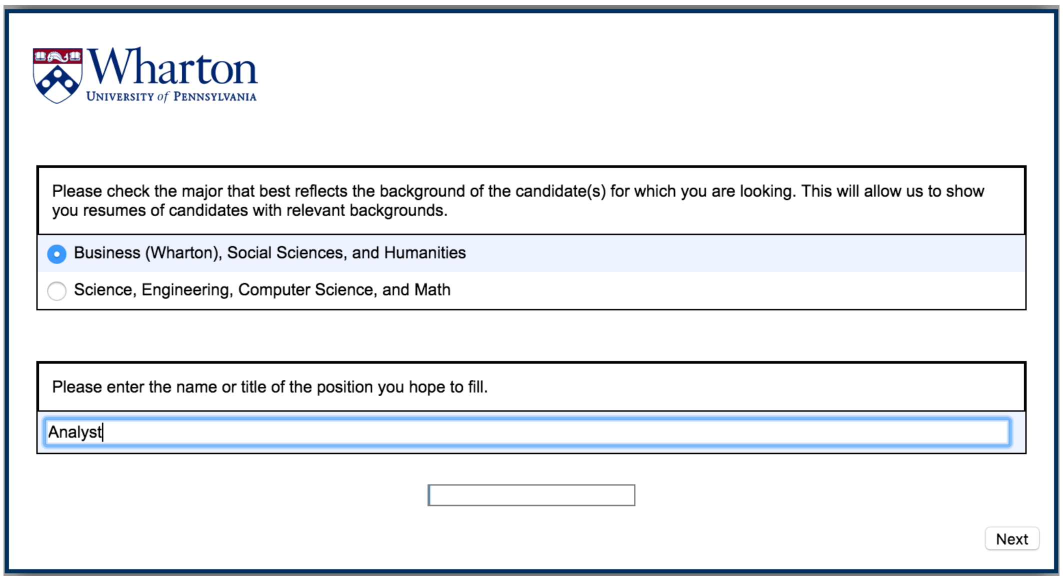
In the field below, please enter the email address where you would like to receive candidate information and then click the "Next" button to continue with the study.

(The tool will take approximately 30 minutes to complete. If you would like to return to complete the survey later, be sure to resume on the same device to avoid losing your data.)

Next

Screenshot of the instructions at the start of the survey tool. This page provided information to subjects and served as instructions. Subjects entered an email address at the bottom of the screen to proceed with the study; the resumes of the 10 real job seekers used as an incentive to participate are sent to this email address.

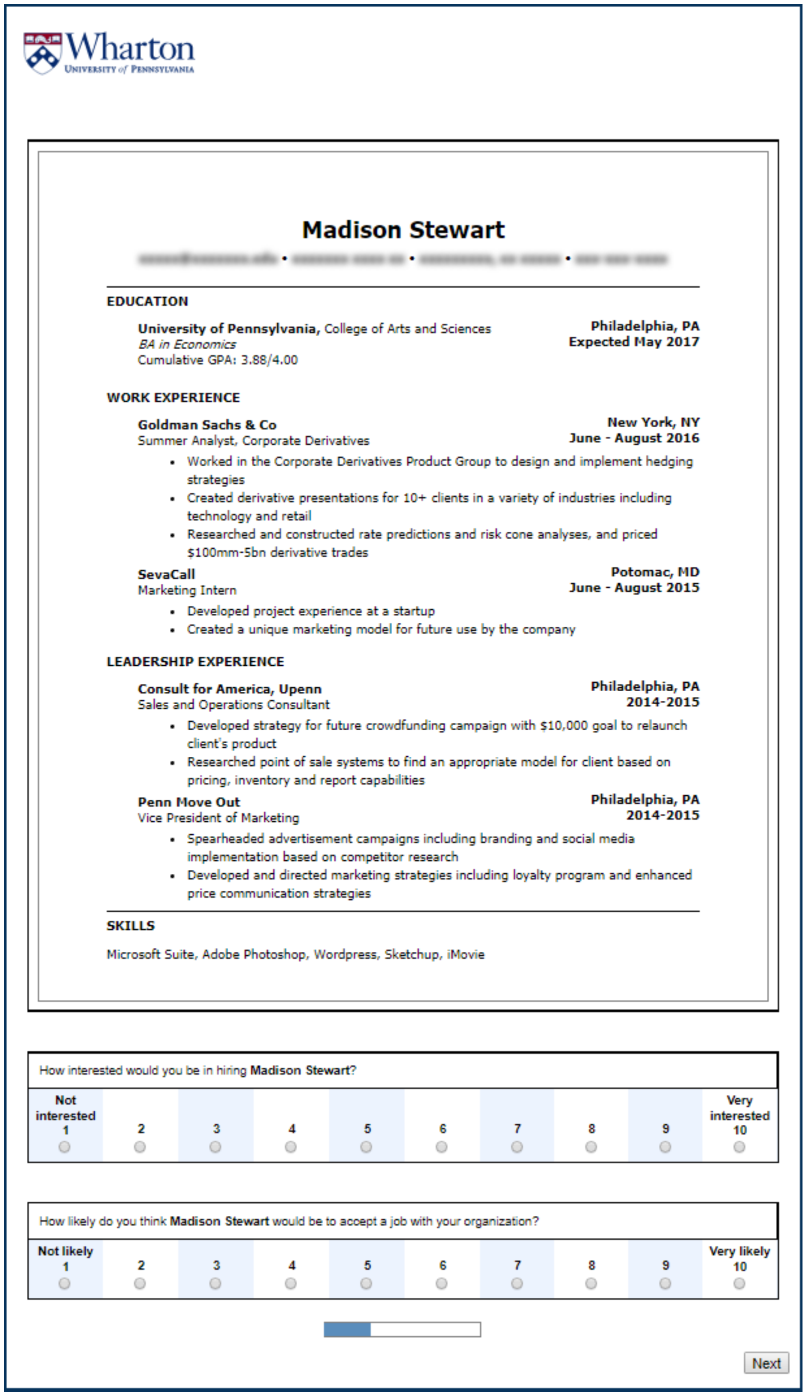
Figure A.4: Major Type Selection



Screenshot of major selection page, as shown to subjects recruiting at the University of Pennsylvania. Subjects must select either Business (Wharton), Social Sciences, and Humanities, or Science, Engineering, Computer Science, and Math. Subjects may also enter the name of the position they wish to fill in the free text box; the information in this box was not used for analysis. Here, we have selected Business (Wharton), Social Sciences, and Humanities and entered "Analyst" as a demonstration only—by default all radio boxes and text boxes were empty for all subjects.

most contained some subset of these sections. Since our main objective with resume variation was to improve realism for each subject rather than to test the effectiveness of different resume formats, we did not vary the order of the resume formats across subjects. In other words, the first resume always had the same font and page layout for each subject, although the content of the resume differed each time. Given that formats are in a fixed order in the 40 hypothetical resumes, the order fixed effects included in most specifications control for any effect of resume format. Resumes templates were built in HTML/CSS for display in a web browser, and populated dynamically in Qualtrics using JavaScript. Randomization occurred for all 40 resumes simultaneously, without replacement, each time a subject completed the instructions and selected their major category of interest.

Figure A.5: Sample Resume



Wharton
UNIVERSITY of PENNSYLVANIA

Madison Stewart

EDUCATION

University of Pennsylvania, College of Arts and Sciences — **Philadelphia, PA**
BA in Economics — **Expected May 2017**
Cumulative GPA: 3.88/4.00

WORK EXPERIENCE

Goldman Sachs & Co — **New York, NY**
Summer Analyst, Corporate Derivatives — **June - August 2016**

- Worked in the Corporate Derivatives Product Group to design and implement hedging strategies
- Created derivative presentations for 10+ clients in a variety of industries including technology and retail
- Researched and constructed rate predictions and risk cone analyses, and priced $100mm-5bn derivative trades

SevaCall — **Potomac, MD**
Marketing Intern — **June - August 2015**

- Developed project experience at a startup
- Created a unique marketing model for future use by the company

LEADERSHIP EXPERIENCE

Consult for America, Upenn — **Philadelphia, PA**
Sales and Operations Consultant — **2014-2015**

- Developed strategy for future crowdfunding campaign with $10,000 goal to relaunch client's product
- Researched point of sale systems to find an appropriate model for client based on pricing, inventory and report capabilities

Penn Move Out — **Philadelphia, PA**
Vice President of Marketing — **2014-2015**

- Spearheaded advertisement campaigns including branding and social media implementation based on competitor research
- Developed and directed marketing strategies including loyalty program and enhanced price communication strategies

SKILLS

Microsoft Suite, Adobe Photoshop, Wordpress, Sketchup, iMovie

How interested would you be in hiring **Madison Stewart**?

Not interested 1	2	3	4	5	6	7	8	9	Very interested 10
○	○	○	○	○	○	○	○	○	○

How likely do you think **Madison Stewart** would be to accept a job with your organization?

Not likely 1	2	3	4	5	6	7	8	9	Very likely 10
○	○	○	○	○	○	○	○	○	○

Next

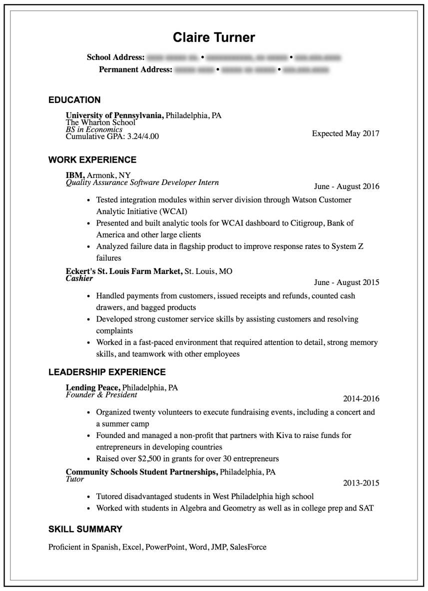
A sample resume rating page from the Incentivized Resume Rating tool. Each resume is dynamically generated when the subject begins the study. Each resume has five sections: Personal Information (including first and last name, and blurred text to represent contact information); Education Information (university, school within university, degree, major, GPA, and expected graduation date); Work Experience (one or two experiences with employer name, location, job title, date, and descriptive bullet points); Leadership Experience (two experiences with organization, location, position title, date, and descriptive bullet points); and Skills. Resume randomization described in detail in Section 2 and Appendix A.2. At the bottom of each resume, subjects must respond to two questions before proceeding: "How interested would you be in hiring [**Name**]?" and "How likely do you think [**Name**] would be to accept a job with your organization?"

A.2.3 Names

A hypothetical candidate name appears as the first element on each resume. Names were generated to be highly indicative of race and gender, following the approach of Fryer and Levitt [2004]. As described in Section 2.3.4, first names were selected from a dataset of all births in the state of Massachusetts between 1989-1996 and in New York City between 1990-1996. These years reflect the approximate birth years of the job seekers in our study. We identified 100 first names with the most indicative race and gender for each of the following race-gender combinations: Asian Female, Asian Male, Black Female, Black Male, Hispanic Female, Hispanic Male, White Female, and White Male. We then eliminated names that were gender-ambiguous in the broad sample even if they might be unambiguous within an ethnic group. We also eliminated names strongly indicative of religion. We followed a similar process for last names, using name and ethnicity data from the 2000 Census. Finally, we paired first and last names together by race and selected 50 names for each race-gender combination for randomization. Names of hypothetical female candidates are shown in Table A.1; names of hypothetical male candidates are shown in Table A.2.

At the point of randomization, names were drawn without replacement according to a distribution of race and gender intended to reflect the US population (50% female, 50% male; 65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian). Gender and race were randomized independently. In other words, we selected either Table A.1 or Table A.2 with equal probability, then selected a column to draw from according to the race probabilities. Finally, names were selected uniformly and without replacement from the appropriate column of the table. We use the variation induced by these names for the analysis variable *Not a White Male*.

Figure A.6: Four Sample Resumes



Claire Turner

School Address:
Permanent Address:

EDUCATION

University of Pennsylvania, Philadelphia, PA
The Wharton School
BS in Economics
Cumulative GPA: 3.24/4.00 Expected May 2017

WORK EXPERIENCE

IBM, Armonk, NY
Quality Assurance Software Developer Intern June - August 2016
- Tested integration modules within server division through Watson Customer Analytic Initiative (WCAI)
- Presented and built analytic tools for WCAI dashboard to Citigroup, Bank of America and other large clients
- Analyzed failure data in flagship product to improve response rates to System Z failures

Eckert's St. Louis Farm Market, St. Louis, MO
Cashier June - August 2015
- Handled payments from customers, issued receipts and refunds, counted cash drawers, and bagged products
- Developed strong customer service skills by assisting customers and resolving complaints
- Worked in a fast-paced environment that required attention to detail, strong memory skills, and teamwork with other employees

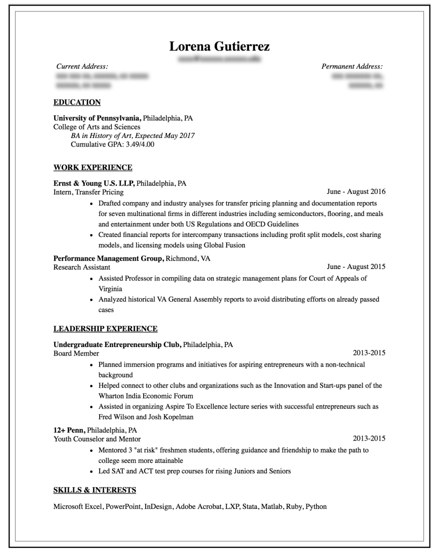
LEADERSHIP EXPERIENCE

Lending Peace, Philadelphia, PA
Founder & President 2014-2016
- Organized twenty volunteers to execute fundraising events, including a concert and a summer camp
- Founded and managed a non-profit that partners with Kiva to raise funds for entrepreneurs in developing countries
- Raised over $2,500 in grants for over 30 entrepreneurs

Community Schools Student Partnerships, Philadelphia, PA
Tutor 2013-2015
- Tutored disadvantaged students in West Philadelphia high school
- Worked with students in Algebra and Geometry as well as in college prep and SAT

SKILL SUMMARY

Proficient in Spanish, Excel, PowerPoint, Word, JMP, SalesForce



Lorena Gutierrez

Current Address: *Permanent Address:*

EDUCATION

University of Pennsylvania, Philadelphia, PA
College of Arts and Sciences
BA in History of Art, Expected May 2017
Cumulative GPA: 3.49/4.00

WORK EXPERIENCE

Ernst & Young U.S. LLP, Philadelphia, PA
Intern, Transfer Pricing June - August 2016
- Drafted company and industry analyses for transfer pricing planning and documentation reports for seven multinational firms in different industries including semiconductors, flooring, and meals and entertainment under both US Regulations and OECD Guidelines
- Created financial reports for intercompany transactions including profit split models, cost sharing models, and licensing models using Global Fusion

Performance Management Group, Richmond, VA
Research Assistant June - August 2015
- Assisted Professor in compiling data on strategic management plans for Court of Appeals of Virginia
- Analyzed historical VA General Assembly reports to avoid distributing efforts on already passed cases

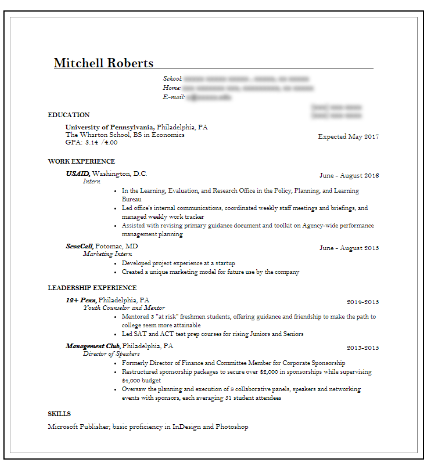
LEADERSHIP EXPERIENCE

Undergraduate Entrepreneurship Club, Philadelphia, PA
Board Member 2013-2015
- Planned immersion programs and initiatives for aspiring entrepreneurs with a non-technical background
- Helped connect to other clubs and organizations such as the Innovation and Start-ups panel of the Wharton India Economic Forum
- Assisted in organizing Aspire To Excellence lecture series with successful entrepreneurs such as Fred Wilson and Josh Kopelman

12+ Penn, Philadelphia, PA
Youth Counselor and Mentor 2013-2015
- Mentored 3 "at risk" freshmen students, offering guidance and friendship to make the path to college seem more attainable
- Led SAT and ACT test prep courses for rising Juniors and Seniors

SKILLS & INTERESTS

Microsoft Excel, PowerPoint, InDesign, Adobe Acrobat, LXP, Stata, Matlab, Ruby, Python



Mitchell Roberts

School:
Home:
E-mail:

EDUCATION

University of Pennsylvania, Philadelphia, PA
The Wharton School, BS in Economics Expected May 2017
GPA: 3.14 /4.00

WORK EXPERIENCE

USAID, Washington, D.C. June - August 2016
Intern
- In the Learning, Evaluation, and Research Office in the Policy, Planning, and Learning Bureau
- Led office's internal communications, coordinated weekly staff meetings and briefings, and managed weekly work tracker
- Assisted with revising primary guidance document and toolkit on Agency-wide performance management planning

SevaCall, Potomac, MD June - August 2015
Marketing Intern
- Developed project experience at a startup
- Created a unique marketing model for future use by the company

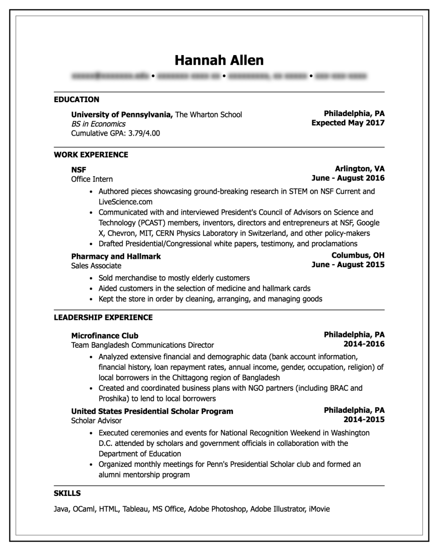
LEADERSHIP EXPERIENCE

12+ Penn, Philadelphia, PA 2014-2015
Youth Counselor and Mentor
- Mentored 3 "at risk" freshmen students, offering guidance and friendship to make the path to college seem more attainable
- Led SAT and ACT test prep courses for rising Juniors and Seniors

Management Club, Philadelphia, PA 2013-2015
Director of Speakers
- Formerly Director of Finance and Committee Member for Corporate Sponsorship
- Restructured sponsorship packages to secure over $2,000 in sponsorships while supervising $4,000 budget
- Oversaw the planning and execution of 8 collaborative panels, speakers and networking events with sponsors, each averaging 31 student attendees

SKILLS

Microsoft Publisher; basic proficiency in InDesign and Photoshop



Hannah Allen

EDUCATION

University of Pennsylvania, The Wharton School **Philadelphia, PA**
BS in Economics **Expected May 2017**
Cumulative GPA: 3.79/4.00

WORK EXPERIENCE

NSF **Arlington, VA**
Office Intern **June - August 2016**
- Authored pieces showcasing ground-breaking research in STEM on NSF Current and LiveScience.com
- Communicated with and interviewed President's Council of Advisors on Science and Technology (PCAST) members, inventors, directors and entrepreneurs at NSF, Google X, Chevron, MIT, CERN Physics Laboratory in Switzerland, and other policy-makers
- Drafted Presidential/Congressional white papers, testimony, and proclamations

Pharmacy and Hallmark **Columbus, OH**
Sales Associate **June - August 2015**
- Sold merchandise to mostly elderly customers
- Aided customers in the selection of medicine and hallmark cards
- Kept the store in order by cleaning, arranging, and managing goods

LEADERSHIP EXPERIENCE

Microfinance Club **Philadelphia, PA**
Team Bangladesh Communications Director **2014-2016**
- Analyzed extensive financial and demographic data (bank account information, financial history, loan repayment rates, annual income, gender, occupation, religion) of local borrowers in the Chittagong region of Bangladesh
- Created and coordinated business plans with NGO partners (including BRAC and Proshika) to lend to local borrowers

United States Presidential Scholar Program **Philadelphia, PA**
Scholar Advisor **2014-2015**
- Executed ceremonies and events for National Recognition Weekend in Washington D.C. attended by scholars and government officials in collaboration with the Department of Education
- Organized monthly meetings for Penn's Presidential Scholar club and formed an alumni mentorship program

SKILLS

Java, OCaml, HTML, Tableau, MS Office, Adobe Photoshop, Adobe Illustrator, iMovie

Four sample resumes generated by the survey tool. Note that the resumes each have a different format, differentiated by elements such as font, boldface type, horizontal rules, location of information, and spacing. All resumes have the same five sections: Personal Information, Education, Work Experience, Leadership Experience, and Skills. Resumes differ in length based on the dynamically selected content, such as the randomized number of work experiences and the (non-randomized) number of description bullet points associated with an experience.

Table A.1: Female Names Populating Resume Tool

Asian Female	**Black Female**	**Hispanic Female**	**White Female**
Tina Zheng	Jamila Washington	Ivette Barajas	Allyson Wood
Annie Xiong	Asia Jefferson	Nathalie Orozco	Rachael Sullivan
Julie Xu	Essence Banks	Mayra Zavala	Katharine Myers
Michelle Zhao	Monique Jackson	Luisa Velazquez	Colleen Peterson
Linda Zhang	Tianna Joseph	Jessenia Meza	Meghan Miller
Anita Zhu	Janay Mack	Darlene Juarez	Meaghan Murphy
Alice Jiang	Nia Williams	Thalia Ibarra	Lindsey Fisher
Esther Zhou	Latoya Robinson	Perla Cervantes	Paige Cox
Winnie Thao	Jalisa Coleman	Lissette Huerta	Katelyn Cook
Susan Huang	Imani Harris	Daisy Espinoza	Jillian Long
Sharon Yang	Malika Sims	Cristal Vazquez	Molly Baker
Gloria Hwang	Keisha James	Paola Cisneros	Heather Nelson
Diane Ngo	Shanell Thomas	Leticia Gonzalez	Alison Hughes
Carmen Huynh	Janae Dixon	Jesenia Hernandez	Bridget Kelly
Angela Truong	Latisha Daniels	Alejandra Contreras	Hayley Russell
Janet Kwon	Zakiya Franklin	Iliana Ramirez	Carly Roberts
Janice Luong	Kiana Jones	Julissa Esparza	Bethany Phillips
Irene Cheung	Ayana Grant	Giselle Alvarado	Kerry Bennett
Amy Choi	Ayanna Holmes	Gloria Macias	Kara Morgan
Shirley Yu	Shaquana Frazier	Selena Zuniga	Kaitlyn Ward
Kristine Nguyen	Shaniqua Green	Maribel Ayala	Audrey Rogers
Cindy Wu	Tamika Jenkins	Liliana Mejia	Jacquelyn Martin
Joyce Vu	Akilah Fields	Arlene Rojas	Marissa Anderson
Vivian Hsu	Shantel Simmons	Cristina Ochoa	Haley Clark
Jane Liang	Shanique Carter	Yaritza Carillo	Lindsay Campbell
Maggie Tsai	Tiara Woods	Guadalupe Rios	Cara Adams
Diana Pham	Tierra Bryant	Angie Jimenez	Jenna Morris
Wendy Li	Raven Brown	Esmeralda Maldonado	Caitlin Price
Sally Hoang	Octavia Byrd	Marisol Cardenas	Kathryn Hall
Kathy Duong	Tyra Walker	Denisse Chavez	Emma Bailey
Lily Vang	Diamond Lewis	Gabriela Mendez	Erin Collins
Helen Trinh	Nyasia Johnson	Jeanette Rosales	Marisa Reed
Sandy Oh	Aliyah Douglas	Rosa Castaneda	Madeleine Smith
Christine Tran	Aaliyah Alexander	Beatriz Rodriguez	Mackenzie King
Judy Luu	Princess Henderson	Yessenia Acevedo	Sophie Thompson
Grace Cho	Shanae Richardson	Carolina Guzman	Madison Stewart
Nancy Liu	Kenya Brooks	Carmen Aguilar	Margaret Parker
Lisa Cheng	Charisma Scott	Yesenia Vasquez	Kristin Gray
Connie Yi	Shante Hunter	Ana Munoz	Michaela Evans
Tiffany Phan	Jada Hawkins	Xiomara Ortiz	Jaclyn Cooper
Karen Lu	Shanice Reid	Lizbeth Rivas	Hannah Allen
Tracy Chen	Chanelle Sanders	Genesis Sosa	Zoe Wilson
Betty Dinh	Shanequa Bell	Stephany Salinas	Caitlyn Young
Anna Hu	Shaniece Mitchell	Lorena Gutierrez	Charlotte Moore
Elaine Le	Ebony Ford	Emely Sandoval	Kaitlin Wright
Sophia Ly	Tanisha Watkins	Iris Villarreal	Holly White
Jenny Vo	Shanelle Butler	Maritza Garza	Kate Taylor
Monica Lin	Precious Davis	Marilyn Arroyo	Krista Hill
Joanne Yoon	Asha Willis	Lourdes Soto	Meredith Howard
Priya Patel	Ashanti Edwards	Gladys Herrera	Claire Turner

Names of hypothetical female candidates. 50 names were selected to be highly indicative of each combination of race and gender. A name drawn from these lists was displayed at the top of each hypothetical resume, and in the questions used to evaluate the resumes. First and last names were linked every time they appeared. For details on the construction and randomization of names, see Section 2.3.4 and Appendix A.2.3.

Table A.2: Male Names Populating Resume Tool

Asian Male	**Black Male**	**Hispanic Male**	**White Male**
Richard Thao	Rashawn Washington	Andres Barajas	Kyle Wood
Samuel Truong	Devonte Jefferson	Julio Orozco	Derek Sullivan
Daniel Cheung	Marquis Banks	Marcos Zavala	Connor Myers
Alan Tsai	Tyree Jackson	Mike Velazquez	Douglas Peterson
Paul Li	Lamont Joseph	Jose Meza	Spencer Miller
Steven Zhang	Jaleel Mack	Alfredo Juarez	Jackson Murphy
Matthew Zheng	Javon Williams	Fernando Ibarra	Bradley Fisher
Alex Vu	Darryl Robinson	Gustavo Cervantes	Drew Cox
Joshua Vo	Kareem Coleman	Adonis Huerta	Lucas Cook
Brandon Lu	Kwame Harris	Juan Espinoza	Evan Long
Henry Dinh	Deshawn Sims	Jorge Vazquez	Adam Baker
Philip Hsu	Terrell James	Abel Cisneros	Harrison Nelson
Eric Liang	Akeem Thomas	Cesar Gonzalez	Brendan Hughes
David Yoon	Daquan Dixon	Alberto Hernandez	Cody Kelly
Jonathan Yu	Tarik Daniels	Elvin Contreras	Zachary Russell
Andrew Trinh	Jaquan Franklin	Ruben Ramirez	Mitchell Roberts
Stephen Yi	Tyrell Jones	Reynaldo Esparza	Tyler Phillips
Ryan Nguyen	Isiah Grant	Wilfredo Alvarado	Matthew Bennett
Aaron Jiang	Omari Holmes	Francisco Macias	Thomas Morgan
Kenneth Zhao	Rashad Frazier	Emilio Zuniga	Sean Ward
Johnny Hwang	Jermaine Green	Javier Ayala	Nicholas Rogers
Tony Choi	Donte Jenkins	Guillermo Mejia	Brett Martin
Benjamin Luong	Donnell Fields	Elvis Rojas	Cory Anderson
Raymond Tran	Davon Simmons	Miguel Ochoa	Colin Clark
Michael Duong	Darnell Carter	Sergio Carillo	Jack Campbell
Andy Hoang	Hakeem Woods	Alejandro Rios	Ross Adams
Alexander Pham	Sheldon Bryant	Ernesto Jimenez	Liam Morris
Robert Yang	Antoine Brown	Oscar Maldonado	Max Price
Danny Xu	Marquise Byrd	Felix Cardenas	Ethan Hall
Anthony Huynh	Tyrone Walker	Manuel Chavez	Eli Bailey
Jason Liu	Dashawn Lewis	Orlando Mendez	Patrick Collins
John Chen	Shamel Johnson	Luis Rosales	Luke Reed
Brian Vang	Reginald Douglas	Eduardo Castaneda	Alec Smith
Joseph Zhou	Shaquille Alexander	Carlos Rodriguez	Seth King
James Cho	Jamel Henderson	Cristian Acevedo	Austin Thompson
Nicholas Lin	Akil Richardson	Pedro Guzman	Nathan Stewart
Jeffrey Huang	Tyquan Brooks	Freddy Aguilar	Jacob Parker
Christopher Wu	Jamal Scott	Esteban Vasquez	Craig Gray
Timothy Ly	Jabari Hunter	Leonardo Munoz	Garrett Evans
William Oh	Tyshawn Hawkins	Arturo Ortiz	Ian Cooper
Patrick Ngo	Demetrius Reid	Jesus Rivas	Benjamin Allen
Thomas Cheng	Denzel Sanders	Ramon Sosa	Conor Wilson
Vincent Le	Tyreek Bell	Enrique Salinas	Jared Young
Kevin Hu	Darius Mitchell	Hector Gutierrez	Theodore Moore
Jimmy Xiong	Prince Ford	Armando Sandoval	Shane Wright
Justin Zhu	Lamar Watkins	Roberto Villarreal	Scott White
Calvin Luu	Raheem Butler	Edgar Garza	Noah Taylor
Edward Kwon	Jamar Davis	Pablo Arroyo	Ryan Hill
Peter Phan	Tariq Willis	Raul Soto	Jake Howard
Victor Patel	Shaquan Edwards	Diego Herrera	Maxwell Turner

Names of hypothetical male candidates. 50 names were selected to be highly indicative of each combination of race and gender. A name drawn from these lists was displayed at the top of each hypothetical resume, and in the questions used to evaluate the resumes. First and last names were linked every time they appeared. For details on the construction and randomization of names, see Section 2.3.4 and Appendix A.2.3.

A.2.4 Education

We randomized two components in the Education section of each resume: grade point average (GPA) and major. We also provided an expected graduation date (fixed to May 2017 for all students), the name of the university (University of Pennsylvania), the degree (BA or BS) and the name of the degree-granting school within Penn to maintain realism.

GPA We selected GPA from a $Unif[2.90, 4.00]$ distribution, rounding to the nearest hundredth. We chose to include GPA on all resumes, although some students omit GPA on real resumes. We decided to avoid the complexity of forcing subjects to make inferences about missing GPAs. The range was selected to approximate the range of GPAs observed on real resumes. We chose a uniform distribution (rather than, say, a Gaussian) to increase our power to identify preferences throughout the distribution. We did not specify GPA in major on any resumes. We use this variation to define the variable *GPA*.

Major Majors for the hypothetical resumes were selected according to a predefined probability distribution intended to balance the realism of the rating experience and our ability to detect and control for the effect of majors. Table A.3 shows each major along with its school affiliation and classification as Humanities & Social Sciences or STEM, as well as the probability assigned to each. We use this variation as the variable *Major* and control for it with fixed effects in most regressions.

A.2.5 Components from Real Resumes

For work experiences, leadership experiences, and skills, we drew on components of resumes of real Penn students. This design choice improved the realism of the study by matching the tone and content of real Penn job seekers. Moreover, it improved the validity of our results by ensuring that our distribution of resume characteristics is close to the true distribution. This also helps us identify the range of interest for the study, since resumes of unrealistically low (or high) quality are unlikely to produce useful variation for identification.

Source resumes came from campus databases (for example, student club resume books) and from seniors who submitted their resumes in order to participate in the

Table A.3: Majors in Generated Penn Resumes

Type	School	Major	Probability
Humanities & Social Sciences	The Wharton School	BS in Economics	0.4
	College of Arts and Sciences	BA in Economics	0.2
		BA in Political Science	0.075
		BA in Psychology	0.075
		BA in Communication	0.05
		BA in English	0.05
		BA in History	0.05
		BA in History of Art	0.025
		BA in Philosophy	0.025
		BA in International Relations	0.025
		BA in Sociology	0.025
STEM	School of Engineering and Applied Science	BS in Computer Engineering	0.15
		BS in Biomedical Science	0.075
		BS in Mechanical Engineering and Applied Mechanics	0.075
		BS in Bioengineering	0.05
		BS in Chemical and Biomolecular Engineering	0.05
		BS in Cognitive Science	0.05
		BS in Computational Biology	0.05
		BS in Computer Science	0.05
		BS in Electrical Engineering	0.05
		BS in Materials Science and Engineering	0.05
		BS in Networked and Social Systems Engineering	0.025
		BS in Systems Science and Engineering	0.025
	College of Arts and Sciences	BA in Biochemistry	0.05
		BA in Biology	0.05
		BA in Chemistry	0.05
		BA in Cognitive Science	0.05
		BA in Mathematics	0.05
		BA in Physics	0.05

Majors, degrees, schools within Penn, and their selection probability by major type. Majors (and their associated degrees and schools) were drawn with replacement and randomized to resumes after subjects selected to view either Humanities & Social Sciences resumes or STEM resumes.

matching process. When submitting resumes, students were informed that components of their resumes could be shown directly to employers. We scraped these resumes using a commercial resume parser (the Sovren Parser). From the scraped data we compiled one list with collections of skills, and a second list of experiences comprising an organization or employer, a position title, a location, and a job description (generally in the form of resume bullet points).

Resume components were selected to be interchangeable across resumes. To that end, we cleaned each work experience, leadership experience, and skills list in the following ways:

- Removed any information that might indicate gender, race, or religion (e.g., "Penn Women's Varsity Fencing Team" was changed to "Penn Varsity Fencing Team" and "Penn Muslim Students Association" was not used)
- Screened out components indicative of a specific major (e.g., "Exploratory Biochemistry Intern" was not used)
- Corrected grammatical errors

Work Experience We designed our resumes to vary both the quality and quantity of work experience. All resumes had a work experience during the summer before the candidate's senior year (June–August 2017). This work experience was either a regular internship (20/40) or a top internship (20/40). In addition, some resumes also had a second work experience (26/40), which varied in quality between a work-for-money job (13/40) or a regular internship (13/40). The job title, employer, description, and location shown on the hypothetical resumes were the same as in the source resume, with the minimal cleaning described above.

Before selecting the work experiences, we defined a *Top Internship* to be a substantive position at a prestigious employer. We chose this definition to both identify prestigious firms and distinguish between different types of jobs at those firms, such as a barista at a local Starbucks and a marketing intern at Starbucks headquarters. We identified a prestigious employer to be one of the 50 firms hiring the most Penn graduates in 2014 (as compiled by our Career Services partners). Since experiences at these firms were much more common among Humanities & Social Sciences majors, we supplemented this list with 39 additional firms hiring most often from Penn's School of Engineering and Applied Science. We extracted experiences at these firms from our full list of scraped experiences, and selected a total of 40 *Top Internship* experiences, with 20 coming from resumes of Humanities & Social Sciences majors and 20 from resumes of STEM majors. All of these *Top Internship* experiences had to be believably interchangeable within a major category. These internships included positions at Bain Capital, Goldman Sachs, Morgan Stanley, Northrop Grumman, Boeing Company, and Google (see Table A.4 for a complete list). This variation identified the variable *Top Internship* in our analysis, which is measured relative to having a regular internship (since all resumes had some job in this position).

Table A.4: Top Internship Employers

Humanities & Social Sciences	**STEM**
Accenture plc	Accenture
Bain Capital Credit	Air Products and Chemicals, Inc
Bank of America Merrill Lynch	Bain & Company
Comcast Corporation	Boeing Company
Deloitte Corporate Finance	Credit Suisse Securities (USA) LLC
Ernst & Young U.S. LLP	Deloitte
Goldman Sachs	Epic Systems
IBM	Ernst & Young
McKinsey & Company	Federal Reserve Bank of New York
Morgan Stanley	Google
PricewaterhouseCoopers	J.P. Morgan
UBS Financial Services Inc.	McKinsey & Company
	Microsoft
	Morgan Stanley Wealth Management
	Northrop Grumman Aerospace Systems
	Palantir Technologies
	Pfizer Inc
	PricewaterhouseCoopers, LLP

Employers of top internships in Humanities & Social Sciences and STEM. A total of 20 *Top Internship* positions were used for each major type; some employers were used multiple times, when they appeared on multiple source resumes. Each firm name was used as provided on the source resume, and may not reflect the firm's official name. The names of some repeat *Top Internship* employers were provided differently on different source resumes (e.g., "Ernst & Young U.S. LLP" and "Ernst & Young"); in this case, we retained the name from the source resume associated with the internship.

We selected 33 regular internships separately for the two major groups: 20 regular internships for randomization in the first work experience position, and 13 for the second position. Regular internships had few restrictions, but could not include employment at the firms who provided top internships, and could not include work-for-money job titles (described below and shown in Table A.5). All jobs had to be believably interchangeable within major category. The regular internships in the second job position defined the variable *Second Internship*, and is measured relative to having no job in the second work experience position. Our dynamically generated resumes automatically adjusted in length when no second job was selected, in order to avoid a large gap on the page.

The remaining 13 jobs in the second work position (the summer after the sophomore year) were identified as *Work for Money*. We identified these positions in the real resume components by compiling a list of job titles and phrases that we thought would be indicative of typical in this category, such as Cashier, Barista, and Waiter or Waitress (see Table A.5 Columns 2–4 for the full list). We extracted components in our full list of scraped experiences that matched these search terms, and selected 13 that could be plausibly interchangeable across any major. During randomization, these 13 jobs were used for both Humanities & Social Sciencesf and STEM majors. The first column of Table A.5 shows the job titles that appeared as *Work for Money* jobs in our hypothetical resumes. Columns 2–4 provide the list of job titles used for identifying work-for-money jobs in the scraped data, and for matching candidates to employer preferences.

Leadership Experience We defined leadership experiences to be those resume components that indicated membership or participation in a group, club, volunteer organization, fraternity/sorority, or student government. We selected leadership experiences from our full list of scraped experience components, requiring that the positions be clearly non-employment, include a position title, organization, and description, be plausibly interchangeable across gender, race, and major type. While many real resumes simply identified a position title and organization, we required that the components for our hypothetical resumes include a description of the activity for use as bullet points. We curated a list of 80 leadership experiences to use

Table A.5: Work for Money Job Titles & Identifying Phrases

Used for Resume Tool	**Used for Identifying Components & Matching**		
Assistant Shift Manager	Assistant coach	Courier	Phone Bank
Barista	Attendant	Custodian	Prep Cook
Cashier	Babysitter	Customer Service	Receptionist
Front Desk Staff	Backroom Employee	Dishwasher	Retail Associate
Host & Cashier	Bag Boy	Doorman	Rug Flipper
Sales Associate	Bagger	Driver	Sales Associate
Salesperson, Cashier	Bank Teller	Employee	Sales Representative
Server	Barback	Front Desk	Salesman
	Barista	Fundraiser	Salesperson
	Bartender	Gardener	Saleswoman
	Bellhop	Host	Server
	Bodyguard	Hostess	Shift Manager
	Bookseller	House Painter	Stock boy
	Bouncer	Instructor	Stockroom
	Bus boy	Janitor	Store Employee
	Busser	Laborer	Temp
	Caddie	Landscaper	Tour Guide
	Caddy	Librarian	Trainer
	Call center	Lifeguard	Tutor
	Canvasser	Line Cook	Valet
	Cashier	Maid	Vendor
	Caterer	Messenger	Waiter
	Cleaner	Mover	Waitress
	Clerk	Nanny	Work Study
	Counselor	Petsitter	Worker

Position titles and relevant phrases used to identify work for money in hypothetical resumes for evaluation and in candidate pool resumes. The first column contains the eight unique positions randomized into hypothetical resumes; position titles Cashier, Barista, Sales Associate, and Server were used more than once and associated with different firms. Columns 2–4 specify the work-for-money positions used to predict hiring interest of potential candidates from the pool of prospective matches. Any position title containing one of these phrases was identified as work for money for the purposes of matching.

for both Humanities & Social Sciences and STEM resumes. Each resume included two randomly selected leadership experiences.

Every leadership position was assigned to the location of Penn's campus, Philadelphia, PA. This was done for consistency and believability, even if some of the leadership positions were held in other locations in the source resume. We randomly selected two ranges of years during a student's career to assign to the experiences, and we ordered the experiences chronologically on the hypothetical resume based on the end year of the experience.

Skills We selected 40 skill sets from STEM resumes and 40 from Humanities & Social Sciences resumes for randomization in the survey tool. We intended for these skill sets to accurately reflect the types of skills common in the resumes we collected, and to be plausibly interchangeable within a major type. For randomization, skill sets were drawn from within a major type. To induce variation for the variable *Technical Skills*, we randomly upgraded a skill set with probability 25% by adding two skills from the set of programming languages {Ruby, Python, PHP, Perl} and two skills from the set of statistical programming packages {SAS, R, Stata, Matlab} in random order. To execute this randomization, we removed any other references to these eight languages from the skill sets. Many display their skills in list format, with the word "and" coming before the final skill; we removed the "and" to make the addition of *Technical Skills* more natural.

A.3 Matching Appendix

A.3.1 Students

For job-seeking study participants, the career services office sent an email to seniors offering "an opportunity to reach more employers" by participating in our pilot study, to be run in parallel with all existing recruiting activities. The full student recruitment email is reproduced in Appendix A.2. After uploading a resume and answering basic questions on their industry and locations of interest, students were entered into the applicant pool, and we did not contact them again. If matched with an employer, we emailed the student's resume to the employer and encouraged the employer to contact the student directly. Students received no other incentive for participating.

A.3.2 Matches with Job Seekers

To match job seeking students with the recruiters in our study, we parsed the student resumes and coded their content into variables describing the candidate's education, work experience, and leadership experience, using a combination of parsing software and manual transcription. We did not include any measure of ethnicity or gender in providing matches, nor did we take into account any employer's revealed ethnic or gender preferences. The full list of variables used for matching is shown in Table A.6.

We ran individual ridge regressions for each completed firm-position survey, merging the responses of multiple recruiters in a company if recruiting for the same position. We ran separate regressions using the hiring interest rating (the response to the question "How interested would you be in hiring [**Name**]?") and the likelihood of acceptance (the response to the question "How likely do you think [**Name**] would be to accept a job with your organization?") as outcome variables. We used cross-validation to select the punishment parameter of the ridge regression by running pooled regressions with a randomly selected hold-out sample, and identifying the punishment parameter that minimized prediction error in the hold-out sample. Repeating this process with 100 randomly selected hold-out samples separately for Humanities & Social Sciences and STEM employerd, we use the average of the best-performing punishment parameters as the punishment parameter for the individual regressions. Based on the individual regression results, we then generated out-of-sample predictions of hiring interest and likelihood of acceptance for the resumes in our match pool that met minimal matching requirements for industry and geographic location. Finally, we generated a "callback index" as a weighted average of the predicted hiring interest and likelihood of acceptance (callback = $\frac{2}{3}$hiring interest + $\frac{1}{3}$likelihood of acceptance). The 10 resumes with the highest callback indices for each employer were their matches.

We emailed each employer a zipped file of these matches (i.e., 10 resumes in PDF format). If multiple recruiters from one firm completed the tool for one hiring position, we combined their preferences and provided a single set of 10 resumes to

Table A.6: Candidate Matching Variables

Variable	Definition
GPA	Overall GPA, if available. If missing, assign lowest GPA observed in the match pool
Engineering	Indicator for Computer Sciences, Engineering, or Math majors (for STEM candidates)
Humanities	Indicator for Humanities majors (for Humanities & Social Sciences Candidates)
Job Count	Linear variable for 1, 2, or 3+ work experiences.
Top Firm	Resume has a work experience at one of the firms hiring the most Penn graduates
Major City	Resume has a work experience in New York, San Francisco, Chicago, or Boston
Work for Money	Resume has a job title including identifying phrase from Table A.5
S&P500 or Fortune 500	Resume has an experience at an S&P 500 or Fortune 500 firm
Leader	Resume has a leadership position as Captain, President, Chair, Chairman, or Chairperson

Variables used to identify individual preferences and recommend matched candidates. Variables were identified in hypothetical resumes and in the candidate resume pool. Subjects were provided with 10 real job seekers from Penn whose qualifications matched their preferences based on predictions from a ridge regression with these features.

the group.[66] This set of candidate resumes was the only incentive for participating in the study.

[66] In cases where multiple recruiters from a firm completed the tool in order to fill different positions, or where a single recruiter completed multiple times for different positions, we treated these as unique completions and provided them with 10 candidate resumes for each position.

B Results Appendix

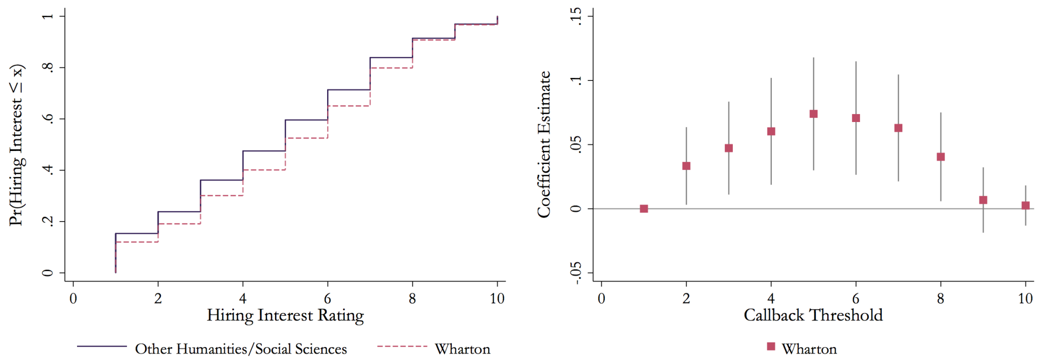
In this section, we describe additional results and robustness checks to validate our main results. In Section B.1, we show additional analysis related to our main human capital results. In Section B.2, we verify our results after reweighting observations to the true distribution of GPAs in actual Penn student resumes. In Section B.3, we break down our evidence of demographic discrimination by gender and race and provide additional results based on subject characteristics. In Section B.4, we discuss preferences over the quality distribution.

B.1 Additional Results on Human Capital

The human capital results in Section 3.2 rely on the independent randomization of work experiences and other resume elements. This randomization leads to some combinations of resume elements that are that are unlikely to arise in practice, despite drawing each variable from a realistic univariate distribution. If employers value a set of experiences that form a cohesive narrative, independent randomization could lead to strange relationships in our data. In Table B.1, we address this concern by showing that the cross-randomization of work experiences does not drive our results. To test this, we had three undergraduate research assistants at the University of Pennsylvania rate all possible combinations of work experiences that could have appeared on our hypothetical resumes.[67] We used their responses to create a dummy—denoted *Narrative*—that is equal to 1 when a resume has a work experience in the summer before junior year that is related to the work experience before senior year, and 0 otherwise. As a result of this process, we identified that 17.5% of the realized resumes in our study (i.e., those resumes actually shown to subjects) had a cohesive work experience narrative. None of these resumes included *Work for Money* because our RA raters did not see these jobs as contributing to a

[67] As Penn students, these RAs were familiar with the type of work experiences Penn students typically have in the summers before their junior and senior years. Each RA rated 1040 combinations (40 work experiences in the summer before senior year $\times$ 26 work experiences in the summer before junior year) for Humanities & Social Sciences majors, and another 1040 combinations (40×26) for the STEM majors blind to our results. They rated each combination on the extent to which the two work experiences had a cohesive narrative on a scale of 1 to 3 where 1 indicated "These two jobs are not at all related," 2 indicated "These two jobs are somewhat related," and 3 indicated "These two jobs are very related." The majority of combinations received a rating of 1 so we introduce a binary variable *Narrative* equal to 1 if the jobs were rated as somewhat or very related, and 0 if the jobs were not at all related.

Figure B.1: Wharton



(a) Empirical CDF

(b) Linear Probability Model

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

narrative. Appendix Table B.1 runs the same regressions as Table 2 but additionally controls for *Narrative*. All results from Table 2 remain similar in size and statistical significance.

In Table B.2, we estimate the value of degrees from more prestigious schools within Penn. We replace the major fixed effects of Table 2 with binary variables for *School of Engineering and Applied Science* and *Wharton*, as well as a binary control for whether the subject has chosen to review Humanities & Social Sciences or STEM resumes (coefficients not reported).[68] We find that employers find degrees from these schools 0.4–0.5 Likert-scale points more desirable than degrees from Penn's College of Arts and Sciences. As shown in Figure B.1, and as discussed in Section 3.3, we also investigate the effect of having a degree from Wharton across the distribution of hiring interest.

[68]Major fixed effects are perfectly multicollinear with the variables for school, since no two schools grant the same degrees in the same major.

Table B.1: Work Experience Narrative

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.127***	2.193***	2.199***	0.892***
	(0.145)	(0.150)	(0.129)	(0.0613)
Top Internship	0.899***	0.897***	0.893***	0.377***
	(0.0944)	(0.0988)	(0.0805)	(0.0397)
Second Internship	0.346**	0.360**	0.315***	0.155***
	(0.142)	(0.150)	(0.122)	(0.0593)
Work for Money	0.112	0.154	0.152*	0.0509
	(0.110)	(0.113)	(0.0914)	(0.0468)
Technical Skills	0.0459	0.0517	-0.0727	0.0115
	(0.104)	(0.108)	(0.0899)	(0.0442)
Not a White Male	-0.121	-0.158	-0.118	-0.0529
	(0.0986)	(0.103)	(0.0842)	(0.0415)
Narrative	0.216	0.240	0.278*	0.0935
	(0.165)	(0.175)	(0.144)	(0.0677)
Observations	2880	2880	2880	2880
R^2	0.129	0.181	0.483	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 1.91, 2.28, 2.64, 2.94, 3.26, 3.6, 4.05, 4.52, and 5.03.

Table shows OLS and ordered probit regressions of hiring interest from Equation (1), with an additional control for *Narrative*. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. *Narrative* is a characteristic of resumes, defined as work experiences that are related in some way. 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-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, likelihood ratio test for ordered probit regressions).

Table B.2: Prestigious Schools

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.128***	2.186***	2.191***	0.886***
	(0.145)	(0.149)	(0.128)	(0.0624)
Top Internship	0.911***	0.918***	0.910***	0.379***
	(0.0942)	(0.0984)	(0.0803)	(0.0394)
Second Internship	0.441***	0.462***	0.448***	0.195***
	(0.112)	(0.118)	(0.0945)	(0.0466)
Work for Money	0.105	0.137	0.139	0.0485
	(0.110)	(0.113)	(0.0918)	(0.0460)
Technical Skills	0.0419	0.0437	-0.0787	0.0102
	(0.103)	(0.107)	(0.0902)	(0.0430)
Not a White Male	-0.121	-0.156	-0.118	-0.0512
	(0.0983)	(0.103)	(0.0842)	(0.0411)
School of Engineering	0.485**	0.427**	0.387**	0.235***
	(0.198)	(0.205)	(0.164)	(0.0861)
Wharton	0.459***	0.501***	0.416***	0.185***
	(0.110)	(0.115)	(0.0935)	(0.0455)
Observations	2880	2880	2880	2880
R^2	0.115	0.167	0.471	
Major FEs	No	No	No	No
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 2.47, 2.84, 3.20, 3.49, 3.81, 4.15, 4.60, 5.06, and 5.57.

Table shows OLS and ordered probit regressions of hiring interest from Equation (1), with effects for school, and a control for whether the employer selected to view Humanities & Social Sciences resumes or STEM resumes (coefficient not displayed). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. *School of Engineering* indicates a resume with a degree from Penn's School of Engineering and Applied Sciences; *Wharton* indicates a resume with a degree from the Wharton School. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression.

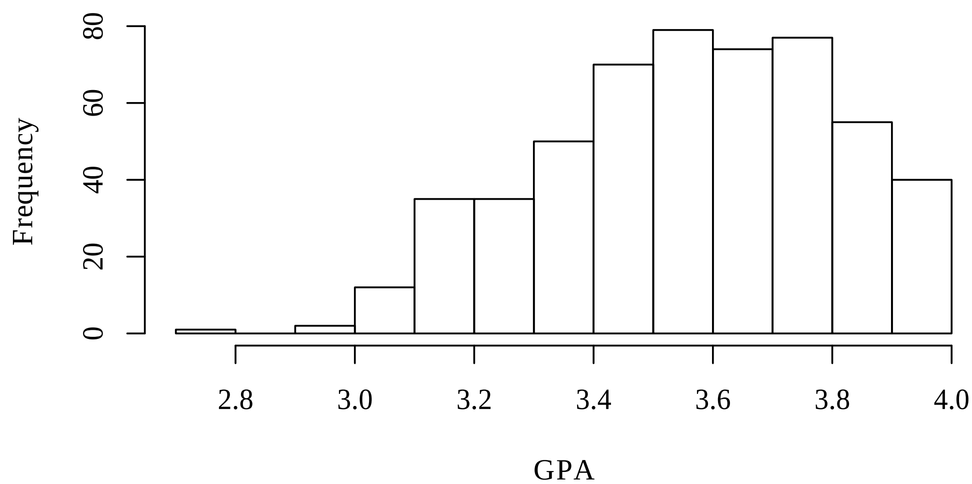
B.2 Re-weighting by GPA

In generating hypothetical resumes, we randomly selected candidate GPAs from $Unif[2.90, 4.00]$, rather than from the true distribution of GPAs among job seekers at Penn, which is shown in Figure B.2.[69] In this section, we demonstrate that this choice does not drive our results. In Tables B.3, B.4, and B.5, we rerun the regressions of Tables 2, 3, and 5 weighted to reflect the naturally occurring distribution of GPA among our Penn senior candidate pool (i.e., the job seekers used for matching, see Appendix A.3). We do not include missing GPAs in the reweighting, though our results are robust to re-weighting with missing GPAs treated as low GPAs.[70] These regressions confirm the results of Tables 2, 3, and 5 in direction and statistical significance.

[69]We parameterized *GPA* to be drawn $Unif[2.90, 4.00]$ to give us statistical power to test the importance of GPA on hiring interest, but this distribution is not exactly the distribution of GPA among Penn seniors engaging in on campus recruiting.

[70]Some students may strategically omit low GPAs from their resumes, and some resume formats were difficult for our resume parser to scrape.

Figure B.2: Distribution of GPA Among Scraped Resumes



Histogram representing the distribution of GPA among scraped resumes in our candidate matching pool. Distribution excludes any resumes for which GPA was not available (e.g., resume did not list GPA, resume listed only GPA within concentration, or parser failed to scrape). GPAs of participating Penn seniors may not represent the GPA distribution at Penn as a whole.

Table B.3: Human Capital Experience—Weighted by GPA

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.270***	2.335***	2.316***	0.961***
	(0.176)	(0.168)	(0.146)	(0.0788)
Top Internship	0.832***	0.834***	0.864***	0.354***
	(0.110)	(0.109)	(0.0881)	(0.0473)
Second Internship	0.488***	0.481***	0.512***	0.215***
	(0.129)	(0.130)	(0.105)	(0.0545)
Work for Money	0.178	0.191	0.196*	0.0757
	(0.129)	(0.125)	(0.100)	(0.0555)
Technical Skills	0.0812	0.0423	-0.102	0.0238
	(0.118)	(0.119)	(0.102)	(0.0507)
Not a White Male	-0.0948	-0.105	-0.0390	-0.0401
	(0.116)	(0.113)	(0.0907)	(0.0497)
Observations	2880	2880	2880	2880
R^2	0.145	0.224	0.504	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 2.30, 2.71, 3.04, 3.34, 3.66, 3.98, 4.48, 4.95, and 5.45.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1), weighted by the distribution of GPA in resumes in the candidate matching pool. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated for each model (F-test for OLS regressions, χ^2 test for ordered probit regression).

Table B.4: Human Capital Experience by Major Type—Weighted by GPA (with Bonferroni-Corrected Significance Levels)

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
	Panel A: Humanities & Social Sciences			
GPA	2.362***	2.452***	2.474***	1.007***
	(0.213)	(0.198)	(0.173)	(0.0967)
Top Internship	0.972***	0.941***	0.982***	0.411***
	(0.127)	(0.125)	(0.102)	(0.0555)
Second Internship	0.477***	0.384**	0.495***	0.217***
	(0.153)	(0.155)	(0.125)	(0.0645)
Work for Money	0.0965	0.0350	0.0880	0.0387
	(0.152)	(0.145)	(0.118)	(0.0654)
Technical Skills	0.0898	0.0260	-0.146	0.0254
	(0.142)	(0.142)	(0.120)	(0.0609)
Not a White Male	0.0528	0.0332	0.0815	0.0214
	(0.135)	(0.131)	(0.106)	(0.0581)
Observations	2040	2040	2040	2040
R^2	0.140	0.242	0.521	
p-value for test of joint significance of Majors	0.112	0.151	0.024	0.149
	Panel B: STEM			
GPA	2.003***	2.119***	1.940***	0.836***
	(0.310)	(0.328)	(0.269)	(0.135)
Top Internship	0.453*	0.509**	0.564***	0.206*
	(0.219)	(0.222)	(0.182)	(0.0928)
Second Internship	0.490*	0.440	0.335	0.205*
	(0.238)	(0.252)	(0.197)	(0.102)
Work for Money	0.353	0.421	0.486**	0.166
	(0.246)	(0.267)	(0.199)	(0.105)
Technical Skills	0.0234	-0.0376	-0.0767	0.00982
	(0.216)	(0.239)	(0.194)	(0.0927)
Not a White Male	-0.475*	-0.477*	-0.458**	-0.197*
	(0.226)	(0.224)	(0.185)	(0.0968)
Observations	840	840	840	840
R^2	0.145	0.402	0.639	
p-value for test of joint significance of Majors	< 0.001	0.003	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints (A): 2.53, 2.89, 3.23, 3.54, 3.86, 4.20, 4.71, 5.18, 5.70.

Ordered probit cutpoints (B): 1.76, 2.28, 2.59, 2.86, 3.17, 3.48, 3.95, 4.40, 4.89.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1), weighted by the distribution of GPA in resumes in the candidate matching pool. Robust standard errors are reported in parentheses. After Bonferroni correction for multiple hypothesis testing, *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, χ^2 test for ordered probit regression).

Table B.5: Likelihood of Acceptance—Weighted by GPA

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.543***	0.551***	0.663***	0.245***
	(0.174)	(0.168)	(0.131)	(0.0738)
Top Internship	0.726***	0.710***	0.696***	0.299***
	(0.110)	(0.108)	(0.0833)	(0.0471)
Second Internship	0.524***	0.455***	0.431***	0.221***
	(0.131)	(0.133)	(0.101)	(0.0556)
Work for Money	0.204	0.148	0.183*	0.0875
	(0.128)	(0.125)	(0.0975)	(0.0543)
Technical Skills	0.0426	-0.0376	-0.115	0.0127
	(0.119)	(0.120)	(0.0971)	(0.0503)
Not a White Male	-0.211*	-0.263**	-0.177**	-0.0915*
	(0.116)	(0.114)	(0.0894)	(0.0489)
Observations	2880	2880	2880	2880
R^2	0.077	0.162	0.509	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: -0.09, 0.29, 0.64, 0.90, 1.26, 1.67, 2.13, 2.65, and 3.01.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1), weighted by the distribution of GPA in resumes in our candidate matching pool. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, χ^2 test for ordered probit regression).

B.3 Candidate Demographics Appendix

In this section, we provide additional analyses for our main results on candidate demographics. In B.3.1, we analyze our findings by the demographics of employers evaluating resumes. In B.3.2, we discuss differential returns to quality by demographic group. In B.3.3, we compare our findings for different demographic groups to results in the prior literature. In B.3.4, we break down the aggregate demographic results to identify the effects of gender and race separately.

B.3.1 Rater Demographics

IRR allows us to collect information about the specific individuals rating resumes at the hiring firm. In Table B.6 we explore our main results by rater gender and race. Raters do not differ significantly by demographic group in their interest in hiring candidates who are not White males.

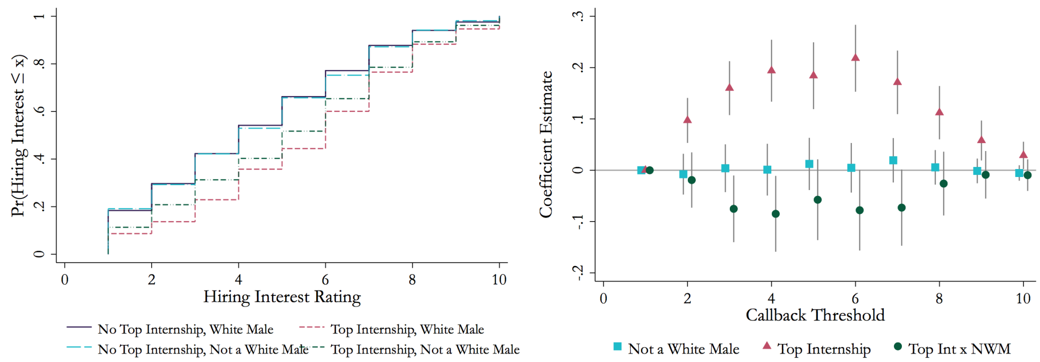
Table B.6: Hiring Interest by Rater Demographics (with Bonferroni-Corrected Significance Levels

	Dependent Variable: Hiring Interest		
	Panel A: Rater Gender		
	All	Female Raters	Male Raters
GPA	2.195***	2.351***	2.087***
	(0.129)	(0.170)	(0.212)
Top Internship	0.902***	0.734***	1.140***
	(0.0806)	(0.105)	(0.140)
Second Internship	0.463***	0.617***	0.192
	(0.0947)	(0.127)	(0.153)
Work for Money	0.149	0.296**	-0.0835
	(0.0913)	(0.120)	(0.155)
Technical Skills	-0.0680	-0.0674	-0.0202
	(0.0900)	(0.122)	(0.151)
Not a White Male	-0.117	-0.184	-0.149
	(0.0842)	(0.111)	(0.142)
Observations	2880	1720	1160
R^2	0.482	0.524	0.555

	Panel B: Rater Race		
	All	Non-White Raters	White Raters
GPA	2.195***	2.186***	2.127***
	(0.129)	(0.375)	(0.146)
Top Internship	0.902***	1.408***	0.773***
	(0.0806)	(0.233)	(0.0913)
Second Internship	0.463***	0.626**	0.453***
	(0.0947)	(0.272)	(0.108)
Work for Money	0.149	-0.127	0.185
	(0.0913)	(0.255)	(0.104)
Technical Skills	-0.0680	-0.121	-0.0114
	(0.0900)	(0.230)	(0.104)
Not a White Male	-0.117	0.0122	-0.143
	(0.0842)	(0.230)	(0.0961)
Observations	2880	600	2280
R^2	0.482	0.588	0.502
Major FEs	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes
Order FEs	Yes	Yes	Yes
Subject FEs	Yes	Yes	Yes

OLS regressions of *Hiring Interest* on candidate characteristics by rater gender and race. Panel A sample includes 29 male and 42 female subjects; Panel B sample includes 57 White and 15 non-White subjects. Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively, after Bonferroni correction for multiple hypothesis testing in which we multiplied p-values by 2. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. R^2 is indicated for each OLS regression. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated.

Figure B.3: Top Internship × Not a White Male



(a) Empirical CDF

(b) Linear Probability Model

Empirical CDF of *Hiring Interest* (Panel B.3a) and difference in counterfactual callback rates (Panel B.3b) for *Top Internship*, *Not a White Male*, and *Top Internship* × *Not a White Male*. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

B.3.2 Interaction of Demographics with Quality

Table B.7 shows that White males gain more from having a *Top Internship* than candidates who are not White males. Figure B.3 looks at the relationship between *Top Internship* and being *Not a White Male* throughout the quality distribution. We find that when a candidate is of sufficiently high quality, a *Top Internship* is equally valuable for White male candidates and those who are not White males. This may suggest that other signals of quality may inoculate candidates from the assumption that an impressive work history is the result of diversity initiatives.

B.3.3 Results in Comparison with Previous Literature

Our results can be compared to other studies of employer preferences. To our knowledge, ours is the first study to isolate employers' preferences for quality by separately eliciting hiring interest and likelihood of acceptance. As a result, our measure of the firms' interest in hiring a candidate is uncontaminated by expected likelihood of acceptance, which is a common confound in studies of labor market

Table B.7: Return to Top Internship by Demographic Group

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.118***	2.184***	2.189***	0.888***
	(0.145)	(0.150)	(0.129)	(0.0613)
Top Internship	1.147***	1.158***	1.155***	0.471***
	(0.167)	(0.175)	(0.145)	(0.0704)
Not a White Male	0.0197	-0.00771	0.0330	0.000886
	(0.127)	(0.133)	(0.106)	(0.0539)
Top Internship × Not a White Male	-0.351*	-0.367*	-0.367**	-0.132
	(0.201)	(0.209)	(0.172)	(0.0843)
Second Internship	0.465***	0.490***	0.465***	0.206***
	(0.112)	(0.118)	(0.0946)	(0.0469)
Work for Money	0.107	0.146	0.144	0.0489
	(0.110)	(0.113)	(0.0912)	(0.0468)
Technical Skills	0.0539	0.0615	-0.0631	0.0149
	(0.104)	(0.108)	(0.0899)	(0.0442)
Observations	2880	2880	2880	2880
R^2	0.129	0.181	0.483	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 1.94, 2.31, 2.68, 2.97, 3.29, 3.63, 4.09, 4.55, and 5.06.

Table shows OLS and ordered probit regressions of hiring interest from Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS, likelihood ratio test for ordered probit).

hiring decisions. Broadly, our discrimination results can be summarized as finding that candidates who are not White males are less preferred by STEM employers, receive a lower return to prestigious internships in all fields, and are expected to be less likely to accept a position if offered. All three results are consistent with lower callback rates for minorities and women in resume audit studies.[71] Despite differences in our employers and candidate pool, the magnitude of our effects are comparable to those in Bertrand and Mullainathan [2004].

To compare results with Bertrand and Mullainathan [2004], we define a counterfactual callback threshold in which employers call back all resumes with a 9 or 10 rating on *Hiring Interest*. This is the threshold where 7.95% of our resumes would be called back, similar to the 8% callback rate in Bertrand and Mullainathan [2004]. We find that having a *Top Internship* increases the chance of receiving a 9 or 10 by 5 percentage points; having a *Second Internship* leads to a 3.7 percentage point increase. In Bertrand and Mullainathan [2004], honors on a resume increased callback by 5 percentage points, and one year of work experience increased callback by 0.7 percentage points. Our demographic results are also similar in magnitude: in Bertrand and Mullainathan [2004], having a Black name decreases callback by 3.2 percentage points. We find that for STEM jobs, candidates who are not White males are 2.8 percentage points less likely to be rated a 9 or 10 (not statistically significant). We find that female and minority candidates receive only 68% of the benefit afforded to White male candidates of having a *Top Internship*. Bertrand and Mullainathan [2004] find that resumes with Black names receive between 30.5% and 67.3% of the return to quality afforded to resumes with White names, depending on which measure of quality is used.

B.3.4 Breakdown of Demographic Results

Our analysis of candidate demographics suggests three main discrimination results: (1) STEM employers rate female and minority candidates as less desirable than White male candidates (see Table 3); (2) returns to *Top Internship* are lower for female and minority candidates than for White male candidates (see Table B.7);

[71] Resume audit studies have consistently shown lower callback rates for minorities. Results on gender have been mixed, as resume audit studies typically target female-dominated occupations, such as clerical or administrative work. See Baert [2018] for a summary. Studies specifically targeting male-dominated occupations have shown discrimination against women [Riach and Rich, 2006].

and (3) employers rate female and minority candidates as less likely to accept a job if offered than their White male counterparts (see Table 5). In this section, we explore these results separately for different demographic groups by defining three new variables: *Female, White*; *Male, Non-White*; and *Female, Non-White*. We replicate the main results of Tables 2, 3, 5, and B.7, replacing the indicator variable for *Not a White Male* with the three new demographic variables.

In Table B.8, we show the regression from Table 2 with theses variables. As before, we find coefficients that are directionally negative but not statistically significant under most specifications. However, restricting the sample to subjects seeking STEM candidates in Table B.9, we find that discrimination we detect in Table 3 is most evident against candidates who are either White females or non-White males. Table B.10 shows the results of Table B.7 with separate interactions between *Top Internship* and each demographic group. We find that the returns to *Top Internship* are diminished for White females candidates by about 0.45-0.50 Likert-scale points, relative to the returns for White male candidates. Results for other demographic groups are smaller and directionally negative but not statistically significantly different from zero.

Figure B.3 shows relationship between *Top Internship* and *Not a White Male* across the quality distribution. Returns to *Top Internship* are higher for White males than for minority and female candidates throughout most of the quality distribution, but this difference is not statistically significant at the highest quality thresholds. In addition, this coefficient estimate is smaller at every threshold than the benefit of the Top Internship.

Finally, in Table B.11, we examine the results of *Likelihood of Acceptance* (see Table 5) by demographic group. Again, we find that the effect is strongest for White female candidates, though we can't reject the equivalence of the coefficients for *Female, White*; *Male, Non-White*; and *Female, Non-White* under any of the specifications.

Table B.8: Demographics

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.125***	2.190***	2.196***	0.891***
	(0.145)	(0.150)	(0.129)	(0.0626)
Female, White	-0.152	-0.215*	-0.161*	-0.0609
	(0.114)	(0.118)	(0.0963)	(0.0478)
Male, Non-White	-0.172	-0.177	-0.169	-0.0754
	(0.136)	(0.142)	(0.115)	(0.0576)
Female, Non-White	-0.00936	-0.0220	0.0281	-0.0144
	(0.137)	(0.144)	(0.120)	(0.0573)
Top Internship	0.902***	0.900***	0.897***	0.378***
	(0.0945)	(0.0989)	(0.0806)	(0.0397)
Second Internship	0.465***	0.490***	0.466***	0.206***
	(0.112)	(0.118)	(0.0947)	(0.0468)
Work for Money	0.116	0.157	0.154*	0.0520
	(0.110)	(0.113)	(0.0914)	(0.0464)
Technical Skills	0.0463	0.0531	-0.0711	0.0120
	(0.104)	(0.108)	(0.0899)	(0.0434)
Observations	2880	2880	2880	2880
R^2	0.129	0.181	0.483	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 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. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Female, White*, *Male, Non-White*, *Female, Non-White*, *Top Internship*, *Second Internship*, *Work for Money*, and *Technical Skills* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, likelihood ratio test for ordered probit regression).

Table B.9: Demographics in STEM (with Bonferroni-Corrected Significance Levels)

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	1.932***	1.885***	1.882***	0.802***
	(0.267)	(0.309)	(0.242)	(0.111)
Female, White	-0.419	-0.612**	-0.545**	-0.171
	(0.215)	(0.249)	(0.208)	(0.0891)
Male, Non-White	-0.567*	-0.617	-0.507*	-0.265**
	(0.271)	(0.318)	(0.257)	(0.116)
Female, Non-White	-0.329	-0.260	-0.0465	-0.142
	(0.264)	(0.301)	(0.261)	(0.110)
Top Internship	0.398*	0.559**	0.545***	0.175*
	(0.191)	(0.216)	(0.173)	(0.0788)
Second Internship	0.242	0.307	0.311	0.111
	(0.208)	(0.246)	(0.189)	(0.0870)
Work for Money	0.151	0.275	0.337	0.0761
	(0.212)	(0.254)	(0.187)	(0.0891)
Technical Skills	-0.0283	-0.113	-0.180	-0.000579
	(0.197)	(0.228)	(0.186)	(0.0812)
Observations	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
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 1.44, 1.90, 2.22, 2.51, 2.80, 3.14, 3.56, 4.05, and 4.48.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1). Robust standard errors are reported in parentheses. After Bonferroni correction for multiple hypothesis testing, *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Female, White*, *Male, Non-White*, *Female, Non-White*, *Top Internship*, *Second Internship*, *Work for Money*, and *Technical Skills* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, likelihood ratio test for ordered probit regression).

Table B.10: Demographics Interactions with Top Internship

	Dependent Variable: Hiring Interest			
	OLS	OLS	OLS	Ordered Probit
GPA	2.119***	2.184***	2.191***	0.889***
	(0.145)	(0.150)	(0.129)	(0.0627)
Top Internship	1.147***	1.160***	1.155***	0.471***
	(0.168)	(0.175)	(0.145)	(0.0704)
Female, White	0.0327	-0.0188	0.0225	0.0118
	(0.146)	(0.152)	(0.121)	(0.0620)
Female, White × Top Internship	-0.464**	-0.492**	-0.459**	-0.181*
	(0.234)	(0.243)	(0.199)	(0.0976)
Male, Non-White	-0.0604	-0.0488	-0.0553	-0.0287
	(0.175)	(0.184)	(0.145)	(0.0748)
Male, Non-White × Top Internship	-0.280	-0.316	-0.276	-0.116
	(0.279)	(0.288)	(0.233)	(0.117)
Female, Non-White	0.0806	0.0685	0.159	0.0104
	(0.182)	(0.191)	(0.156)	(0.0776)
Female, Non-White × Top Internship	-0.229	-0.224	-0.316	-0.0653
	(0.273)	(0.286)	(0.240)	(0.114)
Second Internship	0.468***	0.495***	0.470***	0.208***
	(0.112)	(0.118)	(0.0944)	(0.0468)
Work for Money	0.109	0.151	0.148	0.0496
	(0.110)	(0.113)	(0.0913)	(0.0464)
Technical Skills	0.0494	0.0576	-0.0670	0.0132
	(0.104)	(0.108)	(0.0899)	(0.0435)
Observations	2880	2880	2880	2880
R^2	0.130	0.182	0.484	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: 1.94, 2.31, 2.68, 2.97, 3.29, 3.63, 4.09, 4.55, and 5.06.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Female, White*, *Male, Non-White*, *Female, Non-White*, *Top Internship*, *Second Internship*, *Work for Money*, and *Technical Skills* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, likelihood ratio test for ordered probit regression).

B.4 Distributional Appendix

As discussed in Section 3.3, average preferences for candidate characteristics might differ from the preferences observed in the tails. The stylized example in Figure B.4 shows this concern graphically. Imagine the light (green) distribution shows the expected productivity—based on the content of their resumes—of undergraduate research assistants (RAs) majoring in Economics at the University of Pennsylvania and the dark (gray) distribution shows the expected productivity of undergraduate RAs enrolled at the Wharton School. In this example, the mean Wharton student would make a less productive RA, reflecting a lack of interest in academic research relative to business on average; however, the tails of the Wharton distribution are fatter, reflecting the fact that admission into Wharton is more selective, so a Wharton student who has evidence of research interest on her resume is expected to be better than an Economics student with an otherwise identical resume. Looking across the panels in Figure B.4, we see that as callback thresholds shift from being high (panel (a), where professors are very selective, only calling back around 8% of resumes) to medium (panel (b), where professors are calling back around 16% of resumes) to low (panel (c), where professors are calling back around 28% of resumes), a researcher conducting a resume audit study might conclude that there is an advantage on the RA market of being at Wharton, no effect, or a disadvantage.[72]

A researcher might particularly care about how employers respond to candidate characteristics around the empirically observed threshold (e.g., the researcher may be particularly interested in how employers respond to candidates in a particular market, with a particular level of selectivity, at a particular point in time). Nevertheless, there are a number of reasons why richer information about the underlying distribution of employer preferences for characteristics would be valuable for a researcher to uncover. A researcher might want to know how sensitive estimates are to: (1) an economic expansion or contraction that changes firms' hiring needs or (2) new technologies, such as video conferencing, which may change the callback threshold by changing the costs of interviewing. Similarly, a researcher may be interested in how candidate characteristics would affect callback in different markets

[72]This stylized example uses two normal distributions. In settings where distributions are less well-behaved, the difference in callback rates might be even more sensitive to specific thresholds chosen.

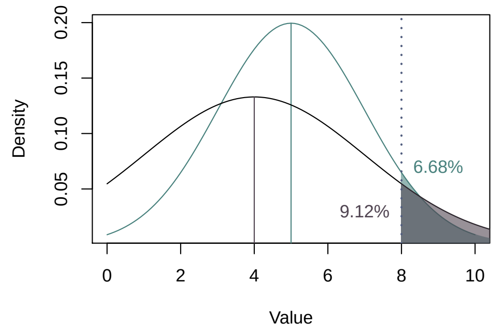
Table B.11: Likelihood of Acceptance: Demographics

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.605***	0.631***	0.734***	0.263***
	(0.144)	(0.150)	(0.120)	(0.0603)
Female, White	-0.231**	-0.294**	-0.258***	-0.0928*
	(0.114)	(0.118)	(0.0935)	(0.0476)
Male, Non-White	-0.125	-0.170	-0.117	-0.0602
	(0.137)	(0.142)	(0.110)	(0.0574)
Female, Non-White	-0.221	-0.236*	-0.162	-0.103*
	(0.135)	(0.142)	(0.112)	(0.0568)
Top Internship	0.683***	0.677***	0.664***	0.285***
	(0.0943)	(0.0979)	(0.0763)	(0.0396)
Second Internship	0.418***	0.403***	0.394***	0.179***
	(0.112)	(0.119)	(0.0911)	(0.0472)
Work for Money	0.197*	0.192*	0.204**	0.0880*
	(0.111)	(0.116)	(0.0896)	(0.0467)
Technical Skills	-0.0508	-0.0594	-0.103	-0.0248
	(0.104)	(0.108)	(0.0861)	(0.0435)
Observations	2880	2880	2880	2880
R^2	0.070	0.124	0.492	
p-value for test of joint significance of Majors	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

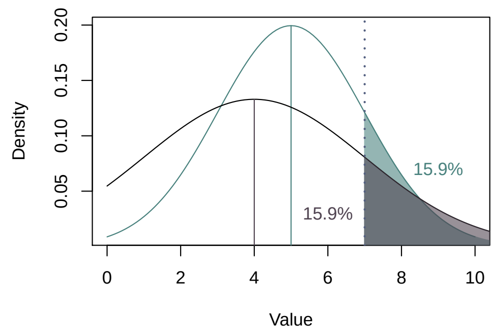
Ordered probit cutpoints: -0.26, 0.13, 0.49, 0.75, 1.12, 1.49, 1.94, 2.46, and 2.83.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1). Robust standard errors are reported in parentheses. *, **, *** denote $p < 0.1$, 0.05, and 0.01, respectively. *GPA*, *Female, White*, *Male, Non-White*, *Female, Non-White*, *Top Internship*, *Second Internship*, *Work for Money*, and *Technical Skills* are characteristics of the hypothetical resume, constructed as described in Section 2.3 and in Appendix A.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p-value of a test of joint significance of major fixed effects is indicated (F-test for OLS regressions, likelihood ratio test for ordered probit regression).

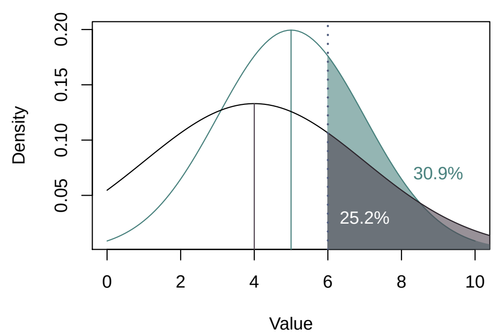
Figure B.4: Callback Thresholds Example



(a) High Threshold



(b) Medium Threshold



(c) Low Threshold

A stylized example where average preferences differ from preferences at the upper tail. The distribution in green has a higher mean and lower variance, leading to thinner tails; the distribution in gray has a lower mean but higher variance, leading to more mass in the upper tail. As the callback threshold decreases from Panel (a) to Panel (c), the share of candidates above the threshold from each distribution changes. Estimating preferences from callbacks following this type of threshold process might lead to spurious conclusions.

(e.g., those known to be more or less selective) than the market where a resume audit was conducted. To conduct these counterfactual analyses, richer preference information would be valuable.

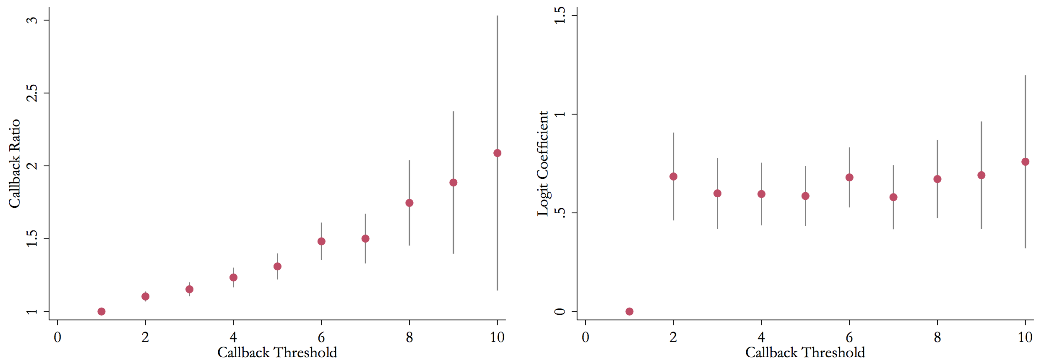
B.4.1 Comparing Results Across the Distribution

Resume audit studies often report differences in callback rates between two types of job candidates, either in a t-test or in a regression. However, as the overall callback rate becomes very large (i.e., almost all candidates get called back) or very small (i.e., few candidates get called back), the differences in callback rates tend toward zero. This is because, as discussed in footnote 45, the maximum possible difference in callback rates is capped by the overall callback rate.

This is not a threat to the internal validity of most resume audit studies executed in a single hiring environment. However, this can cause problems when comparing across studies, or within a study run in different environments. For example, if one wanted to show that there was less racial discrimination in one city versus another, and the underlying callback rates in those cities differed, an interaction between city and race may be difficult to interpret. Note that such an exercise is performed in Kroft et al. [2013] to compare the response to unemployment in cities with high unemployment (and likely low overall callback rates) versus cities with low unemployment rates (and high callback rates). In that particular study, the "bias" caused by comparing across different callback rates does not undermine the finding that high unemployment rate cities respond less to unemployment spells. Nonetheless, researchers should use caution when implementing similar study designs.

In Figures B.5 and B.6, we look at how two different ways of measuring callback differences perform across the distribution compared to the linear probability model. The lefthand side of each figure shows the ratio of the callback rates, another common way of reporting resume audit study results. For the positive effects in our study, this odds ratio tends to be larger at the upper tail, where a small difference in callbacks can result in a large response in the ratio. On the righthand side of each figure, we show effects estimated from a logit specification. We find that in our data, the effects estimated in logistic regression tend to be flatter across the quality distribution.

Figure B.5: Alternative Specifications: Top Internship

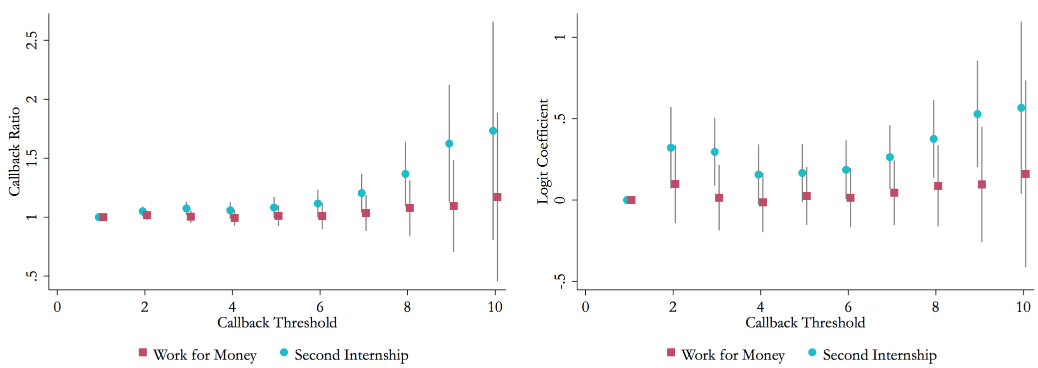


(a) Callback Ratio

(b) Logit

Counterfactual callback ratios (Panel B.5a) and counterfactual logit coefficients (Panel B.5b) for *Top Internship*. Counterfactual callback is an indicator for each value of *Hiring Interest* equal to 1 if *Hiring Interest* is greater than or equal to the value, and 0 otherwise. Callback ratio is defined as the counterfactual callback rate for candidates with the characteristic divided by the counterfactual callback rate for candidates without. 95% confidence intervals are calculated from a linear probability model using the delta method. Logit coefficients are estimated from a logit regression with counterfactual callback as the dependent variable.

Figure B.6: Alternative Specifications: Second Job Type



(a) Callback Ratio

(b) Logit

Counterfactual callback ratios (Panel B.6a) and counterfactual logit coefficients (Panel B.6b) for *Work for Money* and *Second Internship*. Counterfactual callback is an indicator for each value of *Hiring Interest* equal to 1 if *Hiring Interest* is greater than or equal to the value, and 0 otherwise. Callback ratio is defined as the counterfactual callback rate for candidates with the characteristic divided by the counterfactual callback rate for candidates without. 95% confidence intervals are calculated from a linear probability model using the delta method. Logit coefficients are estimated from a logit regression with counterfactual callback as the dependent variable.

C Pitt Appendix

In our replication study at the University of Pittsburgh, we followed a similar approach to that described for our experimental waves at Penn in Section A.2. The tool structure was essentially the same as at Penn, with references to Penn replaced with Pitt in the instructions, and the reference to Wharton removed from the major selection page. Resume structure was identical to that described in Sections A.2.1 and A.2.2. Names were randomized in the same manner as described in Section A.2.3. The education section of each resume at Pitt followed the same structure as that described in Section A.2.4, but had a degree from the University of Pittsburgh, with majors, schools, and degrees randomly drawn from a set of Pitt's offerings. In selecting majors for our Pitt replication, we attempted to match the Penn major distribution as closely as possible, but some majors were not offered at both schools. When necessary, we selected a similar major instead. The majors, schools, classifications, and probabilities for Pitt are shown in Table C.1.

We used a single pool of Pitt resumes for both the hypothetical resume elements and for a candidate pool for Pitt employers, saving significant effort on scraping and parsing. These components were compiled and randomized in much the same way as at Penn, as described in Section A.2.5. For *Top Internship* at Pitt, we collected work experiences from Pitt resumes at one of Pitt's most frequent employers, or at one of the employers used to define *Top Internship* at Penn. Similarly, Pitt *Work for Money* was identified from the same list of identifying phrases shown in Table A.5. *Technical Skills* were randomized in the same way as at Penn, described in A.2.5.

Table C.1: Majors in Generated Pitt Resumes

Type	School	Major	Probability
Humanities & Social Sciences	Dietrich School of Arts and Sciences	BS in Economics	0.4
		BA in Economics	0.2
		BS in Political Science	0.075
		BS in Psychology	0.075
		BA in Communication Science	0.05
		BA in English Literature	0.05
		BA in History	0.05
		BA in History of Art and Architecture	0.025
		BA in Philosophy	0.025
		BA in Social Sciences	0.025
		BA in Sociology	0.025
STEM	Dietrich School of Arts and Sciences	BS in Natural Sciences	0.1
		BS in Molecular Biology	0.075
		BS in Bioinformatics	0.05
		BS in Biological Sciences	0.05
		BS in Chemistry	0.05
		BS in Mathematical Biology	0.05
		BS in Mathematics	0.05
		BS in Physics	0.05
		BS in Statistics	0.025
	Swanson School of Engineering	BS in Computer Engineering	0.15
		BS in Mechanical Engineering	0.075
		BS in Bioengineering	0.05
		BS in Chemical Engineering	0.05
		BS in Computer Science	0.05
		BS in Electrical Engineering	0.05
		BS in Materials Science and Engineering	0.05
		BS in Civil Engineering	0.025

Majors, degrees, schools within Pitt, and their selection probability by major type. Majors (and their associated degrees and schools) were drawn with replacement and randomized to resumes after subjects selected to view either Humanities & Social Sciences resumes or STEM resumes.