An Audit Alternative: Measuring Employer Preferences and Beliefs without Deception

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

July 23, 2018

PRELIMINARY DRAFT. PLEASE DO NOT CITE OR CIRCULATE.

Abstract

Audit studies have been a tremendously valuable tool in labor economics to measure how different characteristics affect the rate at which employers call back candidates. In this paper, we introduce a new, incentive-compatible experimental paradigm that complements audit studies by giving researchers insight into the full distribution of employer preferences over candidate characteristics. The paradigm involves employers rating hypothetical resumes with randomly assigned characteristics so that the can be matched — based on their reported preferences — with real candidates from a large hiring pool. We measure multiple determinants of callback by asking employers to rate candidates on two dimensions: hiring interest and likelihood of offer acceptance. We deploy the experimental paradigm with the career services offices of two major universities and present new results on employers' valuation of an array of measures of human capital, including quality and quantity of work experience. We additionally identify a new channel for discrimination: beliefs that candidates who are not white males will be less likely to accept jobs when offered.

1 Introduction

How do employers' hiring decisions depend on the experience, education, and other human capital investments made by workers? These are fundamental questions in Labor Economics and the Economics of Education (e.g., Autor and Houseman [2010], Pallais [2014]). Similarly, the existence and role of discrimination in labor markets is a key concern for both policy makers and economists (e.g., Altonji and Blank [1999]).

Since these characteristics of job candidates are correlated with each other and with many other traits, researchers aiming to understand employer preferences have turned to experiments to identify causal impacts of characteristics on employers' hiring decisions. The primary experimental method developed for this purpose is the resume audit study, which was used infrequently before being brought into the mainstream by Bertrand and Mullainathan [2004] and is now the workhorse of the literature. Resume audit studies send otherwise identical resumes, with only minor differences associated with a treatment (e.g., different names associated with different races), to randomly selected prospective employers and measure the rate at which candidates are called back by those employers (henceforth the "callback rate"). By comparing callback rates across groups (e.g., those with white names to those with minority names), researchers can identify the existence of discrimination.[1]

Today, correspondence audit studies are pervasive and have become the workhorse method for social scientists interested in discrimination against applicants (e.g., to jobs, to graduate programs) or against requesters (e.g., of services, of in-person meetings).[2] In addition, a number of recent, prominent resume audit studies have focused on the impact of labor market characteristics that do not have to do with belonging to a particular group, including studying the impact of unemployment spells [Kroft et al., 2013, Eriksson and Rooth, 2014, Nunley et al., 2017], for-profit college credentials [Deming et al., 2016], college selectivity [Gaddis, 2015], and military service [Kleykamp, 2009].[3] In these cases, we care not only about whether or not employers respond to these characteristics, but how much employers value them.

One limitation of the resume audit approach, however, is that the binary nature of call back (i.e., whether an employer responds to a resume) gives the researcher only limited insight into the preference of employers.[4] For example, the estimates that come out of a resume audit study are

[1] Resume audit studies were inspired by 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 callbacks 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]. Resume audit studies were designed to overcome these empirical concerns and are an example of a "correspondence audit study," which refers to any audit study that involves randomized correspondence rather than confederates.

[2] Bertrand and Duflo [2016] notes 90 correspondence audit studies focused on discrimination against protected classes between 2005 and 2016. Bertrand and Duflo [2016] note that the literature has largely not evolved past measuring differences in callback means between groups, and has mostly failed to probe mechanisms behind these differences.

[3] Baert [2018] provides a useful summary of the areas that have been explored in correspondence audit studies.

[4] These concerns extend to other correspondence audit studies more generally since they typically rely on a binary

reflective of how employers respond to candidate characteristics at a particular level of selectivity. Note, however, that if the shape of the distribution of employer preferences for candidates depends on a characteristic, then how callback rates respond to a characteristic may differ dramatically — even flipping signs — across settings with different employer selectivity.

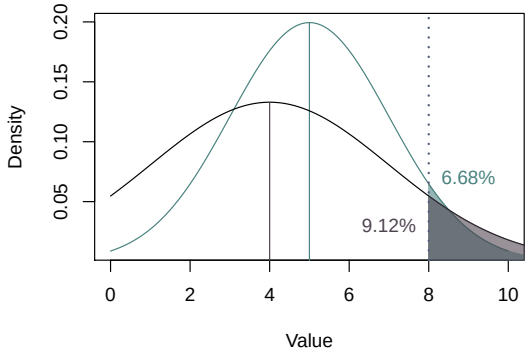
The stylized example in Figure 1 shows this concern graphically. Imagine the light (green) distribution shows the expected productivity — based on the content of their resumes — of undergraduate research assistants (RAs) majoring in Economics at the University of Pennsylvania and the dark (gray) distribution 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 (e.g., the average Wharton student is less likely to go to graduate school in Economics than the average Economics major at the University of Pennsylvania); 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 1, 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.[5]

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 — beyond what can be identified in a resume audit study — 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 (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. Similar concerns might arise in considering how resume characteristics interact with one another. In a particular resume audit study, with certain characteristics randomized and others remaining constant, the effect of a particular characteristic is limited to the marginal effect at that threshold conditional

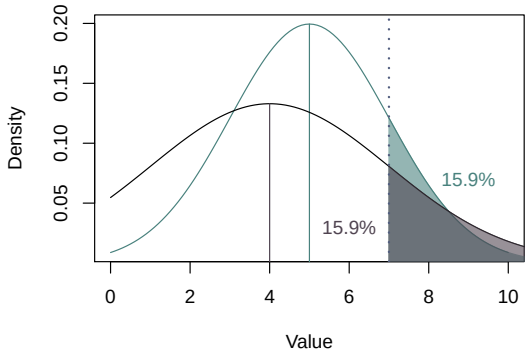
measure of response.

[5]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.

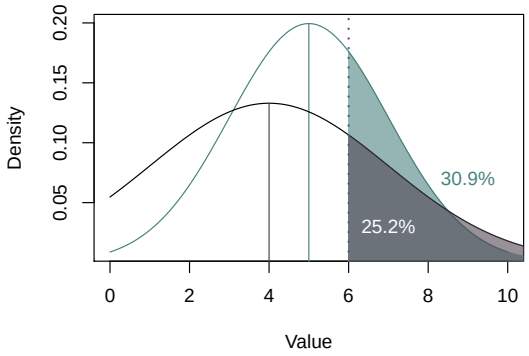
Figure 1: Callback Thresholds Example



(a) High Threshold



(b) Medium Threshold



(c) Low Threshold

on the other resume characteristics that are held constant. Researchers may want to explore how interactions in candidate characteristics may look at various levels of selectivity, rather than just one. Beyond these specific concerns, researchers might find that exploring the full distribution of employer valuations is of broader scientific interest.

To generate richer data on employer preferences in the hiring domain, we pioneer a new experimental paradigm for eliciting employer preferences over candidate characteristics. Our method, an "Incentivized Resume Rating" (IRR) study, combines the incentives of the field with the control of the laboratory. In our paradigm, employers are asked to rate resumes that they know to be hypothetical but are incentivized because their reported preferences are used to match them with a pool of real candidates (e.g., via machine learning [Mullainathan and Spiess, 2017]).

We implement our paradigm to explore preferences of employers who engage in on campus recruiting for graduating seniors at the University of Pennsylvania and replicate the study at the University of Pittsburgh. Across both schools, we recruit 158 real employers through partnerships with career services offices at those universities. In our experiment, each employer rates 40 unique, hypothetical resumes with randomly assigned traits.[6] The employers know that the resumes are hypothetical and we use their responses to match them to a set of 10 real candidates from pools of hundreds of real job seekers, avoiding concerns about deception.[7] We ask employers to evaluate resumes on two dimensions — hiring interest and likelihood of offer acceptance — and elicit a measure for each dimension on a 10-point Likert scale.

Our method generates a number of interesting empirical results regarding how employers at the University of Pennsylvania respond to the quantity and quality of human capital characteristics. Employers value GPA, having a prestigious internship, and having a second internship, but place no value on a typical "work-for-money" job (e.g., barista, server, or cashier).[8] We do not find evidence of discrimination on average in our data, but we do find that recruiters looking for employees in STEM fields do display discrimination against candidates who are not white males (i.e., who are women and/or minorities). In addition, we replicate results from Bertrand and Mullainathan [2004] in that returns to resume "quality" are smaller for candidates who are not white males. In addition, our results suggest a novel channel for discrimination: employers robustly report that candidates who are not white males are less likely to accept positions if offered, which could result in less effort

[6]The hypothetical resumes are constructed dynamically as employers click through our survey, using component pieces (e.g., majors, summer jobs, leadership activities, and skills) from real student resumes from each university and hypothetical names chosen for racial and gender specificity. While randomization of characteristics means there are trillions of possible resumes in our study, each characteristic is used often enough among different employers that we can accurately estimate employer preferences for a characteristic and control for response to that characteristic when estimating other preferences.

[7]Receiving the list of matched candidates is the only incentive for participating in our study, ensuring the employers value the matches and take the exercise seriously (median time spent taking our survey was 28 minutes per employer). Our design builds off of the experimental method pioneered in Low [2017], which used a similar-in-spirit incentive to elicit incentive-compatible evaluations of dating profiles.

[8]Major of study also affects employer preferences while leadership (i.e., extracurricular) activities have only a marginally statistically significant effect.

expended in recruiting female and minority candidates.

The remainder of this paper proceeds as follows. Section 2 presents a simple theoretical framework that highlights what we can learn from correspondence audit studies with a focus on resume audit studies. Section 3 presents the experimental paradigm. Section 4 reports on the results of our experimental deployments of the paradigm. Section 5 discusses the implications of our results and concludes.

2 Theoretical framework

While differential callback rates are a useful measure of labor market access, they are not sufficient for identifying employer preferences, even under the controlled conditions of a resume audit study. We provide a simple framework to illustrate this fact, extending earlier work by Heckman and Siegelman [1992], Heckman [1998], and Neumark [2011], and we generate examples that demonstrate why inferences drawn from callback rates are likely to be biased and inconsistent estimates of employer preferences.

Let V_{ij} represent the subjective expected productivity of candidate j if employed at firm i. We now specify a trait of interest, x, that affects this expected productivity (note, for generality, we could replace x with a vector of traits $\mathbf{X}$. We restrict our attention for now to a single trait for ease of explanation), and represent the systematic preferences for this trait as β.

The meaning of systematic preferences formally is that it is the common preference, that once extracted, makes the distribution of employer's individual preferences for that trait, represented by $\eta_i x_j$, mean 0.

The employer's preferences can therefore be represented as:

$$V_{ij} = \beta x_j + \xi_{ij}, \tag{1}$$

where βx_i represents the common preference of all employers for characteristic x, and ξ_{ij} represents all other elements of employer i's expected productivity from candidate j, including: (i) employer i's specific preferences for characteristic x (ii) all employers' systematic preferences for all other traits of candidate j, and finally (iii) the idiosyncratic match quality between candidate j and employer i, which, for example, includes employer i's specific preference for characteristics other than x.

Thus, we could further break down ξ_{ij} by writing it as:

$$\xi_{ij} = \eta_i x_j + \Omega' \mathbf{Z}_j + \psi_i \mathbf{Z}_j + \nu_{ij} \tag{2}$$

η_i, ψ_i, and ν_{ij} are defined to be mean zero. Let ν_{ij} be distributed according to some distribution

with CDF F. Naturally, without a resume audit framework, if some elements of $\mathbf{Z}$ were correlated with x, it would be impossible to recover an unbiased estimate of β. The implicit promise of a resume audit study is that, by randomly assigning the characteristic x orthogonally to all elements of $\mathbf{Z}$, such that $E(\mathbf{Z_j}|X_j) = \mathbf{0}$, the resulting estimate of β will be unbiased.

However, we note that in a resume audit study, what is observed is not V_{ij} directly, but rather whether the employers valuation is higher than some threshold set for callback rates. We will represent this simply as V_i^*, which may be a function of the firm's selectivity, their costs associated with interviewing and hiring candidates, and other variables.[9]

The researcher observes D_{ij}, a binary variable indicating whether employer j chooses to call candidate i. First, assume that a candidate's acceptance of a position, if offered, is guaranteed, so candidates will be called back as long as V exceeds V^*.

That is,

$$D_{ij} = \left\{ \begin{array}{ll} 1, & \text{if } V_{ij} \geq V_i^*(c_i) \\ 0, & \text{if } V_{ij} < V_i^*(c_i) \end{array} \right\} \tag{3}$$

The researcher is therefore estimating

$$D_{ij} = fn(\beta x_j + \xi_{ij}) \tag{4}$$

For simplicity, suppose that the characteristic of interest is binary, such as proficiency in a particular programming language. If the researcher could observe V_{ij} directly and control all other relevant variables, she could estimate β, the value of that programming skill, directly as the difference in mean desirability between those with and without the skill, $E[V_{ij}|x_j = 1] - E[V_{ij}|x_j = 0]$, while holding all other characteristics constant. Since it has been discussed elsewhere, we will ignore the difficulties of keeping all variables constant and assume that in resume audit studies all productivity-relevant variables have been correctly controlled for. However, the researcher only observes the callback outcome D_{ij}, and would like to estimate preferences from callback rates, $E[D_{ij}|x_j = 1] - E[D_{ij}|x_j = 0]$.

2.1 Correlated Variance of Preferences

The results of resume audit studies are often described as if the difference in callback rates $E[D_{ij}|x_j = 1] - E[D_{ij}|x_j = 0]$ is an estimator for β allowing the researcher to make inferences about employer preferences. However, it is clear that the difference in callback rates is limited to

[9] We assume when a firm chooses to call back an applicant, it incurs some cost $c \geq 0$. This cost represents the time and expense invested in interviewing and screening candidates. Of course, while this is a natural place for employers to incur costs, there may be additional costs incurred other phases of the hiring process (such as during initial resume screening, or in case of making job offers to candidates who do not accept); for simplicity, however, we assume that the cost of interviewing is the most salient for the consideration of resume audit studies. Moreover, this one cost is sufficient to derive the main intuition from our model, which would hold in models with more complicated cost structures.

the support $[-1, 1]$, and cannot be an unbiased or consistent estimator for a parameter defined on the support of real numbers. The best that a resume audit can hope for is to interpret the difference in callback rates as an estimator for the local change of productivity quantile around the callback threshold. Unfortunately, even this interpretation is not reliable: Proposition 1 below demonstrates that even in an ideal resume audit study where all groups have equal means on all characteristics aside from the characteristic of interest, callback rates may have the opposite sign from β, depending on the upper end of the error distribution. In particular, if the error distributions are heteroskedastic in the characteristic of interest, the researcher may draw opposite conclusions about the direction of employer preferences.

Note that if the idiosyncratic error term ξ_{ij} is distributed according to a standard Gaussian distribution, this is identical to the latent variable formulation of a probit model, where productivity is the unobserved latent variable and callback decisions are the dichotomous observed outcome. Similarly, if the error term is distributed according to a Type I Extreme Value distribution, this model reflects a logit model. As is commonly recognized in the logit and probit models, heteroskedasticity in the errors may lead to biased and inconsistent estimates of the relationship between candidate characteristics and employer preferences.

Proposition 1 *Assume no omitted variable bias ($E[Z_{ij}|x] = 0$) and a callback threshold V^* that is above the median of V_{ij}. Then the estimate of the sign of β obtained from the regression $D_{ij} \sim \beta x_j + \varepsilon$ will not be unbiased or consistent if $Var(\xi|x) \neq Var(\xi)$.*

Proof.

■

As a concrete—if extreme—example, suppose that employer preferences are increasing in GPA, but with larger variance at the low end of the GPA distribution, as simulated in Figure 2a. Depending on the cutoff level c, estimates based only on callback rates could vary from the correct direction and magnitude to its opposite. The observations from one such cutoff ($c = 10$) and the linear probability model estimated from the resulting observations are shown in Figure 2b.

As this example shows, it is not possible to make inferences about the mean of unobserved employer preferences from the number of observations exceeding a threshold without making strong assumptions about the higher moments of the error distribution.

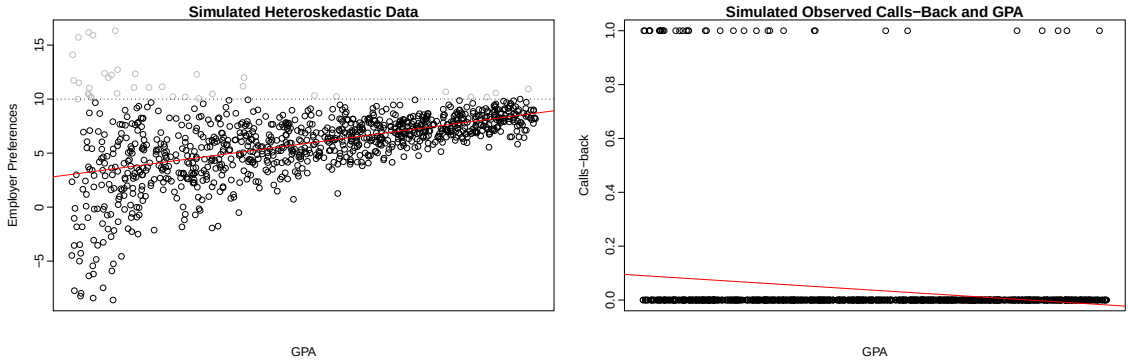
The extended version of the error term,

$$\xi_{ij} = \eta_i x_j + \Omega' \mathbf{Z}_j + \psi_i \mathbf{Z}_j + \nu_{ij} \tag{5}$$

helps illustrate why this heteroskedasticity is not just possibly, but likely to occur.

First, the firm specific tastes for the characteristic of interest, x_j. If the intensity of firms' preferences vary with x_j, ξ_{ij} will be heteroskedastic. For example, if the heterogeneity in tastes for

Figure 2: Simulations of Preferences and Callback Outcomes with Heteroskedasticity



(a) True Preferences

(b) Observed Calls

Preferences simulated from data generating process $y = \alpha + \beta X + \varepsilon$, with $\alpha = 3$, $\beta = 5$, and $\varepsilon \sim \mathcal{N}(0, 1/X)$. Cutoff set at $c = 10$. Observations above the call-back threshold are shown in gray in (a), and are observed as 1 in (b); observations below the threshold are shown in black, and are observed as 0 in (b). The red line indicates the regression line estimating the relationship between GPA and true preferences in (a) and the relationship between GPA and observed calls in (b). Observing calls results in biased and inconsistent estimates of employer preferences in this case.

x_j is increasing in x_j. For example, imagine that lack of a particular programming language, $x_j = 0$, is neutral to everyone. However, the presence of that language, $x_j = 1$ is a strong positive signal to firms who use that language, but a strong negative signal to firms looking for "well-rounded candidates".

Second, the generalized taste for other characteristics, $\Omega' \mathbf{Z}_j$. If firms care less or more about other characteristics depending on the value of x_j, this will create heteroskedasticty in ξ_{ij}. For example, if a certain programming language is a *necessary* condition for a firm to hire a candidate, in which case the firm will not even closely evaluate other characteristics when the programming language is absent. Then, there will be very little variation in ξ_{ij}, stemming from flat tastes for $\mathbf{Z}_j$, for $x_j = 0$, but substantial variation when $x_j = 1$.

Third, the firm-specific tastes for other characteristics, which combines the two issues illustrated above. These examples show that it does not require a mysterious error structure, but rather reasonable preferences by heterogeneous firms, to generate substantial heteroskedasticity, and thus estimation bias.

2.2 Different estimates at different thresholds

Naturally, different firms might have different expected productivity thresholds at which they call back candidates. Moreover, different experimental settings might have differing average thresholds. We will think of this as the percentage of candidates overall who are called back. For example,

hiring during a recession might be more selective than hiring during an expansion. Mass hiring for factory workers might be less selective than targeted recruiting to replace a single position. The existence of different overall callback rates illustrates how tempermental $\hat{\beta}$ can be as an estimator.

We first make the general, and perhaps obvious, point that because $\hat{\beta}$ is the level difference in callback rates between groups, the estimate will be mechanically larger the closer the callback rate is to 50%. Thus, comparing experiments, or experimental arms, from areas with high callback rates to those with low callback rates can be misleading.

But more troublingly, under heteroskedasticity (or any correlation between higher-order moments of the error term and the x-variable of interest), $\hat{\beta}$ can actually flip signs if one moves from a more selective hiring environment to a less selective one.

We state this formally in Proposition 2.

Proposition 2 *The estimate $\hat{\beta}$ obtained from the regression $D_{ij} \sim \beta x_j + \varepsilon$ will not generally be constant in the callback threshold V^*, unless $\hat{\beta} = 0$ for all thresholds.*

In particular, if errors are normally distributed and mean zero, and the group with lower mean has higher variance, it is always possible to identify two callback thresholds, V^* and $V^{*\prime}$, far enough apart such that the sign of $\hat{\beta}$ flips.

Proof.

■

2.3 Beliefs about Probability of Acceptance

We now turn to another issue that will bias estimates from callback rates. If employers include a candidate's probability of accepting a job in their decision to follow up on an application, then it is clear that researchers cannot distinguish between the two factors by observing callback rates alone. To demonstrate, set aside the issue of heteroskedasticity and assume that $\xi_{ij} \sim \mathcal{N}(0, 1)$.

We now extend the model to allow firms to consider whether a candidate will accept a job offer. Let $\pi_{ij} \in [0, 1]$ be the firm's belief over whether candidate j would accept a job at firm i if offered. This belief may be a function of applicant characteristics observed on the resume. For instance, applicants with a very high GPA might be less likely to accept a job offer at some firms, due to the higher value of their outside options. Assuming the firm incurs a cost at the time of the call back, the firm's probability of calling will decrease with the perceived probability of acceptance (holding constant the candidates perceived value to the firm). As a result, if the perceived likelihood of acceptance is correlated with productivity-relevant characteristics, preferences inferred from callback rates will suffer from further bias. This point is made more formally in Proposition 3.

To help illustrate this, you can imagine a simple structure to firm decisions. Let's say they receive positive utility when they hire an employee with $V_{ij} > V^*$ and negative when $V_{ij} < V^*$. If

the firm experiences costs c of interviewing and recruiting an employee, their utility from making an offer to candidate j is a function of: $\pi_{ij}(V_{ij} - V^* - c) - (1 - \pi_{ij})c$. For simplicity, we could assume the firm hires whenever that equation is positive, reducing to $\pi_{ij}(V_{ij} - V^*) - c > 0$. Depending on c, the expected probability of the applicant accepting the offer plays more or less of a role in whether an offer is made. In other words, for an offer to be made, we need $\pi > \frac{c}{V_{ij} - V^*}$.

Proposition 3 *Assume no omitted variable bias and a homoskedastic error distribution. Then if the probability a candidate accepts an offer, $\pi_(x_j)$ is correlated with V_{ij}, and the employer experiences costs associated with interviewing candidates, then the estimate of β obtained from the regression $D_{ij} \sim \beta x_j + \varepsilon$ will not be unbiased or consistent. Furthermore, $\hat{\beta}$ may have a different sign from β.*

Proof. ■

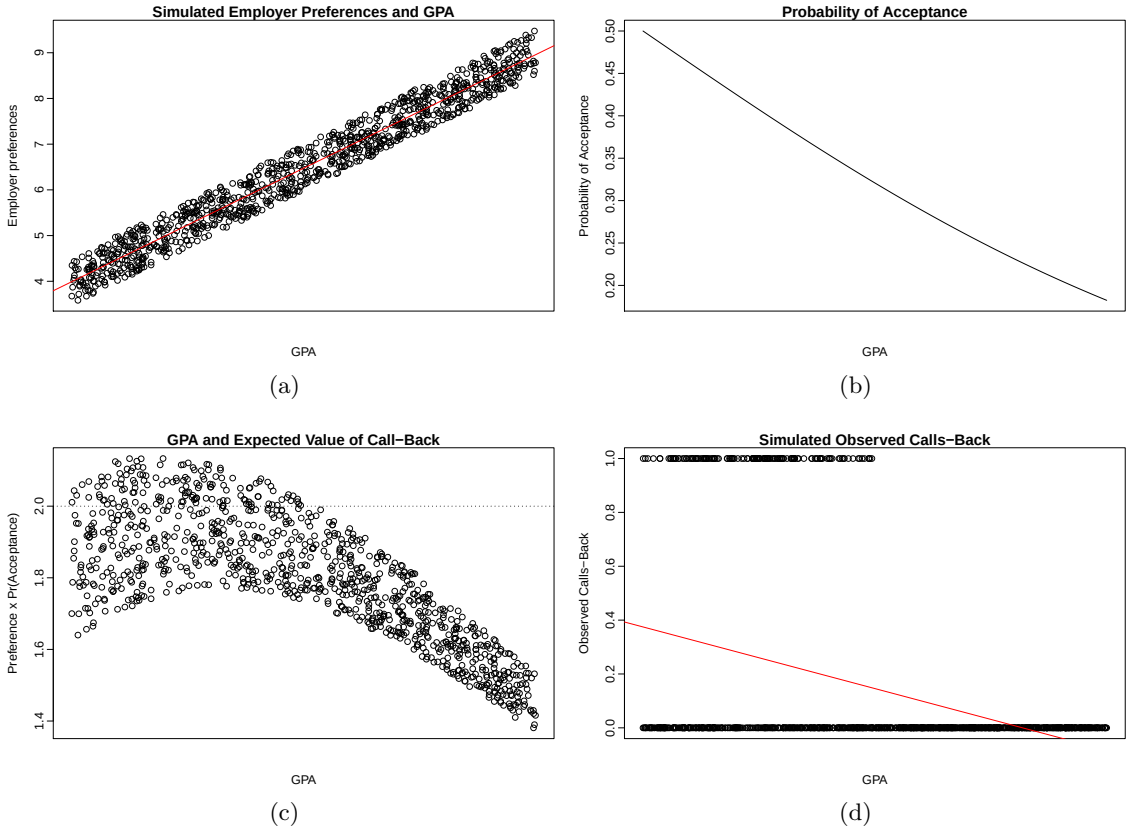
To illustrate this issue using our GPA example, suppose employers believe that candidates with high GPAs are less likely to accept a job offer than candidates with low GPAs. While all candidates under review have applied for the job, this does not imply that all candidates value the job equally, or have comparable outside options. In this scenario, an employer would prefer not to waste resources interviewing candidates who are unlikely to accept an offer. This tradeoff between the quality of the candidate and the likelihood of acceptance suggests another source of bias and inconsistency for the researcher observing call-back rates. An example is shown in Figure 3.

2.4 Model Implications

Call-back rates are indeed a measure of market access, in the sense that the candidates with call-backs are more likely to be offered jobs. However, without identifying preferences directly, it is impossible disentangle perceived candidate quality from the likelihood a given candidate will accept a job offer. Differential call-back rates for candidates with different demographic characteristics is not necessarily evidence of a difference in average perceived quality among employers. Rather, it could suggest either a higher variance in quality among one group, or a difference in the expected probability a candidate from that group would accept a job offer. Of course, it is difficult to convince employers to reveal their preferences directly, particularly on sensitive characteristics such as race or gender. These concerns suggest the need for an experimental design that allows us to randomize candidate characteristics, incentivizes employers to reveal their true preferences about candidates, allows us to separate perceived candidate quality from perceived probability of acceptance, and that does not rely on call-back rates as the outcome of interest.

Can callback rates be informative about preferences? And if so, what exactly can the difference in callback rates tell us? This simple model highlights two embedded assumptions in most resume audit studies that make it possible to make inferences about employer preferences from callback rates:

Figure 3: Simulated Preferences and Calls with Probability of Acceptance



Preferences simulated from data generating process $y = \alpha + \beta X + \varepsilon$, with $\alpha = 3$, $\beta = 5$, and $\varepsilon \sim \mathcal{N}(0, 1)$, shown in 3a. Probability of acceptance represented by the function $1 - \frac{e^{\frac{3}{2}GPA}}{1+e^{\frac{3}{2}GPA}}$, shown in 3b. Figure 3c shows the expected benefit of a callback, preference × probability of acceptance. Figure 3d shows the calls the researcher would observe with cutoff set to $c = 2$, and the linear probability regression she might estimate.

1. ξ_{ij} must be homoskedastic in x_j
2. Probability of acceptance must be constant with respect to the candidate characteristic of interest.

Under these conditions, the difference in callback rates will at least have the same sign as β; however, the magnitude of the estimate will depend on the cutoff threshold V^*, with higher V^* leading to estimates that are closer to 0. The difference in callback rates can be interpreted as firms' marginal preferences toward x_j at threshold V^*.

Callback rates are a useful measure of labor market access, being a necessary step in the hiring process. In addition, as Neumark [2011] points out, employers are prohibited from making assumptions about unobserved characteristics of candidates based on race or sex, so differences in callback rates along these measures are sufficient as legal evidence of a combination of statistical and taste discrimination. However, as described previously (Heckman and Siegelman [1992], Heckman [1998], Neumark [2011]) and illustrated below, callback rates are not a sufficient statistic for measuring preferences toward a characteristic. There are many reasons to prefer to measure preferences directly. In the case of discrimination, distinguishing between statistical and taste-based discrimination allows us to consider policies to correct the causes of inappropriate differential treatment (Gneezy et al. [2012]). This also opens the door to counterfactual policy analysis. For instance, will increasing the number of female graduates with STEM majors lead to increases in female hiring in the sciences? The answer depends not only on the mean productivity of female scientists, but also on the variance and higher moments. In addition, if callbacks follow the threshold model described below, then a change in the threshold could change callback rates for different groups. In other words, a change in threshold (such as a hiring surge) could lead the econometrician to make drastically different conclusions about preferences, though the underlying preferences remain unchanged. As described in Proposition 2, a change in thresholds alone could lead employers to pursue the less-preferred group more often.

The issues illustrated by this simple theoretical framework are essentially a problem of mismeasurement. If the researcher could observe preferences directly, and randomize candidate attributes, she could estimate the relationship between attributes and preferences without bias or inconsistency.

Our approach doesn't attempt to estimate an employee's productivity at a given firm, which should theoretically provide the basis for firm preferences. However, we are interested in subjective preferences since they may act as a blocking point in access to the labor market. In addition, identifying preferences for characteristics in our carefully controlled setting allows us to measure the subjective valuations and tradeoffs that recruiters use when evaluating candidates. This permits a more flexible analysis of elasticities and relative values of employee characteristics.

3 Study Design

The previous section raised concerns about the inferences that can be drawn about employer preferences from both audit studies and resume audit studies. In this section, we describe our experimental design and outline how it allows us to achieve three goals: (1) measure employer preferences over a wide variety of candidate characteristics (including grade point average, previous work experience, skills, major, as well as race and gender), (2) provide incentives for reporting preferences without introducing deception, and (3) recruit employers at highly selective firms who are unreachable for audit and resume audit studies. We call our experimental method an "incentivized resume rating" study. As we elucidate below, it combines the incentives and ecological validity of the field and the control of the laboratory.

The following subsections further explain our experimental design. In section 3, we provide more details about our experimental subjects and how we recruited and incentivized them. In section 3.2, we outline how we created the hypothetical resumes and describe the extensive variation in candidate characteristics that we included in the experiment. In section 3.3, we highlight the two questions that we asked subjects about each hypothetical resume and discuss how our elicitation method addresses the concerns raised in section 6. Finally, in section 3.4, we remark more broadly on our method and its potential use in future studies.

3.1 Employers, Recruitment and Incentives

3.1.1 Employers and Recruitment

Our subject pool was comprised of employers who engage in on-campus recruiting for graduating seniors at our two partner universities, the University of Pennsylvania (Penn) and the University of Pittsburgh (Pitt). At each school, we partnered with the office in charge of on-campus recruiting.[10] The career services offices sent recruitment emails to firms they were in contact with for recruiting (e.g., those that regularly attended career fairs, recruited on campus, or hired students). At Penn, this happened in two waves, around the time of Fall on-campus recruiting in 2016, and around the time of Spring career-fair recruiting in 2017. At Pitt, where the Fall season is less significant, we also recruited employers in two waves, in the Spring of 2017 and the Spring of 2018. The recruitment email advertised: "a new tool that can help you to identify potential job candidates" and invited employers to use the tool to identify candidates.[11] While the recruitment email and the information that employers received at the start of the tool (provided in Appendix 6.1.1) noted

[10] We initially partnered with Penn Career Services since all the authors were affiliated with the University of Pennsylvania. We then reached out to the career services departments of other Pennsylvania schools that differed from Penn on a number of dimensions. University of Pittsburgh's career services department was the most willing to participate, and provided a different environment than Penn for us to run our study.

[11] The full employer recruitment email for Penn is provided in Appendix 6.2.1. The Pitt recruitment email was similar.

that anonymized data from their responses would be used for research purposes, the experience was designed so that — from the perspective of the employers — taking the survey tool resembled using new recruitment software much more than participating in a research study. 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 as will be described below in section 3.1.2 — each employer's one and only incentive for participating in the study is 10 resumes of job seekers that match the preferences they report in the survey tool.[13]

Partnering with career services at Penn and Pitt gave us access to a large number of potential subjects. A key additional advantage of partnering with these career services offices is that we are able to recruit selective employers who are hard — or impossible — to study in audit or resume audit studies because they seek out particular institutions to source applicants and their recruitment methods do not involve screening unsolicited resumes.

3.1.2 Incentives

The main innovation of our experimental design is utilizing a method of incentivized preference elicitation — previously pioneered by Low [2017] in a different context. The method asks subjects to evaluate objects, which are known to be hypothetical, with the understanding that more accurate evaluations will maximize the value of their participation incentive. In our case, employers evaluate 40 hypothetical candidate resumes and, as their only incentive for their participation, we provide each employer with the resumes of 10 real job seekers. The 10 real job seekers come from a larger pool for the study, and we inform employers that we will select the job seekers that match their preferences as reported in their evaluations.[14] Consequently, the participation incentive becomes more valuable as employers' evaluations of candidates better reflect their true preferences for candidates.[15] While this type of non-monetary incentive is new to the labor economics literature, it has many features in common with monetary incentives employed in many laboratory experimental paradigms. In those settings, subjects make choices and they earn more money when their choices are more accurate (or earn higher utility when their choices more accurately reflect their

[12] Indeed, the only thing that differentiates us from the "natural field experiment" definition as defined by Harrison and List [2004] is that subjects know that academic research is ostensibly taking place, even though it is framed as an afterthought 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 their meeting of the potential candidates. For the first three of our four recruitment waves, 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 as researchers described the survey software. They could submit questions via chat box. Attendance at these sessions was low, and so we did not offer them in the last wave.

[14] The recruitment materials stated that: "the tool uses a newly developed machine-learning algorithm to identify candidates who would be a particularly good for for your job based on your evaluations." The process by which we match job seekers to employers based on employer preferences is described in detail in Appendix **??**.

[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.

preferences).

One key design decision that ensures the efficacy of this incentive structure is that we provided no additional incentive besides the 10 resumes of job seekers for participating in the study, which took 27.6 minutes (median) to complete. Limiting the incentive to the 10 resumes has two direct benefits. The first is that it allows us to be confident that participants value the incentive (since they have no other reason to participate in the study). Since subjects care about the incentive and the incentive becomes more valuable as preferences are reported more accurately, we can be confident that subjects have good reason to report their preferences accurately. The second is that it helps to mitigate against potential concerns about social desirability or experimenter demand effects driving any of our results, since subjects who may particularly care about the experimenter's opinions of them — and thus may distort their preferences in the survey tool — can simply choose not to participate in the study.[16] The latter potential concern is also mitigated by the framing of the survey tool as helping employers identify candidates (i.e., rather than about helping researchers learn about employer preferences).

3.2 Resume Creation and Variation

Our survey tool dynamically randomized 40 resumes for each employer who participated in the study. In particular, when the employers began the survey tool, each of the 40 resumes were populated with independently drawn, randomly selected candidate characteristics. As shown in Table 1, and as described below, we randomly varied a set of candidate characteristics related to education; a set related to work, leadership and skills; and the candidate's name to signal race and gender.

A key advantage of our design in identifying preferences for candidate characteristics is that employers opt into the study knowing that they will be asked to evaluate 40 hypothetical resumes. This gives us flexibility to vary multiple candidate characteristics simultaneously and independently across many different resumes. In a traditional audit or resume audit study, researchers are somewhat limited in: (i) the number of candidates or resumes that they can show to any particular employer (i.e., sending too many fake resumes to the same firm might cause employer skepticism and have other unintended consequences) and (ii) how candidate characteristics can covary across resumes (e.g., in a resume audit study, a resume with a very low GPA but very prestigious internships might come off as unusual or suspicious).[17] In our study, however, employers know the resumes they are rating are hypothetical — so employers need not be surprised when our independent random variation generates "unusual" resumes in which candidates appear to be high quality

[16] That is, rather than receive 10 resumes that are unlikely to be good fits for the firm — since they would be based on distorted preferences — the employer can simply choose not to participate in the study.

[17] For example, in their seminal resume audit study paper, Bertrand and Mullainathan [2004] only send two resumes to each firm and only use two quality levels of resumes (a high quality resume and a low quality resume, in which various candidate characteristics vary together).

on some dimensions and low quality on others. In addition, employers evaluate 40 resumes, so we can learn a great deal about their preferences as they face various trade-offs in candidate quality.

We made a number of additional design decisions to both improve the quality of responses and to increase the realism of the study. First, we built the hypothetical resumes using real majors and real "components" (e.g., work experiences and leadership experiences) from real resumes of seniors at the target school (i.e., we used components of real Penn resumes to build the hypothetical Penn resumes and components of real Pitt resumes to build the hypothetical Pitt resumes). Second, we asked the employers to choose the type of candidates that they were interested in hiring, based on major of study. They could choose either "Business, Social Sciences, and Humanities" or "Science, Engineering, Computer Science, and Math." They were then shown hypothetical resumes focused on the set of majors they selected. As is described below, this choice affects a wide range of candidate characteristics — major, work experiences, leadership experiences, and skills — that appeared on all the hypothetical resumes.[18] Third, to enhance realism, and to make the evaluation of the resumes less tedious, we used used 10 different resume templates, which we populated with the candidate characteristics and component pieces described below, to generate the 40 hypothetical resumes. We based these templates on real student resume formats. A sample resume is shown in Appendix 6.1.3.[19]

3.2.1 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.[20] Majors are chosen from a list of majors available at the relevant university, with higher probability put on more common majors. Each major was associated with a degree (e.g., a B.A. or a B.S.) and with the name of the group or school granting the degree within Penn (e.g., "College of Arts and Sciences" or "The Wharton School") and Pitt (e.g., "Dietrich School of Arts and Sciences" or "Swanson School of Engineering").[21]

[18] 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.

[19] We blurred the text in place of a phone number and email address in the templates, since were were not interested in inducing variation in those candidate characteristics.

[20] We suspect that some students omit GPA from the resume rather than display a low GPA. We chose not to test the value of omitted compared to low GPA for this experiment, but it would be an easy modification to our experimental design for future researchers.

[21] Tables 9 and 10 show the list of majors by school and major category and the probability that the major was drawn for each group. As can be seen on the table, most majors are associated with a single degree and are from a single school at the University. In cases where there were overlap in the names of majors we treated majors at different schools within a university as different (e.g., the "Economics" major in the Economics department at Penn always appeared with an indication that the candidate was scheduled to receive a B.A. from Penn's College of Arts and Sciences, this major was treated as distinct from the "Economics" major at Wharton, which always appeared with an indication that the candidate was scheduled to receive a B.S. from the The Wharton School.

Table 1: Randomization of Resume Components

Resume Component	Variable	Description
Personal Information		
Name	*Race*	Race drawn from U.S. distribution (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian)
	Gender	Gender drawn at random (50% male, 50% female)
	First and last name	Drawn from list of 50 possible names for selected race/gender combination (names shown in Table 7)
Education Information		
GPA	*GPA*	Drawn from $Unif(2.90, 4.00)$ to second decimal place
Degree	*Major*	Drawn from a list of majors at the school (majors and weights shown in Tables 9 & 10)
	Degree type	BA, BS fixed to randomly drawn major
	School within university	Fixed to randomly drawn major
	Graduation date	Fixed to upcoming spring (May 2017 or May 2018)
Work Experience		
First job	*Top Internship*	Drawn independently from curated list (50% top internship, 50% regular internship)
	First job location	Fixed to randomly drawn job
	First job dates	Summer after candidate's junior year
	First job description	Bullet points fixed to randomly drawn job
Second job	*Second Internship* *Work For Money*	Drawn from no job (probability 14/40), second internship (13/40), work-for-money job (13/40, see examples in Table 8)
	Second job location	Fixed to randomly drawn job
	Second job dates	Summer after candidate's sophomore year
	Second job description	Bullet points fixed to randomly drawn job
Leadership Experience		
First and second leadership	Leadership location	Fixed to school location for all experiences
	Leadership dates	Start and end years randomized within candidate's college career, with more recent experience coming first
	Leadership description	Bullet points fixed to randomly drawn experience
Skills		
Skills	Skills list	Drawn from curated list.
	Technical skills	Four high-tech skills drawn from {Ruby, Python, SAS, R, Stata, PHP, Perl, Matlab} shuffled and added to skills list with probability 25%.

Italicized variables were randomized for analysis. For degree, first job, second job, and skills, content differs between humanities/social science resumes and science resumes. For name, GPA, and leadership, randomized elements are drawn from the same lists for both resume types.

3.2.2 Quality and Quantity of Work Experience

To vary work experience on the hypothetical resumes, we needed to generate realistic resume components. We generated these by scraping a total of 1,200 real resumes submitted by Penn and Pitt graduating seniors. We then followed a process described in the Appendix to select and lightly sanitize work experience component pieces so that they could be randomly assigned to different resume templates without generating conflicts or inconsistencies on the resume (e.g., we eliminated references to particular majors or to gender or race).[22] We did this procedure twice at each school, once for each set of major types. The logic of doing it separately for each major type was based on the intuition that work experiences could be interchangeable within each group of majors (e.g., internships from "Business, Social Sciences, and Humanities" resumes would not be that unusual to see on any other resume from that group) but were unlikely to be interchangeable across major groups (e.g., internships from "Business, Social Sciences, and Humanities" resumes would be unusual to see on "Science, Engineering, Computer Science, and Math" resumes and vice versa).

Our goal in randomly assigning these work experience components to resumes 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 student had an internship in the summer before their junior year, or only an internship in the most recent summer, the summer before their Senior year. The latter scenario can naturally occur, since students might travel, relax, or even take classes during the summer rather than pursuing an internship in their field. Thus, students in the "more experience" category had two jobs components on their resume, while students in the "less" experience category had only one.

To introduce random variation in *quality* of work experience we grouped work experience components into three categories: (1) "top internships" (defined as the substantive jobs at firms hiring the most students from that school, which tended to be prestigious firms); (2) "work-for-money" jobs, which were paid jobs that seemed unlikely to develop future career prospects (e.g., barista, cashier, waiter, etc.); and (3) standard internships, which comprised all other work experiences.[23]

By sorting internships into those at a top firm, while leaving all the other components unchanged, we were also selecting for higher quality work experience "blurbs" that tended to have more substantive work experience. We thus consider this a good proxy for quality of work experience. The first level of quality randomization was to assign each hypothetical resume to have either a "top" employer or a standard internship in the first job slot (the summer before Senior year). This allows us to detect the impact of having a better quality internship.

[22] For realism, the component pieces included the associated bullet points listed on the real resumes from which the components were drawn. For example, a summer internship component would include a job title, the location of the job, and a few bullet points with a description of the tasks completed while in that internship.

[23] See Appendix Table 8 for a list of work-for-money job titles.

Since many students may have an economic need to earn money during the summer, and thus be unable to take an unpaid internship that could be better for building experience, we wanted to measure whether these "work-for-money" jobs had any value to employers, and if so, to which kinds of employers. We thus additionally assigned resumes that did have a second job to have either a standard job or a "work-for-money" job in the second job slot (the summer before junior year).[24] Resumes were assigned to have one-third no second job, one-third "work-for-money," and one-third standard in this slot. When holding number of jobs constant, this enables us to measure the comparative value of having an internship rather than a work-for-money position.[25]

3.2.3 Leadership and skills

The procedure followed to extract and clean leadership and skills components was identical to that of work experiences. Leadership positions were allowed to be interchangeable across all resumes, however. A scraped leadership experience component includes a title and a few bullet points with a description of the experience (the city of the campus was given as the location of all leadership experiences). A skills component includes a list of all skills on the resume.

Leadership experiences were included primarily to realistically reflect 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]

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. With 75% probability, the hypothetical resume included only this list of skills. With 25% probability, we appended to this list two randomly drawn advanced programming languages and two randomly drawn statistical programs, drawn from {Ruby, Python, SAS, R, Stata, PHP, Perl, Matlab}. This variation allows us to measure how employers value candidates having more technical skills listed on their resume.

3.2.4 Names Indicating Gender and Race

We randomly assigned gender and race by assigning each hypothetical resume a name that would be indicative of gender (male or female) and race / ethnicity (black, white, Hispanic, or Asian).[27]

[24] The resume was constructed dynamically so that if there were fewer jobs listed on the resume this made the work experience section of the resume shorter rather than introducing empty space.

[25] Note that this variation also allows us to measure the relationship between employer evaluations and a candidate having an internship in a big city or with a Fortune 500 company, but these relationships should be interpreted with more caution — in particular, we are not randomly assigning these variable and they are likely correlated with other features of a work experience that employers might care about.

[26] Note that this variation allows us to measure the relationship between employer evaluations and a candidate being a president of a club or the captain of a team, but again these relationships should be interpreted with more caution — these titles were not randomly assigned and are likely correlated with other features of a leadership experience that employers might care about.

[27] For ease of explanation, we will refer to race / ethnicity as "race" from hereafter.

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 (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 racial-indicative last names, under the assumption that last names should not be informative of gender. We used last name data from the 1990 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 combined first and last names within race.[28] The full lists of names are given in Appendix 7.

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). The benefit of this approach is a more subtle variation in race, making it less likely that subjects might be surprised by the racial distribution of the (albeit hypothetical) resume pool and become suspicious that we were studying their racial preferences. The drawback, however, is that this limits our power to detect race-based discrimination since minority names are relatively rare in the data.

3.3 Two Questions

As described in section 2, a limitation of audit and resume audit studies is that the data the researcher observes is a binary measure of call back, which might be influenced by both employer preferences for candidates and employer beliefs that candidates will accept a job if offered. A major advantage of our experimental design is that we can get an independent measure of each of these two factors that might influence call backs. In particular, for each resume we ask employers to answer two questions (see a sample at the bottom of Appendix Figure 6.1.3):

1. "How interested would you be in hiring **[NAME]**?"

2. "How likely do you think **[NAME]** would be to accept a job with your organization?"

In the instructions (shown in Appendix 6.1.1), employers were specifically told to focus only on the candidate's quality when answering the first question (i.e., they were instructed to assume the

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

candidate would accept an offer if given one) and were told to focus on the likelihood a candidate would accept a job offer when answering the second question (i.e., they were instructed to assume they candidate had been given an offer and to assess the likelihood they would accept it).

By asking these two questions, we can both focus on employers' beliefs about how resume characteristics affect perceptions of candidate quality — ostensibly the object of interest in audit and resume audit studies — and beliefs that the candidate would accept a job offer. We can also use the data generated in our experiment to consider the sources of bias discussed in section 2, which might affect the relationship between estimates of candidate characteristics on call back. In particular, we use responses to our two questions to assess whether estimates generated from resume audit studies might be biased (as well as the direction and magnitude of that bias) in section **??**.

3.4 Incentivized Resume Rating Study: Benefits and Costs

The previous sections highlight many of the benefits of using our incentivized resume rating design. The design allows us to estimate preferences for a large range of candidate characteristics, to recruit employers from selective firm who might not be accessible using audit or resume audit methods (because they do not accept unsolicited resumes), to observe employer ratings of 40 resumes, to elicit a measure of quality and of likelihood of accepting the job for each resume, and to do all of it without introducing deception into the study. We believe these advantages are large enough to justify using incentivized resume rating studies to explore employer preferences over candidate characteristics in other settings.[29] Researchers who have access to a subject pool of employers, a pool of real job seekers, and a pool of resumes that can be used to create the hypothetical resumes can potentially run an incentivized resume rating study of their own.[30]

Of course, these advantages also carry with them some costs. We devoted significant time and resources to building the survey tool software, constructing the hypothetical resumes (e.g., scraping and sanitizing the resume components), and developing a process to match employer preferences to candidates. Thankfully, however, our experience has shown that while many of these costs are fixed costs (some borne by us that can be leveraged by other researchers, others that are fixed costs per research team), the marginal cost of building the survey tool for a new setting (e.g., at Pitt after running it at Penn) was much more manageable. Once the resume tool was up and running at Penn, the marginal cost for Pitt was mostly limited to constructing Pitt-specific hypothetical resumes. We therefore believe other research teams will find it fruitful to conduct incentivized resume rating studies of their own.

Additionally, because part of the benefit of the incentivized resume rating design is that it does

[29] In addition, the general methodology of incentivizing subjects can also be used in other domains as shown by Low [2017].

[30] Depending on the timing of recruiting real job seekers and soliciting employers as subjects, the resumes of the real job seekers can also be used to construct the hypothetical resumes.

Table 2: Experimental Waves

Location	Timing	Subjects	Observations
University of Pennsylvania	Fall, 2016	46	1,840
University of Pennsylvania	Spring, 2017	26	1,040
University of Pittsburgh	Spring, 2017	50	2,000
University of Pittsburgh	Spring, 2018	36	1,440

not deceive unwitting subjects, it also cannot be inherently absent of all experimenter demand effects. We believe this impact on end results is small, since evidence for things like racial and gender discrimination have been shown in non-incentivized resume rating studies done in the lab, where other incentives for truth-telling are entirely absent. We discuss this and other considerations around using incentivized resume rating to examine discrimination more thoroughly in Section **??**.

4 Results

4.1 Data

We recruited a total of 158 employers through our partnerships with career services offices at the University of Pennsylvania and the University of Pittsburgh. As described in section 3, each employer rates 40 unique, hypothetical resumes with randomly assigned candidate characteristics. These employers were recruited in four waves listed in Table 2. We ran waves at Penn (in Fall 2016 and Spring 2017) and two waves at Pitt (in Spring 2017 and Spring 2018).

As shown in Table 3, the employers who participated in our study as subjects were primarily female and primarily white. They worked at firms of a variety of sizes with a slight tilt towards very large firms that were over 1000 employees. Respondents at Pitt were marginally more likely to be female ($p = 0.08$), but were from organizations of approximately the same size as at Penn ($p = 0.14$).

4.2 Candidate Desirability

We start by restricting our analysis to the first question, "How interested would you be in hiring [name]," which we call V_{ij}. V_{ij} is an employer i's rating of a resume j on the 1-10 Likert scale for either candidate quality or likelihood of acceptance, depending on the regression being run. We begin by showing OLS results, which tell you the impact on employer preferences at the mean,

Table 3: Employer Characteristics

	Pitt	Penn	Pooled
Subject Gender			
Male	23	29	52
Female	62	43	105
Subject Ethnicity			
Asian	1	11	12
Caucasian	74	57	131
Black, Hispanic, Other	11	4	15
Organization Size			
1-99 employees	15	29	44
100-999 employees	28	16	44
1,000 or more employees	42	27	69
N	85	72	157

using the following specification:

$$V_{ij} = \beta_0 + \beta_1 GPA + \beta_2 TopInt + \beta_3 SecondInt + \beta_4 WFM + \beta_5 TechSkills + \beta_6 NotWhiteMale + \alpha_i + \gamma_j + \xi_{ij},$$

where α_i are rater fixed effects and γ_j includes leadership and major fixed effects.

Table 4 shows that employers respond strongly to the candidate's human capital experience. An increase in GPA of one point increases ratings on the Likert scale by 2.2 points, suggesting employers find grade point average to be a particularly relevant metric of candidate quality. The standard deviation of quality ratings is 2.81, suggesting that a full point improvement in GPA (going from a 3.0 to 4.0, for example) moves quality ratings almost a standard deviation.

We now turn to the effect of quantity and quality of work experiences on employers' perceptions of candidate quality. The coefficient on Top Internship (which stands in for higher-quality internships) reveals that the quality of the work experience a candidate has in the summer before senior year has a large impact on quality ratings. This effect is 0.9 Likert scale points. By using GPA as a "numeraire," we can interpret this result. An upgrade from a standard to a top internship is almost equivalent in employer ratings to a half-point improvement in GPA, meaning a candidate with a 3.5 and a high-powered internship would be viewed equivalently to a candidate with a 4.0 but a lower-quality internship.

Having a Second Internship, relative to no listed work experience for the summer before junior year, provides a 0.5 Likert scale point boost to ratings. In others words, a second summer of work experience is worth half as much as better quality experience in the final summer before graduation. On the other hand, listing Work for Money does not provide any statistically significant increase in improved quality. We can also reject that work-for-money jobs and second internships are treated

Table 4: Desirability

	All	Humanities & Social Sciences	Science & Math
GPA	2.195***	2.300***	1.852***
	(0.129)	(0.153)	(0.243)
Top Internship	0.902***	1.039***	0.530***
	(0.0806)	(0.0944)	(0.173)
Second Internship	0.463***	0.514***	0.291
	(0.0947)	(0.114)	(0.187)
Work for Money	0.149	0.114	0.319*
	(0.0913)	(0.109)	(0.185)
Technical Skills	-0.0680	-0.0492	-0.171
	(0.0900)	(0.106)	(0.186)
Not White Male	-0.117	-0.0110	-0.399**
	(0.0842)	(0.0998)	(0.188)
Observations	2880	2040	840
F-test p-value for Majors	< 0.001	0.0036	< 0.001
F-test p-value for Leadership	0.0649	0.0246	< 0.001

Standard errors in parentheses

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

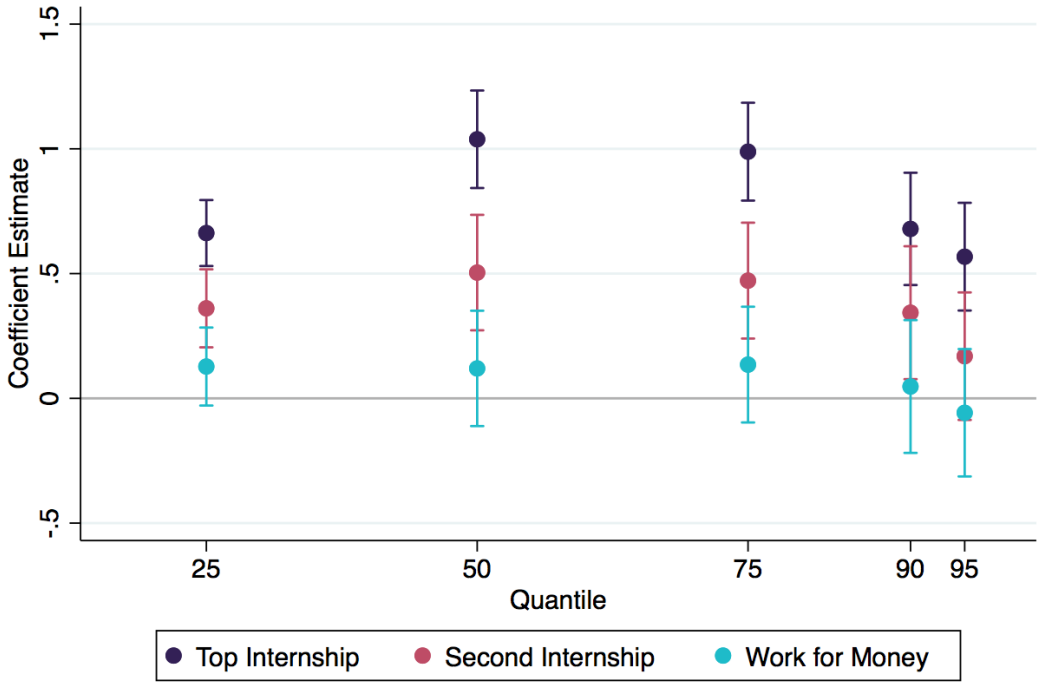
equivalently, at the 5% level. We see no effect from increased Technical Skills, and no impact of Not White Male in aggregate.

This latter effect, of no aggregate discrimination on the hiring desirability rating contrasts somewhat with existing literature, from both resume audit studies and unincentivized laboratory exercises, showing discrimination. However, we should perhaps not be surprised, since our pool of employers is substantially different than those usually targeted through resume audit studies, and may in fact have positive tastes for diversity.

Major fixed effects do matter significantly, showing that horizontal differentiation is important to employers. As one example, the effect of getting a Wharton degree, for humanities employers, is worth 0.4 Likert points relative to the average Arts & Sciences major we present. Leadership fixed effects are marginally significant.

Given that we asked employers to choose between two major groupings, we also break these results down to see if the different fields have differing preferences. We find that employers in STEM, in column 3, seem to respond less strongly to the quality and quantity of candidate work experience than employers in humanities. This could indicate that these employers are more concerned with horizontal fit. In column 3 we also see that STEM employers do rate female and minority candidates 0.4 points lower on average. This result provides some reassurance that employers do not simply ignore the name on the resume when deciding on their ratings. Whether through implicit bias or explicit preference, STEM employers respond that they are less interested in candidates who are

Figure 4: Human Capital Characteristics at Various Quantiles



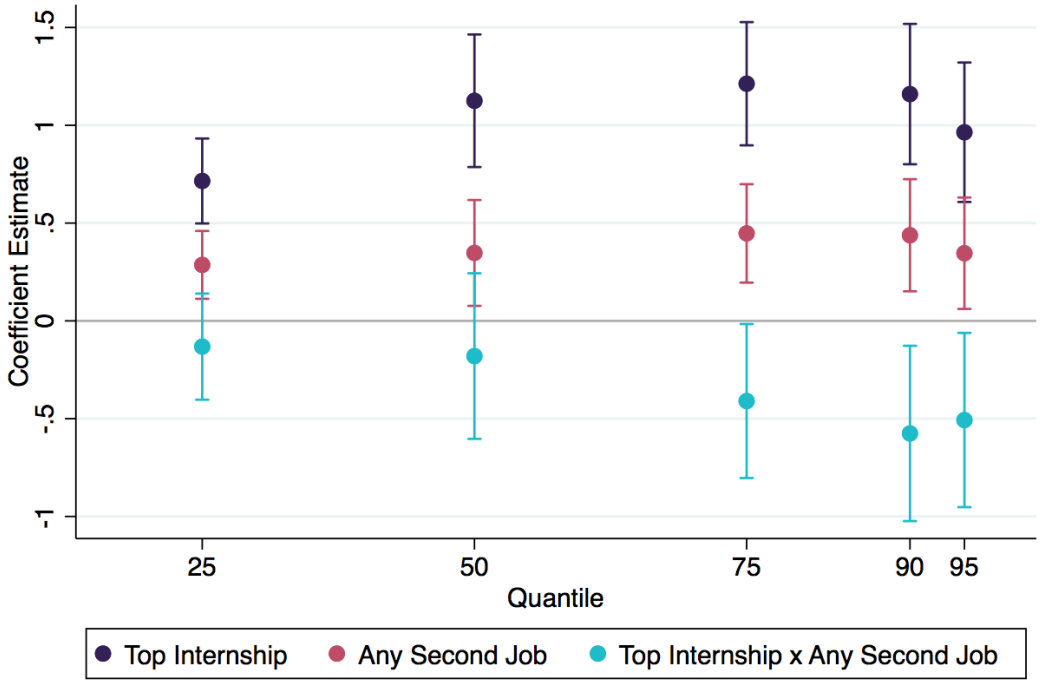
Bars indicate 95% Confidence Intervals.

not white and male.

We now examine the effects of our key human capital variables, Top Internship, Second Internship, and Work for Money, at different quantiles of employer preferences. In Figure 4, we show the regression coefficients on each variable from quantile regressions at the 25th, 50th, 75th, 90th and 95th percentiles. Starting with Top Internship, we see that it is significantly positive throughout all quantiles. The coefficient size is much larger for the 50th and 75th percentile of employer ratings, though, than it is at either the bottom or the top of the distribution. Similarly, Work for Money is statistically zero at all percentiles. Having a Second Internship, though, is statistically significantly positive for most quantiles, but not differentiable from zero at the 95th percentile. This provides an example of where resume audit studies with very low callback rates might draw different conclusions from those with higher callback rates.

Next, in Figure 5, we examine how the first and second job slot on the resume interact between one another. In this case, we combine Second Internship and Work for Money as one "Any Second Job" parameter, which is equal to one for either type of summer experience. Both having a higher

Figure 5: Interaction Between Work Experiences

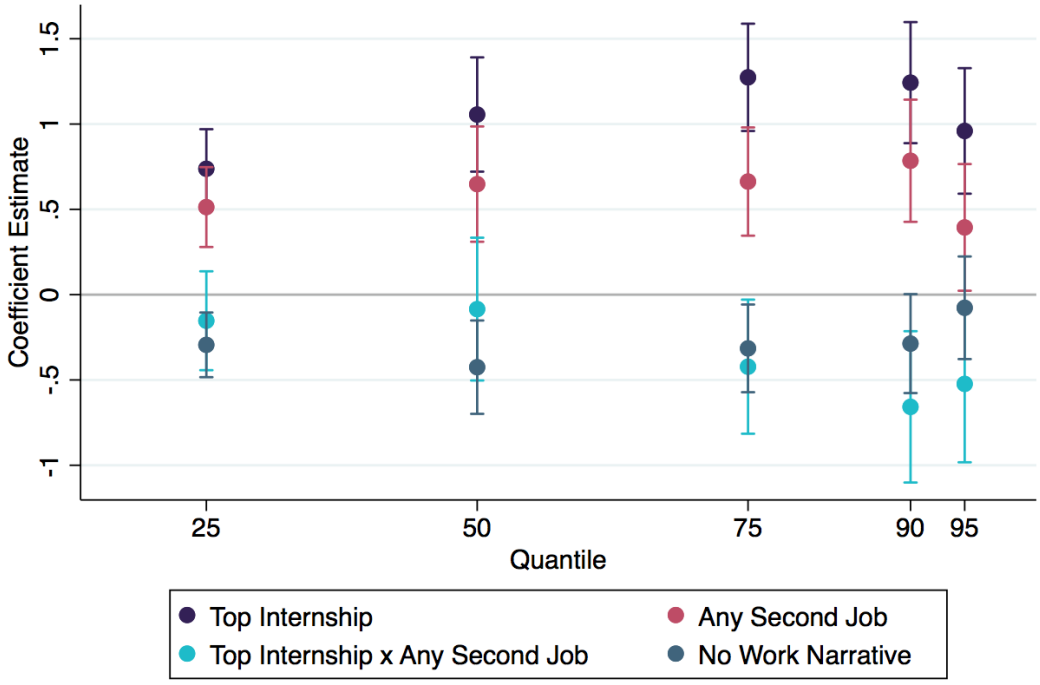


quality first work experience and having any second work experience are significantly positive. However, the interaction between the two is negative at the 75th and higher percentiles. This can be interpreted as saying that either having a Top Internship makes the second experience matter less, or having a second work experience that is less illustrious than the first "takes the shine off" the first high-quality experience.

One could also be concerned that a reason for the negative interaction is that the random assignment inherent in our design creates mismatched internships, which may not form a cohesive "narrative" of work experience. We test this directly. First, we have undergraduate RAs (at the University of Pennsylvania) rate every single possible combination of work experiences for how strongly they form a common narrative. In Figure 6, we also include a control for having "No Work Narrative"—two pieces of work experience that are not related to one another. We see that having a lack of narrative to one's work experience is penalized, but that this does not reduce the negative interaction between Top Internship and Any Second Job.

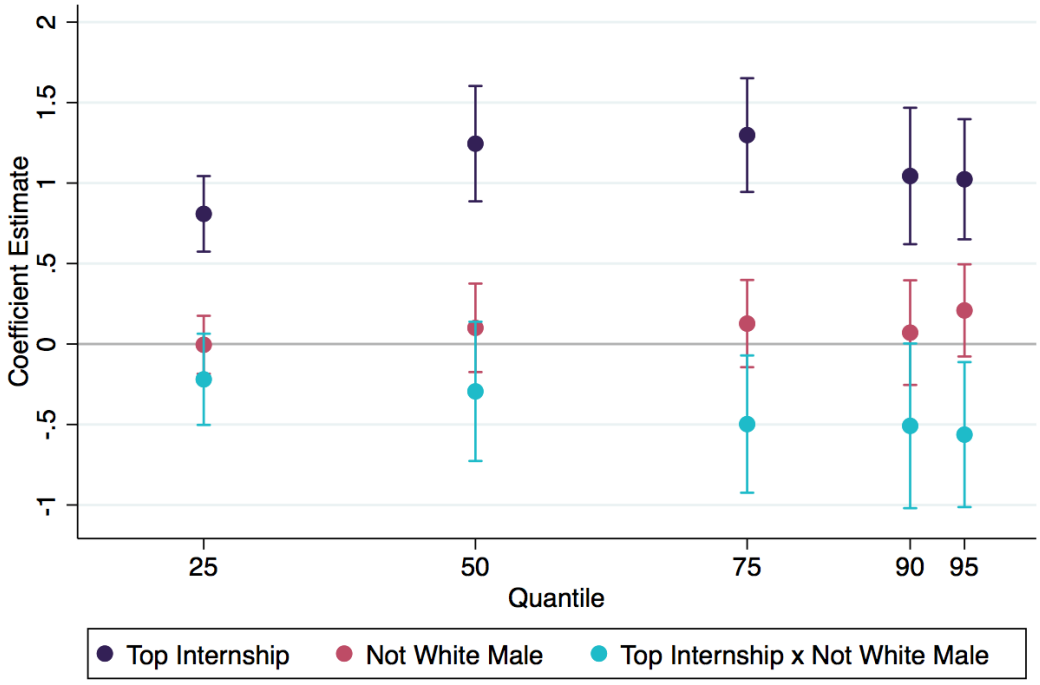
We now transition to examining the impact of demographic characteristics more closely. We

Figure 6: Effect of Narrative



Bars indicate 95% Confidence Intervals.

Figure 7: Interaction Between Demographics and Work Experience

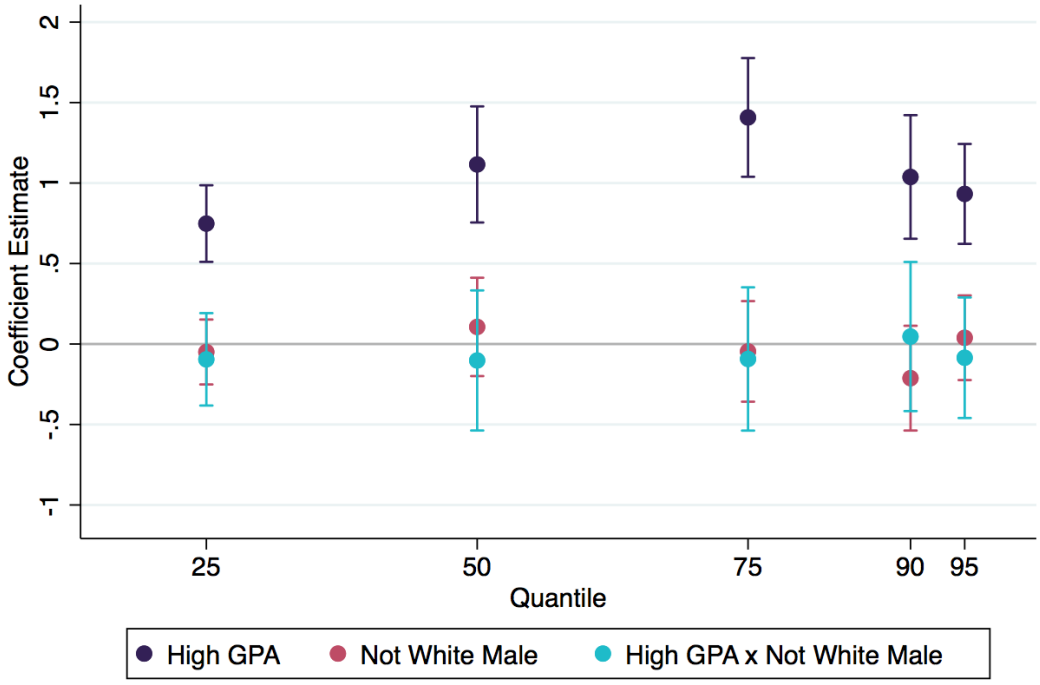


Bars indicate 95% Confidence Intervals.

have already shown that STEM employers express a lower preference for female or minority candidates. To further validate that our methodology elicits a response by employers to demographic characteristics, we test a hypothesis from **?**, who show that resumes with black names also receive a lower "return to quality," in that callback rates are more similar between low and high quality resumes for black names. In Figure 7, we show the interaction between a candidate being a Not White Male and having a Top Internship, as well as the main effects. The interaction is statistically significantly negative at the 75th, 90th, and 95th percentile, confirming that in our experiment, under-represented groups also receive a lower benefit from quality improvements. This penalty is present at the relevant selectivity in the **?** experiment, since callback rates were under 10% in that paper. (The results also appear in an OLS specification.)

We repeat this exercise by upgrading multiple characteristics at once, comparing having a Top Internship, Second Internship, and above-median GPA to resumes without any of those characteristics, and again find a statistically significantly negative interaction. In Figure 8, we show the results from interacting with above-median GPA alone. In this case, we do not see a negative interaction

Figure 8: Interaction Between Demographics and GPA



effect. This difference between Top Internship and GPA interacted with Not White Male begins to suggest a possible mechanism for the discrimination. Perhaps, if an employer has positive tastes for diversity (or believes they do — since implicit bias may cancel out this taste), they believe other employers do as well. Therefore, securing a high quality internship may be a less strong signal of candidate quality if one is from an under-represented group. GPA, on the other hand, may not have this asymmetry in signaling effect between different types of candidates.

4.2.1 Candidate Likelihood of Acceptance

We now turn to our second measure, "How likely do you think [name] would be to accept a job with your organization?" One might wonder whether employers respond to this second measure independently from the first. Our data confirms there is additional information in the second measure. Although the two measures are positively correlated, indicating that employers believe candidates they think are a good fit for their organization are also likely to be interested in the firm, we can examine the residual variation in Likelihood of Acceptance after controlling for Hiring

Desirability. When we regress the Likelihood measure on dummies for each Desirability rating, we find that the residuals have a strong negative relationship with our most objective measure of quality, GPA. In other words, holding constant how much I like a candidate, I think a candidate with a higher GPA is less likely to accept an offer with my firm. This means I would prefer idiosyncratic fit to universal quality, which also makes the candidate more appealing to others. This measure also interacts in expected ways with firm characteristics, for example, showing a more negative relationship with GPA for less prestigious firms.

Table 5 then regresses this second measure on our key variables, repeating the specification in Table 4 with Likelihood of Acceptance as the dependent variable. We find that firms believe that female or minority candidates are less likely to accept a position with their firm, by 0.2 Likert points.

Table 5: Likelihood of Acceptance

	All	Desirability < 5	Desirability ≥ 5
GPA	0.734***	-0.341**	-0.133
	(0.120)	(0.140)	(0.144)
Top Internship	0.666***	0.435***	0.0632
	(0.0763)	(0.0910)	(0.0880)
Second Internship	0.393***	0.293***	0.194*
	(0.0910)	(0.105)	(0.104)
Work for Money	0.200**	0.0895	0.136
	(0.0895)	(0.0991)	(0.106)
Technical Skills	-0.105	0.00508	-0.119
	(0.0862)	(0.0982)	(0.0962)
Not White Male	-0.197**	-0.0664	-0.208**
	(0.0805)	(0.0913)	(0.0919)
Observations	2880	1367	1513

Standard errors in parentheses

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

Our Career Services partners interpreted this result as reflecting reports they have heard from firms that highly qualified female and minority candidates receive many offers, and because they can only accept one job, therefore end up turning down many jobs. To gather very suggestive evidence as to whether this perception could indeed drive our result, we split our sample by whether the candidate is rated above or below 5 on the Hiring Desirability Likert scale. We find the result indeed appears to be driven by the highly rated candidate.

Note that this belief could produce (or exacerbate) lower callback rates for female or minority candidates, under the weak assumption that employers experience costs to calling, interviewing, and recruiting candidates. In this case, the expected payoff of the firm will depend on the candidate's likelihood of accepting the job, and they may call back candidates with a low likelihood less, even

if they rate these candidates highly on quality. In other words, firms with limited interview slots may engage in "top-coding" to ensure candidates with the best chance of success. This essentially creates an omitted variable problem in resume audit studies, but it is one not solved with randomization, since the appeal of a trait and its impact on "gettability" are assigned simultaneously. To our knowledge, this is the first evidence that firms' beliefs on Likelihood of Acceptance differ by demographic groups.

4.2.2 Firm-Level Heterogeneity

Finally, given that we believe one of the advantages of our experimental paradigm is accessing employers who have not been studied by resume audits, we present some results on firm-level differences in our findings.

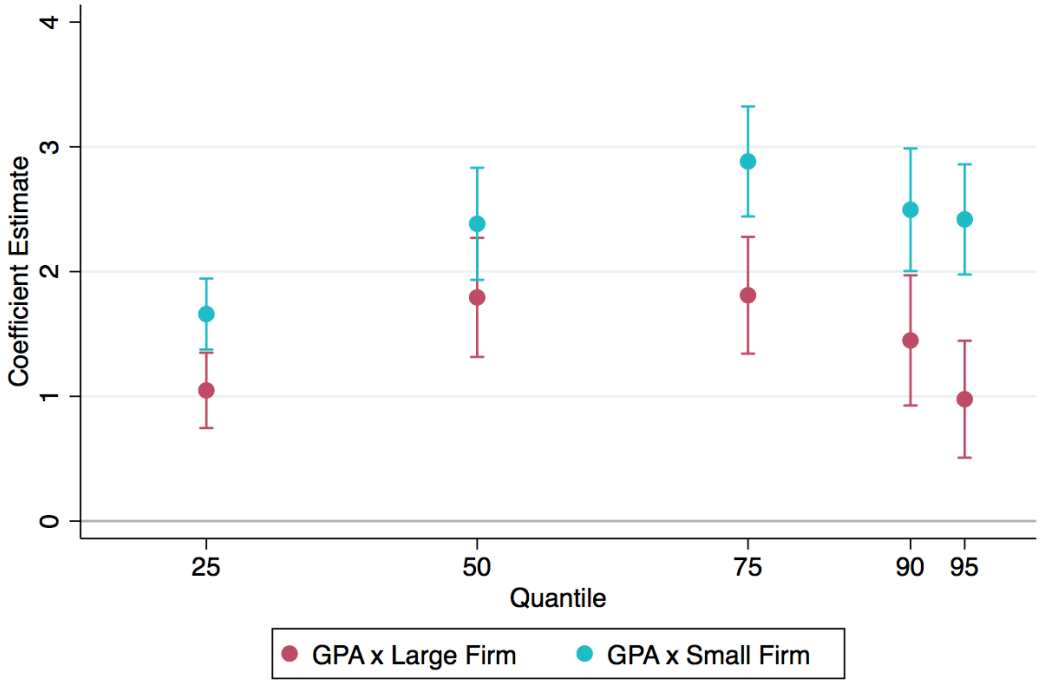
First, we show the movement on our estimates for Top Internship based on organization size in Figure 10, splitting our sample into firms smaller than 500 people ("Small Firm"), roughly the median in our data, and those larger. We find that smaller firms have a statistically significantly higher taste for high-quality internships than larger firms. This result is consistent for GPA as well, shown in Figure 9, although the difference is only statistically significant at some quantiles. Why would smaller firms have an apparently stronger preference for "vertical" quality? One could speculate that at small firms, workers may be asked to wear multiple hats, and thus firms seek "best athlete" generalists. On the other hand, larger organizations have more specialized roles, and so firms may respond more to horizontal fit than these more universal measures of quality.

We also repeated our experiment at the University of Pittsburgh, to test if our results had external validity outside the Ivy League, as well as show how firms recruiting at a slightly less prestigious university may differ in their preferences. Our results are highly consistent. However, interesting differences emerge. For example, GPA appears to be a less strong indicator of candidate quality in the eye of recruiters at Pitt. Secondly, while Second Internship and Work for Money were statistically different at Penn, these two types of jobs elicit a statistically identical response at Pitt. This is consistent with recruiter reports that Work for Money jobs might reflect valuable work ethic and customer service experience that Pitt employers valued.

5 Conclusion

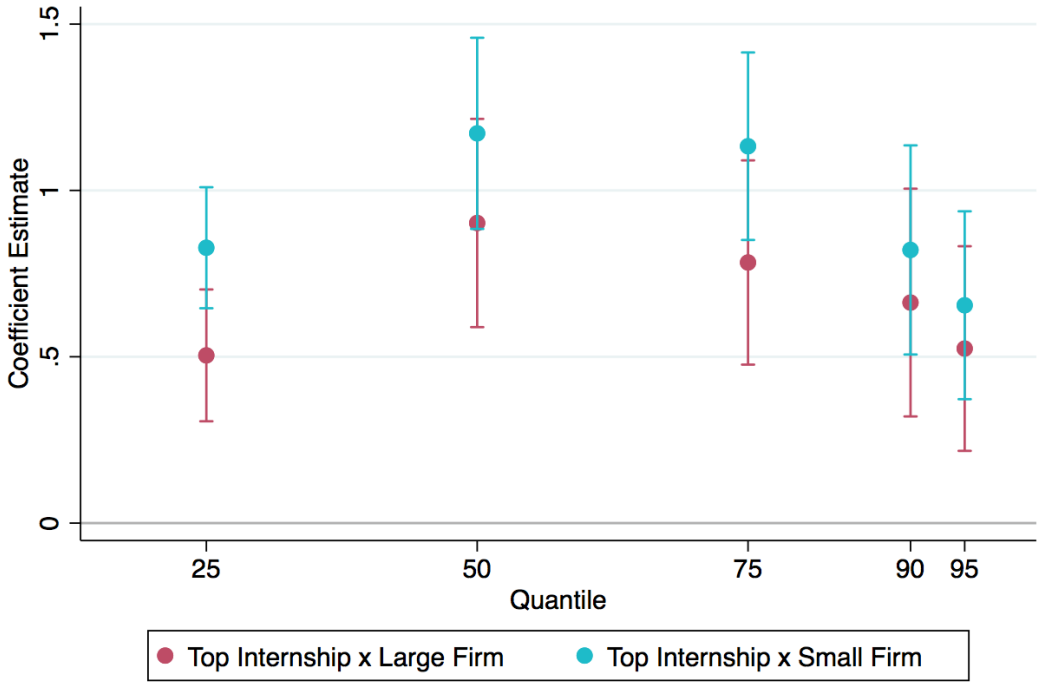
Note that while here we explore incentivized resume rating, what we are pioneering is really part of a broader class of incentivized preference elicitation. Economics has had some success with eliciting preferences for things such as risk aversion (choices between lottery compensation and cash payments) and willingness to pay (Becker DeGroot Marschak) in incentivized way, but there has been little innovation in the domain of eliciting truthful responses about who one would like to date

Figure 9: Interaction Between Firm Size and GPA



Bars indicate 95% Confidence Intervals.

Figure 10: Interaction Between Firm Size and Work Experience



Bars indicate 95% Confidence Intervals.

Table 6: Desirability at Penn and Pitt

	Penn	Pitt
GPA	2.195***	0.263**
	(0.129)	(0.113)
Top Internship	0.902***	0.222***
	(0.0806)	(0.0741)
Second Internship	0.463***	0.212**
	(0.0947)	(0.0844)
Work for Money	0.149	0.154*
	(0.0913)	(0.0807)
Technical Skills	-0.0680	0.107
	(0.0900)	(0.0768)
Not White Male	-0.117	0.00297
	(0.0842)	(0.0710)
Observations	2880	3440
F-test p-value for Majors	< 0.001	< 0.001
F-test p-value for Leadership	0.0649	0.937

Standard errors in parentheses

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

or who one would like to hire. Our methodology, first applied to dating in Low (2017), could be expanded for preference elicitation in other settings where matches with preferred outcomes could be provided as an incentive.

Labor economists are interested in understanding how both human capital and demographic factors are treated by economists. Audit studies and correspondence audit studies have been one powerful tool in the arsenal, due the advantage of allowing characteristics to be randomly assigned.

However, we demonstrate that while correspondence studies solve some of the issues of traditional audit studies, they also introduce new issues by identifying off of callback rates, which we show can be a biased estimate of underlying employer preferences for multiple reasons. First, because callback rate reflects being above a certain threshold of expected productivity, measurement of underlying preferences can be biased not only if the mean of all other variables is not zero (presumably solved by randomization) but also if the variance (or other moments) of other characteristics is related to the X-variable of interest, i.e., heteroskedasticity. Secondly, since not all applicants will accept a job if offered, and employers incur costs from interviewing and recruiting applicants, callback decisions may also be affected by an applicant's likelihood of acceptance. If this acceptance probability is correlated with the variable of interest, it will again bias measurement of employer preferences.

We simulate distributions to demonstrate the potential bias that could arise at various callback thresholds.

We then introduce a new methodology that solves these problems, while remaining incentive compatible, and not requiring deception. We call our methodology incentivized resume rating. By deploying our incentivized tool with the career services offices of two major universities, we demonstrate its ability to measure the returns to human capital characteristics that would be difficult to capture in an audit study, for example, the benefits of GPA, additional work experience, paid summer jobs, and field of study.

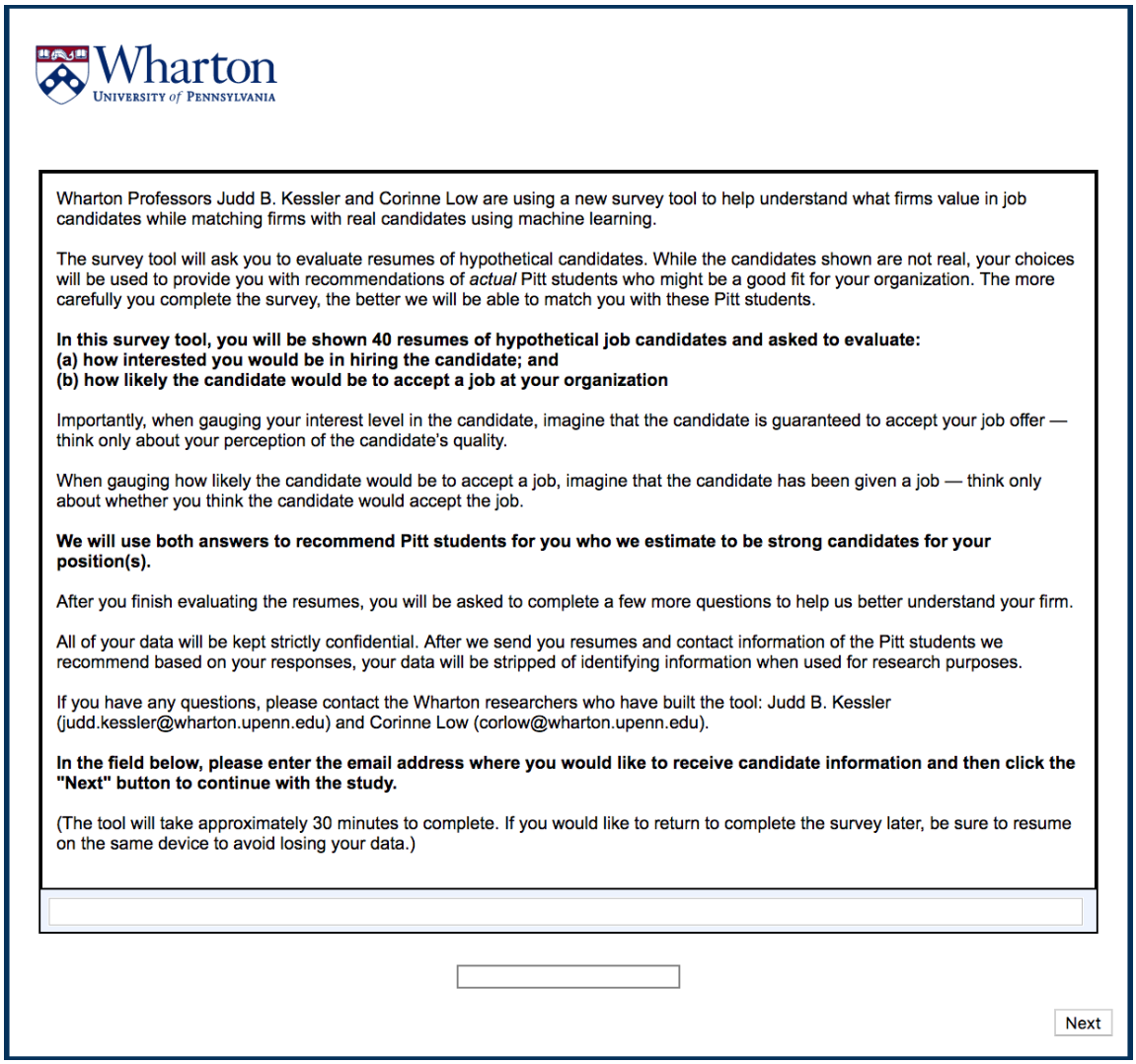
We also show a novel channel for gender discrimination: that competitive employers believe female candidates are less likely to accept jobs if offered. If this is the case in other samples, employers might rationally expend less energy recruiting female candidates, since they believe they would not accept. In a resume audit study, this mechanism would be indistinguishable from employers having a lower preference for female candidates.

Resume audit studies have produced notable insights about how job seekers of different demographic characteristics are treated differently than employers. But, if we wish to go beyond knowing that certain traits are not ignored (though legally they should be) and actually understand the shape and drivers of these preferences, more sophisticated tools are needed. Moreover, if we wish to avoid over-using a tool that has ethical as well as statistical concerns, replacement methodologies should be in high demand. We provide one possible methodology that can be used to measure employer preferences over a variety of job seeker characteristics in multiple contexts. The tool is easy to use, cheap to implement, and flexible to incorporate adjustments and improvements for other research questions.

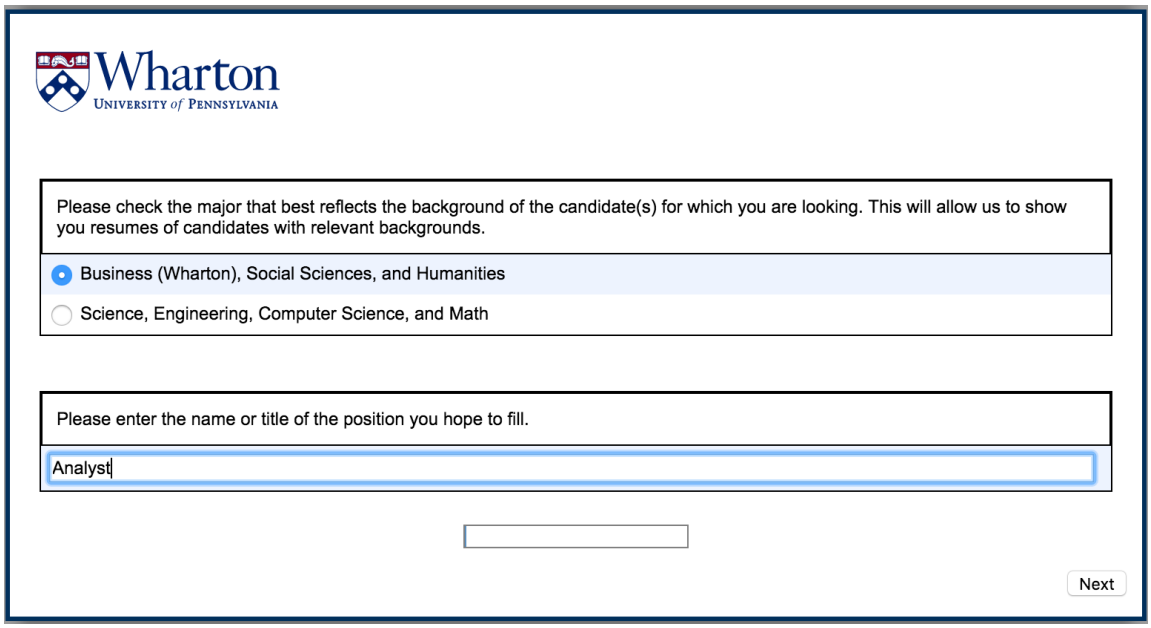
6 Appendix

6.1 Survey Tool Sample

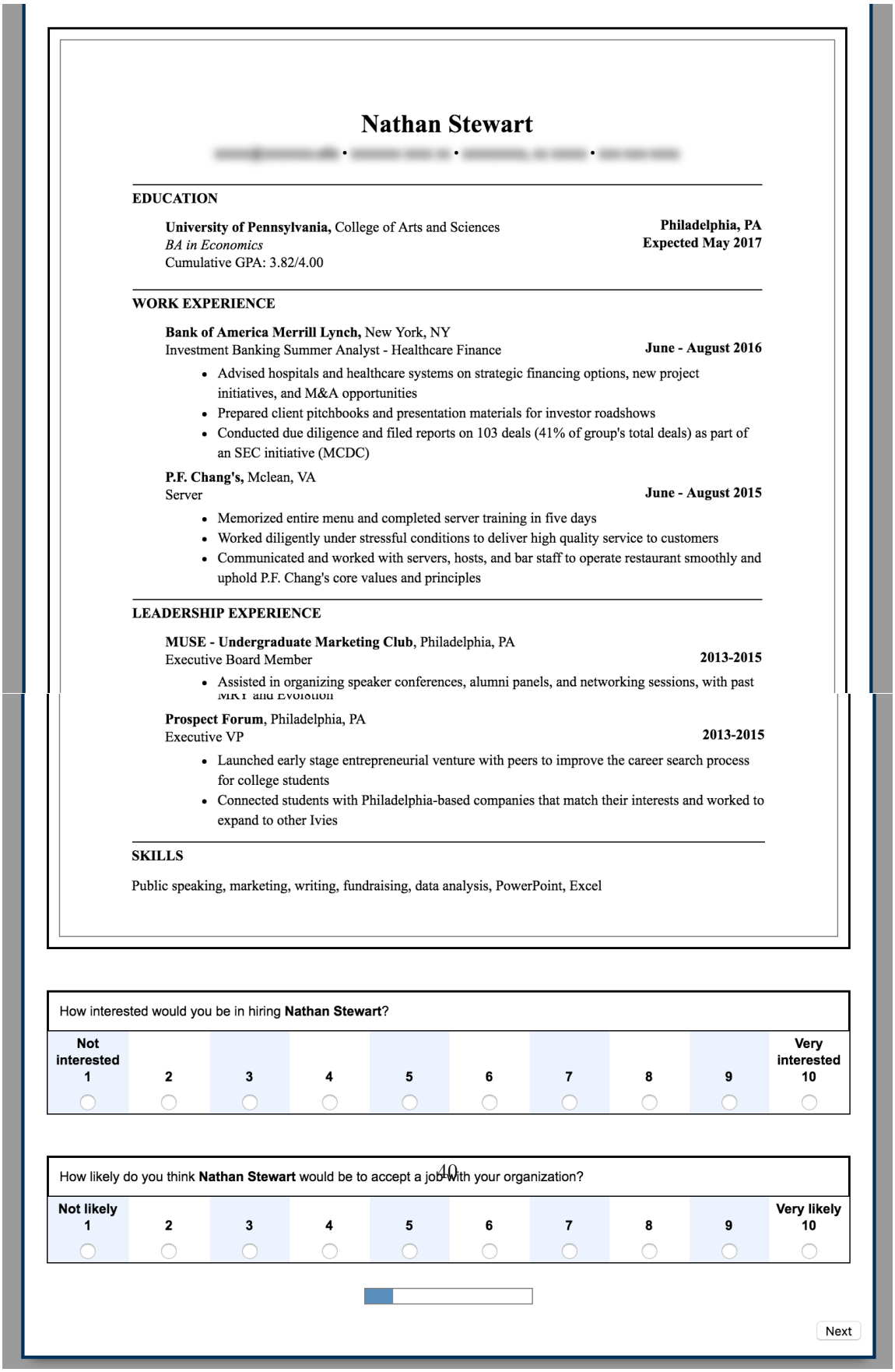
6.1.1 Consent



6.1.2 Major Type Selection



6.1.3 Sample Resume



Nathan Stewart

[redacted]

EDUCATION

University of Pennsylvania, College of Arts and Sciences — **Philadelphia, PA**
BA in Economics — **Expected May 2017**
Cumulative GPA: 3.82/4.00

WORK EXPERIENCE

Bank of America Merrill Lynch, New York, NY
Investment Banking Summer Analyst - Healthcare Finance — **June - August 2016**

- Advised hospitals and healthcare systems on strategic financing options, new project initiatives, and M&A opportunities
- Prepared client pitchbooks and presentation materials for investor roadshows
- Conducted due diligence and filed reports on 103 deals (41% of group's total deals) as part of an SEC initiative (MCDC)

P.F. Chang's, Mclean, VA
Server — **June - August 2015**

- Memorized entire menu and completed server training in five days
- Worked diligently under stressful conditions to deliver high quality service to customers
- Communicated and worked with servers, hosts, and bar staff to operate restaurant smoothly and uphold P.F. Chang's core values and principles

LEADERSHIP EXPERIENCE

MUSE - Undergraduate Marketing Club, Philadelphia, PA
Executive Board Member — **2013-2015**

- Assisted in organizing speaker conferences, alumni panels, and networking sessions, with past MKT and Evolution

Prospect Forum, Philadelphia, PA
Executive VP — **2013-2015**

- Launched early stage entrepreneurial venture with peers to improve the career search process for college students
- Connected students with Philadelphia-based companies that match their interests and worked to expand to other Ivies

SKILLS

Public speaking, marketing, writing, fundraising, data analysis, PowerPoint, Excel

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

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

How likely do you think **Nathan 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

6.2 Recruitment Materials

6.2.1 Email to Recruiting Firms

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

6.2.2 Email to Graduating Seniors

Subject: An Opportunity To Reach More Employers

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. Nearly 50 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!

The deadline to submit your resume is Saturday, September 3rd at 11:59 pm. Uploading your resume takes less than a minute, and only those who upload their resume through this link can participate in the pilot.

To upload your resume, click here: {Survey Link}

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.

Best, {Name of University Career Services Representative}

6.3 Resume Materials

6.3.1 Names Indicative of Race & Gender

Table 7: Names Populating Resume Tool

Asian Female	Asian Male	Black Female	Black Male	Hispanic Female	Hispanic Male	White Female	White Female
Tina Zheng	Richard Thao	Jamila Washington	Rashawn Washington	Ivette Barajas	Andres Barajas	Allyson Wood	Kyle Wood
Annie Xiong	Samuel Truong	Asia Jefferson	Devonte Jefferson	Nathalie Orozco	Julio Orozco	Rachael Sullivan	Derek Sullivan
Julie Xu	Daniel Cheung	Essence Banks	Marquis Banks	Mayra Zavala	Marcos Zavala	Katharine Myers	Connor Myers
Michelle Zhao	Alan Tsai	Monique Jackson	Tyree Jackson	Luisa Velazquez	Mike Velazquez	Colleen Peterson	Douglas Peterson
Linda Zhang	Paul Li	Tianna Joseph	Lamont Joseph	Jessenia Meza	Jose Meza	Meghan Miller	Spencer Miller
Anita Zhu	Steven Zhang	Janay Mack	Jaleel Mack	Darlene Juarez	Alfredo Juarez	Meaghan Murphy	Jackson Murphy
Alice Jiang	Matthew Zheng	Nia Williams	Javon Williams	Thalia Ibarra	Fernando Ibarra	Lindsey Fisher	Bradley Fisher
Esther Zhou	Alex Vu	Latoya Robinson	Darryl Robinson	Perla Cervantes	Gustavo Cervantes	Paige Cox	Drew Cox
Winnie Thao	Joshua Vo	Jalisa Coleman	Kareem Coleman	Lissette Huerta	Adonis Huerta	Katelyn Cook	Lucas Cook
Susan Huang	Brandon Lu	Imani Harris	Kwame Harris	Daisy Espinoza	Juan Espinoza	Jillian Long	Evan Long
Sharon Yang	Henry Dinh	Malika Sims	Deshawn Sims	Cristal Vazquez	Jorge Vazquez	Molly Baker	Adam Baker
Gloria Hwang	Philip Hsu	Keisha James	Terrell James	Paola Cisneros	Abel Cisneros	Heather Nelson	Harrison Nelson
Diane Ngo	Eric Liang	Shanell Thomas	Akeem Thomas	Leticia Gonzalez	Cesar Gonzalez	Alison Hughes	Brendan Hughes
Carmen Huynh	David Yoon	Janae Dixon	Daquan Dixon	Jesenia Hernandez	Alberto Hernandez	Bridget Kelly	Cody Kelly
Angela Truong	Jonathan Yu	Latisha Daniels	Tarik Daniels	Alejandra Contreras	Elvin Contreras	Hayley Russell	Zachary Russell
Janet Kwon	Andrew Trinh	Zakiya Franklin	Jaquan Franklin	Iliana Ramirez	Ruben Ramirez	Carly Roberts	Mitchell Roberts
Janice Luong	Stephen Yi	Kiana Jones	Tyrell Jones	Julissa Esparza	Reynaldo Esparza	Bethany Phillips	Tyler Phillips
Irene Cheung	Ryan Nguyen	Ayana Grant	Isiah Grant	Giselle Alvarado	Wilfredo Alvarado	Kerry Bennett	Matthew Bennett
Amy Choi	Aaron Jiang	Ayanna Holmes	Omari Holmes	Gloria Macias	Francisco Macias	Kara Morgan	Thomas Morgan
Shirley Yu	Kenneth Zhao	Shaquana Frazier	Rashad Frazier	Selena Zuniga	Emilio Zuniga	Kaitlyn Ward	Sean Ward
Kristine Nguyen	Johnny Hwang	Shaniqua Green	Jermaine Green	Maribel Ayala	Javier Ayala	Audrey Rogers	Nicholas Rogers
Cindy Wu	Tony Choi	Tamika Jenkins	Donte Jenkins	Liliana Mejia	Guillermo Mejia	Jacquelyn Martin	Brett Martin
Joyce Vu	Benjamin Luong	Akilah Fields	Donnell Fields	Arlene Rojas	Elvis Rojas	Marissa Anderson	Cory Anderson
Vivian Hsu	Raymond Tran	Shantel Simmons	Davon Simmons	Cristina Ochoa	Miguel Ochoa	Haley Clark	Colin Clark
Jane Liang	Michael Duong	Shanique Carter	Darnell Carter	Yaritza Carillo	Sergio Carillo	Lindsay Campbell	Jack Campbell
Maggie Tsai	Andy Hoang	Tiara Woods	Hakeem Woods	Guadalupe Rios	Alejandro Rios	Cara Adams	Ross Adams
Diana Pham	Alexander Pham	Tierra Bryant	Sheldon Bryant	Angie Jimenez	Ernesto Jimenez	Jenna Morris	Liam Morris
Wendy Li	Robert Yang	Raven Brown	Antoine Brown	Esmeralda Maldonado	Oscar Maldonado	Caitlin Price	Max Price
Sally Hoang	Danny Xu	Octavia Byrd	Marquise Byrd	Marisol Cardenas	Felix Cardenas	Kathryn Hall	Ethan Hall
Kathy Duong	Anthony Huynh	Tyra Walker	Tyrone Walker	Denisse Chavez	Manuel Chavez	Emma Bailey	Eli Bailey
Lily Vang	Jason Liu	Diamond Lewis	Dashawn Lewis	Gabriela Mendez	Orlando Mendez	Erin Collins	Patrick Collins
Helen Trinh	John Chen	Nyasia Johnson	Shamel Johnson	Jeanette Rosales	Luis Rosales	Marisa Reed	Luke Reed
Sandy Oh	Brian Vang	Aliyah Douglas	Reginald Douglas	Rosa Castaneda	Eduardo Castaneda	Madeleine Smith	Alec Smith
Christine Tran	Joseph Zhou	Aaliyah Alexander	Shaquille Alexander	Beatriz Rodriguez	Carlos Rodriguez	Mackenzie King	Seth King
Judy Luu	James Cho	Princess Henderson	Jamel Henderson	Yessenia Acevedo	Cristian Acevedo	Sophie Thompson	Austin Thompson
Grace Cho	Nicholas Lin	Shanae Richardson	Akil Richardson	Carolina Guzman	Pedro Guzman	Madison Stewart	Nathan Stewart
Nancy Liu	Jeffrey Huang	Kenya Brooks	Tyquan Brooks	Carmen Aguilar	Freddy Aguilar	Margaret Parker	Jacob Parker
Lisa Cheng	Christopher Wu	Charisma Scott	Jamal Scott	Yesenia Vasquez	Esteban Vasquez	Kristin Gray	Craig Gray
Connie Yi	Timothy Ly	Shante Hunter	Jabari Hunter	Ana Munoz	Leonardo Munoz	Michaela Evans	Garrett Evans
Tiffany Phan	William Oh	Jada Hawkins	Tyshawn Hawkins	Xiomara Ortiz	Arturo Ortiz	Jaclyn Cooper	Ian Cooper
Karen Lu	Patrick Ngo	Shanice Reid	Demetrius Reid	Lizbeth Rivas	Jesus Rivas	Hannah Allen	Benjamin Allen
Tracy Chen	Thomas Cheng	Chanelle Sanders	Denzel Sanders	Genesis Sosa	Ramon Sosa	Zoe Wilson	Conor Wilson
Betty Dinh	Vincent Le	Shanequa Bell	Tyreek Bell	Stephany Salinas	Enrique Salinas	Caitlyn Young	Jared Young
Anna Hu	Kevin Hu	Shaniece Mitchell	Darius Mitchell	Lorena Gutierrez	Hector Gutierrez	Charlotte Moore	Theodore Moore
Elaine Le	Jimmy Xiong	Ebony Ford	Prince Ford	Emely Sandoval	Armando Sandoval	Kaitlin Wright	Shane Wright
Sophia Ly	Justin Zhu	Tanisha Watkins	Lamar Watkins	Iris Villarreal	Roberto Villarreal	Holly White	Scott White
Jenny Vo	Calvin Luu	Shanelle Butler	Raheem Butler	Maritza Garza	Edgar Garza	Kate Taylor	Noah Taylor
Monica Lin	Edward Kwon	Precious Davis	Jamar Davis	Marilyn Arroyo	Pablo Arroyo	Krista Hill	Ryan Hill
Joanne Yoon	Peter Phan	Asha Willis	Tariq Willis	Lourdes Soto	Raul Soto	Meredith Howard	Jake Howard
Priya Patel	Victor Patel	Ashanti Edwards	Shaquan Edwards	Gladys Herrera	Diego Herrera	Claire Turner	Maxwell Turner

Table 8: Work-for-Money Job Titles

Cashier	Busser	Trainer
Barista	Dishwasher	Instructor
Server	Barback	Tutor
Waiter	Bartender	Valet
Waitress	Bouncer	Doorman
Host	Caterer	Stock boy
Hostess	Messenger	Stockroom
Attendant	Courier	Librarian
Lifeguard	Bank teller	Worker
Receptionist	Front desk	Tour guide
Counselor	Bellhop	House painter
Babysitter	Maid	Landscaper
Nanny	Custodian	Gardener
Petsitter	Janitor	Assistant coach
Caddy	Cleaner	Work-study
Caddie	Bagger	Work study
Retail associate	Bag boy	Backroom Employee
Sales associate	Bookseller	Employee
Sales representative	Rug flipper	Customer Service
Salesperson	Laborer	Bodyguard
Salesman	Vendor	Canvasser
Saleswoman	Temp	Fundraiser
Store employee	Clerk	Phone bank
Prep cook	Mover	Call center
Line cook	Driver	Shift Manager
Bus boy		

Table 9: Majors in Generated Penn Resumes

School	Type	Major	Probability
Penn	Humanities & Social Sciences	BS in Economics	0.4
		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
Penn	Science	BS in Computer Engineering	0.15
		BS in Biomedical Science	0.075
		BS in Mechanical Engineering and Applied Mechanics	0.075
		BA in Biochemistry	0.05
		BS in Bioengineering	0.05
		BA in Biology	0.05
		BA in Chemistry	0.05
		BS in Chemical and Biomolecular Engineering	0.05
		BA in Cognitive Science	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
		BA in Physics	0.05
		BA in Mathematics	0.05
		BS in Networked and Social Systems Engineering	0.025
		BS in Systems Science and Engineering	0.025

Table 10: Majors in Generated Pitt Resumes

School	Type	Major	Probability
Pitt	Humanities & Social 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
Pitt	Science	BS in Computer Engineering	0.15
		BS in Natural Sciences	0.1
		BS in Mechanical Engineering	0.075
		BS in Molecular Biology	0.075
		BS in Bioengineering	0.05
		BS in Bioinformatics	0.05
		BS in Biological Sciences	0.05
		BS in Chemical Engineering	0.05
		BS in Chemistry	0.05
		BS in Computer Science	0.05
		BS in Electrical Engineering	0.05
		BS in Materials Science and Engineering	0.05
		BS in Mathematical Biology	0.05
		BS in Mathematics	0.05
		BS in Physics	0.05
		BS in Civil Engineering	0.025
		BS in Statistics	0.025

6.3.2 Work-for-Money Job Titles

6.4 Majors

6.5 Resume Creation

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 matching process. When submitting their resumes, they were informed that components of their resumes could be shown directly to employers. We scraped these resumes using Sovren resume scraping tool.

Each work and leadership experience "grouping" (a job/activity and its component bullet points) was sorted and cleaned in the following ways:

- Cleaned for major grammar errors, since this would be inconsistent as components were randomly paired together
- Work experience was sorted by major type, to ensure experience was relevant, whereas leadership experience was screened to be interchangeable across all resumes
- Cleansed of any information that might indicate gender or race (e.g., women's basketball)
- Sorted to eliminate components that were overly specific to a certain major
- Work locations were maintained as-is, leadership locations were set to the location of the school
- Dates of work experience were standardized to be the summers after junior and (optionally) sophomore years
- Dates of leadership experience were randomly generated time ranges over college career

Finally, skills were cleaned to be interchangeable within major type, cleaned of grammar errors, and restricted to a certain word count.

6.6 Matching

6.6.1 Students

For each school, we designed a unique database of resume fields to populate the resume tool based on real student resumes. We used a combination of resumes collected from clubs, resume books, campus job applicants, and job seekers who joined our study. For job-seeking study participants, the career services offices 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

Figure 11: Value of Humanities and Social Science Majors at Penn

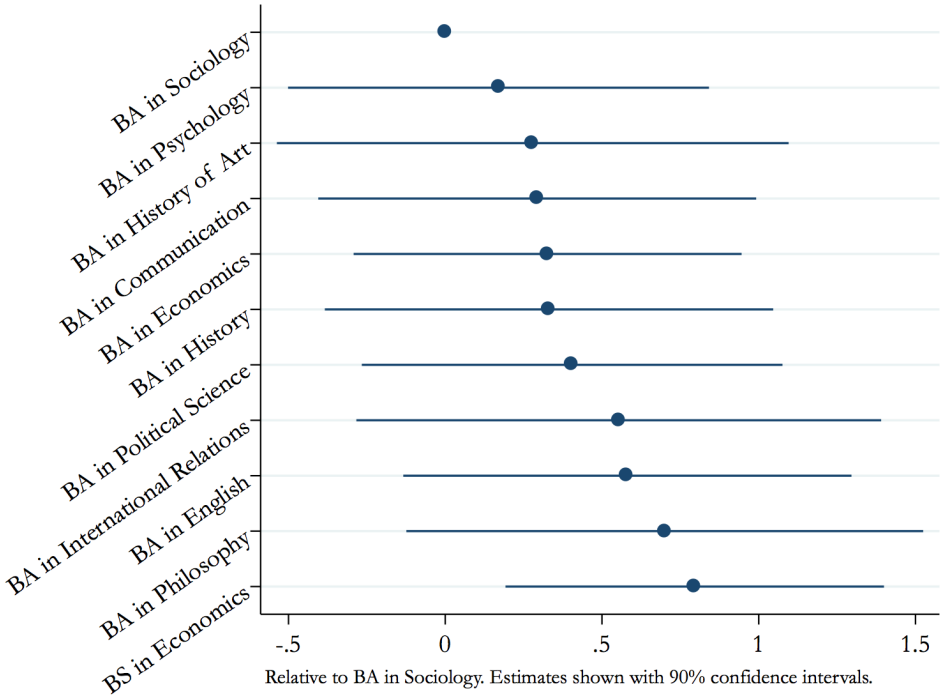


Figure 12: Value of Humanities and Social Science Majors at Pitt

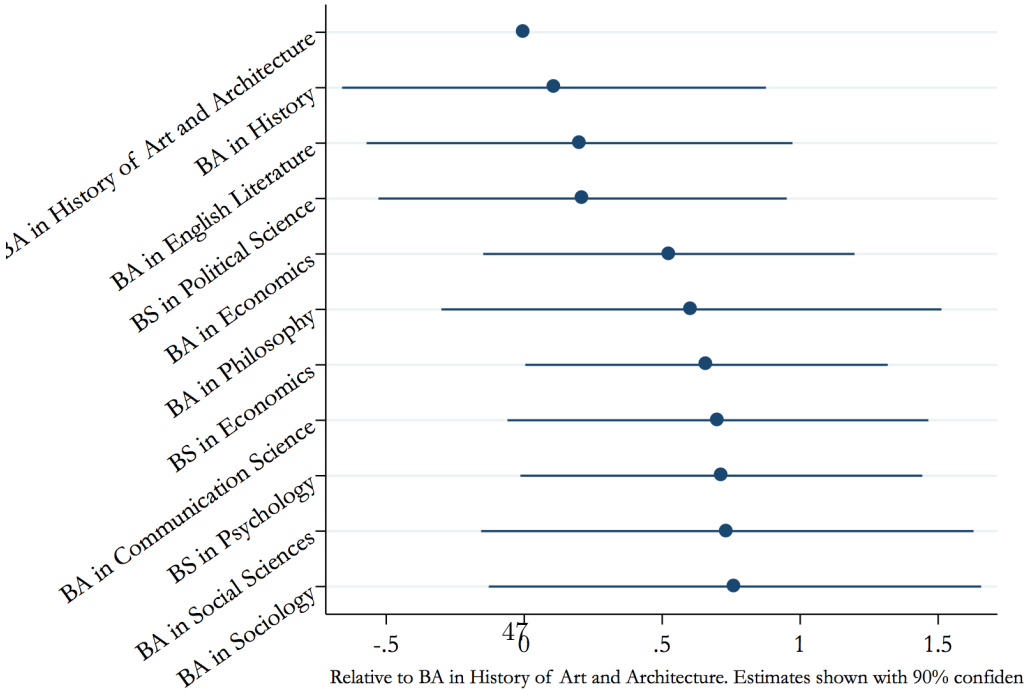


Figure 13: Value of Science Majors at Penn

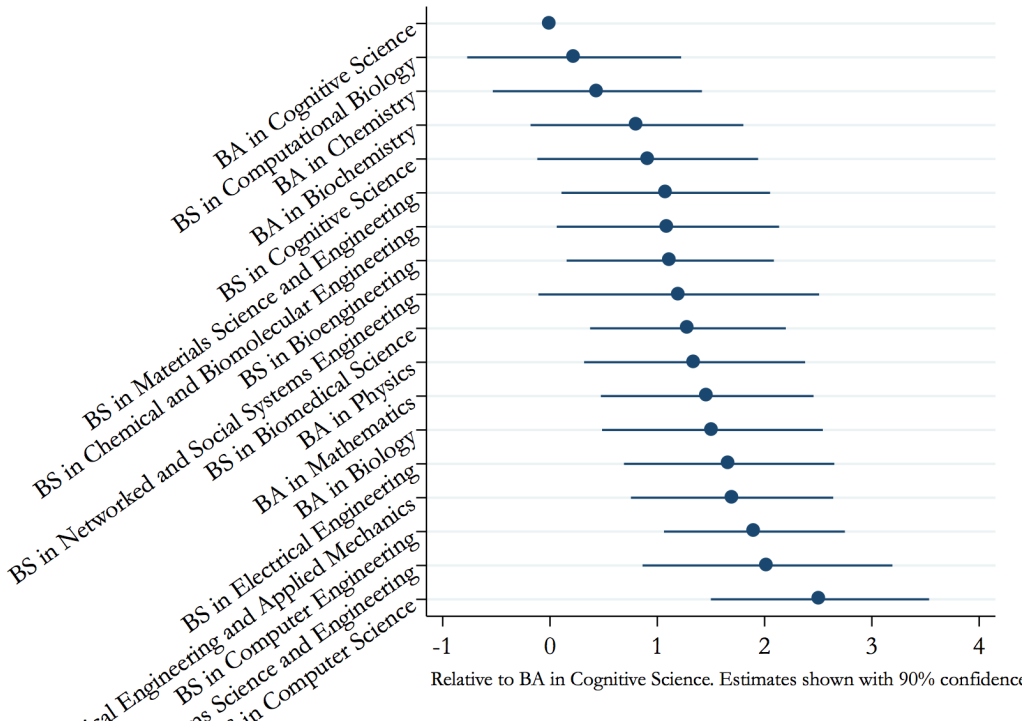


Figure 14: Value of Science Majors at Pitt

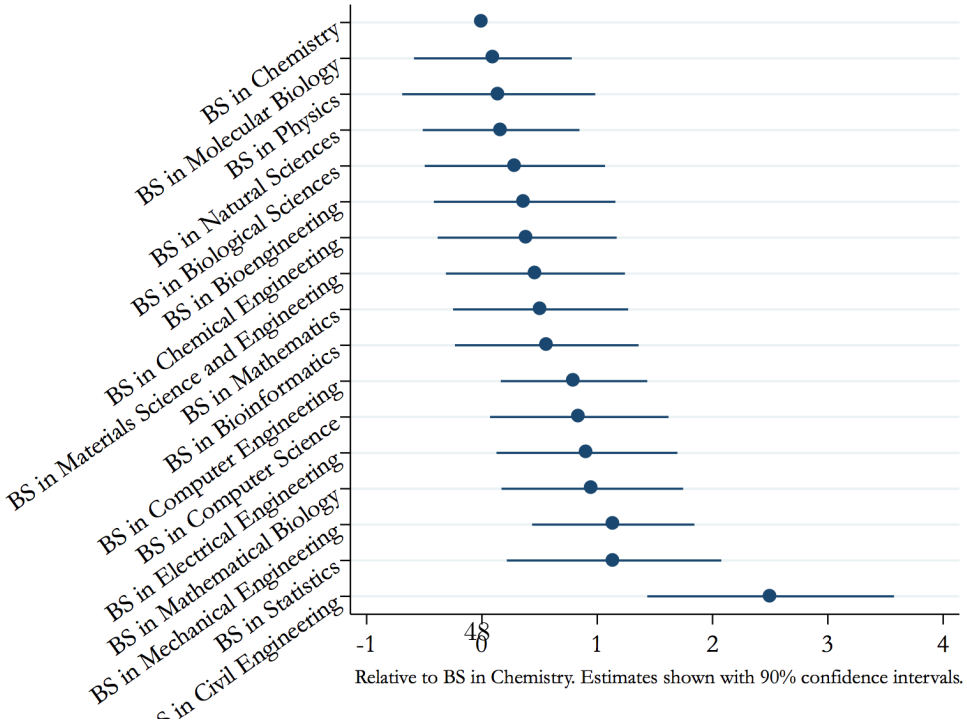


Table 11: Candidate Matching Variables

full student recruitment email is reproduced in Appendix 6.2.2. 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.

6.6.2 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 ethic or gender preferences. The full list of variables used for matching is shown in Table **??**.

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 desirability rating (the response to the question "How interested would you be in hiring [Candidate Name]?") and the likelihood of acceptance (the response to the question "How likely do you think [Candidate 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 science subjects and humanities/social science subjects, 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 desirability 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 desirability and likelihood of acceptance ($\text{callback} = \frac{2}{3}\text{desirability} + \frac{1}{3}\text{likelihood of acceptance}$). We collected the 10 resumes with the highest callback indices for each subject and emailed them to the subject.

After a subject completed the survey tool, we emailed them a zipped file of 10 resumes in PDF format corresponding to the 10 student resumes best matching their beliefs and preferences as identified by the machine learning algorithm. 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 the group.[31] This set of candidate resumes was the only incentive for participating in the study.

[31] 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

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. doi: 10.1007/978-3-319-71153-9_3.

Marianne Bertrand and Esther Duflo. Field experiments on discrimination. NBER Working Papers 22014, National Bureau of Economic Research, Inc, Feb 2016. URL https://ideas.repec.org/p/nbr/nberwo/22014.html.

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

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. doi: 10.1257/aer.20141757. URL http://www.aeaweb.org/articles?id=10.1257/aer.20141757.

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. URL http://www.aeaweb.org/articles?id=10.1257/aer.104.3.1014.

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. doi: 10.1093/sf/sou111. URL http://dx.doi.org/10.1093/sf/sou111.

Uri Gneezy, John A. List, and Michael K. Price. Toward and understanding of why people discriminate: Evidence from a series of natural field experiments. Technical report, National Bureau of Economic Research, 2012.

Glenn W. Harrison and John A. List. Field experiments. *Journal of Economic Literature*, 42(4): 1009–1055, December 2004. doi: 10.1257/0022051043004577. URL http://www.aeaweb.org/articles?id=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.

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. URL https://doi.org/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. URL http://dx.doi.org/10.1093/qje/qjt015.

Corinne Low. Pricing the biological clock: Reproductive capital on the us marriage market. *Manuscript, Wharton School of Business*, 2017.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

David Neumark. Detecting discrimination in audit and correspondence studies. *Journal of Human Resources*, 47(4):1128–1157, 2011.

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. doi: 10.1177/0019793916654686. URL https://doi.org/10.1177/0019793916654686.

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

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