



2019

Essays In Matching Markets

Colin D. Sullivan

University of Pennsylvania, cdsulliv@gmail.com

Follow this and additional works at: <https://repository.upenn.edu/edissertations>



Part of the [Economics Commons](#)

Recommended Citation

Sullivan, Colin D., "Essays In Matching Markets" (2019). *Publicly Accessible Penn Dissertations*. 3336.
<https://repository.upenn.edu/edissertations/3336>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/3336>
For more information, please contact repository@pobox.upenn.edu.

Essays In Matching Markets

Abstract

I present two experiments exploring failures in matching markets.

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

In the second experiment, I examine out-of-equilibrium truth-telling in strategic matching markets. In two-sided settings, market designers tend to advocate for deferred acceptance (DA) over priority mechanisms, even though theory tells us that both types of mechanisms can yield unstable matches in incomplete information equilibrium. However, if match participants on the proposed-to side deviate from equilibrium by truth-telling, then DA yields stable outcomes. In a novel experimental setting, I find out-of-equilibrium truth-telling under DA but not under a priority mechanism, which could help to explain the success of DA in preventing unraveling in the field. I then attempt to explain the difference in behavior across mechanisms by estimating an experience-weighted learning model adapted to this complex strategic environment. I find that initial cognition and willingness to explore new strategies drive the difference in agents' ability to find strategic equilibria.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Applied Economics

First Advisor

Judd B. Kessler

Keywords

Labor Economics, Matching Markets

Subject Categories

Economics

ESSAYS IN MATCHING MARKETS

Colin D. Sullivan

A DISSERTATION

in

Applied Economics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Judd B. Kessler

Associate Professor of Business Economics and Public Policy

Graduate Group Chairperson

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee:

Clayton Featherstone, Assistant Professor of Business Economics and Public Policy

Corinne Low, Assistant Professor of Business Economics and Public Policy

ESSAYS IN MATCHING MARKETS

© COPYRIGHT

2019

Colin D. Sullivan

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

ABSTRACT

ESSAYS IN MATCHING MARKETS

Colin D. Sullivan

Judd B. Kessler

I present two experiments exploring failures in matching markets.

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

In the second experiment, I examine out-of-equilibrium truth-telling in strategic matching markets. In two-sided settings, market designers tend to advocate for deferred acceptance (DA) over priority mechanisms, even though theory tells us that both types of mechanisms can yield unstable matches in incomplete information equilibrium. However, if match participants on the proposed-to side deviate from equilibrium by truth-telling, then DA yields stable outcomes. In a novel experimental setting, I find out-of-equilibrium truth-telling under DA but not under a priority mechanism, which could help to explain the success of DA in preventing unraveling in the field. I then attempt to explain the difference in behavior across mechanisms by estimating an experience-weighted learning model adapted to this complex strategic environment. I find that initial cognition and willingness to explore new strategies drive the difference in agents' ability to find strategic equilibria.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF TABLES	vii
LIST OF ILLUSTRATIONS	viii
CHAPTER 1 : Incentivized Resume Rating: Eliciting Employer Preferences without Deception (with Corinne Low and Judd B. Kessler)	1
1.1 Introduction	1
1.2 Study Design	6
1.3 Results	15
1.4 Pitt Replication: Results and Lessons	36
1.5 Conclusion	40
CHAPTER 2 : Learning to Manipulate: Experimental Evidence on Out-of-Equilibrium Truth-Telling (with Clayton R. Featherstone and Eric Mayefsky)	45
2.1 Introduction	45
2.2 Two markets	51
2.3 Experimental setup	54
2.4 Experimental Results	58
2.5 Learning Model	62
2.6 Conclusion	70
APPENDIX	74
CHAPTER A : Appendices to Chapter 1	74
CHAPTER B : Appendices to Chapter 2	125

BIBLIOGRAPHY	140
------------------------	-----

LIST OF TABLES

TABLE 1 :	Randomization of Resume Components	10
TABLE 2 :	Human Capital Experience	19
TABLE 3 :	Effects by Major Type	28
TABLE 4 :	Likelihood of Acceptance	33
TABLE 5 :	Likelihood of Acceptance by Major Type	34
TABLE 6 :	Hiring Interest at Penn and Pitt	39
TABLE 7 :	Experimental treatments	54
TABLE 8 :	Payoff table	56
TABLE 9 :	Truth-telling rates (all periods)	58
TABLE 10 :	Truth-telling rates (last 10 periods)	59
TABLE 11 :	Non-truncation rates	59
TABLE 12 :	Number of Blocking Pairs per Period	60
TABLE 13 :	Percentage of <i>Ms</i> and <i>Ws</i> Unmatched	61
TABLE 14 :	Parameter Estimates by Treatment	69
TABLE 15 :	Female Names Populating Resume Tool	84
TABLE 16 :	Male Names Populating Resume Tool	85
TABLE 17 :	Majors in Generated Penn Resumes	87
TABLE 18 :	Top Internship Employers	90
TABLE 19 :	Work for Money Job Titles & Identifying Phrases	92
TABLE 20 :	Candidate Matching Variables	95
TABLE 21 :	Work Experience Narrative	99
TABLE 22 :	Prestigious Schools	100
TABLE 23 :	Human Capital Experience—Weighted by GPA	103
TABLE 24 :	Human Capital Experience by Major Type—Weighted by GPA	104

TABLE 25 : Likelihood of Acceptance—Weighted by GPA	105
TABLE 26 : Hiring Interest by Rater Demographics	111
TABLE 27 : Implicit Bias	113
TABLE 28 : Return to Top Internship by Demographic Group	115
TABLE 29 : Likelihood of Acceptance with Hiring Interest Controls	119
TABLE 30 : Majors in Generated Pitt Resumes	121
TABLE 31 : Effects by Major Type at Pitt	122
TABLE 32 : Likelihood of Acceptance at Pitt	123
TABLE 33 : Likelihood of Acceptance by Major Type at Pitt	124
TABLE 34 : Table of cases	131
TABLE 35 : Payoffs for the cases	132

LIST OF ILLUSTRATIONS

FIGURE 1 :	Value of Quality of Experience Over Selectivity Distribution . . .	24
FIGURE 2 :	Demographics by Major Type Over Selectivity Distribution . . .	30
FIGURE 3 :	Expected payoff versus number of <i>Ms</i> truncated (empirical) . . .	56
FIGURE 4 :	Best response frequency CDFs	63
FIGURE 5 :	Employer Recruitment Email	76
FIGURE 6 :	Email Announcement to Graduating Seniors	77
FIGURE 7 :	Survey Tool Instructions & Contact Information	78
FIGURE 8 :	Major Type Selection	79
FIGURE 9 :	Sample Resume	81
FIGURE 10 :	Four Sample Resumes	83
FIGURE 11 :	Wharton	101
FIGURE 12 :	Distribution of GPA Among Scraped Resumes	102
FIGURE 13 :	Callback Thresholds Example	107
FIGURE 14 :	Alternative Specifications: Top Internship	109
FIGURE 15 :	Alternative Specifications: Second Job Type	110
FIGURE 16 :	Top Internship \times Not a White Male	116

CHAPTER 1 : Incentivized Resume Rating: Eliciting Employer Preferences without Deception (with Corinne Low and Judd B. Kessler)

1.1. Introduction

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

Despite the strengths of this workhorse methodology, however, resume audit studies are subject to two major concerns. First, they use deception, generally considered problematic within economics (Ortmann and Hertwig, 2002; Hamermesh, 2012). Employers in resume

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

audit studies waste time evaluating fake resumes and pursuing non-existent candidates. If fake resumes systematically differ from real resumes, employers could become wary of certain types of resumes sent out by researchers, harming both the validity of future research and real job seekers whose resumes are similar to those sent by researchers. These concerns about deception become more pronounced as the method becomes more popular.² To our knowledge, audit and correspondence audit studies are the only experiments within economics for which deception has been permitted, presumably because of the importance of the underlying research questions and the absence of a method to answer them without deception.

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

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

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

respond to candidate characteristics at other points in the distribution of candidate quality.

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

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

We find that employers value higher grade point averages as well as the quality and quantity of summer internship experiences. Employers place extra value on prestigious and substan-

tive internships but do not appear to value summer jobs that Penn students typically take for a paycheck, rather than to develop human capital for a future career, such as barista, server, or cashier. This result suggests a potential benefit on the post-graduate job market for students who can afford to take unpaid or low-pay internships during the summer rather than needing to work for an hourly wage.

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

The employers in our study report having a positive preference for diversity in hiring.⁴ While we do not find that employers are more or less interested in female and minority candidates on average, we find some evidence of discrimination against white women and minority men among employers looking to hire candidates with Science, Engineering, and Math majors.⁵ In addition, employers report that white female candidates are less likely to accept job offers than their white male counterparts, suggesting a novel channel for discrimination.

Of course, the IRR method also comes with some drawbacks. First, while we attempt to directly identify employer interest in a candidate, our Likert-scale measure is not a step in the hiring process and thus—in our implementation of IRR—we cannot draw a direct link between our Likert-scale measure and hiring outcomes. However, we imagine future

⁴In a survey employers complete after evaluating resumes in our study, over 90% of employers report that both “seeking to increase gender diversity / representation of women” and “seeking to increase racial diversity” factor into their hiring decisions, and 82% of employers rate both of these factors at 5 or above on a Likert scale from 1 = “Do not consider at all” to 10 = “This is among the most important things I consider.”

⁵We find suggestive evidence that discrimination in hiring interest is due to implicit bias by observing how discrimination changes as employers evaluate multiple resumes. In addition, consistent with results from the resume audit literature finding lower returns to quality for minority candidates (see Bertrand and Mullainathan (2004)), we also find that—relative to white males—other candidates receive a lower return to work experience at prestigious internships.

IRR studies could make advances on this front (e.g., by asking employers to guarantee interviews to matched candidates). Second, because the incentives in our study are similar but not identical to those in the hiring process, we cannot be sure that employers evaluate our hypothetical resumes with the same rigor or using the same criteria as they would real resumes. Again, we hope future work might validate that the time and attention spent on resumes in the IRR paradigm is similar to resumes evaluated as part of standard recruiting processes.

Our implementation of IRR was the first of its kind and thus left room for improvement on a few fronts. For example, as discussed in detail in Section 1.4, we attempted to replicate our study at the University of Pittsburgh to evaluate preferences of employers more like those traditionally targeted by resume audit studies. We underestimated how much Pitt employers needed candidates with specific majors and backgrounds, however, and a large fraction of resumes that were shown to Pitt employers were immediately disqualified based on major. This mistake resulted in highly attenuated estimates. Future implementations of IRR should more carefully tailor the variables for their hypothetical resumes to the needs of the employers being studied. We emphasize other lessons from our implementation in Section 2.6.

Despite the limitations of IRR, our results highlight that the method can be used to elicit employer preferences and suggest that it can also be used to detect discrimination. Consequently, we hope IRR provides a path forward for those interested in studying labor markets without using deception. The rest of the paper proceeds as follows: Section 1.2 describes in detail how we implement our IRR study; Section 1.3 reports on the results from Penn and compares them to extant literature; Section 1.4 describes our attempted replication at Pitt; and Section 2.6 concludes.

1.2. Study Design

In this section, we describe our implementation of IRR, which combines the incentives and ecological validity of the field with the control of the laboratory. In Section 1.2.1, we outline how we recruit employers who are in the market to hire elite college graduates. In Section 1.2.2, we describe how we provide employers with incentives for reporting preferences without introducing deception. In Section 1.2.3, we detail how we created the hypothetical resumes and describe the extensive variation in candidate characteristics that we included in the experiment, including grade point average and major (see 1.2.3), previous work experience (see 1.2.3), skills (see 1.2.3), and race and gender (see 1.2.3). In Section 1.2.4, we highlight the two questions that we asked subjects about each hypothetical resume, which allowed us to get a granular measure of interest in a candidate without a confound from the likelihood that the candidate would accept a job if offered.

1.2.1. Employers and Recruitment

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

Penn Career Services sent invitation emails (see Appendix Figure 5 for recruitment email) in two waves during the 2016-2017 academic year to employers who historically recruited Penn seniors (e.g., firms that recruited on campus, regularly attended career fairs, or otherwise hired students). The first wave was around the time of on-campus recruiting in the fall of 2016. The second wave was around the time of career-fair recruiting in the spring of 2017. In both waves, the recruitment email invited employers to use “a new tool that can help you to identify potential job candidates.” While the recruitment email and the information that

employers received before rating resumes (see Appendix Figure 7 for instructions) noted that anonymized data from employer responses would be used for research purposes, this was framed as secondary. The recruitment process and survey tool itself both emphasized that employers were using new recruitment software. For this reason, we note that our study has the ecological validity of a field experiment.⁶ As was outlined in the recruitment email (and described in detail in Section 1.2.2), each employer’s one and only incentive for participating in the study is to receive 10 resumes of job seekers that match the preferences they report in the survey tool.

1.2.2. Incentives

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

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

⁷The recruitment email (see Appendix Figure 5) stated: “the tool uses a newly developed machine-learning algorithm to identify candidates who would be a particularly good fit for your job based on your evaluations.” We did not use race or gender preferences when suggesting matches from the candidate pool. The process by which we identify job seekers based on employer evaluations is described in detail in Appendix A.1.3.

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

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

1.2.3. Resume Creation and Variation

Our implementation of IRR asked each employer to evaluate 40 unique, hypothetical resumes, and it varied multiple candidate characteristics simultaneously and independently across resumes, allowing us to estimate employer preferences over a rich space of baseline candidate characteristics.⁹ Each of the 40 resumes was dynamically populated when a subject began the survey tool. As shown in Table 1 and described below, we randomly varied a set of candidate characteristics related to education; a set of candidate characteristics related to work, leadership, and skills; and the candidate’s race and gender.

We made a number of additional design decisions to increase the realism of the hypothetical resumes and to otherwise improve the quality of employer responses. First, we built the hypothetical resumes using components (i.e., work experiences, leadership experiences, and skills) from real resumes of seniors at Penn. Second, we asked the employers to choose the type of candidates that they were interested in hiring, based on major (see Appendix Figure 8). In particular, they could choose either “Business (Wharton), Social Sciences, and Humanities” (henceforth “Humanities & Social Sciences”) or “Science, Engineering,

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

Computer Science, and Math” (henceforth “STEM”). They were then shown hypothetical resumes focused on the set of majors they selected. As described below, this choice affects a wide range of candidate characteristics; majors, internship experiences, and skills on the hypothetical resumes varied across these two major groups. Third, to enhance realism, and to make the evaluation of the resumes less tedious, we used 10 different resume templates, which we populated with the candidate characteristics and component pieces described below, to generate the 40 hypothetical resumes (see Appendix Figure 9 for a sample resume). We based these templates on real student resume formats (see Appendix Figure 10 for examples).¹⁰ Fourth, we gave employers short breaks within the study by showing them a progress screen after each block of 10 resumes they evaluated. As described in Section 1.3.4 and Appendix A.2.4, we use the change in attention induced by these breaks to construct tests of implicit bias.

Education Information

In the education section of the resume, we independently randomized each candidate’s grade point average (GPA) and major. GPA is drawn from a uniform distribution between 2.90 and 4.00, shown to two decimal places and never omitted from the resume. Majors are chosen from a list of Penn majors, with higher probability put on more common majors. Each major was associated with a degree (BA or BS) and with the name of the group or school granting the degree within Penn (e.g., “College of Arts and Sciences”). Appendix Table 17 shows the list of majors by major category, school, and the probability that the major was used in a resume.

Work Experience

We included realistic work experience components on the resumes. To generate the components, we scraped more than 700 real resumes of Penn students. We then followed a process

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

Table 1: Randomization of Resume Components

Resume Component	Description	Analysis Variable
Personal Information		
First & last name	Drawn from list of 50 possible names given selected race and gender (names in Tables 15 & 16)	<i>Female, White</i> (32.85%) <i>Male, Non-White</i> (17.15%)
	Race drawn randomly from U.S. distribution (65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian)	<i>Female, Non-White</i> (17.15%)
	Gender drawn randomly (50% male, 50% female)	<i>Not a White Male</i> (67.15%)
Education Information		
GPA	Drawn <i>Unif</i> [2.90, 4.00] to second decimal place	<i>GPA</i>
Major	Drawn from a list of majors at Penn (Table 17)	<i>Major</i> (weights in Table 17)
Degree type	BA, BS fixed to randomly drawn major	<i>Wharton</i> (40%)
School within university	Fixed to randomly drawn major	<i>School of Engineering and</i>
Graduation date	Fixed to upcoming spring (i.e., May 2017)	<i>Applied Science</i> (70%)
Work Experience		
First job	Drawn from curated list of top internships and regular internships	<i>Top Internship</i> (20/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate’s junior year (i.e., 2016)	
Second job	Left blank or drawn from curated list of regular internships and work-for-money jobs (Table 19)	<i>Second Internship</i> (13/40) <i>Work for Money</i> (13/40)
Title and employer	Fixed to randomly drawn job	
Location	Fixed to randomly drawn job	
Description	Bullet points fixed to randomly drawn job	
Dates	Summer after candidate’s sophomore year (i.e., 2015)	
Leadership Experience		
First & second leadership	Drawn from curated list	
Title and activity	Fixed to randomly drawn leadership	
Location	Fixed to Philadelphia, PA	
Description	Bullet points fixed to randomly drawn leadership	
Dates	Start and end years randomized within college career, with more recent experience coming first	
Skills		
Skills list	Drawn from curated list, with two skills drawn from {Ruby, Python, PHP, Perl} and two skills drawn from {SAS, R, Stata, Matlab} shuffled and added to skills list with probability 25%.	<i>Technical Skills</i> (25%)

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

described in Appendix A.1.2 to select and lightly sanitize work experience components so that they could be randomly assigned to different resumes without generating conflicts or inconsistencies (e.g., we eliminated references to particular majors or to gender or race). Each work experience component included the associated details from the real resume from which the component was drawn, including an employer, position title, location, and a few descriptive bullet points.

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

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

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

¹¹See Appendix Table 18 for a list of top internship employers and Table 19 for a list of work-for-money job titles. As described in Appendix A.1.2, different internships (and top internships) were used for each major type but the same work-for-money jobs were used for both major types. The logic of varying internships by major type was based on the intuition that internships could be interchangeable within each group of majors (e.g., internships from the Humanities & Social Sciences resumes would not be unusual to see on any other resume from that major group) but were unlikely to be interchangeable across major groups (e.g., internships from Humanities & Social Sciences resumes would be unusual to see on STEM resumes and vice versa). We used the same set of work-for-money jobs for both major types, since these jobs were not linked to a candidate’s field of study.

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

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

Leadership Experience and Skills

Each resume included two leadership experiences as in typical student resumes. A leadership experience component includes an activity, title, date range, and a few bullet points with a description of the experience (Philadelphia, PA was given as the location of all leadership experiences). Participation dates were randomly selected ranges of years from within the four years preceding the graduation date. For additional details, see Appendix A.1.2.

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

Names Indicating Gender and Race

We randomly varied gender and race by assigning each hypothetical resume a name that would be indicative of gender (male or female) and race (Asian, Black, Hispanic, or White).¹³ To do this randomization, we needed to first generate a list of names that would clearly randomize the elements of work experience.

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

indicate both gender and race for each of the groups. We used birth records and Census data to generate first and last names that would be highly indicative of race and gender, and combined names within race.¹⁴ The full lists of names are given in Appendix Tables 15 and 16 (see Appendix A.1.2 for additional details).

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

1.2.4. Rating Candidates on Two Dimensions

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

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

likelihood that a candidate accepts a job when offered is decreasing in the candidate’s quality (e.g., if higher quality candidates have better outside options), employers’ actual effort spent pursuing candidates may be non-monotonic in candidate quality. Consequently, concerns about a candidate’s likelihood of accepting a job may be a confound in interpreting callback as a measure of interest in a candidate.¹⁵

An advantage of the IRR methodology is that researchers can ask employers to provide richer, more granular information than a binary measure of callback. We leveraged this advantage to ask two questions, each on a Likert scale from 1 to 10. In particular, for each resume we asked employers to answer the following two questions (see an example at the bottom of Appendix Figure 9):

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

In the instructions (see Appendix Figure 7), employers were specifically told that responses to both questions would be used to generate their matches. In addition, they were told to focus only on their interest in hiring a candidate when answering the first question (i.e., they were instructed to assume the candidate would accept an offer if given one). We denote responses to this question “hiring interest.” They were told to focus only on the likelihood a candidate would accept a job offer when answering the second question (i.e., they were instructed to assume they candidate had been given an offer and to assess the likelihood they would accept it). We denote responses to this question a candidate’s “likelihood of acceptance.” We asked the first question to assess how resume characteristics affect hiring interest. We asked the second question both to encourage employers to focus only on hiring interest when answering the first question and to explore employers’ beliefs about

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

the likelihood that a candidate would accept a job if offered.

The 10-point scale has two advantages. First, it provides additional statistical power, allowing us to observe employer preferences toward characteristics of inframarginal resumes, rather than identifying preferences only for resumes crossing a binary callback threshold in a resume audit setting. Second, it allows us to explore how employer preferences vary across the distribution of hiring interest, an issue we explore in depth in Section 1.3.3.

1.3. Results

1.3.1. Data and Empirical Approach

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

As described in Section 1.2, each employer rated 40 unique, hypothetical resumes with randomly assigned candidate characteristics. For each resume, employers rated hiring interest and likelihood of acceptance, each on a 10-point Likert scale. Our analysis focuses initially on hiring interest, turning to how employers evaluate likelihood of acceptance in Section 1.3.5. Our main specifications are ordinary least squares (OLS) regressions. These specifications make a linearity assumption with respect to the Likert-scale ratings data. Namely, they assume that, on average, employers treat equally-sized increases in Likert-scale ratings

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

equivalently (e.g., an increase in hiring interest from 1 to 2 is equivalent to an increase from 9 to 10). In some specifications, we include subject fixed effects, which account for the possibility that employers have different mean ratings of resumes (e.g., allowing some employers to be more generous than others with their ratings across all resumes), while preserving the linearity assumption. To complement this analysis, we also run ordered probit regression specifications, which relax this assumption and only require that employers, on average, consider higher Likert-scale ratings more favorably than lower ratings.

In Section 1.3.2, we examine how human capital characteristics (e.g., GPA, major, work experience, and skills) affect hiring interest. These results report on the mean of preferences across the distribution; we show how our results vary across the distribution of hiring interest in Section 1.3.3. In Section 1.3.4, we discuss how employers’ ratings of hiring interest respond to demographic characteristics of our candidates. In Section 1.3.5, we investigate the likelihood of acceptance ratings and identify a potential new channel for discrimination. In Section 1.3.6, we compare our results to prior literature.

1.3.2. Effect of Human Capital on Hiring Interest

Employers in our study are interested in hiring graduates of the University of Pennsylvania for full-time employment, and many recruit at other Ivy League schools and other top colleges and universities. This labor market has been unexplored by resume audit studies, in part because the positions employers aim to fill through on-campus recruiting at Penn are highly unlikely to be filled through online job boards or by screening unsolicited resumes. In this section, we evaluate how randomized candidate characteristics—described in Section 1.2.3 and Table 1—affect employers’ ratings of hiring interest.

We denote an employer i ’s rating of a resume j on the 1–10 Likert scale as V_{ij} and estimate variations of the following regression specification (1.1). This regression allows us to

investigate the average response to candidate characteristics across employers in our study.

$$\begin{aligned}
V_{ij} = & \beta_0 + \beta_1 \textit{GPA} + \beta_2 \textit{Top Internship} + \beta_3 \textit{Second Internship} + \beta_4 \textit{Work for Money} + \\
& \beta_5 \textit{Technical Skills} + \beta_6 \textit{Female, White} + \beta_7 \textit{Male, Non-White} + \\
& \beta_8 \textit{Female, Non-White} + \mu_j + \gamma_j + \omega_j + \alpha_i + \varepsilon_{ij}
\end{aligned} \tag{1.1}$$

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

Table 2 shows regression results where V_{ij} is *Hiring Interest*, which takes values from 1 to 10. The first three columns report OLS regressions with slightly different specifications. The first column includes all candidate characteristics we varied to estimate their impact on ratings. The second column adds leadership dummies γ and resume order dummies ω . The third column also adds subject fixed effects α . As expected, results are robust to

¹⁷Coefficient estimates on these variables report comparisons to white males, which is the excluded group. While we do not discuss demographic results in this section, we include controls for this randomized resume component in our regressions and discuss the results in Section 1.3.4 and Appendix A.2.4.

the addition of these controls. The fourth column, labeled *GPA-Scaled OLS*, rescales all coefficients from the third column by the coefficient on GPA (2.196) so that the coefficients on other variables can be interpreted in GPA points. These regressions show that employers respond strongly to candidate characteristics related to human capital.

GPA is an important driver of hiring interest. An increase in GPA of one point (e.g., from a 3.0 to a 4.0) increases ratings on the Likert scale by 2.1–2.2 points. The standard deviation of quality ratings is 2.81, suggesting that a point improvement in GPA moves hiring interest ratings by about three quarters of a standard deviation.

As described in Section 1.2.3, we created *ex ante* variation in both the quality and quantity of candidate work experience. Both affect employer interest. The quality of a candidate’s work experience in the summer before senior year has a large impact on hiring interest ratings. The coefficient on *Top Internship* ranges from 0.9–1.0 Likert-scale points, which is roughly a third of a standard deviation of ratings. As shown in the fourth column of Table 2, a top internship is equivalent to a 0.41 improvement in GPA.

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

¹⁸These results are consistent with a penalty for working-class candidates. In a resume audit study of

Table 2: Human Capital Experience

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.125 (0.145)	2.190 (0.150)	2.196 (0.129)	1 (.)	0.891 (0.0626)
Top Internship	0.902 (0.0945)	0.900 (0.0989)	0.897 (0.0806)	0.409 (0.0431)	0.378 (0.0397)
Second Internship	0.465 (0.112)	0.490 (0.118)	0.466 (0.0947)	0.212 (0.0446)	0.206 (0.0468)
Work for Money	0.116 (0.110)	0.157 (0.113)	0.154 (0.0914)	0.0703 (0.0416)	0.0520 (0.0464)
Technical Skills	0.0463 (0.104)	0.0531 (0.108)	-0.0711 (0.0899)	-0.0324 (0.0410)	0.0120 (0.0434)
Female, White	-0.152 (0.114)	-0.215 (0.118)	-0.161 (0.0963)	-0.0733 (0.0441)	-0.0609 (0.0478)
Male, Non-White	-0.172 (0.136)	-0.177 (0.142)	-0.169 (0.115)	-0.0771 (0.0526)	-0.0754 (0.0576)
Female, Non-White	-0.00936 (0.137)	-0.0220 (0.144)	0.0281 (0.120)	0.0128 (0.0546)	-0.0144 (0.0573)
Observations	2880	2880	2880	2880	2880
R^2	0.129	0.181	0.483		
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

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

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. The p -values of tests of joint significance of major fixed effects are indicated (F -test for OLS, likelihood ratio test for ordered probit).

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

Table 2 also reports the p -value of a test of whether the coefficients on the major dummies are jointly different from zero. Results suggest that the randomly assigned major significantly affects hiring interest. While we do not have the statistical power to test for the effect of each major, we can explore how employers respond to candidates being from more prestigious schools at the University of Pennsylvania. In particular, 40% of the Humanities & Social Sciences resumes are assigned a BS in Economics from Wharton and the rest have a BA major from the College of Arts and Sciences. In addition, 70% of the STEM resumes are assigned a BS from the School of Engineering and Applied Science and the rest have a BA major from the College of Arts and Sciences. As shown in Appendix Table 22, in both cases, we find that being from the more prestigious school—and thus receiving a BS rather than a BA—is associated with an increase in hiring interest ratings of about 0.4 Likert-scale points or 0.18 GPA points.¹⁹

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

law firms, Rivera and Tilcsik (2016) found that resume indicators of lower social class (such as receiving a scholarship for first generation college students) led to lower callback rates.

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

specifications described above.²⁰

As discussed in Section 1.2, we made many design decisions to enhance realism. However, one might be concerned that our independent cross-randomization of various resume components might lead to unrealistic resumes and influence the results we find. We provide two robustness checks in the appendix to address this concern. First, our design and analysis treat each work experience as independent, but, in practice, candidates may have related jobs over a series of summers that create a work experience “narrative.” In Appendix A.2.1 and Appendix Table 21, we describe how we construct a measure of work experience narrative, we test its importance, and find that while employers respond positively to work experience narrative ($p = 0.054$) our main results are robust to its inclusion. Second, the GPA distribution we used for constructing the hypothetical resumes did not perfectly match the distribution of job seekers in our labor market. In Appendix A.2.2, we re-weight our data to match the GPA distribution in the candidate pool of real Penn job seekers and show that our results are robust to this re-weighting. These exercises provide some assurance that our results are not an artifact of how we construct hypothetical resumes.

1.3.3. Effects Across the Distribution of Hiring Interest

The regression specifications described in Section 1.3.2 identify the average effect of candidate characteristics on employers’ hiring interest. As pointed out by Neumark (2012), however, these average preferences may differ in magnitude—and even direction—from differences in callback rates, which derive from whether a characteristic pushes a candidate above a specific quality threshold (i.e., the callback threshold). For example, in the low callback rate environments that are typical of resume audit studies, differences in callback rates will be determined by how employers respond to a candidate characteristic in the right

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

tail of their distribution of preferences.²¹ To make this concern concrete, Appendix A.2.3 provides a simple graphical illustration in which the average preference for a characteristic differs from the preference in the tail of the distribution. In practice, we may care about preferences in any part of the distribution for policy. For example, preferences at the callback threshold may be relevant for hiring outcomes, but those thresholds may change with a hiring expansion or contraction.

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

While the theoretical concerns raised by Neumark (2012) may be relevant in other settings, the average results we find in Section 1.3.2 are all consistent across the distribution of hiring interest, including in the tails (except for a preference for Wharton students, which we discuss below). The top half of Figure 1 shows that *Top Internship* is positive and statistically significant at all levels of selectivity. Panel (a) reports the empirical CDF of hiring interest ratings for candidates with and without a top internship. Panel (b) shows the difference in callback rates that would arise for *Top Internship* at each counterfactual callback threshold. The estimated difference in callback rates is positive and significant everywhere, although it is much larger in the midrange of the quality distribution than at

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

either of the tails.²² The bottom half of Figure 1 shows that results across the distribution for *Second Internship* and *Work for Money* are also consistent with the average results from Section 1.3.2. *Second Internship* is positive everywhere and almost always statistically significant. *Work for Money* consistently has no impact on employer preferences throughout the distribution of hiring interest.

As noted above, our counterfactual callback threshold exercise suggests that a well-powered audit study would likely find differences in callback rates for most of the characteristics that we estimate as statistically significant on average in Section 1.3.2, regardless of employers' callback threshold. This result is reassuring both for the validity of our results and in considering the generalizability of results from the resume audit literature. However, even in our data, we observe a case where a well-powered audit study would be unlikely to find a result, even though we find one on average. Appendix Figure 11 mirrors Figure 1 but focuses on having a Wharton degree among employers seeking Humanities & Social Sciences candidates. Employers respond to Wharton in the middle of the distribution of hiring interest, but preferences seem to converge in the right tail (i.e., at hiring interest ratings of 9 or 10), suggesting that the best students from the College of Arts and Sciences are not evaluated differently than the best students from Wharton.

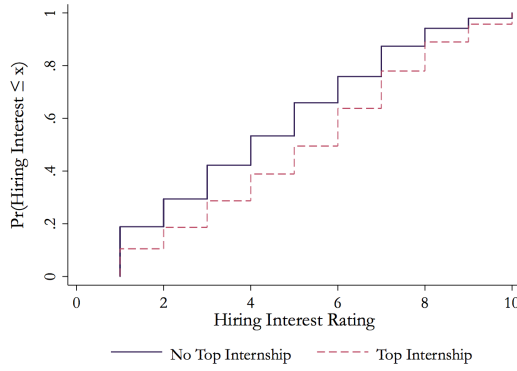
1.3.4. Demographic Discrimination

In this section, we examine how hiring interest ratings respond to the race and gender of candidates. As described in Section 1.2 and shown in Table 1, we use our variation in names to create the variables: *Female, White*; *Male, Non-White*; and *Female, Non-White*.

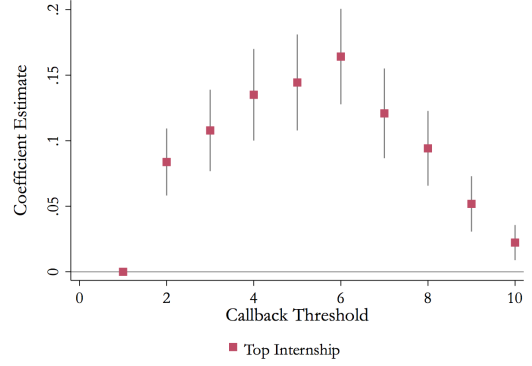
As shown in Table 2, the coefficients on the demographic variables are not significantly

²²This shape is partially a mechanical feature of low callback rate environments: if a threshold is set high enough that only 5% of candidates with a desirable characteristic are being called back, the difference in callback rates can be no more than 5 percentage points. At lower thresholds (e.g., where 50% of candidates with desirable characteristics are called back), differences in callback rates can be much larger. In Appendix A.2.3, we discuss how this feature of difference in callback rates could lead to misleading comparisons across experiments with very different callback rates.

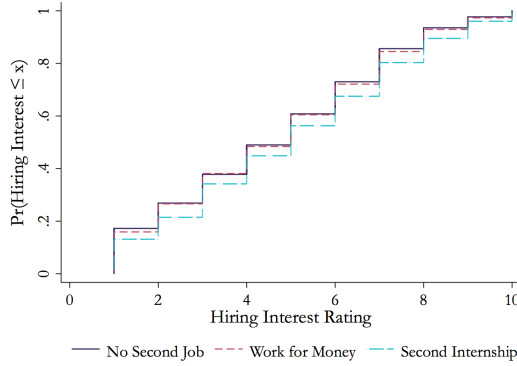
Figure 1: Value of Quality of Experience Over Selectivity Distribution



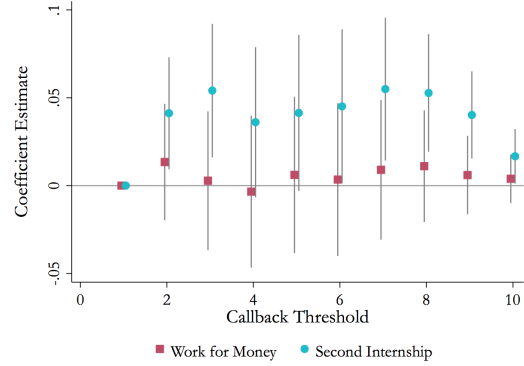
(a) Empirical CDF for Top Internship



(b) Linear Probability Model for Top Internship



(c) Empirical CDF for Second Job Type



(d) Linear Probability Model for Second Job Type

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

different from zero, suggesting no evidence of discrimination on average in our data.²³ This null result contrasts somewhat with existing literature—both resume audit studies (e.g., Bertrand and Mullainathan (2004)) and laboratory experiments (e.g., Bohnet et al. (2015)) generally find evidence of discrimination in hiring. Our differential results may not be surprising given that our employer pool is different than those usually targeted through resume audit studies, with most reporting positive tastes for diversity.

While we see no evidence of discrimination on average, a large literature addressing diversity in the sciences (e.g., Carrell et al. (2010); Goldin (2014)) suggests we might be particularly likely to see discrimination among employers seeking STEM candidates. In Table 3, we estimate the regression in Equation (1.1) separately by major type. Results in Columns 5-10 show that employers looking for STEM candidates display a large, statistically significant preference for white male candidates over white females and non-white males. The coefficients on *Female, White* and *Male, Non-White* suggest that these candidates suffer a penalty of 0.5 Likert-scale points—or about 0.27 GPA points—that is robust across our specifications. These effects are at least marginally significant even after multiplying our p -values by two to correct for the fact that we are analyzing our results within two subgroups (uncorrected p -values are: $p = 0.009$ for *Female, White*; $p = 0.049$ for *Male, Non-White*). Results in Columns 1-5 show no evidence of discrimination in hiring interest among Humanities & Social Sciences employers.

As in Section 1.3.3, we can examine these results across the hiring interest rating distribution. Figure 2 shows the CDF of hiring interest ratings and the difference in counterfactual callback rates. For ease of interpretation and for statistical power, we pool female and minority candidates and compare them to white male candidates in these figures and in some analyses that follow. The top row shows these comparisons for employers interested in Humanities & Social Sciences candidates and the bottom row shows these comparisons for employers interested in STEM candidates. Among employers interested in Humanities &

²³In Appendix Table 26, we show that this effect does not differ by the gender and race of the employer rating the resume.

Social Sciences candidates, the CDFs of *Hiring Interest* ratings are nearly identical. Among employers interested in STEM candidates, however, the CDF for white male candidates first order stochastically dominates the CDF for candidates who are not white males. At the point of the largest counterfactual callback gap, employers interested in STEM candidates would display callback rates that were 10 percentage points lower for candidates who were not white males than for their white male counterparts.

One might be surprised that we find any evidence of discrimination, given that employers may have (correctly) believed we would not use demographic tastes in generating their matches and given that employers may have attempted to override any discriminatory preferences to be more socially acceptable. One possibility for why we nevertheless find discrimination is the role of implicit bias (Greenwald et al., 1998; Nosek et al., 2007), which Bertrand et al. (2005) has suggested is an important channel for discrimination in resume audit studies. In Appendix A.2.4, we explore the role of implicit bias in driving our results.²⁴ In particular, we leverage a feature of implicit bias—that it is more likely to arise when decision makers are fatigued (Wigboldus et al., 2004; Govorun and Payne, 2006; Sherman et al., 2004)—to test whether our data are consistent with employers displaying an implicit racial or gender bias. As shown in Appendix Table 27, employers spend less time evaluating resumes both in the latter half of the study and in the latter half of each set of 10 resumes (after each set of 10 resumes, we introduced a short break for subjects), suggesting evidence of fatigue. Discrimination is statistically significantly larger in the latter half of each block of 10 resumes, providing suggestive evidence that implicit bias plays a role in our findings, although discrimination is not larger in the latter half of the study.

Race and gender could also subconsciously affect how employers view other resume components. We test for negative interactions between race and gender and desirable candidate

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

characteristics, which have been found in the resume audit literature (e.g., minority status has been shown to lower returns to resume quality (Bertrand and Mullainathan, 2004)). Appendix Table 28 interacts *Top Internship*, our binary variable most predictive of hiring interest, with our demographic variables. These interactions are all directionally negative, and the coefficient $Top\ Internship \times Female, White$ is negative and significant, suggesting a lower return to a prestigious internships for white females. One possible mechanism for this effect is that employers believe that other employers exhibit positive preferences for diversity, and so having a prestigious internship is a less strong signal of quality if one is from an under-represented group. This aligns with the findings shown in Appendix Figure 16, which shows that the negative interaction between *Top Internship* and demographics appears for candidates with relatively low ratings and is a fairly precisely estimated zero when candidates receive relatively high ratings.

Table 3: Effects by Major Type

	Humanities & Social Sciences					Dependent Variable: Hiring Interest					STEM		
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit	OLS	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit	OLS	Ordered Probit
GPA	2.208 (0.173)	2.304 (0.179)	2.296 (0.153)	1 (.)	0.933 (0.0735)	1.932 (0.267)	1.885 (0.309)	1.882 (0.242)	1 (.)	1.882 (0.242)	0.802 (0.112)	1.882 (0.242)	0.802 (0.112)
Top Internship	1.075 (0.108)	1.043 (0.116)	1.033 (0.0945)	0.450 (0.0500)	0.452 (0.0461)	0.398 (0.191)	0.559 (0.216)	0.545 (0.173)	0.289 (0.0997)	0.545 (0.173)	0.175 (0.0784)	0.545 (0.173)	0.175 (0.0784)
Second Internship	0.540 (0.132)	0.516 (0.143)	0.513 (0.114)	0.224 (0.0514)	0.240 (0.0555)	0.242 (0.208)	0.307 (0.246)	0.311 (0.189)	0.165 (0.103)	0.307 (0.246)	0.111 (0.0881)	0.307 (0.246)	0.111 (0.0881)
Work for Money	0.0874 (0.129)	0.107 (0.134)	0.116 (0.110)	0.0504 (0.0477)	0.0371 (0.0555)	0.151 (0.212)	0.275 (0.254)	0.337 (0.187)	0.179 (0.102)	0.337 (0.187)	0.0761 (0.0881)	0.337 (0.187)	0.0761 (0.0881)
Technical Skills	0.0627 (0.122)	0.0841 (0.130)	-0.0502 (0.106)	-0.0219 (0.0463)	0.0132 (0.0522)	-0.0283 (0.197)	-0.113 (0.228)	-0.180 (0.186)	-0.0959 (0.0998)	-0.180 (0.186)	-0.000579 (0.0831)	-0.180 (0.186)	-0.000579 (0.0831)
Female, White	-0.0466 (0.134)	-0.117 (0.142)	-0.0545 (0.117)	-0.0237 (0.0510)	-0.0154 (0.0566)	-0.419 (0.215)	-0.612 (0.249)	-0.545 (0.208)	-0.290 (0.115)	-0.545 (0.208)	-0.171 (0.0895)	-0.545 (0.208)	-0.171 (0.0895)
Male, Non-White	-0.0293 (0.158)	-0.0100 (0.169)	-0.0259 (0.137)	-0.0113 (0.0595)	-0.00691 (0.0664)	-0.567 (0.271)	-0.617 (0.318)	-0.507 (0.257)	-0.270 (0.136)	-0.507 (0.257)	-0.265 (0.111)	-0.507 (0.257)	-0.265 (0.111)
Female, Non-White	0.0852 (0.160)	0.101 (0.171)	0.0909 (0.137)	0.0396 (0.0599)	0.0245 (0.0680)	-0.329 (0.264)	-0.260 (0.301)	-0.0465 (0.261)	-0.0247 (0.138)	-0.0465 (0.261)	-0.142 (0.111)	-0.0465 (0.261)	-0.142 (0.111)
Observations	2040	2040	2040	2040	2040	840	840	840	840	840	840	840	840
R^2	0.128	0.196	0.500			0.119	0.323	0.593		0.119	0.323	0.593	
<i>p-value for test of joint significance of Majors</i>	0.021	0.027	0.007	0.007	0.030	< 0.001	0.035	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	No
Order FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	No
Subject FEs	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	Yes	No

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

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

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 and Column 8 divided by the Column 3 and Column 8 coefficients on GPA, with standard errors calculated by delta method. The p -values of tests of joint significance of major fixed effects are indicated (F -test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

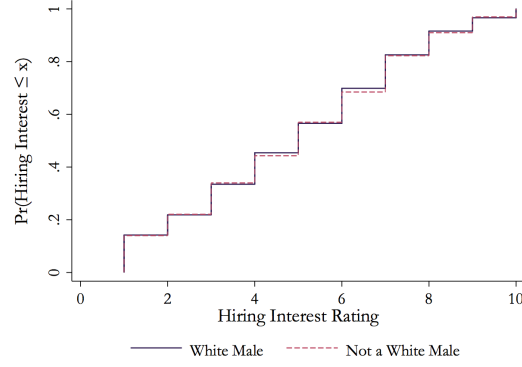
1.3.5. *Candidate Likelihood of Acceptance*

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

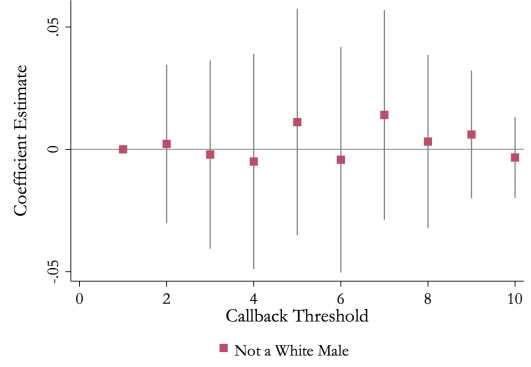
As described in Section 1.2.4, for each resume we asked employers “How likely do you think [Name] would be to accept a job with your organization?” Asking this question helps ensure that our measure of hiring interest is unconfounded with concerns that a candidate would accept a position when offered. However, the question also allows us to study this second factor, which also affects callback decisions.

Table 4 replicates the regression specifications from Table 2, estimating Equation (1.1) when V_{ij} is *Likelihood of Acceptance*, which takes values from 1 to 10. Employers in our sample view high quality candidates as *more likely* to accept a job with their firm than low quality candidates. This suggests that employers in our sample believe candidate fit at their firm outweighs the possibility that high quality candidates will be pursued by many other firms. In Appendix A.2.5, we further consider the role of horizontal fit and vertical quality and find that—holding hiring interest in a candidate constant—reported likelihood of acceptance falls as evidence of vertical quality (e.g., GPA) increases. This result highlights that there is independent information in the likelihood of acceptance measure.

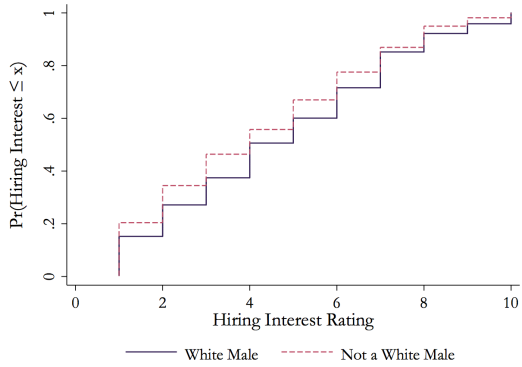
Figure 2: Demographics by Major Type Over Selectivity Distribution



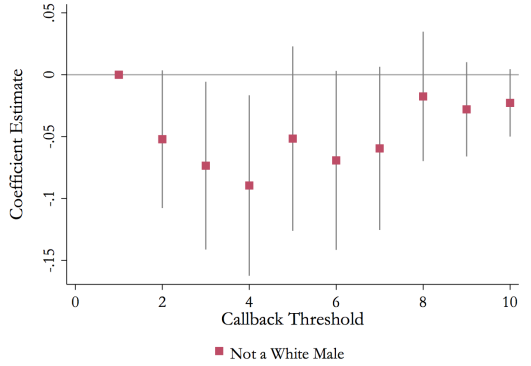
(a) Empirical CDF: Not a White Male, Humanities & Social Sciences



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



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



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

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

Table 4 shows that employers report female and minority candidates are less likely to accept a position with their firm, by 0.2 points on the 1–10 Likert scale (or about one tenth of a standard deviation). This effect is robust to the inclusion of a variety of controls, and it persists when we hold hiring interest constant in Appendix Table 29. Table 5 splits the sample and shows that while the direction of these effects is consistent among both groups of employers, the negative effects are particularly large among employers recruiting STEM candidates.

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

1.3.6. Comparing our Demographic Results to Previous Literature

Qualitative comparison

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

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

different preferences than the typical firms audited in prior literature.

Resume audit studies have consistently shown lower callback rates for minorities. We see no evidence of lower ratings for minorities on average, but we do see lower ratings of minority male candidates by STEM employers. Results on gender in the resume audit literature have been mixed. In summarizing results from 11 studies conducted between 2005 and 2016, (Baert, 2018) finds four studies with higher callback rates for women, two with lower callback rates, and five studies with no significant difference. None of these studies found discrimination against

Table 4: Likelihood of Acceptance

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

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

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for *major*, *leadership experience*, *resume order*, and *subject* included in some specifications as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of *major* fixed effects are indicated (F -test for OLS, likelihood ratio test for ordered probit).

Table 5: Likelihood of Acceptance by Major Type

	Dependent Variable: Likelihood of Acceptance									
	Humanities & Social Sciences					STEM				
	OLS	OLS	OLS	OLS	Ordered Probit	OLS	OLS	OLS	Ordered Probit	Ordered Probit
GPA	0.581 (0.176)	0.610 (0.186)	0.694 (0.142)	0.688 (0.251)	0.251 (0.0719)	0.724 (0.287)	0.813 (0.237)	0.314 (0.110)		
Top Internship	0.786 (0.111)	0.773 (0.118)	0.754 (0.0892)	0.391 (0.178)	0.316 (0.0458)	0.548 (0.199)	0.527 (0.171)	0.190 (0.0782)		
Second Internship	0.481 (0.136)	0.422 (0.148)	0.424 (0.109)	0.254 (0.198)	0.201 (0.0553)	0.324 (0.230)	0.301 (0.187)	0.119 (0.0880)		
Work for Money	0.206 (0.135)	0.173 (0.144)	0.187 (0.108)	0.155 (0.194)	0.0845 (0.0553)	0.346 (0.239)	0.350 (0.186)	0.0923 (0.0878)		
Technical Skills	-0.0942 (0.125)	-0.103 (0.134)	-0.106 (0.104)	0.0495 (0.190)	-0.0460 (0.0521)	0.000154 (0.217)	-0.116 (0.179)	0.0316 (0.0830)		
Female, White	-0.175 (0.139)	-0.211 (0.148)	-0.170 (0.116)	-0.365 (0.198)	-0.0615 (0.0564)	-0.572 (0.236)	-0.577 (0.194)	-0.177 (0.0892)		
Male, Non-White	-0.0691 (0.161)	-0.0756 (0.172)	-0.0462 (0.130)	-0.269 (0.259)	-0.0296 (0.0662)	-0.360 (0.302)	-0.289 (0.246)	-0.147 (0.110)		
Female, Non-White	-0.244 (0.162)	-0.212 (0.175)	-0.163 (0.130)	-0.200 (0.243)	-0.107 (0.0679)	-0.108 (0.278)	-0.0103 (0.245)	-0.105 (0.110)		
Observations	2040	2040	2040	840	2040	840	840	840		
R^2	0.040	0.107	0.516	0.090		0.295	0.540			
<i>p-value for test of joint significance of Majors</i>	0.798	0.939	0.785	< 0.001	0.598	0.001	< 0.001	< 0.001		
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No	No	No	Yes	Yes	No	No
Order FEs	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No
Subject FEs	No	No	Yes	No	No	No	Yes	Yes	No	No

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

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

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1). *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of major fixed effects are indicated (F -test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

women in a U.S. setting. This may be due to resume audit studies targeting female-dominated occupations, such as clerical or administrative work. Riach and Rich (2006), which specifically targets male-dominated occupations, shows lower callback rates for women. Outside the labor market, Bohren et al. (2018) and Milkman et al. (2012) found evidence of discrimination against women using audit-type methodology. We find that firms recruiting STEM candidates give lower ratings to white women, demonstrating the importance of being able to reach new subject pools with IRR. We also find that white women receive a lower return to prestigious internships. This result matches a type of discrimination—lower return to quality—seen in Bertrand and Mullainathan (2004), but we find it for gender rather than race.

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

Quantitative comparison using GPA as a numeraire

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

In studies that randomize GPA, we can divide the observed effect due to race or gender by the effect due to GPA to compare with our GPA-scaled estimates. For example, exploiting the random variation in GPA and gender from Quadlin (2018), we calculate that being female leads to a decrease in callback equivalent to 0.23 GPA points.²⁶ Our results (shown

²⁶Quadlin (2018) reports callback rate in four GPA bins. The paper finds callback is lower in the highest GPA bin than the second highest bin, which may be due to concerns about likelihood of acceptance. Looking at the second and third highest bins (avoiding the non-monotonic bin), we see that an increase in GPA from the range [2.84, 3.20] to [3.21, 3.59]—an average increase of 0.38 GPA points—results in a callback

in Tables 2 and 3) suggest that being a white female, as compared to a white male, is equivalent to a decrease of 0.073 GPA points overall and 0.290 GPA points among employers recruiting for STEM.

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

1.4. Pitt Replication: Results and Lessons

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

rate increase of 3.5 percentage points. Dividing 0.38 by 3.5 suggests that each 0.11 GPA points generates 1 percentage point difference in callback rates. Quadlin (2018) also finds a callback difference of 2.1 percentage points between male (14.0%) and female (11.9%) candidates. Thus, applicant gender has about the same effect as a 0.23 change in GPA.

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

²⁸Unlike at Penn, there is no major fall recruiting season with elite firms at Pitt. We recruited employers in the spring semester only, first in 2017 and again in 2018. The Pitt recruitment email was similar to that used at Penn (Figure 5), and originated from the Pitt Office of Career Development and Placement Assistance. For the first wave at Pitt we offered webinars, as described in Appendix A.1.1, but since attendance at these

Pitt would have given us additional insight into the extent to which Penn employers differed from employers traditionally targeted by audit studies.

Instead, we learned that we were insufficiently attuned to how recruiting differences between Penn and Pitt employer populations should influence IRR implementation. Specifically, we observed significant attenuation over nearly all candidate characteristics in the Pitt data. Table 6 shows fully controlled OLS regressions highlighting that our effects at Pitt (shown in the second column) are directionally consistent with those at Penn (shown in the first column for reference), but much smaller in size. For example, the coefficient on GPA is one-tenth the size in the Pitt data. We find similar attenuation on nearly all characteristics at Pitt for both *Hiring Interest* and *Likelihood of Acceptance*, in the pooled sample and separated by major type. We find no evidence of Pitt employers responding to candidate demographics. (Appendix A.3 provides details for our experimental implementation at Pitt and Tables 31, 32, and 33 display the full results.)

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

sessions was low, we did not offer them in the second wave. We collected resume components to populate the tool at Pitt from real resumes of graduating Pitt seniors. Rather than collect resumes from clubs, resume books, and campus job postings as we did at Penn, we used the candidate pool of job-seeking seniors both to populate the tool and to suggest matches for employers. This significantly eased the burden of collecting and scraping resumes. At Pitt, majors were linked to either the “Dietrich School of Arts and Sciences” or the “Swanson School of Engineering”. Table 30 lists the majors, associated school, major category, and the probability that the major was drawn. We collected top internships at Pitt by identifying the firms hiring the most Pitt graduates, as at Penn. Top internships at Pitt tended to be less prestigious than the top internships at Penn.

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

among the most important considerations during recruitment, compared to only 15.3% at Penn.

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

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

This was a limitation in our implementation at Pitt rather than in the IRR methodology itself. That is, one could design an IRR study specifically for employers interested in hiring registered nurses, or employers interested in hiring mobile software developers, or employers interested in hiring electrical engineers. Our failure at Pitt was in showing all of these employers resumes with the same underlying components. We recommend that researchers using IRR either target employers that specifically recruit high quality generalists, or construct resumes with appropriate variation within the employers’ target areas. For example, if we ran our IRR study again at Pitt, we would ask the *Target Majors* question first and then only generate hypothetical resumes from those majors.

Table 6: Hiring Interest at Penn and Pitt

	Dependent Variable: Hiring Interest			
	Penn	Pitt	Pitt, Wave 2 Non-Target Major	Pitt, Wave 2 Target Major
GPA	2.196 (0.129)	0.265 (0.113)	-0.196 (0.240)	0.938 (0.268)
Top Internship	0.897 (0.0806)	0.222 (0.0741)	0.0199 (0.142)	0.0977 (0.205)
Second Internship	0.466 (0.0947)	0.212 (0.0845)	0.0947 (0.165)	0.509 (0.220)
Work for Money	0.154 (0.0914)	0.153 (0.0807)	0.144 (0.164)	0.378 (0.210)
Technical Skills	-0.0711 (0.0899)	0.107 (0.0768)	0.125 (0.149)	-0.0354 (0.211)
Female, White	-0.161 (0.0963)	0.0279 (0.0836)	-0.0152 (0.180)	-0.151 (0.212)
Male, Non-White	-0.169 (0.115)	-0.0403 (0.0982)	0.00154 (0.185)	-0.331 (0.251)
Female, Non-White	0.0281 (0.120)	-0.000197 (0.100)	0.182 (0.197)	-0.332 (0.256)
Observations	2880	3440	642	798
R^2	0.483	0.586	0.793	0.596
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	0.120	0.850
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes	Yes
Order FEs	Yes	Yes	Yes	Yes
Subject FEs	Yes	Yes	Yes	Yes

Table shows OLS regressions of hiring interest from Equation (1.1). Sample differs in each column as indicated by the column header. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included in all specifications. R^2 is indicated for each OLS regression. The p -value of an F -test of joint significance of major fixed effects is indicated for all models.

1.5. Conclusion

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

We deploy IRR to study employer preferences for candidates graduating from an Ivy League university. We find that employers highly value both more prestigious work experience the summer before senior year and additional work experience the summer before junior year. We use our ten-point rating data to demonstrate that preferences for these characteristics are relatively stable throughout the distribution of candidate quality. We find no evidence that employers are less interested in female or minority candidates on average, but we find evidence of discrimination among employers recruiting STEM candidates. Moreover, employers report that white female candidates are less likely to accept job offers than their white male counterparts, a novel channel for discrimination.

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

A key advantage of the IRR methodology is that it avoids the use of deception. We speculate that economics has tolerated the use of deception in correspondence audit studies in part because of the absence of a deception-free alternative. We developed IRR to provide such an alternative. The availability of an alternative is particularly important given the recent proliferation of deceptive audit studies both within labor economics and into settings beyond labor markets. As discussed in the Introduction, the increasing use of audit studies within labor markets risks contaminating the subject pool—biasing estimates from future audit studies and harming real applicants whose profiles look like fake candidates created by researchers.

Extending deception in new settings may have additional unintended consequences. As prominent examples, researchers have recently audited college professors requesting in-person meetings (Milkman et al., 2012, 2015) and politicians requesting information (Butler and Broockman, 2011; Distelhorst and Hou, 2017). Professors are likely to learn about audit studies ex post and may take the existence of such studies as an excuse to ignore emails from students in the future. Audits of politicians’ responses to correspondence from putative constituents might distort politicians’ beliefs about the priorities of the populations they serve, especially when researchers seek a politician-level audit measure, which requires sending many fake requests to the same politician.

We hope that further development of the IRR method will lead to stricter standards for when deception can be used in economics research and that it will be a welcome change even among researchers who run audit studies, since reducing the number of deceptive audit studies limits contamination of the subject pool.

A second advantage of the IRR method is that it elicits richer preference information than binary callback decisions.³⁰ In our implementation, we elicit granular measures of employers’ hiring interest and of employers’ beliefs about the likelihood of job acceptance. We also see the potential for improvements in preference elicitation by better mapping these metrics into hiring decisions, by collecting additional information from employers, and by raising the stakes, which we discuss below.

The IRR method has other advantages. IRR can access subject populations that are inaccessible with audit or resume audit methods. IRR allows researchers to gather rich data from a single subject—each employer in our implementation rates 40 resumes—which is helpful for power and makes it feasible to identify preferences for characteristics within

³⁰Bertrand and Duflo (2016) argues that the literature has generally not evolved past measuring differences in callback means between groups, and that it has been less successful in illuminating mechanisms driving these differences. That said, there have been some exceptions, like Bartoš et al. (2016), which uses emails containing links to learn more about candidates to show that less attention is allocated to candidates who are discriminated against. Another exception is Bohren et al. (2018), which uses evaluations of answers posted on an online Q&A forum—which are not conflated with concerns about likelihood of acceptance—to test a dynamic model of mistaken discriminatory beliefs.

individual subjects. IRR allows researchers to randomize many candidate characteristics independently and simultaneously, which can be used to explore how employers respond to interactions of candidate characteristics. Finally, IRR allows researchers to collect supplemental data about research subjects, which can be correlated with subject-level preference measures and allows researchers to better understand their pool of employers.

A final advantage of IRR is that it may provide direct benefits to subjects and other participants in the labor market being studied; this advantage stands in stark contrast to using subject time without consent, as is necessary in audit studies. We solicited subject feedback at numerous points throughout the study and heard very few concerns.³¹ Instead, many employers reported positive feedback. Positive feedback also came by way of the career services offices at Penn and Pitt, which were in more direct contact with our employer subjects. Both offices continued the experiment for a second wave of recruitment and expressed interest in making the experiment a permanent feature of their recruiting process. In our meetings, the career services offices reported seeing value in IRR to improve their matching process and to learn how employers valued student characteristics (thus informing the advice they could give to students about pursuing summer work and leadership experience and how to write their resumes). While we did not solicit feedback from student participants in the study, we received hundreds of resumes from students at each school, suggesting that they valued the prospect of having their resumes sent to employers.³²

Naturally, IRR also has some limitations. Because the IRR method informs subjects that responses will be used in research, it may lead to experimenter demand effects (see, e.g., de Quidt et al. (2018)). We believe the impact of any experimenter demand effects is

³¹First, we solicited feedback in an open comments field of the survey itself. Second, we invited participants to contact us with questions or requests for additional matches when we sent the 10 resumes. Third, we ran a follow-up survey in which we asked about hiring outcomes for the recommended matches (unfortunately, we offered no incentive to complete the follow-up survey and so its participation was low).

³²Student involvement only required uploading a resume and completing a short preference survey. We did not notify students when they were matched with a firm, in order to give the firms freedom to choose which students to contact. Thus, most students were unaware of whether or not they were recommended to a firm. We recommended 207 unique student resumes over the course of the study, highlighting the value to students.

likely small, as employers appeared to view our survey tool as a way to identify promising candidates, rather than as being connected to research (see discussion in Section 1.2). For this reason, as well as others highlighted in Section 1.3.4, IRR may be less well equipped to identify explicit bias than implicit bias. More broadly, we cannot guarantee that employers treat our hypothetical resumes as they would real job candidates. As discussed in the Introduction, however, future work could help validate employer attention in IRR studies.³³ In addition, because the two outcome measures in our study are hypothetical objects rather than stages of the hiring process, in our implementation of IRR we cannot draw a direct link between our findings and hiring outcomes. Below, we discuss how this might be improved in future IRR implementations.

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

Future research using IRR can certainly improve upon our implementation. First, as discussed at length in Section 1.4, our failed attempt to replicate at Pitt highlights that future researchers must take care to effectively tailor the content of resumes to match the hiring needs of their subjects. Second, we suggest developing a way to translate Likert-scale responses to the callback decisions typical in correspondence audit studies. One idea is to ask employers to additionally answer, potentially for a subset of resumes, a question of the form: “Would you invite **[Candidate Name]** for an interview?” By having the Likert-scale responses and this measure, researchers could identify what combination of the hiring interest and likelihood of acceptance responses translates into a typical callback decision

³³The time employers spent evaluating resumes in our study at Penn had a median of 18 seconds and a mean that was substantially higher (and varies based on how outliers are handled). These measures are comparable to estimates of time spent screening real resumes (which include estimates of 7.4 seconds per resume (Dishman, 2018) and a mean of 45 seconds per resume (Culwell-Block and Sellers, 1994)).

(and, potentially, how the weight placed on each component varies by firm). Researchers could also explore the origin and accuracy of employer beliefs about likelihood of acceptance by asking job candidates about their willingness to work at participating firms. Third, researchers could increase the stakes of IRR incentives (e.g., by asking employer subjects to guarantee interviews to a subset of the recommended candidates) and gather more information on interviews and hiring outcomes (e.g., by building or leveraging an existing platform to measure employer and candidate interactions).³⁴

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

³⁴An additional benefit of collecting data on interviews and hiring is that it would allow researchers to better validate the value of matches to employers (e.g., researchers could identify 12 potential matches and randomize which 10 are sent to employers, identifying the effect of sending a resume to employers on interview and hiring outcomes). If employers do respond to the matches, one could imagine using IRR as an intervention in labor markets to help mitigate discrimination in hiring, since IRR matches can be made while ignoring race and gender.

CHAPTER 2 : Learning to Manipulate: Experimental Evidence on Out-of-Equilibrium Truth-Telling (with Clayton R. Featherstone and Eric Mayefsky)

2.1. Introduction

Why do some two-sided matching mechanisms continue to be used from year to year while others are abandoned? Although the usual distinction concerns whether a mechanism is stable with respect to the reported preferences,¹ such an explanation is incomplete without also considering whether preferences are truthfully revealed.² Previous theoretical literature has looked at large markets to do this; however, we take a different tack by observing strategic preference revelation in the lab. Our evidence suggests that out-of-equilibrium truth-telling under the deferred acceptance mechanism can lead to matches that are more stable than theory predicts.

Two-sided matching mechanisms are widely used in the field. The most well-known example is the National Resident Matching Program (NRMP) which every year makes about 25,000 matches between newly-minted doctors and residency programs in the United States (NRMP, 2009). Once participants have formed their preferences, they submit rank-order lists of acceptable match partners to the NRMP clearinghouse, which then runs those lists through an algorithm, outputting a match. Other examples of two-sided matching include the Association of Psychology Post-doctoral and Internship Centers (APPIC) match (about 2,800 clinical psychologists matched to internship programs per year (APPIC, 2009)), and the New York City Department of Education public high school match (about 90,000 high school students per year (NYC-DOE, 2009)).³

¹One might also bypass truthful preference revelation entirely and simply look at whether a mechanism yields a stable allocation in equilibrium. See, for instance, Roth (1984b), Ergin and Sönmez (2006) and Pathak and Sönmez (2008).

²Roth (1982) shows that any mechanism that is stable with respect to reported preferences cannot admit truth-telling as a dominant strategy for all players.

³For papers on these matches, see Roth (1984a, 1996, 2003); Roth and Peranson (1999); Roth and Xing (1997); Abdulkadiroğlu et al. (2005); Abdulkadiroğlu et al. (2009).

When deciding which mechanism to use in a matching market, the literature has consistently come back to the idea of stability. A *stable* match has no agents who would prefer to remain unmatched (individual rationality) and no *blocking pairs* (pairwise stability), where a blocking pair is two agents who prefer each other to their assigned matches. If agents are free to recontract ex post, it is not too hard to see how instabilities might render the match moot, but even if agents must abide by the match, they can sidestep it by anticipating blocking pairs and either formally contracting early or informally prearranging a match.⁴ This has been shown both theoretically (Sönmez, 1999; Roth, 1991) and in the lab (Kagel and Roth, 2000). If too many agents leave the match or prearrange, then the clearinghouse will fail to achieve its purpose, and will likely be abandoned. Of course, a stable matching mechanism does not necessarily prevent unraveling,⁵ but in many real world markets, whether or not a stable mechanism is used seems to make the difference.

Most matching schemes we see in the field can be classified as either *priority* mechanisms or *deferred acceptance (DA)* mechanisms.⁶ DA mechanisms are based on the Gale-Shapley algorithm. One such mechanism, *M-Proposing DA*, is implemented in the following way, denoting the members of the two sides of the market *Ms* and *Ws* (Gale and Shapley, 1962):

M-Proposing DA

Step 1: All *Ms* make an offer to their first-choice *W*; *Ws* hold their favorite acceptable offer, rejecting all others.

Step t: Rejected *Ms* make an offer to their favorite acceptable *W* that hasn't rejected them

⁴Usually a pair can do this by agreeing to rank each other first to the clearinghouse. Most mechanisms guarantee that two partners who rank each other first will be matched.

⁵Other causes of early contracting include: insuring over states of the world before payoff relevant information is revealed (Roth and Xing, 1994; Li and Rosen, 1998; Li and Suen, 2000; Suen, 2000), the presence of market power (Roth and Xing, 1994), similar preferences (Halaburda, 2010), arrival of new agents (Du and Livne, 2010), excess supply of workers combined with insufficient supply of high quality workers (Niederle et al., 2009), cultural norms concerning exploding offers (Niederle and Roth, 2009), information transmission through a social network (Fainmesser, 2013), and costs of participation (Damiano et al., 2005).

⁶Another important class of mechanisms, based on linear programming optimization, is not considered here. See Ünver (2001) and Ünver (2005).

yet; W s hold their favorite acceptable offer from this round and previous rounds, rejecting all others.

STOP: The algorithm stops in the first round where no new offers are made. All held offers become finalized matches.

Priority mechanisms instead use the preferences submitted by the participants to order the set of all possible match pairs. They then try to implement those match pairs in that order, skipping those that are not feasible due to previously implemented matches (Roth and Sotomayor, 1990). For concreteness, consider the *M-Proposing Priority* mechanism implemented by the following algorithm:⁷

M-Proposing Priority

Step 1: All M s make an offer to their first-choice W ; W s are permanently matched to their favorite acceptable M who made an offer, rejecting all other offers.

Step t : Rejected M s make an offer to their favorite acceptable W that has not yet rejected them; matched W s reject all offers; and unmatched W s are permanently matched to their favorite acceptable M who made an offer.

STOP: The algorithm stops in the first round where no new offers are made.

A key difference between the *M-Proposing DA* and *M-Proposing Priority* algorithms is that DA mechanisms yield matches that are stable with respect to the reported preferences, while priority mechanisms generally do not. Since the literature looks for stable mechanisms, it has tended to look to DA, a preference which seems to be empirically justified. Unlike in the U.S., residency matches in the United Kingdom are organized at the regional level. Policy variation across regions then provides a natural experiment that is exploited by Roth

⁷The priority ordering for this mechanism ranks potential match pairs in the order of M s' preferences, with ties broken by W s' preferences.

(1991), which finds that regions that adopted DA mechanisms tended to keep using them, while regions that adopted priority mechanisms tended to abandon them after a few years.⁸

Unfortunately, the simple fact that DA is stable relative to the true preferences cannot explain why it outlasts priority mechanisms. Under DA, only participants on the proposing side have incentive to truthfully report. The receiving side often fails to truthfully reveal in Bayes-Nash equilibrium (Roth and Rothblum, 1999; Coles, 2009).⁹ Furthermore, equilibrium predicts that, under incomplete information, neither DA nor priority mechanisms should yield matches that are stable relative to true preferences. Why then does DA persist where Priority fails? Several contributing causes have been considered, but there are still some markets where these explanations are not fully satisfactory.

It could be that preferences are near perfectly correlated on one or both sides of the market. This would push the market toward a unique stable match, thereby removing the incentive to deviate from truth-telling under DA.¹⁰ Although it is intuitive to expect some correlation in preferences, we might also expect a lack of correlation in preferences across matches that are commonly perceived to be of similar quality.

Another possibility is that agents find being unmatched extremely distasteful. Potentially profitable manipulations take a gamble at being unmatched in exchange for a higher probability of matching to a more preferred partner (Roth and Rothblum, 1999). If being unmatched is bad enough, no agent will take this gamble. Even so, in many situations, it is unclear how bad being unmatched is. For instance, in the NRMP match, where hospitals are on the receiving side of the market, unmatched positions can still be filled in the centrally organized aftermarket, known as the “Scramble”.

⁸An interesting nuance of the U.K. study is that, due to the nationalization of healthcare in that country, doctors and hospitals had no choice but to go through the regional match clearinghouses. Unraveling seems to have been enacted through informal prearrangement.

⁹Similar results holds for priority mechanisms (Ehlers, 2008).

¹⁰The simplest way to see this is in the one-to-one case, where it is a straightforward application of the Blocking Lemma and the fact that no individually rational matching can make all the members of one side of the market strictly better off than the unique stable match (Roth and Sotomayor 1990, Lemma 3.5 and Theorem 2.27).

A third option is that the number of stable matches gets small as the market gets large, as established theoretically in Immorlica and Mahdian (2005) and significantly extended in Kojima and Pathak (2009). Although these papers lay out an intuitive mechanism by which core convergence might occur, they do so in the context of a very slowly converging asymptotic (Kadam, 2011); for example, if agents are allowed to list five acceptable members on the other side of the market, as is the case in our experiment, then the Kojima and Pathak bound on the fraction of agents who could proviably deviate from a truth-telling equilibrium does not go below 1 until the market has in excess of 10^{34} agents.¹¹ Because of the extreme looseness of this upper bound at more reasonable market sizes, we must instead rely on computational work to give us an idea of how “big” a market must be for large market results to kick in.

Fortunately, Roth and Peranson (1999) provides just such a benchmark. They show that there is little leeway for manipulation relative to submitted preferences in the NRMP match, although, as they mention, this could be because the submitted preferences had already been manipulated to an equilibrium. To evaluate this possibility, they then look at large simulated markets, finding that markets the size of the NRMP have little room for manipulation, while smaller ones do.¹² Unfortunately, such computational work merely tells us that there is likely a much better bound than the one derived in Kojima Pathak. How much better remains an open question.

Hence, previous research leaves us reasonably confident that very large markets, such as the NRMP (around 20,000 agents), have very small cores, but leaves us less certain about smaller markets. And there are many such markets; in addition to the small regional matches in the UK (about 150 agents) there are many smaller fellowship matches run by the NRMP where DA also seems to halt unraveling, most of which have fewer than 100

¹¹Specifically, the asymptotic states that the upper bound equals $\frac{16 \cdot \bar{q} \cdot k}{\log(\bar{q} \cdot n)}$, where \bar{q} is the maximum capacity of any hospital, k is the number of hospitals that each doctor is allowed to list, and n is the number of hospitals. We set $\bar{q} = 1$ and $k = 5$, and solve for the n that makes the bound equal to 1.

¹²See Figure 2 in Roth and Peranson (1999). Further, note that its simulations involve are for one-to-one markets. The asymptotic mentioned in Footnote 11 implies that there is more leeway for manipulation in many-to-one markets, as \bar{q} and k must increase.

fellowship programs represented, some with multiple positions for each program (Roth, 1991; NRMP, 2009).

A new cause for the empirical success of DA, which we pursue in this paper, is that match participants on the receiving side of a DA mechanism might truth-tell in an out-of-equilibrium manner, leading to truly stable matches. To confirm this intuition, we will look at strategies used by experimental participants on the receiving side of DA and *M*-Proposing Priority both in an environment where they should truth-tell and in an environment where they should deviate from truth-telling. We find that truth-telling rates are similarly high in both environments under DA, but that truth-telling rates are both economically and statistically different under Priority. The first result supports our story of out-of-equilibrium truth-telling, while the second demonstrates that the truth-telling is unlikely to be a mere artifact of the lab.

To understand what drives the differences in strategic play, we estimate a flexible Experience-Weighted Attraction (EWA) learning model that decomposes initial beliefs about successful strategies from willingness to explore new strategies, learning from past play, and learning from counterfactual play. We find major differences between treatments only in players' initial beliefs, suggesting that correcting these beliefs—for instance, by instructing players on the benefits of strategic play, or setting defaults that increased strategic play—could increase best response rates and improve individual players' outcomes, but lead to market unraveling.

We would like to emphasize that we think of the out-of-equilibrium truth-telling explanation put forward by this paper as a complement of, rather than a replacement for, the other explanations we have mentioned. The persistence of DA even in small markets implies that there might be something else going on besides the core convergence explanations which have previously been put forward, and we primarily seek to address this gap in understanding.

Before proceeding, we briefly mention how the current paper fits into the previous experi-

mental matching literature. The first two-sided matching experiments date to the early 90's (Sondak and Bazerman, 1991; Harrison and McCabe, 1996). An experiment that explicitly compares priority and DA mechanisms is described in Kagel and Roth (2000), although their paper focuses more on unraveling behavior than on strategic preference revelation. They do, however, provide a nice demonstration of the intuitive link between stability and persistence. Ünver (2005) runs a similar experiment that also includes linear programming mechanisms. Other different, but related experiments include Haruvy and Ünver (2007) and Echenique and Yariv (2010), which look at repeated decentralized markets, and Nalbantian and Schotter (1995), which looks at several mechanisms that involve matching with money. Our experiment is perhaps most closely related to Echenique et al. (2010), which also looks at strategies in a two-sided matching market. Their design allows agents to go through the DA algorithm as an extensive form game, and their main finding is that agents on the proposing side tend to skip over proposals sub-optimally. Our design treats the DA algorithm as a normal form game, and we focus on the strategies of the receiving side of the algorithm, finding some sub-optimal truth-telling. To our knowledge, we are the first paper to focus on the strategies of the receiving side explicitly. Finally, we mention several other experiments that focus on strategies used by the proposing side, mainly in the context of school choice, such as Chen and Sönmez (2006); Pais and Pintér (2008); Calsamiglia et al. (2009), and Featherstone and Niederle (2008).

2.2. Two markets

In our experiment, we will use *M*-Proposing DA and *M*-Proposing Priority in conjunction with two different market structures.¹³ Under one structure, theory predicts that the receiving side will deviate from truth-telling in a particular way under both mechanisms, while under the other structure, theory predicts truth-telling. Note that our experimental design will constrain the *M*s to truth-tell, focusing on the behavior of the *W*s. Because of this design feature, our equilibrium characterizations concern how the *W*s respond to the

¹³See the Introduction for definitions of these mechanisms.

truth-telling from the M s and whether truth-telling can be sustained in equilibrium for the M s.

Throughout this section, we will only present results specific to our experimental markets, but in the Appendix, we show that there are a broad class of symmetric environments in which we expect similar results.¹⁴ Symmetric environments can be thought of as representing realistic situations where match participants have little information about others' preferences. In such settings, the kinds of manipulations that we expect to see in the lab (truncations) are, in the sense of Roth and Rothblum (1999) and Ehlers (2008), fundamental.¹⁵

2.2.1. The uncorrelated market

Consider a small matching market with 5 M s and 5 W s. The true ordinal preferences of each participant are drawn independently from the uniform distribution over rank-order lists that rank \emptyset (the outcome of being unmatched) last. Cardinal payoffs are a decreasing function of ordinal rank only. We call this the *uncorrelated market*.

Before proceeding to characterize equilibrium, we must first introduce a few definitions. A *revelation strategy* is a mapping from true preferences to reported preferences. Now, due to the symmetry of the problem, any equilibrium in which some agent used a strategy that depended only on a match partner's label would seem unnatural. Therefore, think of an agent's true preferences as a six element vector with the outcome of being unmatched, \emptyset , as its last entry, and define an *anonymous strategy* to be a revelation strategy that always reports the same permutation of the true preference vector.¹⁶ Further, define a *truncation strategy* to be an anonymous strategy where the permutation simply switches the sixth element and some other element of the true preference. We will also consider it

¹⁴The results in the Appendix are also of some independent interest because they extend the results of Roth and Rothblum (1999) and Ehlers (2008) to show how truncation strategies are not just best-responses to symmetric beliefs, but are also the strategies used in equilibria in which agents use anonymous strategies.

¹⁵Also see Day and Milgrom (2008) on how such strategies also appear in core selecting auctions.

¹⁶Note that there is some redundancy in this definition, as the ordering of agents ranked as unacceptable does not matter in any of the mechanisms we consider.

a truncation if the permutation is the identity, that is, truth-telling is also a truncation strategy.

Under M -Proposing DA, the characterization of equilibrium is quite simple, extending the main result of Roth and Rothblum (1999).¹⁷

Proposition 1. *In the uncorrelated market, under M -Proposing DA, any equilibrium in anonymous, weakly undominated strategies involves truth-telling for each $m \in M$ and truncation for each $w \in W$.*

Under M -Proposing Priority, the best-response of the W s when the M s are constrained to truth-tell is similar, extending the main result from Ehlers (2008).

Proposition 2. *In the uncorrelated market, under M -Proposing Priority, if all agents play anonymous, weakly undominated strategies, and all $m \in M$ truth-tell, then all $w \in W$ best-respond to the other agents by playing truncations.*

In the uncorrelated market, then, the unifying principle is that, under both mechanisms, we expect to see the members of W playing truncation strategies.¹⁸

2.2.2. The correlated market

Now, instead of drawing preferences independently for the members of M , draw only one preference and give it to all members of M . Continue to draw a new preference for each member of W . We call this the *correlated market*. A few propositions demonstrate that we expect truth-telling for the members of W under both mechanisms.

Proposition 3. *In the correlated market, under M -Proposing DA, the unique equilibrium in anonymous, weakly undominated strategies entails truth-telling by all agents.*

Proposition 4. *In the correlated market, under M -Proposing Priority, if all members of*

¹⁷Roth and Rothblum (1999) concerns best response to a certain class of beliefs; our theorem concerns strategies used in a certain class of equilibria.

¹⁸We might be worried that an experiment that constrains the M s to truth-tell doesn't have much external validity if such behavior cannot be supported in equilibrium. To this critique, we can provide two statements which are proven in the Appendix. The first is that, at any symmetric equilibrium, the M s must truth-tell. The second is that the strategic problem of the W s is the same, regardless of what anonymous, weakly undominated strategies the M s play, since filtering a uniform distribution through a permutation yields a uniform distribution.

Table 7: Experimental treatments

	Truncation (uncorrelated market)	Truth-telling (correlated market)
Priority	9 groups	8 groups
DA	9 groups	8 groups

M have the same anonymous, weakly undominated strategy, then all members of W best respond by truthfully revealing.

Proposition 3 follows from realizing that if the members of M must truth-tell, then there is a unique stable match relative to the reported preferences. With a unique stable match, there is no reason to deviate from truth-telling.¹⁹ Proposition 4 follows from realizing that if all members of M play the same revelation strategy, then they will all submit the same reported preferences, which means that a member of W receives all offers in the same round of the M -Proposing Priority algorithm.

To conclude, we might worry that it is unrealistic that all members of M should use the same revelation strategy. The next proposition addresses this concern.

Proposition 5. *In the correlated environment, there exist cardinal payoffs that rationalize an equilibrium where all Ms and Ws truthfully reveal their preferences.*

Intuitively, we know this is so by thinking of a case where the payoff for getting a first-ranked W is more than 5 times the payment for getting a second-ranked W , which in turn is more than 4 times the payment for getting a third-ranked W , etc.

2.3. Experimental setup

Table 7 shows the four treatments which comprise the experiment's 2×2 design. We switch the profitability of truncation on and off by switching between the correlated and uncorrelated markets. If our hypothesis holds, we would see no significant difference across these markets under M -Proposing DA. It could then be, however, that experimental participants

¹⁹See Footnote 10 for the sketch of the proof.

always tell the truth in the lab. To control for this, we also observe participant behavior under *M*-Proposing Priority, where the rationale for deviating from truth-telling seems more straightforward. If we observe a difference in truth-telling across markets under Priority, but not under DA, then we will have shown a real effect.

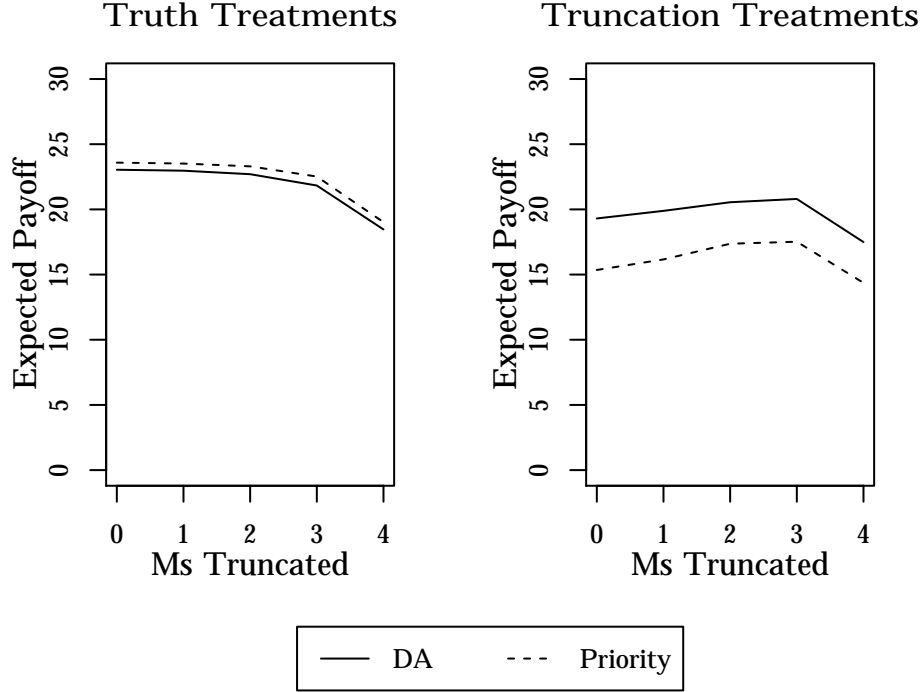
In the experiment, only *Ws* will be played by human participants; the *Ms* will be played by the computer and constrained to truthfully reveal their preferences. Obviously, in real life two-sided matching markets, the proposing side's report to the matching mechanism is not automatic. Under Priority, proposers do not necessarily have dominant strategy incentives to report their preferences truthfully (although as discussed in the theory section, this behavior can occur in equilibrium), and under DA, truthful reporting is a dominant strategy, but there is some experimental evidence that proposing side agents may not propose to all agents in order in an extensive form matching market without frictions (Echenique et al., 2010). We nevertheless use automated proposers playing fixed strategies so that we can focus on the previously unexamined behavior of the receiving side under DA. Using automated *Ms* reduces the complexity and noise in the decision the participants face. If, as we anticipate, subjects have difficulty learning to successfully manipulate the mechanism in this simplified environment, we are confident they will also have trouble in the more complicated real world markets of interest.

In the lab, each participant plays the same market for 40 rounds. In every repetition, each *W* privately learns their new preferences and submits a ranking of some, all or none of the *Ms*. The computer then generates a match outcome according to the rules of the appropriate mechanism to the treatment. *Ws* then learn their match outcome, as well as the outcomes of all other *Ws*. They gain points based on where their match partner appeared in their true preference list for that round, according to payoffs given in Table 8. When designing these payoffs, our goal was to find a payoff scheme which provided behavioral incentives that were as comparable as possible between treatments. In Figure 3, we show

Table 8: Payoff table

Match	1 st choice	2 nd	3 rd	4 th	5 th	No match
Payoff	32 points	16	8	4	2	0

Figure 3: Expected payoff versus number of M s truncated (empirical)



that we succeeded, relative to the actual behavior observed in the lab.²⁰

Finally, we address the design choice to allow for repetition, even though most individuals participate in a matching process in the field only once (or perhaps a handful of times in some applications). In the lab, we can adequately mimic neither the stakes faced by participants in real matching markets nor can we realistically allow experimental participants as much time to consider their prospects as they would have in the field. Instead, by having them participate in repeated trials, we allow for participants to learn about the environment and possibly alter their strategy as they progress. One could argue that this makes

²⁰Note that a simple reinforcement learning model would predict that the slopes of the curves are much more important than the levels.

participants better able to understand the mechanism and behave strategically than in real world markets; however, if this is the case, and, as we anticipate, subjects nonetheless have difficulty successfully manipulating effectively, we can be confident that manipulation is even more difficult in the field.

Briefly, we mention the symmetry of our experimental environments. Non-truncation strategies are not profitable in our setup, but in the field, they might be. Even so, such strategies require much information to implement. Also, though preferences in real-world markets might not look much like those in our experiment, preferences are often tiered. One set (tier) of match partners is clearly preferred to another set, which is preferred to yet another set, but over each tier, preferences are idiosyncratic. In this context, the setup of our experiment can be interpreted as an approximation of at least a sector of the matching market.²¹

All treatments were run at Stanford University during the Spring of 2009. Each session consisted of one or two groups of 5 participants. In sessions with two groups, groups were not mixed during the session, and participants were not informed which other participants were in their group. At the start of each session, participants were read detailed instructions²² and had to successfully work through the steps of the appropriate mechanism for an example set of reported preferences. Actual play commenced only after all participants completed the exercise and indicated they understood the mechanism rules. Nothing was done to overtly suggest what the treatment variables were, i.e., there was no mention of matching mechanisms or preference distributions other than the ones in use in that particular treatment.

During the experimental session, participants could see their preferences for a given round on

²¹Additionally, since interview constraints often prevent match participants from evaluating all potential match partners, we might think that pre-match sorting would lead to market segmentation, to similar effect. For more on modeling the interview process, see Lee and Schwarz (2007), Lee and Schwarz (2009), and Coles et al. (2010).

²²In the lab, we provide a specific context in the hopes of making understanding easier for participants. Proposing side agents (referred to here as *Ms*) are referred to as “Schools” and the agents receiving offers (here, *Ws*) are referred to as “Students.”

Table 9: Truth-telling rates (all periods)

	DA		Priority
Truth-Telling	66.0%	$\leftrightarrow(0.372)$	58.4%
	$\uparrow(0.200)$		$\uparrow(0.002)^{**}$
Truncation	56.6%	$\leftrightarrow(0.001)^{**}$	25.3%

Numbers in parentheses are p -values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

their computer screen and were reminded of payments for all possible match outcomes. They were then directed to click on radio buttons to rank each of the M s.²³ After all participants submitted rankings, a results screen showing the participant’s match for that round, their point accrual for that round and their total cumulative points would be displayed. At all times, a participant had the ability to see, for all prior rounds, the match outcomes for all participants, her own true preferences, and the rank list she submitted in that round.

2.4. Experimental Results

2.4.1. Overall Truth-telling Rates

We are most interested in the rate of truth-telling over all periods across the four primary treatments. This value is significantly higher in the DA truncation treatment than in the Priority truncation treatment; however, for the two truth-telling treatments, the differences between the DA and Priority treatments are not statistically significant. Furthermore, the rate difference between the two DA treatments is not statistically significant, while the difference between the two Priority treatments is highly significant.

When we restrict attention to the last ten periods, focusing on the behavior of subjects when they are more experienced, we find qualitatively similar effects. Statistically, there is a mildly significant difference between the two DA treatments, as well as the high significance between the Priority treatments and the truncation treatments seen in the data for all 40

²³We did this so that participants would have to click the same number of times regardless of what preference they wished to report. If declaring all M s unacceptable were too easy, some participants might choose to do this in order to save time and effort.

Table 10: Truth-telling rates (last 10 periods)

	DA		Priority
Truth-Telling	70.2%	$\leftrightarrow(0.340)$	60.8%
	$\uparrow(0.046)^{**}$		$\uparrow(0.002)^{**}$
Truncation	54.7%	$\leftrightarrow(0.003)^{**}$	19.3%

Numbers in parentheses are p -values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

Table 11: Non-truncation rates

	DA		Priority
Truth-Telling	16.3%	$\leftrightarrow(0.226)$	11.1%
	$\uparrow(0.673)$		$\uparrow(0.210)$
Truncation	14.3%	$\leftrightarrow(0.508)$	17.9%

Numbers in parentheses are p -values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

periods.

Note that for DA, truth-telling rates are slightly lower in the last 10 periods (2% lower) in the truncation treatment, but also 4% higher in the truth-telling treatment. Thus, the significance of the difference in truth-telling rates between the two groups is in some sense as much due to participants in the truth-telling treatment learning to tell the truth as it is those in the truncation treatment learning to truncate. In sum, we only see a significant deviation from the benchmark truth-telling rate under the Priority truncation treatment. Under DA, participants do not respond to the truncation treatment by deviating from truth-telling.

Of course, failure to tell the truth is not synonymous with truncation, and although truncation weakly dominates other non-truth-telling strategies, we do observe some portion of suspects employing “switching” or “dropping” strategies in some rounds. Frequency of this behavior, however, is not significantly different between any of the treatments.

Table 12: Number of Blocking Pairs per Period

	DA		Priority
Truth-Telling	0.47	$\leftrightarrow(0.574)$	0.59
	$\updownarrow(0.809)$		$\updownarrow(0.001)^{**}$
Truncation	0.49	$\leftrightarrow(0.000)^{**}$	1.87

Numbers in parentheses are p -values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

2.4.2. Blocking Pairs and Overall Match Stability

For practical market design, we may be primarily concerned not with the rate at which participants tell the truth, but rather with how successfully a mechanism generates desirable (i.e., stable) match outcomes. One measure of this is the number of blocking pairs present in any given assignment. Since the outcome is never 100% stable in any treatment at any time, the number of blocking pairs is one measure of the degree of stability of a match outcome: a mechanism which generates an outcome that is stable for most participants may still work well enough to be persistent.

Blocking pairs were found to occur significantly more often in the Priority truncation treatment than in the DA truncation treatment or the Priority truth-telling treatment. The two DA treatments were not significantly different in blocking pair frequency; nor were the two truth-telling treatments.

Note that the same M or W can be involved in multiple blocking pairs if there is more than one attainable match partner that they prefer to their actual match partner. However, we do not observe any interesting asymmetries in terms of which unique agents are involved in multiple blocking pairs: the number of unique M s involved in blocking pairs is not significantly different than the number of unique W s for any treatment, and the between-treatment differences are similar qualitatively and in terms of statistical significance when the number of unique M s and W s in blocking pairs are considered separately. The total probability of an M or W being unmatched thus follows a similar pattern across treatments.

Table 13: Percentage of M s and W s Unmatched

	DA		Priority
Truth-Telling	2.7%	$\leftrightarrow(0.065)^*$	4.9%
	$\downarrow(0.311)$		$\downarrow(0.030)^{**}$
Truncation	3.7%	$\leftrightarrow(0.010)^{**}$	11.1%

Numbers in parentheses are p -values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

2.4.3. Best Response Frequencies

Truth-telling rates establish how apt participants are to manipulate, and low non-truth, non-truncation rates²⁴ establish that these manipulations are, for the most part, some sort of truncation. However, participants who truncate are not automatically maximizing their expected payoff: they may be truncating too much or too little. For the set of payoffs used in the experiment, we can find an equilibrium where all agents truncate symmetrically; however, as out-of-equilibrium strategies may be a best response to other out of equilibrium strategies, we would not necessarily expect sophisticated participants to truncate as if in equilibrium. We instead look at the ability of participants to find the strategy which is a best response to the environment in which they find themselves. If a significant proportion of subjects are able to achieve this in a significant portion of sessions for a certain mechanism, we might reach different conclusions as to their sophistication than we would looking strictly at truth-telling rates (or looking at the frequency of play consistent with theoretical equilibrium, for that matter). Also, we might wonder if there is a great deal of heterogeneity in participant sophistication, or if all participants reported optimal truncations about the same fraction of the time.

However, simply comparing subjects' behavior in an individual round to the optimal behavior possible in that period *ex post* fails to capture the uncertainty which is inherent in truncation strategies—it can be optimal *ex ante* to truncate in each period, even though

²⁴The characterization of this other behavior as “non-truthful, non-truncation” is redundant, as truth-telling is one extreme of the set of truncation strategies for participants. We nevertheless use the terminology to ensure clarity.

it may be suboptimal *ex post*. Thus, we consider the participant to be playing optimally in their “environment” if they play the truncation strategy which generates the highest expected utility across some set of rounds they played, given the actual behavior of other participants and generated proposer preferences.

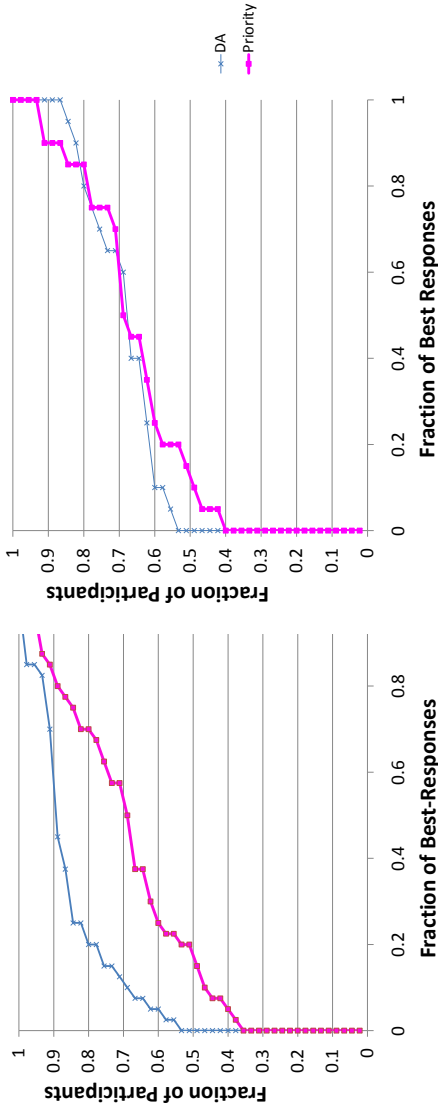
Figure 4a indicates the proportion of participants playing an overall best response at most the indicated proportion of the time for the truncation treatments. For example, approximately 36% of Priority participants never played a best response (compared with about 52% for DA), and 50% of participants played a best response no more than 20% of the time (compared with around 75% for DA). Note that the Priority treatment first order stochastically dominates the DA treatment: for any level of frequency of best response play we consider, more participants best respond at least that frequently in the Priority treatment than in the DA treatment. However, this gap closes when only the last 20 periods are considered, as seen in Figure 4b. Note that this closing of the gap simply implies that under both mechanisms, participants have converged to similarly bad distributions of sub-optimal play.

In the truth-telling treatments (Figures 4c and 4d), truthful reporting is always the unique best response, and much as there was no significant difference in the overall truth-telling rates between DA and Priority in these treatments, there is no noticeable difference in the frequency with which individual subjects play this best response, either in the whole sample or restricting attention to the last 20 periods.

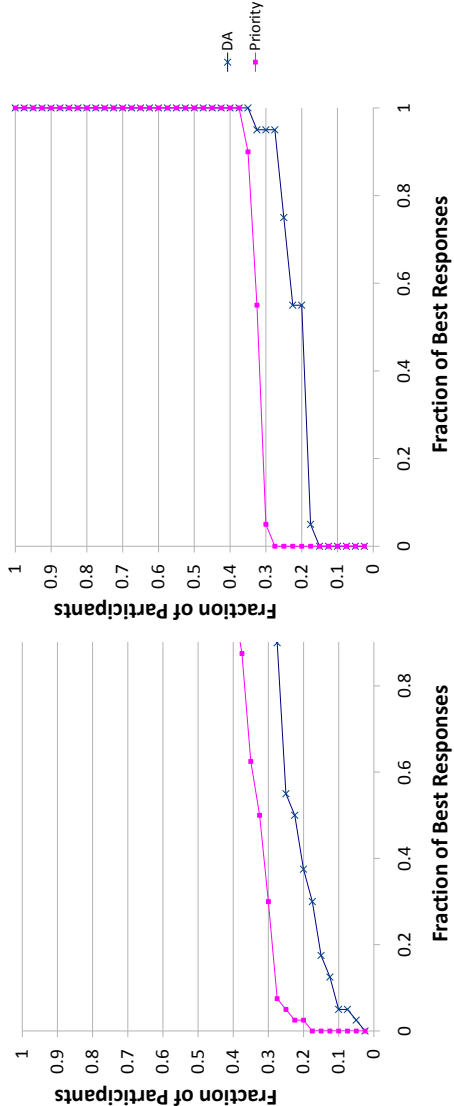
2.5. Learning Model

We have shown that subjects learn to manipulate reported preferences advantageously under the Priority mechanism but not under DA, despite theoretical predictions. A recent body of literature has developed comparing predicted and actual play under different allocation mechanisms (e.g., Li (2017); Rees-Jones (2017); Duflo (2017); Zhang and Levin (2017)). However, there has been little research on the learning process itself, describing

Figure 4: Best response frequency CDFs



(b) Truncation treatments, last 20 periods.



(c) Truth-telling treatments, all periods.



(d) Truth-telling treatments, last 20 periods.



how players approach new mechanisms and learn strategic play over time. These insights could improve the design of mechanisms suggest relationships between properties of mechanisms and characteristics of human players. In this section, we estimate a structural model and describe the dynamics of learning in a repeated game. We lay out a reparametrized version of the Experience-Weighted Attraction model in 2.5.1, and describe our approach to estimation in 2.5.2. We provide estimation results in 2.5.3.

2.5.1. Parameters and Model Dynamics

To understand how subjects determine strategies under the different mechanisms and conditions, we estimate a reparametrized Experience-Weighted Attraction (EWA) learning model introduced by Camerer and Ho (1999). EWA is a flexible model incorporating elements of belief-based and choice reinforcement models.²⁵ We estimate a reparametrized EWA that separately identifies initial cognition and interactive learning.

In the original EWA, the key objects in the model are attractions to strategies. Each agent i begins the game with an initial attraction A to each strategy j , denoted $A_i^j(0)$, derived from pre-game analysis or prior experience. Let s_i^j represent strategy j for agent i , and $s(t)$ represent the set of strategies played in period t . Additionally, define $\pi_i(s_i^j(t), s_{-i}(t))$ as the round t payoffs for player i , which depend on player i 's strategy ($s_i^j(t)$) and all other player's strategies ($s_{-i}(t)$).

After each round of play, each agent updates the previous round's attractions using a weighted combination of their prior attraction, and the payoff from playing the strategy, according to the recursive formula:

$$A_i^j(t) = \varphi \cdot A_i^j(t-1) + \left[\delta + (1-\delta) \cdot \mathbb{1}_{\{s_i^j\}}(s_i(t)) \right] \cdot \pi_i(s_i^j(t), s_{-i}(t)). \quad (2.1)$$

The parameter φ represents a discount factor, and determines how quickly previous at-

²⁵EWA nests belief-based models, where players form expectations about other players' strategies and choose a best response, and choice reinforcement models, in which past payoffs reinforce successful strategies.

tractions decay; the parameter δ is an introspection factor, dictating how much the new attractions depend on realized payoffs from the previous round relative to counterfactual payoffs from unplayed strategies.

In this model, attractions map to probabilities of play in each round according to a power form:

$$P_i^j(t+1) = \frac{\left(A_i^j(t)\right)^\lambda}{\sum_{k=1} \left(A_i^k(t)\right)^\lambda}. \quad (2.2)$$

In this equation, the “exploitation factor” λ determines how often a player chooses her more attractive strategies, relative to the probability of exploring less attractive strategies. This dictates the amount of randomness in a player’s sequence of strategies: when $\lambda = 0$, the player plays all strategies with equal probability, and as λ increases, the probability of playing the most attractive strategy increases.²⁶

Thus, learning dynamics are determined by initial attractions, the weight of previous attractions relative to updating from recent payoffs, and the relative weight of actual and counterfactual payoffs. However, this parametrization fails to fully flesh out the distinction between initial cognition and interactive learning, as well as how these two forces relate.

We define initial cognition to be the process of thinking through a game absent any chance to learn by actually playing it. The culmination of initial cognition is the set of play probabilities for each possible action j , for each individual or type i , in the first round of play, $\{P_i^j(1)\}_{i,j}$. Although these are encoded by the initial attractions, $\{A_i^j(0)\}_{i,j}$, and the exploitation factor, λ , the mapping from these parameters to the initial play probabilities is not one-to-one, since the power-form probability function is invariant to multiplying all initial attractions by a common factor.

It is instructive to consider what other information is codified in the initial attractions and

²⁶Camerer and Ho (1999) refer to λ as the “exploration” factor. We have changed the name to match the intuition behind the model: it is more likely that the player “exploits” its most attractive strategies (rather than “exploring” new strategies) as λ increases.

the exploitation factor. Towards that end, let $\|A_i(0)\|$ denote the λ -norm of the vector of initial attractions. That is,

$$\|A_i(0)\| \equiv \left(\sum_{k=1}^{m_i} \left(A_i^k(0) \right)^\lambda \right)^{1/\lambda}.$$

This norm of the vector of initial attractions contains the extra information: we can now encode our model in terms of initial probabilities of play and $\|A_i(0)\|$. In this reparametrization, the initial attractions are no longer free parameters; instead, they are determined by

$$A_i^j(0) = \|A_i(0)\| \cdot \left(P_i^j(1) \right)^{1/\lambda}.$$

The free parameters of this reparametrized learning model are now $\{P_i^j(1)\}_{i,j}$, $\|A_i(0)\|$, λ , ϕ , and δ . To endow these parameters with simple interpretations, we must first discuss the intuition behind the interactive learning component of the model.

Essentially, each attraction is the net present value of the stream of payoffs associated with a strategy. The parameter φ represents the discount rate, while the parameter δ represents how much counterfactual payoffs are weighted relative to realized payoffs. Agents choose an action randomly according to the power-form probability function discussed above and the exploitation factor λ . All of this is sensible, but we have yet to discuss where these discounted sums should start in the first round of play. The initial play probabilities constrain these initial attractions, but don't completely pin them down.

This is the role of $\|A_i(0)\|$. Intuitively, it is the natural way to sum up all of the payoff streams that have been aggregated across the different actions.²⁷ It tells us how initial cognition will be weighted relative to interactive learning in terms of payoffs from the game.

²⁷Mathematically, the λ -norm is the NPV required to yield the same probability weight while concentrating the NPVs from all the actions into just one. As such, it is, in some sense, the norm that weights entries in a way that corresponds to probability of play. For instance, note that as the exploitation factor λ grows large, $\|A_i(0)\|$ approaches $\max_j A_i^j(0)$, which makes sense as the maximum attraction is the only one that matters as $\lambda \rightarrow \infty$.

In other words, if the average payoff in a game is \$1 per round, then (very roughly), $\|A_i(0)\|$ tells us how initial cognition is weighted in terms of discounted rounds of interactive play.

2.5.2. Model Estimation

In order to estimate the parameters of the model through maximum likelihood estimation (MLE), we first need to simplify the parameter space. Many previous papers estimating the EWA model have done so in games with a small strategy space. With more strategies, it becomes computationally challenging to estimate the initial attraction to each strategy. In our setting, each player chooses between a computationally intractable 325 strategies in each round.²⁸

However, most (225 of 325) strategies are never played in any round of play, and only 20 strategies are played in the first round of any session. Moreover, only 11 strategies are played more than once in an initial round, suggesting that initial probabilities of play are concentrated across a small number of strategies. Rather than estimate initial probabilities for each strategy, we estimate initial probabilities for these 11 strategies, and a single initial probability shared uniformly across all other strategies. This drastically reduces the parameter space, while maintaining flexibility to explain a wide range of observed behaviors.²⁹

With this setup, we can now estimate 15 parameters for each treatment condition: 11 initial probabilities $P_i^j(1)$ describing the initial cognition process, three scalar parameters (ϕ , δ , and λ) to describe the learning process, and $\|A_i(0)\|$ identifying the relative weight of initial cognition and learning.³⁰

²⁸The strategy space for each player during each round of play includes any permutation of preferences over all 5 outcomes, and permutations of any set of truncated preferences (as long as at least one preference is listed). The number of possible strategies in a round is $5! + 4 \times \binom{5}{1} + 3! \times \binom{5}{3} + 2! \times \binom{5}{2} + 1! \times \binom{5}{1} = 325$.

²⁹The 11 estimated strategies include all truncation strategies and six permutation strategies that are not predicted by theory. For a list of all estimated probabilities, see Table 14.

³⁰Note that we only need to estimate 11 probabilities, since the sum of initial probabilities must be one. The probability of playing one of the non-estimated strategies is pinned down by the other estimates. For more details on model estimation, see Appendix B.3.

2.5.3. Structural Model Results

Differences in both initial cognition and learning dynamics help explain subjects' failure to manipulate reported preferences under DA. To summarize the results of the structural estimation clearly in Table 14, we pool initial probabilities of play into three categories: truth-telling, non-truthful truncation, and permutation strategies.

Initial probabilities of truth telling are similar across the *DA Truth* (55.4%), *DA Truncation* (51.9%), and *Priority Truth* (50.8%) treatments, but much lower under *Priority Truncation* (27.4%). This suggests that before play begins, players in the *Priority Truncation* believe there are profitable deviations from truth-telling. The estimates for initial probabilities of playing non-truthful truncation strategies bear out this finding: subjects under *Priority Truncation* are much more likely to truncate (46.7%) than under any other treatment. Under all treatment treatments, permutation strategies are approximately equally likely (between 15.7% and 19.4%) and are not driving differences in the truth-telling rate.

In addition, the weight of initial cognition $\|A_i(0)\|$ is higher under DA treatments, indicating that subjects rely more heavily on pre-game analysis when determining their strategies under DA. This reliance on analysis compounds the errors that subjects make in determining their initial probabilities of play in the *DA Truncation* treatment. Subjects under *DA Truncation* play as if they had about 30% more pre-game experience than their counterparts under *Priority Truncation*.

Three parameters in our model— ϕ , λ , and δ —determine the dynamics of the interactive learning process. We find that λ is significantly lower under Priority than under DA, suggesting that Priority players are more inclined to explore new strategies, while under DA players prefer to exploit their most preferred strategy. This difference may explain why gaps in truncation rates persist after many rounds of play.

Differences between some treatments of the parameters ϕ and δ are also statistically significant, but the differences are economically less significant and unlikely to explain differences

Table 14: Parameter Estimates by Treatment

Parameter	Interpretation	DA Truth	DA Trunc	Priority Truth	Priority Trunc
ϕ	Discount Factor	0.9207 (0.0118)	0.8808 (0.0097)	0.8496 (0.0173)	0.8913 (0.009)
λ	Exploitation Factor	1.5885 (0.1069)	1.5886 (0.1169)	1.016 (0.0764)	1.2836 (0.0784)
δ	Introspection Factor	0.004 (0.0016)	0.005 (0.002)	0.0002 (0.0001)	0.0015 (0.0006)
$\ A_0\ $	Payoff-Weight of Initial Cognition	64.8797 (10.1432)	112.3189 (17.1641)	35.1234 (8.0008)	85.8873 (11.6596)
$P_r^{\{12345\}}(1)$	Initial Probability of Truth-Telling	0.5541 (0.0451)	0.5188 (0.0363)	0.5077 (0.0529)	0.2742 (0.0319)
$P_r^{\{1234\emptyset\}}(1)$	Initial Probability of One-Point Truncation	0.0338 (0.0154)	0.0552 (0.0158)	0.0357 (0.0169)	0.0556 (0.0155)
$P_r^{\{123\emptyset\emptyset\}}(1)$	Initial Probability of Two-Point Truncation	0.0597 (0.0206)	0.1276 (0.0233)	0.1356 (0.0332)	0.1905 (0.0273)
$P_r^{\{12\emptyset\emptyset\emptyset\}}(1)$	Initial Probability of Three-Point Truncation	0.0412 (0.0169)	0.0572 (0.0164)	0.0873 (0.0262)	0.1868 (0.0267)
$P_r^{\{1\emptyset\emptyset\emptyset\emptyset\}}(1)$	Initial Probability of Four-Point Truncation	0.0088 (0.0087)	0.0154 (0.0085)	0.0298 (0.0163)	0.034 (0.0128)
$P_r^{\{21345\}}(1)$	Initial Probability of $\{21345\}$	0.1077 (0.0265)	0.077 (0.0184)	0.0899 (0.0274)	0.086 (0.0187)
$P_r^{\{213\emptyset\emptyset\}}(1)$	Initial Probability of $\{213\emptyset\emptyset\}$	0.0064 (0.0064)	0.0158 (0.0085)	0.0085 (0.0084)	0.0098 (0.0061)
$P_r^{\{21435\}}(1)$	Initial Probability of $\{21435\}$	0.0136 (0.0094)	0.017 (0.0086)	0.0146 (0.0106)	0.0138 (0.007)
$P_r^{\{12354\}}(1)$	Initial Probability of $\{12354\}$	0.0331 (0.0149)	0.0124 (0.0077)	0.046 (0.0199)	0.0205 (0.0091)
$P_r^{\{23145\}}(1)$	Initial Probability of $\{23145\}$	0.0078 (0.0072)	0.0018 (0.0032)	0.0082 (0.0082)	0.017 (0.0081)
$P_r^{\{13245\}}(1)$	Initial Probability of $\{13245\}$	0.0265 (0.0138)	0.0326 (0.0121)	0.0199 (0.0129)	0.0339 (0.0118)
$P_r^{\{\text{other}\}}(1)$	Initial Probability of Other Strategies	0.0003 (0.0001)	0.0002 (0.0001)	0.0001 (0.0001)	0.0002 (0.0001)

Maximum likelihood estimates of reparametrized EWA model, estimated separately by treatment group. Standard errors shown in parenthesis.

in truncation rates. The introspection factor δ describes how much players are able to learn from unplayed strategies. We find that δ is precisely estimated to be between 0.0002 and 0.005 for all treatments, suggesting that more than 99.5% of learning from any round is from realized payoffs rather than counterfactual learning.

Estimates for discount factor ϕ range between 0.850 for the *Priority Truth* treatment and 0.921 for the *DA Truth* treatment. The parameter ϕ dictates how the influence of previous attractions persist over time. To interpret these values, we calculate the half-life of the attraction—the number of periods required to halve the influence of the attraction. The half-life of attractions is about 8.4 periods under *DA Truth*, 5.5 periods under *DA Truncation*, 4.3 periods under *Priority Truth*, and 6.0 under *Priority Truncation*.³¹ These figures provide some insight into the learning process, but they do not explain the systematic differences in learning to manipulate reported preferences.

2.6. Conclusion

Participants in matching markets might not truncate under DA, even when doing so would be significantly profitable. We show this in a simple experimental environment where participants were trained on the mechanism, given ample opportunity to learn through feedback, and were not subject to any randomness that might come from non-straightforward play on the proposing side. Even in this simple setting, players use very little counterfactual analysis, and learning dynamics vary only in players' initial assessment of the game. In the field, where things are more complicated and information is more sparse, we have little reason to think that match participants would be more likely to learn to truncate. These results also suggest that the persistence of DA clearinghouses may rely on participant misoptimization, and that interventions designed to improve understanding could lead to unravelling.

In addition to understanding the persistence of DA in the field, we also think an experiment such as ours feeds into the broader concerns of market design. Whenever a matching

³¹The half-life is given by $t_{\frac{1}{2}} = -\frac{\log(2)}{\log(\phi)}$.

mechanism is strategy-proof, it is straightforward for designers to predict agent behavior in the field, since both focality and optimality push towards truth-telling. Sometimes though, strategy-proofness is either not desired or cannot be achieved due to other design goals. Consider the job of a market designer who has been tasked with creating a two-sided matching mechanism that persists. We can view the current paper as an experiment that would help inform our theoretical designer. Persistence can be intuitively linked to ex post stability, so DA is a natural candidate. Unfortunately, under DA, truth-telling is generally not an equilibrium. Theory provides a set of strategies which could outperform truthful preference revelation: the question is then whether our designer should expect market participants to use these deviations from truth-telling, which is a clear candidate for a focal strategy. If agents use these profitable deviations from truth-telling, then DA will not yield an ex post stable outcome, but if they don't, then it will. To determine which is the more likely outcome, the present lab experiment becomes very informative.

In demonstrating that agents learn to play some deviations from truth-telling, but not others, we bring up the idea that not all equilibria are equal in their predictive power. Depending on the mechanism and environment, agents are sometimes very close to equilibrium play and sometimes not. Some intuitive factors that seem like they should be important for whether a theoretical equilibrium will be realized in the field are focality of truth-telling, obviousness that some deviation from truth-telling will be profitable, difficulty of finding the optimal such deviation, and the profitability of that deviation. Unfortunately, although these factors may guide us intuitively, there is no formal theory for how they might trade off in determining the accuracy of an equilibrium prediction; in fact, most of them are difficult even to define. This is where lab experiments can prove most useful for design. The current paper, for instance, implies that truth-telling is more strongly focal for the receiving side under DA than under Priority. It also shows that under both mechanisms, equilibrium predictions might not hold: under DA, participants truth-tell when they shouldn't, while under Priority, they deviate from truth-telling, but in a sub-optimal way. In short, although the main contribution of this experiment is to show how out-of-equilibrium truth-telling could

lead to ex post stability of DA in the field, we also feel that the experiment is the sort of inquiry that should be used in practical market design.

APPENDIX

Appendices to Chapter 1

We provide three appendices. In Appendix A.1, we describe the design of our experiment in detail, including recruitment materials (A.1.1), survey tool construction (A.1.2), and the candidate matching process (A.1.3). In Appendix A.2, we present additional analyses and results, including human capital results (A.2.1), regressions weighted by GPA (A.2.2), a discussion of our discrimination results (A.2.4), and a discussion of preferences over the quality distribution (A.2.3). In Appendix A.3, we discuss additional details related to replicating our experiment at Pitt.

A.1. Experimental Design Appendix

A.1.1. Recruitment Materials

University of Pennsylvania Career Services sent recruitment materials to both recruiting firms and graduating seniors to participate in the study. All materials marketed the study as an additional tool to connect students with firms, rather than a replacement for any usual recruiting efforts. The recruitment email for employers, shown in Figure 5, was sent to a list of contacts maintained by Career Services and promised to use a “newly developed machine-learning algorithm to identify candidates who would be a particularly good fit for your job based on your evaluations.” In our replication at the University of Pittsburgh, a similar email was sent from the Pitt Office of Career Development and Placement Assistance.

Penn Career Services recruited graduating seniors to participate as part of the candidate matching pool through their regular newsletter called the “Friday Flash.” The relevant excerpt from this email newsletter is shown in Figure 6.

We timed recruitment so that employers would receive their 10 resume matches around the time they were on campus in order to facilitate meeting the job seekers. In addition, we offered webinars for employers who were interested in learning about the survey screening experience before they participated. Employers could anonymously join a call where they

viewed a slideshow about the survey software and could submit questions via chat box. Attendance at these webinars was low.

A.1.2. Survey Tool Design

In this appendix, we describe the process of generating hypothetical resumes. This appendix should serve to provide additional details about the selection and randomization of resume components, and as a guide to researchers wishing to implement our methodology. In Section A.1.2, we describe the structure of the IRR survey tool and participant experience. In Section A.1.2, we describe the structure of our hypothetical resumes. In Section A.1.2, we detail the randomization of candidate gender and race through names. Section A.1.2 details the randomization of educational background. Section A.1.2 describes the process we used to collect and scrape real resume components to randomize work experience, leadership experience, and skills.

Survey Tool Structure

We constructed the survey tool using Qualtrics software for respondents to access from a web browser. Upon opening the survey link, respondents must enter an email address on the instructions page (see Figure 7) to continue. Respondents then select the type of candidates they will evaluate for their open position, either “Business (Wharton), Social Sciences, and Humanities” or “Science, Engineering, Computer Science, and Math.” In addition, they may enter the position title they are looking to fill. The position title is not used in determining the content of the hypothetical candidate resumes. The major selection page is shown in Figure 8. After this selection, the randomization software populates 40 resumes for the respondent to evaluate, drawing on different content by major type. The subject then evaluates 40 hypothetical resumes. After every 10 resumes, a break page encourages subjects to continue.

Figure 5: Employer Recruitment Email

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

Sincerely,

Barbara Hewitt, Senior Associate Director, Career Services

Email sent to firms recruiting at Penn originating from the Senior Associate Director of Career Services at the University of Pennsylvania. Subjects who followed the link in the email were taken to the instructions (Figure 7).

Figure 6: Email Announcement to Graduating Seniors

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

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

[OTHER TEXT APPEARED HERE]

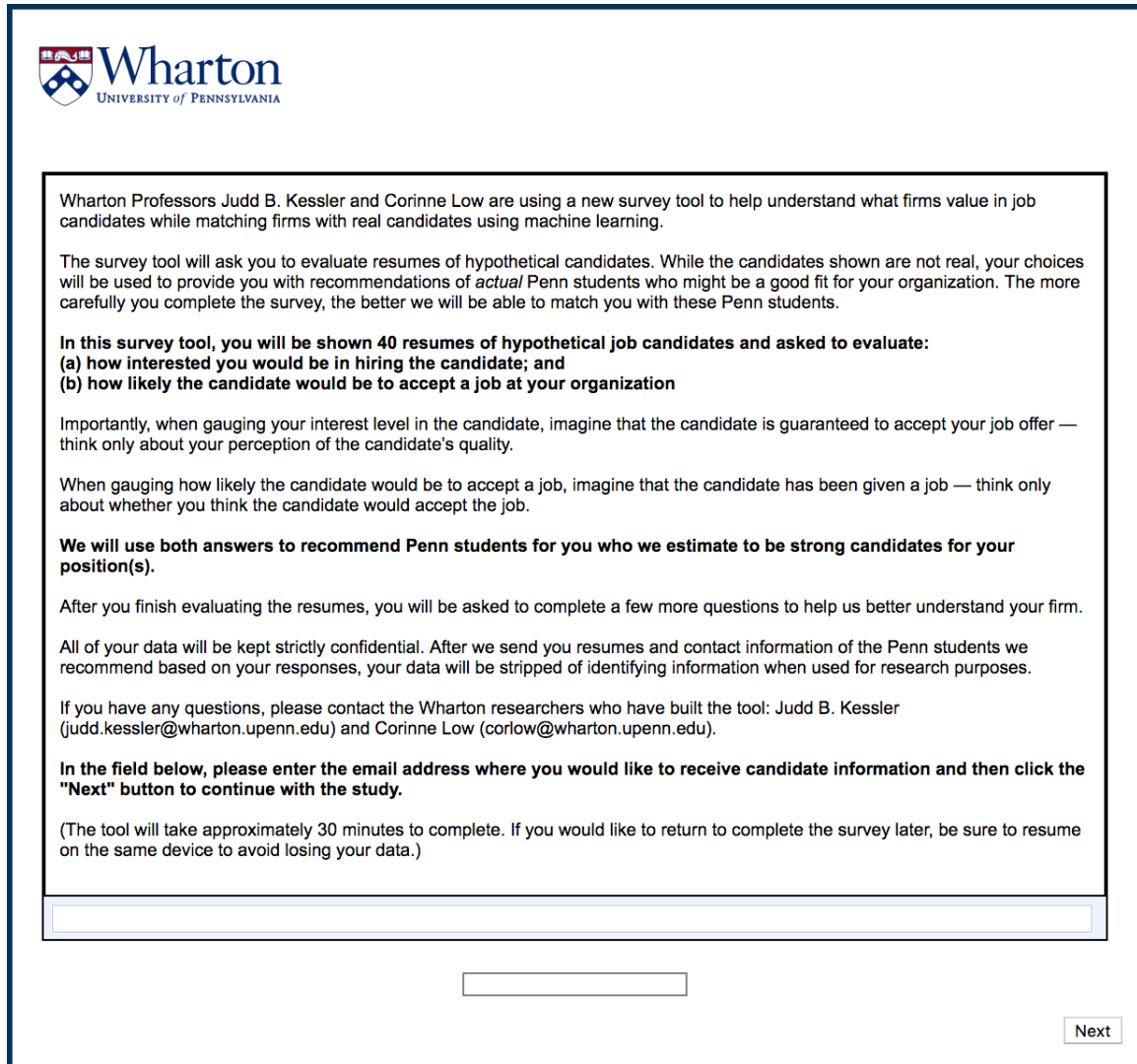
Announcements


An Opportunity To Reach More Employers

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

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

Figure 7: Survey Tool Instructions & Contact Information





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

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

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

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

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

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

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

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

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

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

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

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

Figure 8: Major Type Selection

Wharton
UNIVERSITY of PENNSYLVANIA

Please check the major that best reflects the background of the candidate(s) for which you are looking. This will allow us to show you resumes of candidates with relevant backgrounds.

☒ Business (Wharton), Social Sciences, and Humanities

☐ Science, Engineering, Computer Science, and Math

Please enter the name or title of the position you hope to fill.

Analyst

Next


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

Resume Structure

We designed our resumes to combine realism with the requirements of experimental identification. We designed 10 resume templates to use as the basis for the 40 resumes in the tool. Each template presented the same information, in the same order, but with variations in page layout and font. Figures 9 and 10 show sample resume templates. All resumes contained five sections, in the following order: Personal Information (including name and blurred contact information); Education (GPA, major, school within university); Work Experience; Leadership Experience; and Skills.¹ While the real student resumes we encountered varied in content, most contained some subset of these sections. Since our main objective with resume variation was to improve realism for each subject rather than to test the effectiveness of different resume formats, we did not vary the order of the resume formats across subjects. In other words, the first resume always had the same font and page layout for each subject, although the content of the resume differed each time. Given that formats are in a fixed order in the 40 hypothetical resumes, the order fixed effects included in most specifications control for any effect of resume format. Resumes templates were built in HTML/CSS for display in a web browser, and populated dynamically in Qualtrics using JavaScript. Randomization occurred for all 40 resumes simultaneously, without replacement, each time a subject completed the instructions and selected their major category of interest. Each resume layout was flexible enough to accommodate different numbers of bullet points for each experience, and different numbers of work experiences. If only one job was listed on the resume, for instance, the work experience section of the resume appeared shorter rather than introducing empty space.

¹These sections were not always labelled as such on candidate resumes. Personal Information was generally not identified, though each resume contained a name and blurred text in place of contact information. Skills were also marked as “Skills & Interests” and “Skill Summary”.

Figure 9: Sample Resume



Madison Stewart

blurred text • blurred text • blurred text • blurred text

EDUCATION

University of Pennsylvania, College of Arts and Sciences

Philadelphia, PA

BA in Economics

Expected May 2017

Cumulative GPA: 3.88/4.00

WORK EXPERIENCE

Goldman Sachs & Co

New York, NY

Summer Analyst, Corporate Derivatives

June - August 2016

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

SevaCall

Potomac, MD

Marketing Intern

June - August 2015

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

LEADERSHIP EXPERIENCE

Consult for America, Upenn

Philadelphia, PA

Sales and Operations Consultant

2014-2015

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

Penn Move Out

Philadelphia, PA

Vice President of Marketing

2014-2015

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

SKILLS

Microsoft Suite, Adobe Photoshop, Wordpress, Sketchup, iMovie

How interested would you be in hiring Madison Stewart?

Not interested	2	3	4	5	6	7	8	9	Very interested
1									10
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

Not likely	2	3	4	5	6	7	8	9	Very likely
1									10
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

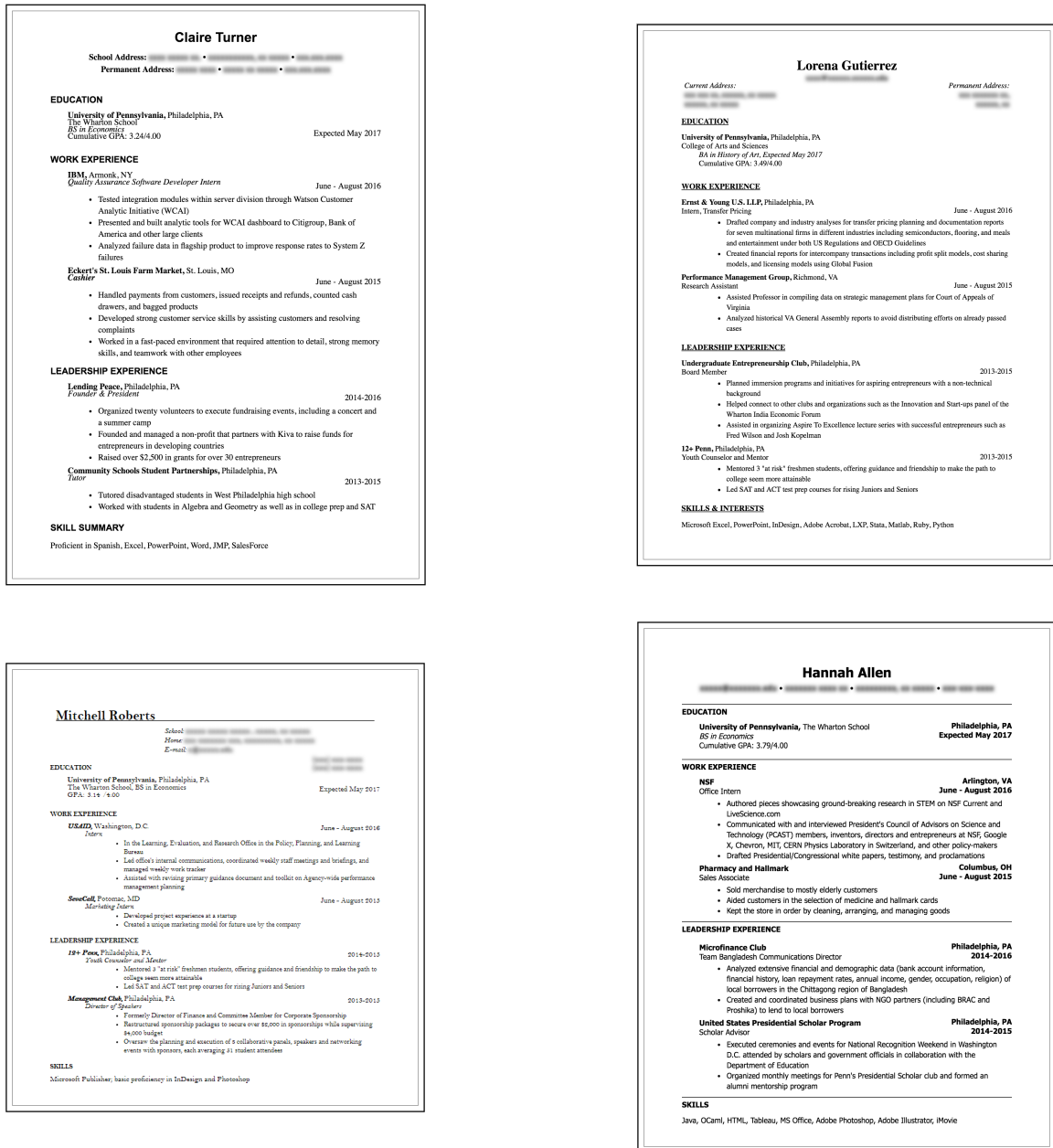
A sample resume rating page from the Incentivized Resume Rating tool. Each resume is dynamically generated when the subject begins the study. Each resume has five sections: Personal Information (including first and last name, and blurred text to represent contact information); Education Information (university, school within university, degree, major, GPA, and expected graduation date); Work Experience (one or two experiences with employer name, location, job title, date, and descriptive bullet points); Leadership Experience (two experiences with organization, location, position title, date, and descriptive bullet points); and Skills. Resume randomization described in detail in Section 1.2 and Appendix A.1.2. At the bottom of each resume, subjects must respond to two questions before proceeding: “How interested would you be in hiring [Name]?” and “How likely do you think [Name] would be to accept a job with your organization?”

Names

A hypothetical candidate name appears as the first element on each resume. Names were generated to be highly indicative of race and gender, following the approach of Fryer and Levitt (2004). As described in Section 1.2.3, first names were selected from a dataset of all births in the state of Massachusetts between 1989-1996 and in New York City between 1990-1996. These years reflect the approximate birth years of the job seekers in our study. We identified 100 first names with the most indicative race and gender for each of the following race-gender combinations: Asian Female, Asian Male, Black Female, Black Male, Hispanic Female, Hispanic Male, White Female, and White Male. We then eliminated names that were gender-ambiguous in the broad sample even if they might be unambiguous within an ethnic group. We also eliminated names strongly indicative of religion. We followed a similar process for last names, using name and ethnicity data from the 2000 Census. Finally, we paired first and last names together by race and selected 50 names for each race-gender combination for randomization. Names of hypothetical female candidates are shown in Table 15; names of hypothetical male candidates are shown in Table 16.

At the point of randomization, names were drawn without replacement according to a distribution of race and gender intended to reflect the US population (50% female, 50% male; 65.7% White, 16.8% Hispanic, 12.6% Black, 4.9% Asian). Gender and race were randomized independently. In other words, we selected either Table 15 or Table 16 with equal probability, then selected a column to draw from according to the race probabilities. Finally, names were selected uniformly and without replacement from the appropriate column of the table. We use the variation induced by these names for the analysis variables *Female*, *White*; *Male*, *Non-White*; *Female*, *Non-White*; and *Not a White Male*.

Figure 10: Four Sample Resumes



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

Table 15: Female Names Populating Resume Tool

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

Names of hypothetical female candidates. 50 names were selected to be highly indicative of each combination of race and gender. A name drawn from these lists was displayed at the top of each hypothetical resume, and in the questions used to evaluate the resumes. First and last names were linked every time they appeared. For details on the construction and randomization of names, see Section 1.2.3 and Appendix A.1.2.

Table 16: Male Names Populating Resume Tool

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

Names of hypothetical male candidates. 50 names were selected to be highly indicative of each combination of race and gender. A name drawn from these lists was displayed at the top of each hypothetical resume, and in the questions used to evaluate the resumes. First and last names were linked every time they appeared. For details on the construction and randomization of names, see Section 1.2.3 and Appendix A.1.2.

Education

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

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

Major Majors for the hypothetical resumes were selected according to a predefined probability distribution intended to balance the realism of the rating experience and our ability to detect and control for the effect of majors. Table 17 shows each major along with its school affiliation and classification as Humanities & Social Sciences or STEM, as well as the probability assigned to each. We use this variation as the variable *Major* and control for it with fixed effects in most regressions.

Components from Real Resumes

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

Table 17: Majors in Generated Penn Resumes

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

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

low (or high) quality are unlikely to produce useful variation for identification.

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

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

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

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

Before selecting the work experiences, we defined a *Top Internship* to be a substantive position at a prestigious employer. We chose this definition to both identify prestigious firms and distinguish between different types of jobs at those firms, such as a barista at a local Starbucks and a marketing intern at Starbucks headquarters. We identified a prestigious employer to be one of the 50 firms hiring the most Penn graduates in 2014 (as compiled by our Career Services partners). Since experiences at these firms were much more common among Humanities & Social Sciences majors, we supplemented this list with 39 additional firms hiring most often from Penn's School of Engineering and Applied Science. We extracted experiences at these firms from our full list of scraped experiences, and selected a total of 40 *Top Internship* experiences, with 20 coming from resumes of Humanities & Social Sciences majors and 20 from resumes of STEM majors. All of these *Top Internship* experiences had to be believably interchangeable within a major category. These internships included positions at Bain Capital, Goldman Sachs, Morgan Stanley, Northrop Grumman, Boeing Company, and Google (see Table 18 for a complete list). This variation identified the variable *Top Internship* in our analysis, which is measured relative to having a regular internship (since all resumes had some job in this position).

We selected 33 regular internships separately for the two major groups: 20 regular internships for randomization in the first work experience position, and 13 for the second position. Regular internships had few restrictions, but could not include employment at the firms who provided top internships, and could not include work-for-money job titles (described below and shown in Table 19). All jobs had to be believably interchangeable within major category. The regular internships in the second job position defined the variable *Second Internship*, and is measured relative to having no job in the second work experience position. Our dynamically generated resumes automatically adjusted in length when no second job was selected, in order to avoid a large gap on the page.

The remaining 13 jobs in the second work position (the summer after the sophomore year) were identified as *Work for Money*. We identified these positions in the real resume com-

Table 18: Top Internship Employers

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

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

ponents by compiling a list of job titles and phrases that we thought would be indicative of typical in this category, such as Cashier, Barista, and Waiter or Waitress (see Table 19 Columns 2–4 for the full list). We extracted components in our full list of scraped experiences that matched these search terms, and selected 13 that could be plausibly interchangeable across any major. During randomization, these 13 jobs were used for both Humanities & Social Sciences and STEM majors. The first column of Table 19 shows the job titles that appeared as *Work for Money* jobs in our hypothetical resumes. Columns 2–4 provide the list of job titles used for identifying work-for-money jobs in the scraped data, and for matching candidates to employer preferences.

Leadership Experience We defined leadership experiences to be those resume components that indicated membership or participation in a group, club, volunteer organization, fraternity/sorority, or student government. We selected leadership experiences from our full list of scraped experience components, requiring that the positions be clearly non-employment, include a position title, organization, and description, be plausibly interchangeable across gender, race, and major type. While many real resumes simply identified a position title and organization, we required that the components for our hypothetical resumes include a description of the activity for use as bullet points. We curated a list of 80 leadership experiences to use for both Humanities & Social Sciences and STEM resumes. Each resume included two randomly selected leadership experiences. We used the same leadership positions for both major types under the assumption that most extracurricular activities at Penn could plausibly include students from all majors; however, this required us to exclude the few leadership experiences that were too revealing of field of study (e.g., “American Institute of Chemical Engineers”).

Every leadership position was assigned to the location of Penn’s campus, Philadelphia, PA. This was done for consistency and believability, even if some of the leadership positions

Table 19: Work for Money Job Titles & Identifying Phrases

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

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

were held in other locations in the source resume. We randomly selected two ranges of years during a student’s career to assign to the experiences, and we ordered the experiences chronologically on the hypothetical resume based on the end year of the experience.

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

A.1.3. Matching Appendix

Students

For job-seeking study participants, the career services office sent an email to seniors offering “an opportunity to reach more employers” by participating in our pilot study, to be run in parallel with all existing recruiting activities. The full student recruitment email is reproduced in Appendix 6. 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.

Matches with Job Seekers

To match job seeking students with the recruiters in our study, we parsed the student resumes and coded their content into variables describing the candidate’s education, work experience, and leadership experience, using a combination of parsing software and manual transcription. We did not include any measure of ethnicity or gender in providing matches, nor did we take into account any employer’s revealed ethnic or gender preferences. The full list of variables used for matching is shown in Table 20.

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

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

Table 20: Candidate Matching Variables

Variable	Definition
GPA	Overall GPA, if available. If missing, assign lowest GPA observed in the match pool
Engineering	Indicator for Computer Sciences, Engineering, or Math majors (for STEM candidates)
Humanities	Indicator for Humanities majors (for Humanities & Social Sciences Candidates)
Job Count	Linear variable for 1, 2, or 3+ work experiences.
Top Firm	Resume has a work experience at one of the firms hiring the most Penn graduates
Major City	Resume has a work experience in New York, San Francisco, Chicago, or Boston
Work for Money	Resume has a job title including identifying phrase from Table 19
S&P500 or Fortune 500	Resume has an experience at an S&P 500 or Fortune 500 firm
Leader	Resume has a leadership position as Captain, President, Chair, Chairman, or Chairperson

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

their preferences and provided a single set of 10 resumes to the group.² This set of candidate resumes was the only incentive for participating in the study.

²In 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.

A.2. Results Appendix

In this section, we describe additional results and robustness checks to validate our main results. In Section A.2.1, we show additional analysis related to our main human capital results. In Section A.2.2, we verify our results after reweighting observations to the true distribution of GPAs in actual Penn student resumes. In Section A.2.3, we discuss preferences over the quality distribution. In Section A.2.4, we provide additional results on candidate demographics. Finally, in Section A.2.5, we discuss the relationship between *Likelihood of Acceptance* and *Hiring Interest*.

A.2.1. Additional Results on Human Capital

The human capital results in Section 1.3.2 rely on the independent randomization of work experiences and other resume elements. This randomization leads to some combinations of resume elements that are unlikely to arise in practice, despite drawing each variable from a realistic univariate distribution. If employers value a set of experiences that form a cohesive narrative, independent randomization could lead to strange relationships in our data. If employers value combinations of work experiences, narrative might be an omitted variable that could introduce bias (e.g., if our *Top Internships* are more likely to generate narratives than regular internships, we may misestimate their effect on hiring interest). In Table 21, we address this concern by showing that the cross-randomization of work experiences does not drive our results. To test this, we had three undergraduate research assistants at the University of Pennsylvania rate all possible combinations of work experiences that could have appeared on our hypothetical resumes.³ We used their responses to create a dummy—denoted *Narrative*—that is equal to 1 when a resume has a work experience in the summer

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

before junior year that is related to the work experience before senior year, and 0 otherwise. As a result of this process, we identified that 17.5% of the realized resumes in our study (i.e., those resumes actually shown to subjects) had a cohesive work experience narrative. None of these resumes included *Work for Money* because our RA raters did not see these jobs as contributing to a narrative. Appendix Table 21 runs the same regressions as Table 2 but additionally controls for *Narrative*. All results from Table 2 remain similar in size and statistical significance.

In Table 22, we estimate the value of degrees from more prestigious schools within Penn. We replace the major fixed effects of Table 2 with binary variables for *School of Engineering and Applied Science* and *Wharton*, as well as a binary control for whether the subject has chosen to review Humanities & Social Sciences or STEM resumes (coefficients not reported).⁴ We find that employers find degrees from these schools 0.4–0.5 Likert-scale points more desirable than degrees from Penn’s College of Arts and Sciences. As shown in Figure 11, and as discussed in Section 1.3.3, we also investigate the effect of having a degree from Wharton across the distribution of hiring interest.

A.2.2. Re-weighting by GPA

In generating hypothetical resumes, we randomly selected candidate GPAs from *Unif*[2.90, 4.00], rather than from the true distribution of GPAs among job seekers at Penn, which is shown in Figure 12.⁵ In this section, we demonstrate that this choice does not drive our results. In Tables 23, 24, and 25, we rerun the regressions of Tables 2, 3, and 4 weighted to reflect the naturally occurring distribution of GPA among our Penn senior candidate pool (i.e., the job seekers used for matching, see Appendix A.1.3). We do not include missing GPAs in the reweighting, though our results are robust to re-weighting with missing GPAs

⁴Major fixed effects are perfectly multicollinear with the variables for school, since no two schools grant the same degrees in the same major.

⁵We parameterized *GPA* to be drawn *Unif*[2.90, 4.00] to give us statistical power to test the importance of GPA on hiring interest, but this distribution is not exactly the distribution of GPA among Penn seniors engaging in on campus recruiting.

Table 21: Work Experience Narrative

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.128 (0.145)	2.194 (0.150)	2.200 (0.129)	1 (.)	0.892 (0.0613)
Top Internship	0.896 (0.0945)	0.892 (0.0989)	0.888 (0.0806)	0.404 (0.0428)	0.375 (0.0397)
Second Internship	0.349 (0.142)	0.364 (0.150)	0.319 (0.122)	0.145 (0.0560)	0.156 (0.0593)
Work for Money	0.115 (0.110)	0.160 (0.114)	0.157 (0.0914)	0.0714 (0.0416)	0.0518 (0.0468)
Technical Skills	0.0424 (0.104)	0.0490 (0.108)	-0.0759 (0.0898)	-0.0345 (0.0409)	0.0102 (0.0442)
Female, White	-0.149 (0.114)	-0.213 (0.118)	-0.159 (0.0963)	-0.0725 (0.0441)	-0.0597 (0.0478)
Male, Non-White	-0.174 (0.137)	-0.181 (0.142)	-0.175 (0.115)	-0.0794 (0.0524)	-0.0761 (0.0569)
Female, Non-White	-0.0108 (0.137)	-0.0236 (0.144)	0.0261 (0.120)	0.0119 (0.0545)	-0.0150 (0.0578)
Narrative	0.214 (0.165)	0.237 (0.175)	0.278 (0.144)	0.126 (0.0656)	0.0930 (0.0678)
Observations	2880	2880	2880	2880	2880
R^2	0.130	0.181	0.484		
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

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

Table shows OLS and ordered probit regressions of hiring interest from Equation (1.1), with an additional control for *Narrative*. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. *Narrative* is a characteristic of resumes, defined as work experiences that are related in some way. Fixed effects for *major*, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. The *p*-value of a test of joint significance of major fixed effects is indicated (*F*-test for OLS regressions, likelihood ratio test for ordered probit regressions).

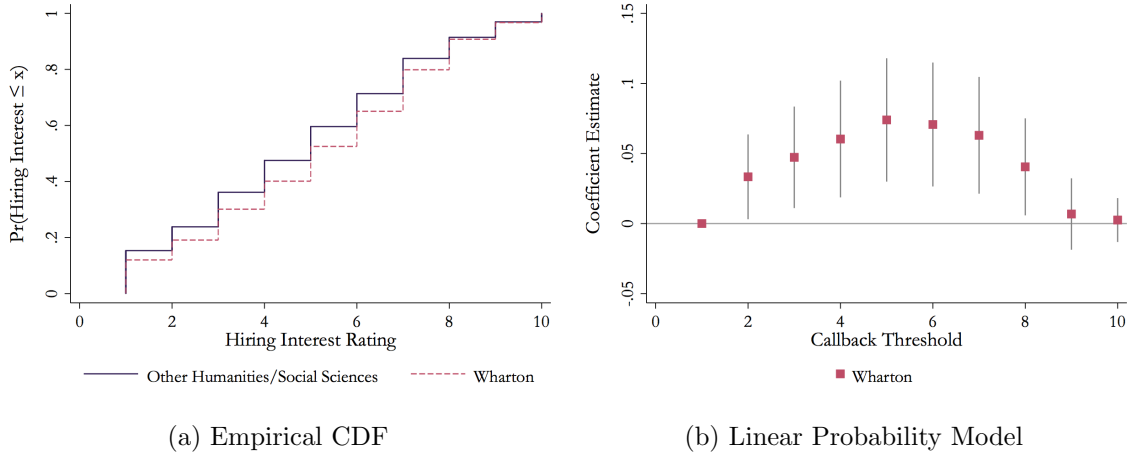
Table 22: Prestigious Schools

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.129 (0.145)	2.187 (0.149)	2.192 (0.128)	1 (.)	0.887 (0.0624)
Top Internship	0.908 (0.0943)	0.913 (0.0984)	0.905 (0.0804)	0.413 (0.0431)	0.378 (0.0395)
Second Internship	0.443 (0.112)	0.465 (0.118)	0.451 (0.0945)	0.206 (0.0446)	0.195 (0.0466)
Work for Money	0.108 (0.110)	0.141 (0.113)	0.143 (0.0918)	0.0654 (0.0419)	0.0493 (0.0461)
Technical Skills	0.0378 (0.103)	0.0404 (0.107)	-0.0820 (0.0901)	-0.0374 (0.0411)	0.00871 (0.0430)
Female, White	-0.146 (0.113)	-0.207 (0.118)	-0.160 (0.0962)	-0.0730 (0.0442)	-0.0573 (0.0473)
Male, Non-White	-0.189 (0.137)	-0.196 (0.142)	-0.181 (0.115)	-0.0828 (0.0527)	-0.0801 (0.0573)
Female, Non-White	-0.0000775 (0.137)	-0.0107 (0.144)	0.0371 (0.120)	0.0169 (0.0549)	-0.00885 (0.0570)
School of Engineering	0.497 (0.199)	0.441 (0.206)	0.403 (0.164)	0.184 (0.0758)	0.239 (0.0863)
Wharton	0.459 (0.110)	0.502 (0.115)	0.417 (0.0934)	0.190 (0.0435)	0.184 (0.0455)
Observations	2880	2880	2880	2880	2880
R^2	0.115	0.168	0.472		
Major FEs	No	No	No	Yes	No
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

Ordered probit cutpoints: 2.48, 2.84, 3.20, 3.49, 3.81, 4.15, 4.60, 5.06, and 5.57.

Table shows OLS and ordered probit regressions of hiring interest from Equation (1.1), with effects for school, and a control for whether the employer selected to view Humanities & Social Sciences resumes or STEM resumes (coefficient not displayed). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. *School of Engineering* indicates a resume with a degree from Penn's School of Engineering and Applied Sciences; *Wharton* indicates a resume with a degree from the Wharton School. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. R^2 is indicated for each OLS regression.

Figure 11: Wharton



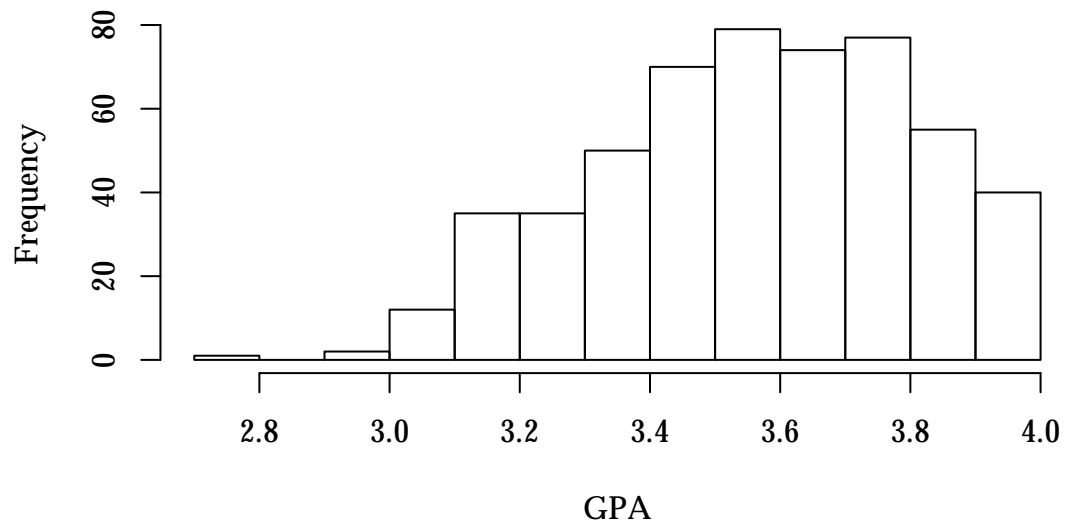
Empirical CDF of *Hiring Interest* (Panel 11a) and difference in counterfactual callback rates (Panel 11b) for *Wharton* and *Other Humanities & Social Sciences*. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

treated as low GPAs.⁶ These regressions confirm the results of Tables 2, 3, and 4 in direction and statistical significance.

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

⁶Some students may strategically omit low GPAs from their resumes, and some resume formats were difficult for our resume parser to scrape.

Figure 12: Distribution of GPA Among Scraped Resumes



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

Table 23: Human Capital Experience—Weighted by GPA

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.274 (0.175)	2.339 (0.168)	2.320 (0.146)	1 (.)	0.963 (0.0785)
Top Internship	0.831 (0.110)	0.832 (0.109)	0.862 (0.0882)	0.372 (0.0428)	0.353 (0.0474)
Second Internship	0.488 (0.129)	0.482 (0.130)	0.513 (0.105)	0.221 (0.0475)	0.216 (0.0545)
Work for Money	0.178 (0.129)	0.193 (0.125)	0.199 (0.100)	0.0856 (0.0436)	0.0753 (0.0556)
Technical Skills	0.0768 (0.118)	0.0388 (0.119)	-0.106 (0.102)	-0.0455 (0.0439)	0.0224 (0.0507)
Female, White	-0.0572 (0.134)	-0.0991 (0.130)	-0.0382 (0.105)	-0.0165 (0.0453)	-0.0214 (0.0574)
Male, Non-White	-0.239 (0.154)	-0.181 (0.154)	-0.111 (0.123)	-0.0480 (0.0530)	-0.0975 (0.0658)
Female, Non-White	-0.0199 (0.166)	-0.0316 (0.162)	0.0398 (0.134)	0.0171 (0.0577)	-0.0175 (0.0710)
Observations	2880	2880	2880	2880	2880
R^2	0.146	0.224	0.505		
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

Ordered probit cutpoints: 2.30, 2.71, 3.04, 3.34, 3.66, 3.99, 4.49, 4.95, and 5.46.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1.1), weighted by the distribution of GPA in resumes in the candidate matching pool. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for *major*, *leadership experience*, *resume order*, and *subject* included in some specifications as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. The p -value of a test of joint significance of *major* fixed effects is indicated for each model (F -test for OLS regressions, χ^2 test for ordered probit regression).

Table 24: Human Capital Experience by Major Type—Weighted by GPA

	Dependent Variable: Hiring Interest									
	Humanities & Social Sciences					STEM				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.365 (0.212)	2.452 (0.198)	2.476 (0.172)	1 (.)	1.008 (0.0964)	2.028 (0.306)	2.187 (0.325)	2.000 (0.266)	1 (.)	0.848 (0.133)
Top Internship	0.973 (0.127)	0.941 (0.125)	0.982 (0.102)	0.397 (0.0486)	0.412 (0.0557)	0.448 (0.218)	0.526 (0.222)	0.581 (0.182)	0.291 (0.101)	0.204 (0.0927)
Second Internship	0.476 (0.153)	0.384 (0.155)	0.494 (0.125)	0.199 (0.0520)	0.217 (0.0645)	0.529 (0.235)	0.496 (0.252)	0.383 (0.199)	0.192 (0.103)	0.223 (0.102)
Work for Money	0.0914 (0.152)	0.0349 (0.145)	0.0861 (0.118)	0.0348 (0.0477)	0.0366 (0.0653)	0.387 (0.247)	0.459 (0.270)	0.517 (0.201)	0.259 (0.106)	0.182 (0.106)
Technical Skills	0.0893 (0.142)	0.0263 (0.142)	-0.146 (0.120)	-0.0591 (0.0484)	0.0258 (0.0609)	0.0111 (0.217)	-0.0591 (0.240)	-0.0928 (0.193)	-0.0464 (0.0965)	0.00518 (0.0932)
Female, White	0.110 (0.159)	0.0360 (0.153)	0.110 (0.125)	0.0445 (0.0506)	0.0475 (0.0683)	-0.460 (0.251)	-0.637 (0.253)	-0.658 (0.206)	-0.329 (0.110)	-0.183 (0.107)
Male, Non-White	-0.0332 (0.181)	0.0366 (0.183)	0.0377 (0.147)	0.0152 (0.0593)	-0.00558 (0.0767)	-0.799 (0.295)	-0.704 (0.322)	-0.590 (0.260)	-0.295 (0.129)	-0.352 (0.130)
Female, Non-White	0.0356 (0.189)	0.0238 (0.186)	0.0785 (0.154)	0.0317 (0.0623)	0.00129 (0.0819)	-0.180 (0.332)	0.0136 (0.318)	0.0391 (0.264)	0.0196 (0.132)	-0.0743 (0.140)
Observations	2040	2040	2040	2040	2040	840	840	840	840	840
R^2	0.141	0.242	0.522			0.150	0.408	0.644		
<i>p-value for test of joint significance of Majors</i>	0.105	0.152	0.022	0.022	0.138	< 0.001	0.003	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No	No	No	Yes	Yes	No

Ordered probit cutpoints (Column 5): 2.54, 2.89, 3.23, 3.54, 3.86, 4.20, 4.71, 5.18, 5.70.

Ordered probit cutpoints (Column 10): 1.78, 2.31, 2.62, 2.89, 3.20, 3.51, 3.98, 4.44, 4.92.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1). *GPA*: *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS columns present the results of Column 3 and Column 8 divided by the Column 3 and Column 8 coefficient on GPA, with standard errors calculated by delta method. The p -values of tests of joint significance of major fixed effects and demographic variables are indicated (F -test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

Table 25: Likelihood of Acceptance—Weighted by GPA

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.545 (0.174)	0.552 (0.168)	0.663 (0.132)	0.246 (0.0738)
Top Internship	0.725 (0.111)	0.709 (0.108)	0.694 (0.0833)	0.299 (0.0472)
Second Internship	0.524 (0.132)	0.456 (0.133)	0.432 (0.101)	0.220 (0.0556)
Work for Money	0.205 (0.128)	0.150 (0.125)	0.185 (0.0977)	0.0872 (0.0544)
Technical Skills	0.0409 (0.120)	-0.0390 (0.120)	-0.114 (0.0972)	0.0122 (0.0504)
Female, White	-0.209 (0.135)	-0.276 (0.133)	-0.224 (0.103)	-0.0830 (0.0571)
Male, Non-White	-0.248 (0.157)	-0.273 (0.155)	-0.114 (0.120)	-0.113 (0.0660)
Female, Non-White	-0.174 (0.160)	-0.224 (0.156)	-0.155 (0.124)	-0.0856 (0.0684)
Observations	2880	2880	2880	2880
R^2	0.077	0.162	0.509	
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: -0.09, 0.29, 0.64, 0.90, 1.26, 1.67, 2.13, 2.65, and 3.02.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1), weighted by the distribution of GPA in resumes in our candidate matching pool. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p -value of a test of joint significance of major fixed effects is indicated (F -test for OLS regressions, χ^2 test for ordered probit regression).

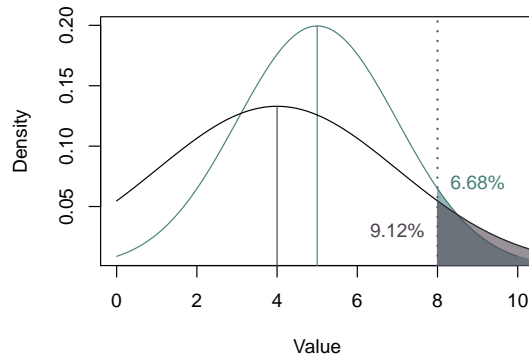
A.2.3. Distributional Appendix

As discussed in Section 1.3.3, average preferences for candidate characteristics might differ from the preferences observed in the tails. The stylized example in Figure 13 shows this concern graphically. Imagine the light (green) distribution shows the expected productivity—based on the content of their resumes—of undergraduate research assistants (RAs) majoring in Economics at the University of Pennsylvania and the dark (gray) distribution shows the expected productivity of undergraduate RAs enrolled at the Wharton School. In this example, the mean Wharton student would make a less productive RA, reflecting a lack of interest in academic research relative to business on average; however, the tails of the Wharton distribution are fatter, reflecting the fact that admission into Wharton is more selective, so a Wharton student who has evidence of research interest on her resume is expected to be better than an Economics student with an otherwise identical resume. Looking across the panels in Figure 13, 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.⁷

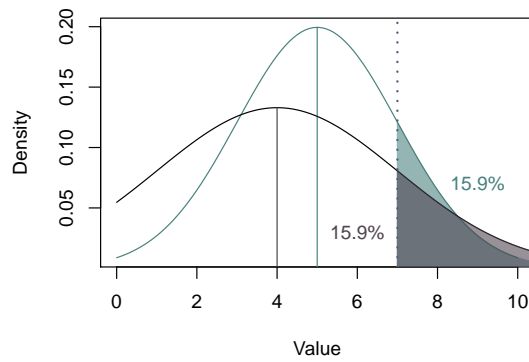
A researcher might particularly care about how employers respond to candidate characteristics around the empirically observed threshold (e.g., the researcher may be particularly interested in how employers respond to candidates in a particular market, with a particular level of selectivity, at a particular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

⁷This 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.

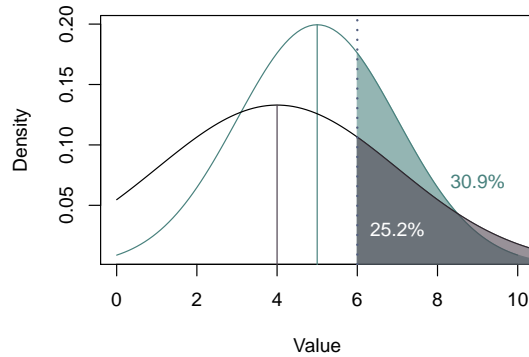
Figure 13: Callback Thresholds Example



(a) High Threshold



(b) Medium Threshold



(c) Low Threshold

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

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.

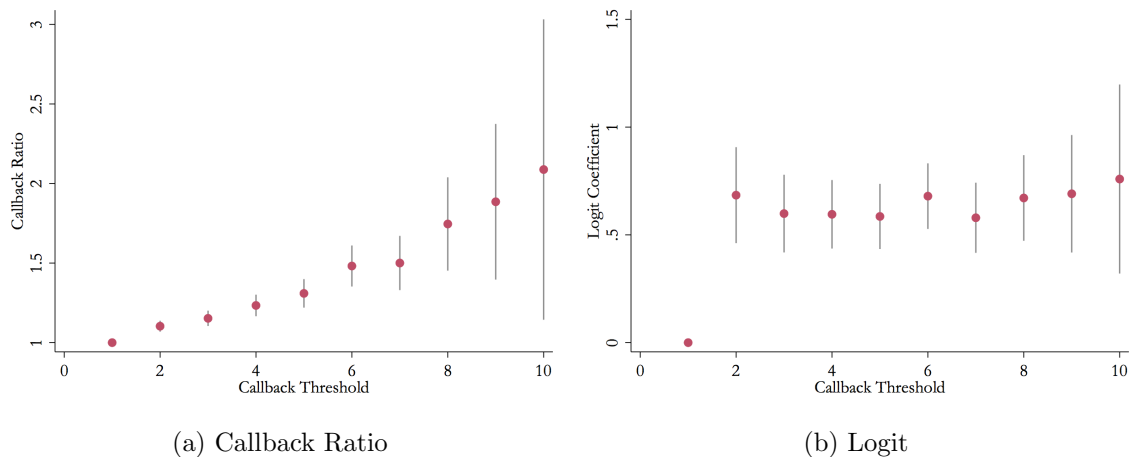
Comparing Results Across the Distribution

Resume audit studies often report differences in callback rates between two types of job candidates, either in a t -test or in a regression. However, as the overall callback rate becomes very large (i.e., almost all candidates get called back) or very small (i.e., few candidates get called back), the differences in callback rates tend toward zero. This is because, as discussed in footnote 22, the maximum possible difference in callback rates is capped by the overall callback rate.

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

In Figures 14 and 15, we look at how two different ways of measuring callback differences perform across the distribution compared to the linear probability model. The lefthand side of each figure shows the ratio of the callback rates, another common way of reporting

Figure 14: Alternative Specifications: Top Internship



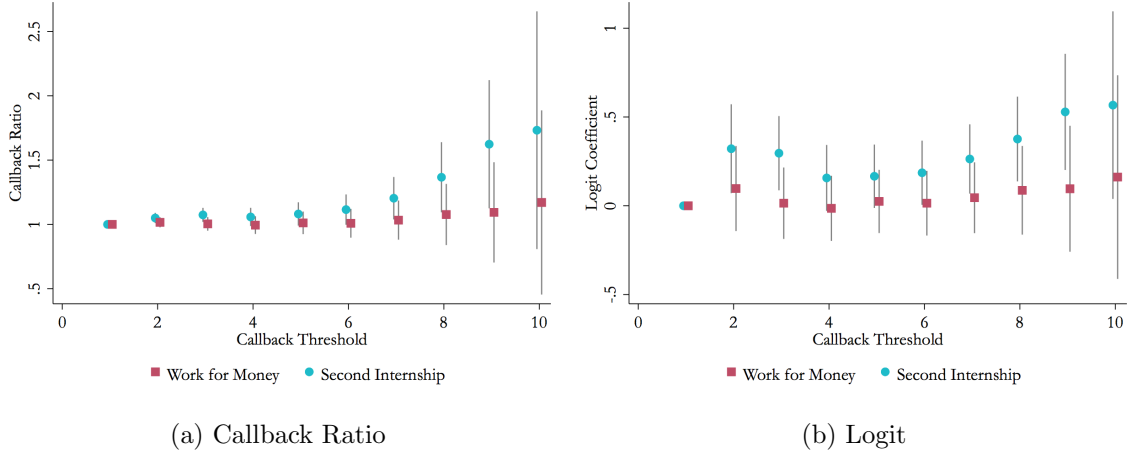
Counterfactual callback ratios (Panel 14a) and counterfactual logit coefficients (Panel 14b) for *Top Internship*. Counterfactual callback is an indicator for each value of *Hiring Interest* equal to 1 if *Hiring Interest* is greater than or equal to the value, and 0 otherwise. Callback ratio is defined as the counterfactual callback rate for candidates with the characteristic divided by the counterfactual callback rate for candidates without. 95% confidence intervals are calculated from a linear probability model using the delta method. Logit coefficients are estimated from a logit regression with counterfactual callback as the dependent variable.

resume audit study results. For the positive effects in our study, this odds ratio tends to be larger at the upper tail, where a small difference in callbacks can result in a large response in the ratio. On the righthand side of each figure, we show effects estimated from a logit specification. We find that in our data, the effects estimated in logistic regression tend to be flatter across the quality distribution.

A.2.4. Candidate Demographics Appendix

In this section, we provide additional analyses for our main results on candidate demographics. In A.2.4, we analyze our findings by the demographics of employers evaluating resumes. In A.2.4 we describe a test for implicit bias. In A.2.4, we discuss differential returns to quality by demographic group.

Figure 15: Alternative Specifications: Second Job Type



Counterfactual callback ratios (Panel 15a) and counterfactual logit coefficients (Panel 15b) for *Work for Money* and *Second Internship*. Counterfactual callback is an indicator for each value of *Hiring Interest* equal to 1 if *Hiring Interest* is greater than or equal to the value, and 0 otherwise. Callback ratio is defined as the counterfactual callback rate for candidates with the characteristic divided by the counterfactual callback rate for candidates without. 95% confidence intervals are calculated from a linear probability model using the delta method. Logit coefficients are estimated from a logit regression with counterfactual callback as the dependent variable.

Rater Demographics

IRR allows us to collect information about the specific individuals rating resumes at the hiring firm. In Table 26 we explore our main results by rater gender and race. White and female raters appear more likely to discriminate against male, non-white candidates than non-white or female raters.

Test for Implicit Bias

We leverage a feature of implicit bias—that it is more likely to arise when decision makers are fatigued (Wigboldus et al., 2004; Govorun and Payne, 2006; Sherman et al., 2004)—to test whether our data are consistent with implicit bias. Appendix Table 27 investigates how employers respond to resumes in the first and second half of the study and to resumes before

Table 26: Hiring Interest by Rater Demographics

	Dependent Variable: Hire Rating				
	All	Rater Gender		Rater Race	
		Female Raters	Male Raters	Non-White Raters	White Raters
GPA	2.196 (0.129)	2.357 (0.170)	2.092 (0.212)	2.187 (0.378)	2.131 (0.146)
Top Internship	0.897 (0.0806)	0.726 (0.105)	1.139 (0.140)	1.404 (0.234)	0.766 (0.0914)
Second Internship	0.466 (0.0947)	0.621 (0.126)	0.195 (0.154)	0.636 (0.273)	0.459 (0.107)
Work for Money	0.154 (0.0914)	0.303 (0.120)	-0.0820 (0.156)	-0.124 (0.255)	0.192 (0.104)
Technical Skills	-0.0711 (0.0899)	-0.0794 (0.122)	-0.0202 (0.151)	-0.123 (0.231)	-0.0164 (0.104)
Female, White	-0.161 (0.0963)	-0.202 (0.128)	-0.216 (0.165)	0.00413 (0.265)	-0.209 (0.109)
Male, Non-White	-0.169 (0.115)	-0.311 (0.149)	-0.105 (0.200)	0.119 (0.285)	-0.241 (0.132)
Female, Non-White	0.0281 (0.120)	0.00110 (0.159)	-0.0648 (0.202)	-0.124 (0.325)	0.0968 (0.137)
Observations	2880	1720	1160	600	2280
R^2	0.483	0.525	0.556	0.588	0.503
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	Yes	Yes	Yes	Yes	Yes
Order FEs	Yes	Yes	Yes	Yes	Yes
Subject FEs	Yes	Yes	Yes	Yes	Yes

OLS regressions of *Hiring Interest* on candidate characteristics by rater gender and race. Sample includes 29 male and 42 female subjects; 57 White and 15 non-White subjects. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. R^2 is indicated for each OLS regression. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated.

and after the period breaks—after every 10 resumes—that we built into the survey tool.⁸ The first and second columns show that subjects spend less time evaluating each resume in the second half of the study and in the latter half of each block of 10 resumes, suggesting evidence of fatigue. The third column reports a statistically significant interaction on *Latter Half of Block* \times *Not a White Male* of -0.385 Likert-scale points, equivalent to about 0.18 GPA points, suggesting more discrimination against candidates who are not white males in the latter half of each block of 10 resumes. The fourth column reports, however, that the bias in the second half of the study is not statistically significantly larger than the bias in the first half. These results provide suggestive, though not conclusive, evidence that the discrimination we detect may indeed be driven by implicit bias.

⁸As described in Section 1.2, after every 10 resumes an employer completed, the employer was shown a simple webpage with an affirmation that gave them a short break (e.g., after the first 10 resumes it read: “You have rated 10 of 40 resumes. Keep up the good work!”). Research suggests that such “micro breaks” can have relatively large effects on focus and attention (Rzeszutarski et al., 2013), and so we compare bias in the early half and latter half of each block of 10 resumes under the assumption that employers might be more fatigued in the latter half of each block of 10 resumes.

Table 27: Implicit Bias

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

Regressions of *Response Time* and *Hiring Interest* on resume characteristics and resume order variables. The first and second columns show *Response Time* regressions; the third and fourth columns show *Hiring Interest* regressions. *Response Time* is defined as the number of seconds before page submission, Winsorized at the 95th percentile (77.9 seconds). Mean of *Response Time*: 23.6 seconds. *GPA*, *Top Internship*, *Second Internship*, *Work for Money*, *Technical Skills*, and *Not a White Male* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. *Latter Half of Block* is an indicator variable for resumes shown among the last five resumes within a 10-resume block. *Second Half of Study* is an indicator variable for resumes shown among the last 20 resumes viewed by a subject. Fixed effects for subjects, majors, and leadership experience included in all specifications. R^2 is indicated for each OLS regression. The p -value of an F -test of joint significance of major fixed effects is indicated for all models.

Interaction of Demographics with Quality

Table 28 shows that white males gain more from having a *Top Internship* than candidates who are not white males. The largest of these coefficients, that for non-white females, nearly halves the benefit of having a prestigious internship. We speculate that this may be due to firms believing that prestigious internships are a less valuable signal of quality if the previous employer may have selected the candidate due to positive tastes for diversity. Figure 16 looks at the relationship between *Top Internship* and being *Not a White Male* throughout the quality distribution. We find that when a candidate is of sufficiently high quality, a *Top Internship* is equally valuable for white male candidates and those who are not white males. This may suggest that other signals of quality may inoculate candidates from the assumption that an impressive work history is the result of diversity initiatives.

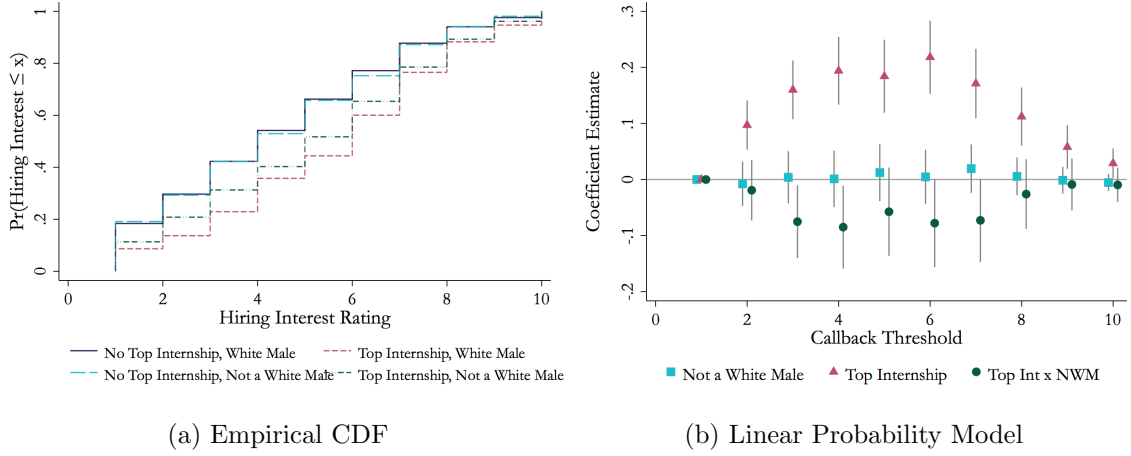
Table 28: Return to Top Internship by Demographic Group

	Dependent Variable: Hiring Interest				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	2.119 (0.145)	2.184 (0.150)	2.191 (0.129)	1 (.)	0.889 (0.0613)
Top Internship	1.147 (0.168)	1.160 (0.175)	1.155 (0.145)	0.527 (0.0736)	0.471 (0.0704)
Second Internship	0.468 (0.112)	0.495 (0.118)	0.470 (0.0944)	0.214 (0.0446)	0.208 (0.0469)
Work for Money	0.109 (0.110)	0.151 (0.113)	0.148 (0.0913)	0.0675 (0.0417)	0.0496 (0.0469)
Technical Skills	0.0494 (0.104)	0.0576 (0.108)	-0.0670 (0.0899)	-0.0306 (0.0411)	0.0132 (0.0442)
Female, White	0.0327 (0.146)	-0.0188 (0.152)	0.0225 (0.121)	0.0103 (0.0554)	0.0118 (0.0617)
Male, Non-White	-0.0604 (0.175)	-0.0488 (0.184)	-0.0553 (0.145)	-0.0253 (0.0659)	-0.0287 (0.0741)
Female, Non-White	0.0806 (0.182)	0.0685 (0.191)	0.159 (0.156)	0.0727 (0.0717)	0.0104 (0.0768)
Top Internship × Female, White	-0.464 (0.234)	-0.492 (0.243)	-0.459 (0.199)	-0.209 (0.0920)	-0.181 (0.0974)
Top Internship × Male, Non-White	-0.280 (0.279)	-0.316 (0.288)	-0.276 (0.233)	-0.126 (0.107)	-0.116 (0.116)
Top Internship × Female, Non-White	-0.229 (0.273)	-0.224 (0.286)	-0.316 (0.240)	-0.144 (0.110)	-0.0653 (0.116)
Observations	2880	2880	2880	2880	2880
R^2	0.130	0.182	0.484		
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No

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

Table shows OLS and ordered probit regressions of hiring interest from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. GPA-Scaled OLS presents the results of Column 3 divided by the Column 3 coefficient on GPA, with standard errors calculated by delta method. The p -value of a test of joint significance of major fixed effects is indicated (F -test for OLS, likelihood ratio test for ordered probit).

Figure 16: Top Internship \times Not a White Male



Empirical CDF of *Hiring Interest* (Panel 16a) and difference in counterfactual callback rates (Panel 16b) for *Top Internship*, *Not a White Male*, and *Top Internship \times Not a White Male*. Empirical CDFs show the share of hypothetical candidate resumes with each characteristic with a *Hiring Interest* rating less than or equal to each value. The counterfactual callback plot shows the difference between groups in the share of candidates at or above the threshold—that is, the share of candidates who would be called back in a resume audit study if the callback threshold were set to any given value. 95% confidence intervals are calculated from a linear probability model with an indicator for being at or above a threshold as the dependent variable.

A.2.5. Relationship Between Likelihood of Acceptance and Human Capital

In evaluating candidates' likelihood of accepting a job offer, the firms in our sample exhibit a potentially surprising belief that candidates with more human capital—indicated by higher GPA, more work experience, and a more prestigious internship—are more likely to accept jobs than candidates with less human capital. This correlation could arise in several ways. First, it is possible that the hiring interest question—which always comes first—creates anchoring for the second question that is unrelated to true beliefs. Second, it is possible that likelihood of acceptance is based on both horizontal fit and vertical quality. Horizontal fit raises both hiring interest and likelihood of acceptance, which would lead to a positive correlation between responses; vertical quality, on the other hand, would be expected to increase hiring interest and decrease likelihood of acceptance, since as it increases hiring interest it also makes workers more desirable for other firms.⁹

⁹It is also possible that respondents deliberately overstate candidates' likelihood of acceptance in order to be sent the best quality candidates. However, firms who are willing to do this likely have a low cost of

If the correlation between *Hiring Interest* and *Likelihood of Acceptance* is driven mostly by horizontal fit, it is important to test whether *Likelihood of Acceptance* is simply a noisy measure of *Hiring Interest*, or whether *Likelihood of Acceptance* contains additional, valuable information. This will help us confirm, for example, that the gender bias we find in *Likelihood of Acceptance* is indeed its own result, rather than a result of bias in *Hiring Interest*. Approaching this is econometrically tricky, since *Hiring Interest* and *Likelihood of Acceptance* are both simultaneous products of the rater's assessment of the randomized resume components. We considered multiple approaches, such as subtracting hiring interest from likelihood of acceptance to capture the difference, regressing likelihood of acceptance on hiring interest and taking residuals, and including controls for hiring interest. All yield similar results, and so we use the latter approach, as it is the most transparent. Despite its econometric issues, we believe this is nonetheless a helpful exercise that can be thought of as akin to a mediation analysis. We want to see if all of the effect on *Likelihood of Acceptance* is mediated through *Hiring Interest*, or if there is independent variation in *Likelihood of Acceptance*.

The first two columns of Table 29 include a linear control for *Hiring Interest*, while Columns 3 and 4 include fixed effect controls for each level of the *Hiring Interest* rating, examining *Likelihood of Acceptance* within each hiring interest band. We find that after controlling for *Hiring interest*, the relationship between GPA and *Likelihood of Acceptance* becomes negative and statistically significant under all specifications. This indicates that the part of *Likelihood of Acceptance* that is uncorrelated with *Hiring Interest* is indeed negatively correlated with one measure of vertical quality. We also find that the coefficients on *Top Internship* and *Second Internship* become statistically indistinguishable from zero.

Under all specifications, the coefficients on *Female*, *White* and *Female, Non-White* remain negative and significant, indicating that employers believe women are less likely to accept

interviewing candidates with a lower probability of acceptance. This is in line with the data, where the firms who consistently rate people a 10 on *Likelihood of Acceptance* are among the most prestigious firms in our sample.

jobs if offered, even controlling for the firm's interest in the candidate.

Thus, we conclude that *Likelihood of Acceptance* does provide some additional information above and beyond *Hiring Interest*. We hope future research will tackle the question of how to measure beliefs about *Likelihood of Acceptance* accurately, how to disentangle them from *Hiring Interest*, and exactly what role they play in hiring decisions.

Table 29: Likelihood of Acceptance with Hiring Interest Controls

	Dependent Variable: Likelihood of Acceptance			
	OLS	Ordered Probit	OLS	Ordered Probit
GPA	-0.812 (0.0820)	-0.638 (0.0641)	-0.823 (0.0815)	-0.660 (0.0646)
Top Internship	0.0328 (0.0535)	0.000290 (0.0406)	0.0313 (0.0534)	0.000698 (0.0408)
Second Internship	0.0656 (0.0634)	0.0511 (0.0477)	0.0680 (0.0634)	0.0491 (0.0480)
Work for Money	0.0951 (0.0611)	0.0824 (0.0475)	0.0954 (0.0610)	0.0868 (0.0477)
Technical Skills	-0.0527 (0.0596)	-0.0572 (0.0449)	-0.0608 (0.0594)	-0.0661 (0.0452)
Female, White	-0.145 (0.0638)	-0.0781 (0.0484)	-0.147 (0.0638)	-0.0820 (0.0486)
Male, Non-White	0.00212 (0.0744)	-0.0162 (0.0577)	0.000650 (0.0744)	-0.00832 (0.0580)
Female, Non-White	-0.182 (0.0741)	-0.154 (0.0587)	-0.185 (0.0737)	-0.159 (0.0591)
Hiring Interest	0.704 (0.0144)	0.478 (0.0104)	FEs	FEs
Observations	2880	2880	2880	2880
R^2	0.766		0.768	
<i>p-value for test of joint significance of Majors</i>	0.025	< 0.001	0.031	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	Yes	No	Yes	No
Order FEs	Yes	No	Yes	No
Subject FEs	Yes	No	Yes	No

Cutpoints (Col 2): -1.82, -1.18, -0.55, -0.11, 0.49, 1.07, 1.71, 2.39, 2.81.

Cutpoints (Col 4): -2.00, -1.26, -0.58, -0.14, 0.45, 1.01, 1.62, 2.28, 2.69.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1), with additional controls for *Hiring Interest*. Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for *major*, *leadership experience*, *resume order*, and *subject* included in some specifications as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of *major* fixed effects and demographic variables are indicated (F -test for OLS, likelihood ratio test for ordered probit).

A.3. Pitt Appendix

In our replication study at the University of Pittsburgh, we followed a similar approach to that described for our experimental waves at Penn in Section A.1.2. The tool structure was essentially the same as at Penn, with references to Penn replaced with Pitt in the instructions, and the reference to Wharton removed from the major selection page. Resume structure was identical to that described in Sections A.1.2 and A.1.2. Names were randomized in the same manner as described in Section A.1.2. The education section of each resume at Pitt followed the same structure as that described in Section A.1.2, but had a degree from the University of Pittsburgh, with majors, schools, and degrees randomly drawn from a set of Pitt’s offerings. In selecting majors for our Pitt replication, we attempted to match the Penn major distribution as closely as possible, but some majors were not offered at both schools. When necessary, we selected a similar major instead. The majors, schools, classifications, and probabilities for Pitt are shown in Table 30.

We used a single pool of Pitt resumes for both the hypothetical resume elements and for a candidate pool for Pitt employers, saving significant effort on scraping and parsing. These components were compiled and randomized in much the same way as at Penn, as described in Section A.1.2. For *Top Internship* at Pitt, we collected work experiences from Pitt resumes at one of Pitt’s most frequent employers, or at one of the employers used to define *Top Internship* at Penn. Similarly, Pitt *Work for Money* was identified from the same list of identifying phrases shown in Table 19. *Technical Skills* were randomized in the same way as at Penn, described in A.1.2.

Table 30: Majors in Generated Pitt Resumes

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

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

Table 31: Effects by Major Type at Pitt

	Dependent Variable: Hiring Interest									
	Humanities & Social Sciences					STEM				
	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit	OLS	OLS	OLS	GPA-Scaled OLS	Ordered Probit
GPA	0.249 (0.189)	0.294 (0.203)	0.249 (0.150)	1 (.)	0.0969 (0.0731)	0.518 (0.245)	0.445 (0.274)	0.340 (0.187)	1 (.)	0.167 (0.0925)
Top Internship	0.267 (0.139)	0.290 (0.150)	0.298 (0.108)	1.196 (0.834)	0.0985 (0.0531)	0.164 (0.156)	0.193 (0.174)	0.174 (0.110)	0.513 (0.419)	0.0579 (0.0602)
Second Internship	0.438 (0.146)	0.496 (0.154)	0.446 (0.112)	1.791 (1.163)	0.169 (0.0567)	-0.0224 (0.184)	-0.0758 (0.204)	-0.0825 (0.133)	-0.243 (0.414)	-0.00184 (0.0718)
Work for Money	0.323 (0.145)	0.354 (0.155)	0.355 (0.109)	1.425 (0.958)	0.121 (0.0569)	-0.0629 (0.186)	-0.0391 (0.207)	-0.0369 (0.129)	-0.109 (0.386)	-0.00114 (0.0720)
Technical Skills	-0.0140 (0.131)	-0.0357 (0.143)	0.0372 (0.103)	0.149 (0.418)	-0.00419 (0.0507)	0.376 (0.179)	0.459 (0.199)	0.283 (0.129)	0.834 (0.611)	0.153 (0.0670)
Female, White	-0.0796 (0.149)	-0.177 (0.160)	-0.0434 (0.113)	-0.174 (0.467)	-0.0211 (0.0579)	-0.0435 (0.184)	0.0334 (0.203)	0.0492 (0.133)	0.145 (0.395)	-0.0126 (0.0720)
Male, Non-White	0.0893 (0.175)	0.0368 (0.189)	-0.155 (0.130)	-0.621 (0.634)	0.0435 (0.0676)	-0.0448 (0.232)	0.0282 (0.259)	0.0835 (0.160)	0.246 (0.481)	-0.0412 (0.0893)
Female, Non-White	-0.196 (0.180)	-0.331 (0.193)	-0.0732 (0.140)	-0.294 (0.592)	-0.0720 (0.0689)	-0.160 (0.225)	-0.0550 (0.258)	0.0906 (0.160)	0.267 (0.482)	-0.0362 (0.0891)
Observations	2000	2000	2000	2000	2000	1440	1440	1440	1440	1440
R^2	0.015	0.078	0.553			0.031	0.109	0.651		
<i>p-value for test of joint significance of Majors</i>	0.713	0.787	0.185	0.185	0.821	0.015	0.023	< 0.001	< 0.001	0.014
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Order FEs	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Subject FEs	No	No	Yes	Yes	No	No	No	Yes	Yes	No

Ordered probit cutpoints (Column 5): -0.38, -0.13, 0.19, 0.42, 0.68, 0.98, 1.40, 1.88, 2.45.

Ordered probit cutpoints (Column 10): 0.40, 0.61, 0.85, 1.02, 1.16, 1.31, 1.58, 1.95, 2.22.

Table shows OLS and ordered probit regressions of *Hiring Interest* from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*: *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of major fixed effects and demographic variables are indicated (F -test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

Table 32: Likelihood of Acceptance at Pitt

	Dependent Variable: Likelihood of Acceptance			
	OLS	OLS	OLS	Ordered Probit
GPA	0.178 (0.148)	0.161 (0.155)	0.0104 (0.101)	0.0710 (0.0572)
Top Internship	0.233 (0.103)	0.245 (0.108)	0.235 (0.0680)	0.0873 (0.0398)
Second Internship	0.224 (0.114)	0.221 (0.119)	0.199 (0.0768)	0.0739 (0.0447)
Work for Money	0.142 (0.114)	0.143 (0.120)	0.130 (0.0738)	0.0504 (0.0443)
Technical Skills	0.195 (0.106)	0.187 (0.110)	0.111 (0.0700)	0.0843 (0.0403)
Female, White	-0.0627 (0.115)	-0.0795 (0.122)	0.0152 (0.0774)	-0.0268 (0.0448)
Male, Non-White	-0.000104 (0.139)	-0.0119 (0.145)	-0.0641 (0.0907)	-0.0111 (0.0539)
Female, Non-White	-0.198 (0.140)	-0.197 (0.147)	-0.0483 (0.0904)	-0.0702 (0.0549)
Observations	3440	3440	3440	3440
R^2	0.037	0.061	0.643	
<i>p-value for test of joint significance of Majors</i>	< 0.001	< 0.001	< 0.001	< 0.001
Major FEs	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	Yes	No
Order FEs	No	Yes	Yes	No
Subject FEs	No	No	Yes	No

Ordered probit cutpoints: -0.10, 0.14, 0.38, 0.58, 0.86, 1.08, 1.42, 1.86, and 2.35.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1). Robust standard errors are reported in parentheses. *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and *major* are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included in some specifications as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of major fixed effects and demographic variables are indicated (F -test for OLS, likelihood ratio test for ordered probit).

Table 33: Likelihood of Acceptance by Major Type at Pitt

	Dependent Variable: Likelihood of Acceptance					
	Humanities & Social Sciences			STEM		
	OLS	OLS	Ordered Probit	OLS	OLS	Ordered Probit
GPA	-0.0641 (0.187)	-0.0437 (0.202)	-0.173 (0.127)	0.499 (0.241)	0.427 (0.268)	0.278 (0.181)
Top Internship	0.261 (0.137)	0.248 (0.149)	0.263 (0.0914)	0.210 (0.155)	0.227 (0.173)	0.214 (0.112)
Second Internship	0.353 (0.146)	0.435 (0.156)	0.373 (0.0955)	0.0433 (0.183)	-0.0259 (0.201)	-0.0205 (0.131)
Work for Money	0.271 (0.144)	0.294 (0.155)	0.303 (0.0949)	-0.0506 (0.184)	-0.0453 (0.205)	-0.0345 (0.126)
Technical Skills	-0.0125 (0.130)	0.00378 (0.140)	-0.00849 (0.0864)	0.521 (0.178)	0.638 (0.195)	0.382 (0.128)
Female, White	-0.0639 (0.148)	-0.149 (0.159)	-0.000568 (0.0969)	-0.0353 (0.183)	-0.00711 (0.204)	-0.0254 (0.136)
Male, Non-White	0.110 (0.173)	0.0600 (0.185)	-0.132 (0.112)	-0.152 (0.232)	-0.0799 (0.259)	0.0216 (0.162)
Female, Non-White	-0.138 (0.180)	-0.258 (0.194)	-0.0954 (0.118)	-0.286 (0.224)	-0.218 (0.258)	-0.0310 (0.158)
Observations	2000	2000	2000	1440	1440	1440
R^2	0.010	0.069	0.666	0.036	0.110	0.654
<i>p-value for test of joint significance of Majors</i>	1.436	1.550	1.061	0.006	0.016	< 0.001
Major FEs	Yes	Yes	Yes	Yes	Yes	Yes
Leadership FEs	No	Yes	No	No	Yes	No
Order FEs	No	Yes	No	No	Yes	No
Subject FEs	No	No	No	No	No	No

Ordered probit cutpoints (Column 4): -0.59, -0.34, -0.11, 0.14, 0.47, 0.76, 1.12, 1.59, 2.37.

Ordered probit cutpoints (Column 8): 0.31, 0.56, 0.78, 0.93, 1.12, 1.25, 1.56, 1.96, 2.26.

Table shows OLS and ordered probit regressions of *Likelihood of Acceptance* from Equation (1.1). *GPA*; *Top Internship*; *Second Internship*; *Work for Money*; *Technical Skills*; *Female, White*; *Male, Non-White*; *Female, Non-White* and major are characteristics of the hypothetical resume, constructed as described in Section 1.2.3 and in Appendix A.1.2. Fixed effects for major, leadership experience, resume order, and subject included as indicated. R^2 is indicated for each OLS regression. The p -values of tests of joint significance of major fixed effects and demographic variables are indicated (F -test for OLS, likelihood ratio test for ordered probit) after a Bonferroni correction for analyzing two subgroups.

Appendices to Chapter 2

B.1. Model and definitions

A **marriage market under incomplete information** is a quadruple $(M, W, \mathcal{P}, \lambda)$, where M and W are sets of agents on the two sides of the market, \mathcal{P} is the set of all possible preference profiles for the agents, and λ is a measure over \mathcal{P} . We require that all agents in the market find at least one match partner acceptable. An element of \mathcal{P} is a vector $(P_i)_{i \in M \cup W}$ of individual **preference profiles**. P_m for some $m \in M$ is an ordering over $W \cup \{\emptyset\}$, where \emptyset represents the outcome of being unmatched; P_w for some $w \in W$ is defined similarly. Hence, we can think of some W 's preference ordering as an $(|M| + 1)$ -vector whose elements are \emptyset and the members of M .¹ A **matching** is a function $\mu : M \cup W \mapsto M \cup W \cup \{\emptyset\}$ such that for any $m \in M$ and $w \in W$, we have $\mu(m) \in W \cup \{\emptyset\}$, $\mu(w) \in M \cup \{\emptyset\}$, and $\mu(m) = w \Leftrightarrow \mu(w) = m$. A **strategy** for an agent i is a function $\sigma_i : \mathcal{P}_i \mapsto \mathcal{P}_i$ where \mathcal{P}_i denotes the projection of \mathcal{P} onto only agent i 's preference profile.

Next, we define an important concept, introduced in Roth and Rothblum (1999), which we use to analyze the information structure of the matching markets used in our experiment. Let the $m \leftrightarrow m'$ operation switch the places of m and m' in the preference of each W and assigns the preferences of m to m' (and vice versa). Let $w \leftrightarrow w'$ be defined analogously. The following definition codifies the idea of a low information environment.

Definition 1. For some $w \in W$, a marriage market (or a distribution over M and W preferences) is **M -symmetric with respect to w** if and only if, for any two $m, m' \in M$, $\lambda(P_{-w}|P_w) = \lambda(P_{-w}^{m \leftrightarrow m'}|P_w)$. If this holds for all $w \in W$, we simply call the market **M -symmetric**. **W -symmetry** is analogously defined. If a marriage market is both W -symmetric and M -symmetric, then we call it **MW -symmetric**.

In such symmetric environments, we want to be able to rule out equilibria where strategies

¹In our context, thinking of preferences as vectors introduces a bit of redundancy since the mechanisms we consider are all individually rational; for example, $(m_1, m_2, m_3, \emptyset, m_4, m_5)$ and $(m_1, m_2, m_3, \emptyset, m_5, m_4)$ are functionally equivalent.

depend on label, as these seem artificial. Formally,

Definition 2. A strategy σ_i is *anonymous* if and only if, for any two preferences, P_i and P'_i , that list the same number of acceptable match partners, there exists some permutation π such that $\sigma_i(P_i) = \pi(P_i)$ and $\sigma_i(P'_i) = \pi(P'_i)$.

Note that this definition allows for different permutations to be used when a different number of match partners are acceptable. Of the set of anonymous strategies, in the low information environments we look at in the lab, we will find that we expect a certain type of strategy in equilibrium.

Definition 3. A *truncation* is an anonymous strategy where the permutation for a given number of acceptable match partners, $k - 1$, is a composition of permutations that first exchanges the k^{th} position (i.e. \emptyset) with the j^{th} position, where $j \leq k$, and then permutes all positions besides k and j in a way that if a position started ranked (above j /between j and k /below k), its permuted position is ranked (above j /between j and k /below k).²

Finally, we introduce a technical condition needed for uniqueness (but not existence) of the types of equilibria we will be looking for.

Definition 4. A distribution over preferences is called *W-thick* if, for any $w \in W$, $m, m' \in M$, and $m'' \in M \setminus \{m, m'\}$ there is a positive probability that m and m' rank w first, while m'' ranks $w'' \neq w$ first and w'' ranks m'' first. *M-thick* is defined analogously. A distribution over preferences is called *MW-thick* if it is both M and W thick.

Thickness is a sufficient condition that prevents an agent from ruling out the possibility that any two potential match partners are her only two stable match partners. Weaker conditions are possible, but thickness itself is quite weak: for instance, it is met when all possible profiles of first choices are drawn with positive probability.

²Note that under this definition, a truthful strategy is a truncations.

B.2. Proofs

Lemma 1. *Under M -Proposing DA, truth-telling is the only weakly undominated strategy for all $m \in M$.*

Proof. Assume that the strategy of some $m \in M$ submits a preference \widetilde{P}_m that is not the true preference, P_m . Dubins and Freedman (1981) show that truth-telling cannot yield a worse outcome than a lie. Let k be the first position in the submitted rank-order list that \widetilde{P}_m differs from the true preference, P_m . Let $w = P_m(k)$ and $w' = \widetilde{P}_m(k)$. If all W s except for w and w' rank m as unacceptable, and w and w' only rank m as acceptable, then m gets w if he submits P_m and w' (which he likes less) if he submits \widetilde{P}_m . Hence, we have shown that truth-telling is never worse than a lie and is strictly better given some profile of strategies for the other agents. \square

Lemma 2 (Roth, 1989). *Under M -Proposing DA, it is weakly dominated for any $w \in W$ to not list her true first choice first.*

Lemma 3. *In a marriage market that is M -symmetric with respect to w , if all agents besides w play anonymous strategies, and all $m \in M$ play the same strategy, then the distribution over submitted preferences, $\widetilde{\lambda}(\cdot)$, is also M -symmetric with respect to w .*

Proof. To prove this, we show why the following equation must hold:

$$\begin{aligned} \widetilde{\lambda}(\sigma_{-w}(P_{-w})|P_w) &= \lambda(P_{-w}|P_w) = \lambda(P_{-w}^{m \leftrightarrow m'}|P_w) \\ &= \widetilde{\lambda}(\sigma_{-w}(P_{-w}^{m \leftrightarrow m'})|P_w) = \widetilde{\lambda}((\sigma_{-w}(P_{-w}))^{m \leftrightarrow m'}|P_w) \end{aligned}$$

The first equality comes from the definition of $\widetilde{\lambda}$, the second from the fact that the true preferences are M -symmetric with respect to w , and the third, again from the definition of $\widetilde{\lambda}$. For the last equality, we must note two things. First, since the $m \leftrightarrow m'$ does not

change the rank of \emptyset for the W 's, the σ_{-w} operator applies the same permutation to $P_{w'}^{m \leftrightarrow m'}$ as it does to $P_{w'}$. Second, since the M 's are all playing the same anonymous strategy, it makes no difference whether we switch the preferences of m and m' before we apply the σ_{-w} operator or after. Hence, σ_{-w} commutes with $m \leftrightarrow m'$.³ \square

Proposition 6. *In an M -symmetric marriage market, under M -Proposing DA, there exists an equilibrium in anonymous, weakly undominated strategies that involves truth-telling for each $m \in M$ and truncation for each $w \in W$. Furthermore, if the market is also W -thick, all equilibria in anonymous, weakly undominated strategies are like this.*

Proof. By Lemma 1, any equilibrium in weakly undominated strategies involves truth-telling by all M 's. By Lemma 3, we then know that, at an equilibrium in weakly undominated, anonymous strategies, the distribution of reported preferences, $\tilde{\lambda}$, is M -symmetric. Then, by the main proposition of Roth and Rothblum (1999), we know that truncation is a best response for all $w \in W$. Furthermore, by Lemma 2, every W must be truthfully ranking her first choice M . Then, by the W -thickness assumption, it is with positive probability that for any $m, m' \in M$, w can only potentially match to m or m' . In these states of the world, we are in Case D of the proof from Roth and Rothblum (1999), which means that truncation strictly dominates non-truncation. \square

Since the uncorrelated market is M -symmetric and W -thick, **Proposition 1 in the main text** is an immediate corollary.

Lemma 4. *Under M -Proposing Priority, it is weakly dominated for any $w \in W$ to not truthfully rank her first choice M .*

Proof. In the first round w gets proposals, she will be permanently matched. Ranking her first choice, $m \in M$ first can not hurt her, but failing to do so can hurt her if she also receives a proposal in that round from an $m' \in M$ that she ranked higher than m , but actually likes

³Note that we are not claiming that permutations commute: our interchange operator references school names and not positions in a rank-order list.

less. Let m and her declared first choice, $\widetilde{P}_w(1)$, both rank w first, and let all other $m'' \in M$ declare w unacceptable. Ranking m first instead of $\widetilde{P}_w(1)$ is an improvement. \square

Proposition 7. *In an M -symmetric marriage market, under M -Proposing Priority, if all agents play anonymous, weakly undominated strategies, and in addition, all $m \in M$ truth-tell, then all $w \in W$ can best-respond to the other agents by truncating. If the market is also W -thick, then all of their best responses are truncations.*

Proof. By Lemma 3, we know that the distribution of reported preferences, $\widetilde{\lambda}$, is M -symmetric with respect to w . Then, by Proposition 3.2 and Remark 3.2 of Ehlers (2008), we know that truncation is a best response for all w . Furthermore, by Lemma 4, every W must be truthfully ranking her first choice M . Then, by the W -thickness assumption, it is true with positive probability that for any $m, m' \in M$, w can only potentially match to m or m' ; hence, Equation A2 from Ehlers (2008) must hold strictly, which means that truncation strictly dominates non-truncation. \square

Since the uncorrelated market is M -symmetric and W -thick, **Proposition 2 in the main text** is an immediate corollary.

Lemma 5. *Under M -Proposing Priority, any report for any $m \in M$ that does not list all and only all truly acceptable $w \in W$ as acceptable is weakly dominated by one that does.*

Proof. Consider an arbitrary $m \in M$ submitting a list L with n acceptable match partners which excludes at least one acceptable $w' \in W$. Now consider L' , a list identical to L for the first n entries with w' listed in the $(n + 1)^{\text{st}}$ position and no acceptable entries thereafter. Under M -Proposing Priority, any set of submissions for other agents resulting in m being matched to a given W when m submits L will also result in M being matched to that W when m submits L' . So L' never generates a worse outcome for m than L . However, consider a set of submissions such that no member of W listed in L ranks m as acceptable, and the submitted preference list of w' lists only m as acceptable. In this case, M -Proposing Priority will match m and w' when L' is submitted and will match m to no one when L is

submitted. Since w' is acceptable to m by construction, m achieves a better result in this case by submitting L' .

Now consider some $m \in M$ who lists a truly unacceptable $w \in W$ as acceptable. Removing this w from his list cannot hurt m , since M -Proposing Priority makes permanent matches after each round. Now, let all $w' \in W \setminus w$ declare m unacceptable, let all $m' \in M \setminus m$ declare w unacceptable and let w declare m acceptable. With this strategy profile, m will match to w which he could have avoided by declaring her unacceptable. \square

Lemma 6. *Under M -Proposing Priority, if the distribution of reported preferences for all agents besides $m \in M$ are W -symmetric with respect to m , then truth-telling is a best-response for m .*

Proof. This proof borrows heavily from Roth and Rothblum (1999). First, we lay out a few of the properties of M -Proposing Priority. Consider, P , $w', w \in W$, $m \in M$, and let $v \in (W \setminus \{w, w'\}) \cup \{\emptyset\}$. Denote the match of m when the submitted preferences are P under M -Proposing Priority as $\text{MPP}[P](m)$. Then,

$$\begin{aligned} \text{MPP}[P](m) = v &\Leftrightarrow \text{MPP}[P^{w \leftrightarrow w'}](m) = v \\ \text{MPP}[P](m) = w &\Leftrightarrow \text{MPP}[P^{w \leftrightarrow w'}](m) = w' \end{aligned}$$

Moreover,

$$\begin{aligned} \text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = v &\Leftrightarrow \text{MPP}[P_m, P_{-m}^{w \leftrightarrow w'}](m) = v \\ \text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w &\Leftrightarrow \text{MPP}[P_m, P_{-m}^{w \leftrightarrow w'}](m) = w' \end{aligned}$$

The first set of logical statements follows immediately from the fact that MPP does not give special treatment to any given label. The fact that applying the $w \leftrightarrow w'$ interchange operator to $(P_m^{w \leftrightarrow w'}, P_{-m})$ yields $(P_m, P_{-m}^{w \leftrightarrow w'})$, implies the second set.

Table 34: Table of cases

		Lie: $\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m)$		
		$= v \notin \{w, w'\}$	$= w$	$= w'$
Truth: $\text{MPP}[P_m, P_{-m}](m)$	$= v \notin \{w, w'\}$	Case A	Case B	Impossible
	$= w$	Impossible	Case C	Impossible
	$= w'$	Case D	Case E	Case F

Now, let $w \prec_m w'$. Then,

$$(\text{MPP}[P](m) = w) \Rightarrow (\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w)$$

Moreover,

$$(\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w') \Rightarrow (\text{MPP}[P](m) = w')$$

Switching w' and w in a submitted ordering means that w is proposed to in an earlier round. If it was available in the later round, it will still be available in the earlier round, and no one else will be proposing to it in that round. This yields the first logical statement. The second follows from a similar line of reasoning.

Now, consider the outcome for some $m \in M$ for whom $w \prec_m w'$ when he submits a preference that truthfully ranks w and w' , $\text{MPP}[P](m)$, and when he submits a preference that switches w and w' , $\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m)$. Using the formulas we just derived, we summarize what can potentially happen in Table 34, while Table 35 tells us what lottery over outcomes m can expect when he truthfully orders w and w' and when he switches their ordering, given that everyone else's preferences are either P_{-m} or $P_{-m}^{w \leftrightarrow w'}$ with equal probability.

Clearly, under every case, if we take symmetry into account, truthfully ordering w and w' either yields an outcome that is equivalent to the outcome achieved with the lie, or weakly stochastically dominates the outcome from the lie. Now, Lemma 5 shows us that an M cannot be hurt by listing all acceptable Ws, so we know that truth-telling is a best response

Table 35: Payoffs for the cases

	Truth		Lie	
	$\text{MPP}[P_m, P_{-m}]$	$\text{MPP}[P_m, P_{-m}^{w \leftrightarrow w'}]$	$\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}]$	$\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}^{w \leftrightarrow w'}]$
Case A	v	v	v	v
Case B	v	w'	w	v
Case C	w	w'	w	w'
Case D	w'	v	v	w
Case E	w'	w'	w	w
Case F	w'	w	w'	w

for M s to W -symmetry.

Note that if we can show that the probability of being in Cases B, D, or E is strictly positive, then we also show that truthfully ordering the W s strictly stochastically dominates any lie, although we would need a further restriction to weakly undominated strategies to get truth-telling as a unique best response. \square

Lemma 7. *In an M -thick, W -symmetric marriage market, under M -Proposing Priority, if all $w \in W$ are playing the same weakly undominated, anonymous strategy, and all $m' \in M \setminus \{m\}$ are playing anonymous strategies, then all best responses for $m \in M$ must truthfully rank his true first choice partner.*

Proof. By similar logic to Lemma 3, the submitted preferences are W -symmetric with respect to m . Consider the argument of Lemma 6 with regard to the true first choice and some other reported first choice. By the M -thickness assumption and Lemma 4, there is some probability that those two W s rank m first, meaning that we are in Case E of Lemma 6, meaning that m does strictly better to truthfully rank his first choice. \square

Lemma 8. *In an M -thick, W -symmetric marriage market, if each $w \in W$ plays the same anonymous, weakly undominated strategy, and each $m' \in M \setminus \{m\}$ truthfully reveals his first choice partner, then under M -Proposing Priority, the only best-response for m' is to truth-tell.*

Proof. By Lemma 2, weakly undominated means that all W s must truthfully rank their first choice partner. Since, by an argument analogous to Lemma 3, reported preferences must be W -symmetric with respect to m , we conclude through Lemma 6 that m cannot do worse than to truthfully reveal. Further, by Lemma 7, m must also best respond by truthfully ranking his first choice partner at equilibrium. From here, the M -thickness assumption allows us to go the rest of the way in showing that, for any two W 's, the probability of being in Case E of Lemma 6 is strictly positive, and that the only best response for m is to truthfully reveal. \square

Formally, a ***symmetric equilibrium*** is one in which any two M s are playing the same strategy, and any two W s are playing the same strategy.

Proposition 8. *In an MW-symmetric marriage market, under M-Proposing Priority, there exists a symmetric equilibrium in anonymous strategies that involves truth-telling by the M s and truncation by the W s. Furthermore, if the market is MW-thick, then all symmetric equilibria in anonymous, weakly undominated strategies are of this form.*

Proof. If every M is playing the same anonymous strategy, and every W is playing an anonymous strategy, then by Lemma 3, the reported preferences are M -symmetric, and by Ehlers (2008), all W s can best-respond with a truncation.

Now, consider the problem of finding the best-response of some $w \in W$ to the symmetric M strategies, σ_M , and a profile σ_{-w} in which all members of $W \setminus \{w\}$ are playing the same mixed strategy over truncations. Call this best response $\sigma_w^*(\sigma_{-w}|\sigma_M)$. Solving for the best response is an optimization problem in which w must choose her mix over truncation levels for each possible number of acceptable M s her preference could hold. The objective is linear in the mixing probabilities,⁴ and the set of possible mixing probabilities is closed and convex. Hence, we know that the solution exists, it is convex, and by the Theorem of the Maximum (Mas-Colell et al. 1991, Theorem M.K.6), it is upper hemicontinuous. Hence,

⁴For a given pure strategy profile, w gets an expected payoff. Her expected payoff from a mixed strategy is just a probability-weighted sum of these expected payoffs from pure strategies.

by Kakutani's Fixed Point Theorem (Mas-Colell et al. 1991, Theorem M.I.2), $\sigma_w^*(\sigma_{-w}|\sigma_M)$ has a fixed point. Hence, for any symmetric σ_M , there is a symmetric σ_W where each W is best responding to the other players.

Now, in any such setup, the M s will not necessarily be best-responding. Since the market is W -symmetric, we know that the reported preferences are W -symmetric, which means that, by Lemma 6, the M s can best-respond by truth-telling. Hence, we have found a symmetric equilibrium of the sort we were looking for.

Now, if strategies are anonymous and weakly undominated, then M -thickness coupled with Lemmas 7 and 8 requires that all such symmetric equilibria involve M s truth-telling. Similarly, W -thickness couples with Lemma 2 requires that all such symmetric equilibria involve W s truncating. \square

This proposition has an immediate corollary, which is referenced in **Footnote 17 of the main text**.

Corollary (to Proposition 8). *In the uncorrelated market, under M -Proposing Priority, there exists a symmetric equilibrium that involves all W s playing the same truncation strategy and all M s truth-telling. Furthermore, all symmetric equilibria in anonymous, weakly undominated strategies are of this form. Also, we can note that so long as the M s use anonymous, weakly undominated strategies, the W s still best-respond with truncation. So long as the M strategies don't key in on a label, the W s view them strategically in the same way as they view truth-telling M s.*

The big implication here is that if an M believes that the equilibrium played will be a symmetric truncation equilibrium, then truth-telling is the best response. This proposition extends work done in Roth and Rothblum (1999) and Ehlers (2008) to conditions that lead to truth-telling for the proposing side under a priority mechanism.⁵ In a broader sense,

⁵Roth and Rothblum (1999) and Ehlers (2008) focus on incentives for the receiving side. These papers also assume that reported preferences are M -symmetric instead of assuming that the true preferences are M -symmetric and backing out sufficient conditions to ensure that the reported preferences inherit M -symmetry as well.

though, it turns out not to matter whether the M s truthfully reveal.

Proposition 9. *In a M -symmetric market, under M -Proposing Priority, for any $w \in W$, if for any distinct $m, m' \in M$, P_m and $P_{m'}$ are conditionally independent given P_w and for any $m' \in M$ and $w' \in W$, $P_{m'}$ and $P_{w'}$ are conditionally independent given P_w , and all agents play anonymous, weakly undominated strategies, then w can best-respond with a truncation. Furthermore, if the market is also W -thick, then any best response must be a truncation.*

Proof. Since the preferences of the M s are all conditionally independent, it must be that for any given number of truly acceptable match partners, all lists with that number of acceptable partners are equally likely. By Lemma 5, the weakly undominated requirement means that the M s must list all acceptable W s. The anonymous requirement then means that these lists must be permutations. Running a uniform distribution through a permutation yields a uniform distribution. Hence, the reported preferences of the M s must be uniformly distributed for each number of acceptable partners, meaning that the reported preferences of the M s are independent of the strategies they use. Looking back to the proof of Lemma 3, the fact that the M s' reported preferences are conditionally independent and uniform for each list length, and that M preferences are conditionally independent of W preferences means that we no longer need that all M s play the same strategy to get the same result. This means, that through a proof very similar to that of Proposition 7, w must best-respond with a truncation. \square

This proposition has an immediate corollary, which is references in **Footnote 17 of the main text**.

Corollary (to Proposition 9). *In the uncorrelated market, under M -Proposing Priority, if all agents play anonymous, weakly undominated strategies, then all W s must best-respond with a truncation.*

Proposition (**Proposition 3 in the main text**). *In the correlated market, under M -Proposing DA, the unique equilibrium in anonymous, weakly undominated strategies entails*

truth-telling by all agents.

Proof. Under M -Proposing DA, weakly undominated strategies require that the M s truthfully reveal (Lemma 1). Hence, under the assumptions, a given W will receive all offers she is going to receive in one round of the algorithm. To see this, first note that the top-ranked W , w_1 , will receive all offers in the first round of the algorithm. She will be matched to her declared favorite M , and since this is a declared top-top match, the algorithm will never break it up. In the next round, the second ranked W , w_2 , will receive offers from all other M s. She will accept her declared favorite M who proposes, and the algorithm will never break this match (since the only potential M that w_2 might defect to is matched to w_1 , who he prefers, and w_1 was given her declared top M). And so on. So at some point in the algorithm, a W 's preference is used to choose a favorite M from a set of M s that higher ranked W s have not yet taken. There is no gain to not truthfully revealing, as our member of W is facing a static decision problem. Since every W has a one-in-five chance of being the last ranked W by all M s, there is always a positive loss to dropping. \square

Proposition (Proposition 4 in the main text). *In the correlated market, under M -Proposing Priority, if all members of M have the same anonymous, weakly undominated strategy, then all members of W best respond by truthfully revealing.*

Proof. Under M -Proposing Priority, weakly undominated for the M s means that all women are listed as acceptable (Lemma 5). Under the assumptions, a member of W will receive all offers in one round of the algorithm. There is no gain to not truthfully revealing then, as our member of W is facing a static decision problem. Since every W has a positive probability of being the last ranked W by all M s, there is always a positive loss to dropping any M . \square

Proposition (Proposition 5 in the main text). *In the correlated environment, there exist cardinal payoffs that rationalize an equilibrium where all M s and W s truthfully reveal their preferences.*

Proof. For each M , consider a payoff vector $\pi = (p_1, p_2, p_3, p_4, p_5)$ which is constructed as $p_5 = 1$, $p_4 = p_5 \cdot |M| + 1$, $p_3 = p_4 \cdot |M| + 1$, etc. In the correlated M -Proposing Priority environment, each M has a $1/|M|$ chance of being the first choice of any W . Thus, from the perspective of an M with payoffs described by π , even in the worst case when all other M s also rank M 's first choice as first, the M would still prefer the $1/|M|$ chance of getting its first choice than a certainty of getting its second choice. Similarly, an M failing to get its first choice would prefer the $1/|M|$ —chance of getting its second choice to a certainty of getting its third choice, and so on. Hence all M s truthfully reveal, and by the previous Proposition, the W s must as well. \square

B.3. Model Estimation

The reparameterized EWA model suggests the need to estimate parameters for the learning process δ , ϕ , and λ ; an initial probability of play for each strategy (with initial probabilities either shared across individuals or estimated separately for each subject); and $\|A_0\|$ (again, either shared across individuals or estimated separately by subject or group), representing the weight of initial cognition in units of payoff amounts. In a Bayesian model, this initial cognition would be akin to pseudo-observations of play from previous rounds.

This suggests a parameter space of at least 329 dimensions, with still higher dimensionality if we allow probabilities of play and the weight of initial cognition to vary across individuals.⁶ This is computationally intractable due to the large number of initial probabilities, even when we assume all initial probabilities are shared by all players in a treatment.

However, most (225 of 325) strategies are never played in any round of any treatment. Moreover, only 20 strategies are ever played in the first round of any session, and only 11 strategies are played more than once in any first round. This suggests that estimating initial probabilities for all 325 strategies is not only computationally infeasible, but also not necessary for us to understand the dynamics of play. Instead, for each of the four treatment groups, we estimate the initial probabilities of play for all strategies played more than once in any initial round, and a single joint attraction toward playing all other strategies. This reduces the search space to 15 dimensions (three learning parameters, 11 probabilities, and the initial cognition weight).⁷

Let us denote strategies by five digits, denoting the true preference ranks of the player's submitted preferences by the digits themselves, and the submitted preferences by the order the digits. Let the symbol \emptyset represent a match listed as unacceptable in the submitted preference list. For instance, the strategy {12345} represents complete truth-telling,

⁶The number of possible strategies in a round is $5! + 4 \times \binom{5}{1} + 3! \times \binom{5}{3} + 2! \times \binom{5}{2} + 1! \times \binom{5}{1} = 325$

⁷Note that we actually want to estimate 12 probabilities that are mutually exclusive and comprehensively exhaustive, and therefore must sum to 1. By estimating 11 of the probabilities directly, we get the 12th for free).

while $\{12354\}$ represents a permutation strategy with the least preferred options listed in reverse order. A truncation strategy such as $\{123\emptyset\emptyset\}$ consists of listing only the most preferred three preferences. These 11 strategies played more than once in an initial round include both complete truth-telling $\{12345\}$ and the four possible truncation strategies: $\{123\emptyset\emptyset\}$, $\{12\emptyset\emptyset\emptyset\}$, $\{1234\emptyset\}$, $\{1\emptyset\emptyset\emptyset\emptyset\}$. The six remaining strategies are permutations, or combinations of permutation and truncation: $\{21345\}$, $\{213\emptyset\emptyset\}$, $\{13245\}$, $\{21435\}$, $\{12354\}$, $\{23145\}$. Thus, this estimation strategy allows us to measure differences in initial probability to truth telling and various truncation strategies, and to use these parameters to follow the trajectory of attractions over the course of the game. For the remaining strategies—those played either once or not at all in an initial round—we estimate a single initial attraction in each treatment. While limiting our estimation of individual attractions to repeated initial strategies requires a *post hoc* justification, we believe this is necessary to make the model tractable, and allows the use of this model in a much more complex space than usual. This approach allows us to capture the differences between truth-telling and truncation that we care about, while significantly simplifying the strategy space. Estimating a joint attraction for all unplayed strategies, also allows the model to scale with payoff values. Thus, this approach is flexible to applications with different payoffs.

B.3.1. Technical Details

To maximize over the rugged likelihood terrain, we implement the stochastic, derivative-free Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) optimizer. CMA-ES is designed to be robust to local optima, ridges, and discontinuities in ill-conditioned and non-separable problems (Hansen, 2016). We estimate standard errors using a numerical approximation of the Hessian, and transform to our reparametrized EWA via the delta method. We executed all maximum likelihood estimation in Java.

For the treatment-level estimation, we directly estimate initial probabilities of 10 of the 11 strategies played initially more than once, and an additional probability shared among all other strategies. We estimate one strategy (truth-telling, $\{12345\}$) indirectly, by taking

one minus the sum of the other probabilities. This decision was merely practical: our optimizer accepts simple boundaries (a minimum and a maximum) for each of the estimated parameters, so we run the risk at each iteration of the optimizer to have the sum of the directly-estimated probabilities sum to more than one. By leaving out the most commonly played strategy, we reduce the frequency of this event. When the sum of the randomly-drawn probability proposal points is greater than one, we instruct the log likelihood function to return an arbitrarily large negative value, encouraging the optimizer to seek elsewhere.

For the practical purposes of estimation, we imposed search boundaries on the estimated parameters as follows:

$$\phi \in [0.00001, 1000]$$

$$\lambda \in [0.00001, 1000]$$

$$\delta \in [0, 1]$$

$$||A_0|| \in [0, 10,000]$$

$$\text{Initial probabilities} \in [1e-9 \text{ and } 1.0]$$

$$\sum \text{Initial probabilities} = 1$$

BIBLIOGRAPHY

- A. Abdulkadiroğlu, P. Pathak, and A. Roth. The new york city high school match. *American Economic Review*, pages 364–367, 2005.
- A. Abdulkadiroğlu, P. Pathak, and A. Roth. Strategy-proofness versus efficiency in matching with indifference: Redesigning the nyc high school match. *The American Economic Review*, 99(5):1954–1978, 2009.
- J. G. Altonji and R. M. Blank. Race and gender in the labor market. *Handbook of Labor Economics*, 3:3143–3259, 1999.
- APPIC. 2009 APPIC Match Statistics. http://www.appic.org/match/5_2_2_1_11_match_about_statistics_general_2009.html, 2009.
- D. H. Autor and S. 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.
- S. Baert. Chapter 3: Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. In M. S. Gaddis, editor, *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, chapter 3, pages 63–77. Springer, 2018.
- V. Bartoš, M. Bauer, J. Chytilová, and F. Matějka. Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–75, June 2016. doi: 10.1257/aer.20140571.
- M. Bertrand and E. Duflo. Field experiments on discrimination. NBER Working Papers 22014, National Bureau of Economic Research, Inc, Feb 2016.
- M. Bertrand and S. Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? *The American Economic Review*, 94(4):991–1013, 2004.
- M. Bertrand, D. Chugh, and S. Mullainathan. Implicit discrimination. *American Economic Review*, 95(2):94–98, 2005.
- I. Bohnet, A. Van Geen, and M. Bazerman. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5):1225–1234, 2015.
- J. A. Bohren, A. Imas, and M. Rosenberg. The dynamics of discrimination: Theory and evidence. *American Economic Review (Forthcoming)*, 2018.
- D. M. Butler and D. E. Broockman. Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*, 55(3):463–477, 2011.
- C. Calsamiglia, G. Haeringer, and F. Klijn. Constrained school choice: An experimental study. *American Economic Review*, 2009.

- C. Camerer and T.-H. Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999. ISSN 1468-0262. doi: 10.1111/1468-0262.00054. URL <http://dx.doi.org/10.1111/1468-0262.00054>.
- S. E. Carrell, M. E. Page, and J. E. West. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144, 2010.
- Y. Chen and T. Sönmez. School choice: an experimental study. *Journal of Economic Theory*, 127(1):202–231, 2006.
- P. Coles. Optimal truncation in matching markets. *Unpublished manuscript*, 2009.
- P. Coles, A. Kushnir, and M. Niederle. Preference signaling in matching markets. Technical report, National Bureau of Economic Research, 2010.
- B. Culwell-Block and J. A. Sellers. Resume content and format - do the authorities agree?, Dec 1994. URL <https://www.questia.com/library/journal/1G1-16572126/resume-content-and-format-do-the-authorities-agree>.
- E. Damiano, H. Li, and W. Suen. Unravelling of dynamic sorting. *Review of Economic Studies*, 72(4):1057–1076, 2005. ISSN 1467-937X.
- R. Darolia, C. Koedel, P. Martorell, K. Wilson, and F. Perez-Arce. Do employers prefer workers who attend for-profit colleges? Evidence from a field experiment. *Journal of Policy Analysis and Management*, 34(4):881–903, 2015.
- R. Day and P. Milgrom. Core-selecting package auctions. *International Journal of Game Theory*, 36(3):393–407, 2008.
- J. de Quidt, J. Haushofer, and C. Roth. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, November 2018. doi: 10.1257/aer.20171330. URL <http://www.aeaweb.org/articles?id=10.1257/aer.20171330>.
- D. J. Deming, N. Yuchtman, A. Abulafi, C. Goldin, and L. F. Katz. The value of post-secondary credentials in the labor market: An experimental study. *American Economic Review*, 106(3):778–806, March 2016.
- L. Dishman. Your resume only gets 7.4 seconds to make an impression here’s how to stand out, Nov 2018. URL <https://www.fastcompany.com/90263970/your-resume-only-gets-7-4-seconds-to-make-an-impression-heres-how-to-stand-out>.
- G. Distelhorst and Y. Hou. Constituency service under nondemocratic rule: Evidence from China. *The Journal of Politics*, 79(3):1024–1040, 2017. doi: 10.1086/690948.
- S. Du and Y. Livne. Chaos and Unraveling in Matching Markets. *Arxiv preprint arXiv:1009.0769*, 2010.

- L. Dubins and D. Freedman. Machiavelli and the Gale-Shapley algorithm. *American Mathematical Monthly*, pages 485–494, 1981.
- E. Dufo. Richard t. ely lecture: The economist as plumber. *American Economic Review*, 107(5):1–26, May 2017. doi: 10.1257/aer.p20171153. URL <http://www.aeaweb.org/articles?id=10.1257/aer.p20171153>.
- F. Echenique and L. Yariv. An experimental study of decentralized matching. Technical report, Discussion paper, Working paper, Caltech, 2010.
- F. Echenique, A. Wilson, and L. Yariv. Clearinghouses for Two-Sided Matching: An Experimental Study. Technical report, Working Paper, 2010.
- L. Ehlers. Truncation Strategies in Matching Markets. *Mathematics of Operations Research*, 33(2):327, 2008.
- H. Ergin and T. Sönmez. Games of school choice under the Boston mechanism. *Journal of Public Economics*, 90(1-2):215–237, 2006.
- S. Eriksson and D.-O. 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.
- M. Ewens, B. Tomlin, and L. C. Wang. Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, 96(1):119–134, 2014.
- I. Fainmesser. Social Networks and Unraveling in Labor Markets. *Journal of Economic Theory*, 148(1):64–103, 2013. ISSN 0022-0531.
- H. S. Farber, C. M. Herbst, D. Silverman, and T. von Wachter. Whom do employers want? The role of recent employment and unemployment status and age. Working Paper 24605, National Bureau of Economic Research, May 2018.
- C. Featherstone and M. Niederle. Ex Ante Efficiency in School Choice Mechanisms: An Experimental Investigation. NBER Working Paper No. 14618. *National Bureau of Economic Research*, 2008.
- R. G. Fryer and S. D. Levitt. The causes and consequences of distinctively black names. *Quarterly Journal of Economics*, 119:767–805, 2004.
- S. M. 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.
- D. Gale and L. Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, pages 9–15, 1962.
- C. Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119, 04 2014.

- O. Govorun and B. K. Payne. Ego-depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, 24(2):111–136, 2006.
- W. Greene. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, 7(1):98–119, 2004.
- A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.
- H. Halaburda. Unravelling in two-sided matching markets and similarity of preferences. *Games and Economic Behavior*, 69(2):365–393, 2010. ISSN 0899-8256.
- D. Hamermesh. Are fake resumes ethical for academic research? *Freakonomics Blog*, 2012.
- N. Hansen. The CMA evolution strategy: A tutorial. *CoRR*, abs/1604.00772, 2016. URL <http://arxiv.org/abs/1604.00772>.
- A. Hanson and Z. Hawley. Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics*, 70(2-3):99–114, 2011.
- G. Harrison and K. McCabe. Stability and preference distortion in resource matching: an experimental study of the marriage problem. *Research in Experimental Economics*, 6: 53–129, 1996.
- G. W. Harrison and J. A. List. Field experiments. *Journal of Economic Literature*, 42(4): 1009–1055, December 2004. doi: 10.1257/0022051043004577.
- E. Haruvy and M. U. Unver. Equilibrium selection and the role of information in repeated matching markets. *Economics Letters*, 94(2):284–289, 2007.
- J. J. Heckman. Detecting discrimination. *Journal of Economic Perspectives*, 12(2):101–116, 1998.
- J. J. Heckman and P. 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.
- N. Immorlica and M. Mahdian. Marriage, Honesty, and Stability. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 53. Society for Industrial Mathematics, 2005.
- S. V. Kadam. Correlation of preferences and honesty in matching markets:.. *working paper*, 2011.

- J. Kagel and A. Roth. The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment. *Quarterly Journal of Economics*, 115(1):201–235, 2000.
- L. J. Kirkeboen, E. Leuven, and M. Mogstad. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111, 2016.
- M. 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.
- F. Kojima and P. Pathak. Incentives and stability in large two-sided matching markets. *The American Economic Review*, 99(3):608–627, 2009.
- K. Kroft, F. Lange, and M. 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.
- K. Lang and J.-Y. K. Lehmann. Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, 50(4):959–1006, 2012.
- R. S. Lee and M. Schwarz. Signalling preferences in interviewing markets. In P. Cramton, R. Müller, E. Tardos, and M. Tennenholtz, editors, *Computational Social Systems and the Internet*, number 07271 in Dagstuhl Seminar Proceedings, Dagstuhl, German, September 2007.
- R. S. Lee and M. Schwarz. Interviewing in two-sided matching markets, 2009.
- H. Li and S. Rosen. Unraveling in matching markets. *American Economic Review*, 88(3):371–387, 1998. ISSN 0002-8282.
- H. Li and W. Suen. Risk sharing, sorting, and early contracting. *Journal of Political Economy*, 108(5):1058–1091, 2000. ISSN 0022-3808.
- S. Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–87, November 2017. doi: 10.1257/aer.20160425. URL <http://www.aeaweb.org/articles?id=10.1257/aer.20160425>.
- C. Low. A “reproductive capital” model of marriage market matching. *Manuscript, Wharton School of Business*, 2017.
- A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic theory*. Oxford University Press, 1991.
- K. L. Milkman, M. Akinola, and D. Chugh. Temporal distance and discrimination: an audit study in academia. *Psychological Science*, 23(7):710–717, 2012.
- K. L. Milkman, M. Akinola, and D. Chugh. What happens before? A field experiment

- exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, 100(6), 2015.
- H. Nalbantian and A. Schotter. Matching and efficiency in the baseball free-agent system: an experimental examination. *Journal of Labor Economics*, 13(1):1–31, 1995.
- D. Neumark. Detecting discrimination in audit and correspondence studies. *Journal of Human Resources*, 47(4):1128–1157, 2012.
- D. Neumark, I. Burn, and P. Button. Is it harder for older workers to find jobs? New and improved evidence from a field experiment. Technical report, National Bureau of Economic Research, 2015.
- M. Niederle and A. Roth. Market Culture: How Rules Governing Exploding Offers Affect Market Performance. *American Economic Journal: Microeconomics*, 1(2):199–219, 2009. ISSN 1945-7669.
- M. Niederle, A. Roth, and M. Ünver. Unraveling Results from Comparable Demand and Supply: An Experimental Investigation. *NBER Working Paper*, 2009.
- B. A. Nosek, F. L. Smyth, J. J. Hansen, T. Devos, N. M. Lindner, K. A. Ranganath, C. T. Smith, K. R. Olson, D. Chugh, A. G. Greenwald, et al. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1):36–88, 2007.
- NRMP. Results and Data: 2009 Main Residency Match. <http://www.nrmp.org/data/resultsanddata2009.pdf>, 2009.
- J. M. Nunley, A. Pugh, N. Romero, and R. A. Seals. Unemployment, underemployment, and employment opportunities: Results from a correspondence audit of the labor market for college graduates. *Auburn University Department of Economics Working Paper Series*, 4, 2014.
- J. M. Nunley, A. Pugh, N. Romero, and R. A. 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.
- NYC-DOE. Statistical summaries: register by grade. <http://schools.nyc.gov/AboutUs/data/stats/Register/CurrentRegisterByGrade/default.htm>, 2009.
- A. Ortmann and R. Hertwig. The costs of deception: Evidence from psychology. *Experimental Economics*, 5(2):111–131, 2002.
- J. Pais and Á. Pintér. School choice and information: an experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303–328, 2008.
- A. Pallais. Inefficient hiring in entry-level labor markets. *American Economic Review*, 104(11):3565–99, 2014.

- P. Pathak and T. Sönmez. Leveling the playing field: Sincere and strategic players in the boston mechanism. *American Economic Review*, 98:1636–1652, 2008.
- D. G. Pope and J. R. Sydnor. What’s in a picture? Evidence of discrimination from prosper.com. *Journal of Human resources*, 46(1):53–92, 2011.
- N. Quadlin. The mark of a womans record: Gender and academic performance in hiring. *American Sociological Review*, 83(2):331–360, 2018. doi: 10.1177/0003122418762291. URL <https://doi.org/10.1177/0003122418762291>.
- A. Rees-Jones. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior*, 2017. ISSN 0899-8256. doi: <https://doi.org/10.1016/j.geb.2017.04.011>. URL <http://www.sciencedirect.com/science/article/pii/S0899825617300751>.
- P. A. Riach and J. Rich. An experimental investigation of sexual discrimination in hiring in the English labor market. *Advances in Economic Analysis & Policy*, 5(2), 2006.
- L. A. Rivera and A. Tilcsik. Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review*, 81(6): 1097–1131, 2016. doi: 10.1177/0003122416668154. URL <https://doi.org/10.1177/0003122416668154>.
- A. Roth. The economics of matching: Stability and incentives. *Mathematics of Operations Research*, pages 617–628, 1982.
- A. Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *The Journal of Political Economy*, 92(6):991–1016, 1984a.
- A. Roth. Misrepresentation and stability in the marriage problem. *Journal of Economic Theory*, 34(2):383–387, 1984b.
- A. Roth. Two-sided matching with incomplete information about others’ preferences. *Games and Economic Behavior*, 1(2):191–209, 1989. ISSN 0899-8256.
- A. Roth. A natural experiment in the organization of entry-level labor markets: Regional markets for new physicians and surgeons in the United Kingdom. *The American Economic Review*, pages 415–440, 1991.
- A. Roth. The nrmp as a labor market. *Journal of the American Medical Association*, 275: 1054–1056, 1996.
- A. Roth. The origins, history, and design of the resident match. *JAMA: The Journal of the American Medical Association*, 289(7):909, 2003.
- A. Roth and E. Peranson. The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *The American Economic Review*, 89(4): 748–780, 1999.

- A. Roth and U. Rothblum. Truncation strategies in matching markets-in search of advice for participants. *Econometrica*, 67(1):21–43, 1999.
- A. Roth and M. Sotomayor. *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1990.
- A. Roth and X. Xing. Jumping the gun: imperfections and institutions related to the timing of market transactions. *The American Economic Review*, 84(4):992–1044, 1994. ISSN 0002-8282.
- A. Roth and X. Xing. Turnaround time and bottlenecks in market clearing: Decentralized matching in the market for clinical psychologists. *Journal of Political Economy*, pages 284–329, 1997.
- J. M. Rzeszotarski, E. Chi, P. Paritosh, and P. Dai. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- J. W. Sherman, F. R. Conrey, and C. J. Groom. Encoding flexibility revisited: Evidence for enhanced encoding of stereotype-inconsistent information under cognitive load. *Social Cognition*, 22(2):214–232, 2004.
- H. Sondak and M. Bazerman. Power balance and the rationality of outcomes in matching markets. *Organizational Behavior and Human Decision Processes*, 50(1):1–23, 1991.
- T. Sönmez. Can Pre-Arranged Matches be Avoided in Two-Sided Matching Markets? *Journal of Economic Theory*, 86(1):148–156, 1999. ISSN 0022-0531.
- W. Suen. A competitive theory of equilibrium and disequilibrium unravelling in two-sided matching. *The Rand journal of economics*, 31(1):101–120, 2000. ISSN 0741-6261.
- M. Turner, M. Fix, and R. J. Struyk. Opportunities denied, opportunities, diminished: racial discrimination in hiring. *Washington, DC: Urban Institute Press*, 1991.
- M. U. Ünver. Backward unraveling over time: The evolution of strategic behavior in the entry level British medical labor markets. *Journal of Economic dynamics and control*, 25(6-7):1039–1080, 2001. ISSN 0165-1889.
- M. U. Ünver. On the survival of some unstable two-sided matching mechanisms. *International Journal of Game Theory*, 33(2):239–254, 2005. ISSN 0020-7276.
- D. H. Wigboldus, J. W. Sherman, H. L. Franzese, and A. v. Knippenberg. Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition*, 22(3): 292–309, 2004.
- L. Zhang and D. Levin. Bounded rationality and robust mechanism design: An axiomatic approach. *American Economic Review*, 107(5):235–39, May 2017. doi: 10.1257/aer.p20171030. URL <http://www.aeaweb.org/articles?id=10.1257/aer.p20171030>.

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