NBER WORKING PAPER SERIES

REASONABLE DOUBT: EXPERIMENTAL DETECTION OF JOB-LEVEL EMPLOYMENT DISCRIMINATION

Patrick M. Kline
Christopher R. Walters

Working Paper 26861
http://www.nber.org/papers/w26861

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2020

This paper previously circulated under the title "Audits as Evidence: Experiments, Ensembles, and Enforcement." We thank Isaiah Andrews, Tim Armstrong, Kerwin Charles, Sendhil Mullainathan, and Andres Santos for helpful conversations related to this project, and Eva Arceo-Gomez, Ray Campos-Vasquez, and John Nunley for providing data. We also thank participants at the UC Berkeley labor and econometrics seminars, the Y-Rise External Validity conference, the 2019 NBER Summer Institute, University of British Columbia, UC Irvine, Harris School of Public Policy, the Tinbergen Institute, the Stockholm School of Economics, the University of Oslo, the 2019 All California Economics Conference, the University of Michigan, Stanford University, and Clemson University for useful feedback. Evan Rose and Benjamin Scuderi provided outstanding research assistance. This project was supported by a Russell Sage Foundation Presidential grant. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination
Patrick M. Kline and Christopher R. Walters
NBER Working Paper No. 26861
March 2020
JEL No. C14,C44,C9,J7,J71,K31,K42

ABSTRACT

This paper develops methods for detecting discrimination by individual employers using correspondence experiments that send fictitious resumes to real job openings. We establish identification of higher moments of the distribution of job-level callback rates as a function of the number of resumes sent to each job and propose shape-constrained estimators of these moments. Applying our methods to three experimental datasets, we find striking job-level heterogeneity in the extent to which callback probabilities differ by race or sex. Estimates of higher moments reveal that while most jobs barely discriminate, a few discriminate heavily. These moment estimates are then used to bound the share of jobs that discriminate and the posterior probability that each individual job is engaged in discrimination. In a recent experiment manipulating racially distinctive names, we find that at least 85% of jobs that contact both of two white applications and neither of two black applications are engaged in discrimination. To assess the potential value of our methods for regulators, we consider the accuracy of decision rules for investigating suspicious callback behavior in various experimental designs under a simple two-type model that rationalizes the experimental data. Though we estimate that only 17% of employers discriminate on the basis of race, we find that an experiment sending 10 applications to each job would enable detection of 7-10% of discriminatory jobs while yielding Type I error rates below 0.2%. A minimax decision rule acknowledging partial identification of the distribution of callback rates yields only slightly fewer investigations than a Bayes decision rule based on the two-type model. These findings suggest illegal labor market discrimination can be reliably monitored with relatively small modifications to existing correspondence designs.

Patrick M. Kline
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720
and NBER
pkline@econ.berkeley.edu

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

1 Introduction

It is illegal to use information on race, sex, or age to make employment decisions in the United States.[1] The voluminous empirical literature on labor market discrimination has focused primarily on establishing whether *markets* discriminate against particular groups of workers on average (Altonji and Blank, 1999; Guryan and Charles, 2013). However, a finding of market-level discrimination provides little guidance to regulators tasked with enforcing anti-discrimination law, who must decide which specific employers to investigate (US Equal Employment Opportunity Commission, 2016). Indeed, classic models of discrimination emphasize that market-level outcomes provide limited guidance regarding the underlying *distribution* of discrimination across employers (Becker, 1957). This paper extends the frontier by developing new tools to characterize discriminatory behavior in markets and detect discrimination by individual employers.

Our approach adapts insights from the literature on empirical Bayes (EB) analysis of large scale testing problems (Efron, 2012) to the study of correspondence experiments that submit fictitious applications with randomly generated characteristics to actual job vacancies (Bertrand and Duflo, 2017 provide a review). Since the influential work of Bertrand and Mullainathan (2004), correspondence experiments typically sample thousands of jobs and send each a few applications with distinctive names that signify race or sex. Our basic insight is that such studies are best viewed as *ensembles* of exchangeable micro-experiments. From the ensemble, one can infer properties of the distribution of discriminatory behavior which can, in turn, be used to form empirical posteriors about the probability that any given job is discriminating.

As in classic EB analyses of count data (Efron and Morris, 1975; Brown, 2008), we treat callback outcomes as independent Bernoulli trials governed by job- and race (or sex)-specific callback probabilities, a modeling choice we show closely approximates callback behavior in correspondence experiments. Because few applications are sent to each job, the distribution of job-specific callback probabilities is under-identified, invalidating standard non-parametric EB approaches (e.g., Efron, 2016). However, we establish identification of a set of moments of the joint distribution of white and black callback probabilities determined by the number of applications sent to each job. To estimate these moments, we propose a Shape-Constrained Generalized Method of Moments (SCGMM) estimator that requires the estimated moments to be consistent with a proper bivariate probability distribution. Applying this estimator to three experimental datasets reveals tremendous heterogeneity across jobs in the extent to which callback probabilities differ by race or sex, along with substantial discriminatory skew: while most jobs barely discriminate, a few discriminate heavily.

Extending classic results on the identification of False Discovery Rates (Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2002), we compute sharp lower bounds on the fraction of jobs engaged in discrimination given the identified moments. In the Bertrand and Mullanaithan experiment, we estimate that at least 13% of jobs discriminate against black applicants. The corresponding estimate in a more recent study by Nunley et al. (2015) is 17%. In a study by Arceo-Gomez

[1] Title VII of the Civil Rights Act of 1964 prohibits employment discrimination on the basis of race and sex, while the Age Discrimination Act of 1975 prohibits certain forms of discrimination on the basis of age.

and Campos-Vasquez (2014), we find that at least 9% of jobs discriminate against women, while at least 19% discriminate against men. These population shares are then used to compute bounds on posterior probabilities that particular jobs are discriminating given their callback patterns. In the Nunley et al. (2015) experiment, we estimate that at least 85% of jobs calling both of two white applicants and neither of two black applicants are engaged in discrimination.

To explore the potential policy implications of our findings, we assess the prospects for systematically detecting discriminators based on callback evidence generated by alternative hypothetical correspondence study designs. This investigation is based on detection/error tradeoffs that arise under a parametric two-type model fit to the Nunley et al. (2015) data. With only two white and two black applications per job, it is difficult to reliably identify discriminating employers. But with only 10 applications per job, we find that a regulator who knows the joint distribution of callback probabilities can correctly identify 7% of discriminating jobs while incurring Type I error rates of less than 0.2%.

Finally, to probe the sensitivity of these conclusions to our modeling assumptions, we consider the problem of a hypothetical regulator who knows only the identified moments of the distribution of callback probabilities. This regulator decides which jobs to investigate using a minimax decision rule that minimizes the maximum risk consistent with the known moments. We develop a tractable approach to estimating the maximum risk function and find that a minimax regulator investigates only slightly fewer jobs than would a Bayesian regulator who knows the joint distribution of callback probabilities. This robustness emerges because the risk function is nearly identified at realistic posterior thresholds that might be used to trigger investigations.

Our results illustrate the potential of experimental methods to assist with regulatory enforcement of anti-discrimination laws. Because employers vary tremendously in their propensity to discriminate against protected groups, regulators face a difficult inferential task. Our findings suggest correspondence experiments can be paired with simple decision rules to reliably identify discriminators. More generally, the methods developed here are applicable to other settings in which analysts seek to make inferences regarding behavioral responses of individual units. Candidate applications include workplace safety audits and laboratory studies of departures from rational choice theory (e.g., Levine et al., 2012 and Halevy et al., 2018).

2 Defining Discrimination

We now develop a formal notion of discrimination tailored to the analysis of correspondence experiments. To simplify exposition we focus on race, which we code as binary ("white" / "black"). Suppose that we have a sample of J jobs with active vacancies. To each of these jobs, we send L_w applications with distinctively white names and L_b applications with distinctively black names as in Bertrand and Mullainathan (2004), for a total of $L = L_w + L_b$ applications. Denote the race associated with the name used in application $\ell \in \{1, ..., L\}$ to job $j \in \{1, ..., J\}$ as $R_{j\ell} \in \{w, b\}$. The function $Y_{j\ell}(r) : \{w, b\} \to \{0, 1\}$ indicates whether job j would call back application ℓ as a function of that applicant's assigned race. Observed callbacks are then given by $Y_{j\ell} = Y_{j\ell}(R_{j\ell})$.

When $Y_{j\ell}(w) \neq Y_{j\ell}(b)$ job j has engaged in racial discrimination with application ℓ. Notably, even if racially distinctive names influence employer behavior only through their role as a proxy for parental background (Fryer and Levitt, 2004), using the names at any point in the hiring process is likely to be viewed by courts as a pretext for discrimination.[2] While courts are typically interested in establishing whether a particular plaintiff experienced discrimination in precisely this sense, we will take the perspective of a regulator tasked with assessing prospectively whether an employer systematically treats applicants differently based upon race. For example, the mission of the US Equal Employment Opportunity Commission (EEOC) is to "***prevent*** and remedy unlawful employment discrimination and advance equal opportunity for all in the workplace" (emphasis added). The following assumption formalizes this prospective notion of discrimination at the employer level.

Assumption 1. *Callbacks are race- and job-specific Bernoulli trials:*

$$Y_{j\ell}(r)\,|R_{j1}...R_{jL} \overset{iid}{\sim} Bernoulli(p_{jr}) \quad for\ r \in \{w, b\}.$$

Note that random assignment of racially distinctive names to applications guarantees independence of $Y_{j\ell}(r)$ from $\{R_{jk}\}_{k=1}^{L}$. The key behavioral restriction in Assumption 1 is that the $\{Y_{j\ell}(r)\}_{\ell=1}^{L}$ are *iid*, which rules out, for example, scenarios in which a job calls back the first qualified applicant and disregards all subsequent applications.[3] We discuss below how to test for such violations. The probability p_{jr} may be interpreted as the callback rate that would emerge in a hypothetical experiment in which a large number of applications of race r are sent to job j.[4]

Letting $C_{jr} = \sum_{\ell=1}^{L} 1\{R_{j\ell} = r\}\, Y_{j\ell}$ denote the number of applications of race r to job j that were called back, Assumption 1 implies the probability $\Pr(C_{jw} = c_w, C_{jb} = c_b | p_{jw}, p_{jb})$ that employer j calls back c_w white applications and c_b black applications is:

$$f(c_w, c_b|p_{jw}, p_{jb}) \quad = \quad \begin{pmatrix} L_w \\ c_w \end{pmatrix} \begin{pmatrix} L_b \\ c_b \end{pmatrix} p_{jw}^{c_w}(1-p_{jw})^{L_w - c_w}\, p_{jb}^{c_b}(1-p_{jb})^{L_b - c_b}. \tag{1}$$

We are now ready to offer a job-level definition of discrimination, which we will henceforth refer to simply as discrimination.

Definition. *Job j engages in discrimination when $p_{jb} \neq p_{jw}$.*

Discriminatory jobs are labeled with the indicator function $D_j = 1\{p_{jb} \neq p_{jw}\}$. This definition is prospective in that an employer with $D_j = 1$ will eventually discriminate against an applicant even if it has not done so yet.

[2]See, e.g., the discussion in U.S. Equal Employment Opportunity Commission v. Target Corporation, 460 F.3d 946, 7th Cir. Wis. 2006 and footnote 27 of Fryer and Levitt (2004).

[3]One could equivalently view such behavior as a violation of our specification of potential outcomes, which builds in the Stable Unit Treatment Value Assumption of Rubin (1980).

[4]If hundreds of applications were sent to a single job the employer would likely be overwhelmed and Assumption 1 would fail. We show below, however, that this assumption provides a suitable approximation to an experiment with 8 applications, which is an unusually large choice of L.

3 Ensembles and Posteriors

The above framework treats each job's callback decisions as a set of race-specific Bernoulli trials. We next consider what can be learned from a collection of experiments conducted at many jobs. This idea is formalized in the following exchangeability assumption on the jobs.

Assumption 2. *Race-specific callback probabilities are independent and identically distributed:*

$$p_{jw}, p_{jb} \overset{iid}{\sim} G(\cdot,\cdot).$$

The distribution function $G(p_w, p_b) : [0,1]^2 \to [0,1]$ describes the population of jobs from which a study samples. In practice, audit studies usually draw small random samples of jobs from online job boards. The *iid* assumption abstracts from the fact that there are a finite number of jobs on these boards. Note that by virtue of random assignment p_{jw} and p_{jb} are independent of the racial mix of applications to job j as well as any other resume characteristics that are randomized.

Assumption 2 implies that the unconditional distribution of callbacks can be expressed as a mixture of binomial trials. We denote the unconditional probability of observing the callback vector (c_w, c_b) by

$$\bar{f}(c_w, c_b) = \int f(c_w, c_b | p_w, p_b)\, dG(p_w, p_b). \tag{2}$$

The distribution $G(\cdot,\cdot)$ will serve as a key object of interest in our analysis. One reason for interest in $G(\cdot,\cdot)$ is that it characterizes both the prevalence and extent of discrimination in a population. For instance, the proportion of jobs that are engaged in discrimination can be written:

$$\bar{\pi} = \Pr(D_j = 1) = \int_{p_w \neq p_b} dG(p_w, p_b).$$

A second reason for interest in $G(\cdot,\cdot)$ lies in its potential forensic value as a tool for identifying which jobs are discriminating. The quantity $\pi(c_w, c_b) = \Pr(D_j = 1 | C_{jw} = c_w, C_{jb} = c_b)$ gives the proportion of jobs with callback vector (c_w, c_b) that are discriminating. Though this quantity has a clear frequentist interpretation as the fraction of discriminators that would be found under repeated sampling, we can also think of it as giving a posterior probability that a job is discriminating given the "evidence" (C_{jw}, C_{jb}). Invoking Bayes' rule, we can write this posterior as a functional of the "prior" $G(\cdot,\cdot)$:

$$\begin{aligned}
\pi(c_w, c_b) &= \frac{\Pr(C_{jw} = c_w, C_{jb} = c_b | D_j = 1)\,\bar{\pi}}{\bar{f}(c_w, c_b)} \\
&= \frac{\bar{\pi}}{\bar{f}(c_w, c_b)} \int_{p_w \neq p_b} f(c_w, c_b | p_w, p_b)\, dG(p_w, p_b) \\
&= \mathcal{P}\Bigg(\underbrace{c_w, c_b}_{\text{direct}}, \underbrace{G(\cdot,\cdot)}_{\text{indirect}}\Bigg).
\end{aligned}$$

The dependence of $\pi(c_w, c_b)$ on $G(\cdot,\cdot)$ is an example of what Efron (2010) refers to as "indirect evidence." To understand the logic of incorporating indirect evidence, suppose $\bar{\pi} = 0$ so that no jobs discriminate. Then $\pi(C_{jw}, C_{jb}) = 0$ with probability one – any seemingly suspicious callback decisions are due to chance. Likewise, if $\bar{\pi} = 1$, all jobs are discriminators and there is no need for direct evidence on the behavior of particular jobs. But in intermediate cases, where some fraction of jobs are discriminators, and some are not, it is rational to blend the direct evidence from a particular job with contextual information on the population from which that job was drawn.

Empirical Bayes approaches seek to form empirical posteriors $\mathcal{P}\left(c_w, c_b, \hat{G}(\cdot,\cdot)\right)$ that substitute the unknown $G(\cdot,\cdot)$ with an estimator $\hat{G}(\cdot,\cdot)$. Important applications of this idea arise in the literature on multiple hypothesis testing, where a key concept is the False Discovery Rate (Benjamini and Hochberg, 1995), which can be thought of as a posterior estimate of the probability that a given null hypothesis is true (Efron et al., 2001; Storey, 2002). In our setting, the False Discovery Rate corresponds to the fraction $1 - \pi(c_w, c_b)$ of jobs with evidence vector (c_w, c_b) that are not discriminating.

4 Identification of G

Each job's realized callback rates $(C_{jw}/L_w, C_{jb}/L_b)$ provide noisy estimates of the latent callback probabilities (p_{jw}, p_{jb}). The binomial structure of this noise is not classical which leads point identification of $G(\cdot,\cdot)$ to fail when the number of applications per job is small.[5] In this section we establish that certain moments of $G(\cdot,\cdot)$ are nonetheless identified by simple linear transformations of unconditional callback probabilities. We then proceed to derive bounds on the posterior probability function $\pi(c_w, c_b)$ consistent with those moments.

Moments

From (2) we can write $\bar{f}(c_w, c_b) = \binom{L_w}{c_w}\binom{L_b}{c_b}\mathbb{E}\left[p_{jw}^{c_w}(1-p_{jw})^{L_w-c_w}p_{jb}^{c_b}(1-p_{jb})^{L_b-c_b}\right]$

$$= \binom{L_w}{c_w}\binom{L_b}{c_b}\sum_{m=0}^{L_w-c_w}\sum_{n=0}^{L_b-c_w}(-1)^{m+n}\binom{L_w-c_w}{m}\binom{L_b-c_b}{n}\mathbb{E}\left[p_{jw}^{c_w+m}p_{jb}^{c_b+n}\right], \tag{3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to $G(\cdot,\cdot)$. Hence, the reduced form callback rates can be written as linear functions of uncentered moments $\mu(m,n) = \mathbb{E}\left[p_{jw}^m p_{jb}^n\right]$ of the latent callback probabilities.

[5] If L were to grow large, one could invoke a normal approximation on each job's sample callback rates and then apply a variance stabilizing transform to make the noise approximately homoscedastic, as in classic EB studies of batting averages (Efron and Morris, 1975; Brown, 2008). With homoscedastic normal estimation error, $G(\cdot,\cdot)$ could then be estimated via deconvolution (e.g., as in Efron, 2016). However, Brown (2008) cautions against using such approximations with 10 or fewer observations per group.

Letting $\bar{f} = \left(\bar{f}(1,0), ..., \bar{f}(L_w,0), ..., \bar{f}(L_w,L_b)\right)'$ denote the vector of frequencies for all possible callback outcomes excluding $(0,0)$ and $\mu = (\mu(1,0), ..., \mu(L_w,0), ..., \mu(L_w,L_b))'$ the corresponding list of moments, we can write the equations in (3) as a linear system $\bar{f} = B\mu$, where B is a known non-singular square matrix of binomial coefficients. Inverting the linear system yields

$$\mu = B^{-1}\bar{f}, \tag{4}$$

which immediately implies the following Lemma.

Lemma 1 (Identification of Moments)**.** *Under Assumptions 1 and 2 and for a given application design (L_w, L_b), all moments $\mu(m,n)$ for $0 \leq m \leq L_w$ and $0 \leq n \leq L_b$ are identified.*

From μ we can compute centered moments of the callback distribution, which are typically easier to interpret. For example, a rudimentary measure of job-level heterogeneity in discriminatory behavior is:

$$\mathbb{V}[p_{jb} - p_{jw}] = \mu(0,2) + \mu(2,0) - 2\mu(1,1) - \mu(0,1)^2 - \mu(1,0)^2 + 2\mu(0,1)\mu(1,0).$$

Lemma 1 implies this variance is identified for any application design that sends at least two resumes per racial group ($\min\{L_w, L_b\} \geq 2$). In experiments where the application design (L_w, L_b) varies randomly across jobs, some moments of $G(\cdot,\cdot)$ will be over-identified. We later exploit these over-identifying restrictions in estimation to improve precision and test our modeling assumptions.

Analytic Bound on Posterior Probabilities

Though the study of moments of the callback distribution $G(\cdot,\cdot)$ can shed light on underlying heterogeneity in callback behavior, the posterior probability $\pi(c_w, c_b)$ need not admit a representation in terms of a finite number of moments. However, a simple analytic bound on the posterior can be derived from an application of Bayes' rule that conditions on the total number of callbacks $C_{jb} + C_{jw}$ to job j. Let $\bar{f}_t(c_w) = \Pr(C_{jw} = c_w, C_{jb} = t - c_w | C_{jb} + C_{jw} = t)$ denote the probability mass function for white callbacks in the stratum of jobs that call back t applicants in total, and let $\bar{f}_t^0(c_w) = \begin{pmatrix} L_w \\ c_w \end{pmatrix}\begin{pmatrix} L_b \\ t - c_w \end{pmatrix} / \begin{pmatrix} L \\ t \end{pmatrix}$ denote the corresponding probability that would arise under Assumptions 1 and 2 if no discrimination were present. Finally, let $\bar{\pi}_t = \Pr(D_j = 1 | C_{jb} + C_{jw} = t)$ denote the share of discriminators among jobs calling back t total applicants.

The following Lemma, which is proved in Appendix A, provides a tractable bound on both the stratum prior $\bar{\pi}_t$ and posterior $\pi(c_w, t - c_w)$.

Lemma 2 (Bounds on Stratum Prior and Posterior)**.**

$$i)\ \bar{\pi}_t \geq \max_{c_w \in \{0,\ldots,t\}} \max \left\{ \frac{\bar{f}_t^0(c_w) - \bar{f}_t(c_w)}{\bar{f}_t^0(c_w)}, \frac{\bar{f}_t(c_w) - \bar{f}_t^0(c_w)}{1 - \bar{f}_t^0(c_w)} \right\},\ t \in \{1, ..., L-1\},$$

$$ii)\ \pi(c_w, t - c_w) \geq 1 - \frac{\bar{f}_t^0(c_w)}{\bar{f}_t(c_w)} \min_{c'_w \in \{0,\ldots,t\}} \min \left\{ \frac{\bar{f}_t(c'_w)}{\bar{f}_t^0(c'_w)}, \frac{1 - \bar{f}_t(c'_w)}{1 - \bar{f}_t^0(c'_w)} \right\},\ t \in \{c_w, ..., L-1\}.$$

Part i) of this Lemma shows that the experiment places a lower bound on the fraction of jobs engaged in discrimination in each callback stratum that is increasing in the discrepancy between the distribution of callback outcomes and the distribution predicted by the non-discrimination null. Part ii) establishes via standard Bayesian updating arguments that the bound on the prior translates into a corresponding lower bound on the posterior.

Sharp Bounds

While the bounds in Lemma 2 are easy to compute, they need not be sharp, as restrictions across callback strata have been ignored. A lower bound on the prior $\bar{\pi}_t$ that exploits all of the logical restrictions in our framework can be written as the solution to the following constrained optimization problem:

$$\min_{G(\cdot,\cdot) \in \mathscr{G}} 1 - \frac{\begin{pmatrix} L \\ t \end{pmatrix}}{\sum_{c'_w=0}^{t} \bar{f}(c'_w, t - c'_w)} \int p^t (1-p)^{L-t} dG(p,p), \tag{5}$$

$$s.t.\ \ \bar{f}(c_w, c_b) = \begin{pmatrix} L_w \\ c_w \end{pmatrix} \begin{pmatrix} L_b \\ c_b \end{pmatrix} \int p_w^{c_w} (1-p_w)^{L_w - c_w} p_b^{c_b} (1-p_b)^{L_b - c_b} dG(p_w, p_b), \tag{6}$$

$$\text{for } (c_w = 0, .., L_w; c_b = 0, .., L_b).$$

To make this problem computationally tractable, we consider a space $\mathscr{G}$ of discretized approximations to the unknown distribution function $G(\cdot,\cdot)$.[6] Because both the objective and constraints are linear in the probability mass function associated with $G(\cdot,\cdot)$, we can apply linear programming (LP) routines to compute bounds given an estimate of the callback probabilities $\left\{\bar{f}(c_w, c_b)\right\}_{c_w, c_b}$.[7] Details of our computational procedure are given in Appendix B.

Note that because the distribution $G(\cdot,\cdot)$ is not indexed by t, the solution to (5) enforces constraints across callback strata. As a result, we may obtain informative bounds on the fraction of discriminatory jobs even among those jobs that call no (or all) applications back.[8]

[6] See Noubiap et al. (2001) for a closely related approach and an asymptotic analysis of the effects of discretization.

[7] Analogous LP formulations can be used to bound any linear functional of $G(\cdot,\cdot)$, including other measures of discrimination. For example, we can bound from below the fraction of employers discriminating against whites by replacing the objective in (5) with $\min_{G(\cdot,\cdot)\in\mathscr{G}} \int_{p_w < p_b} dG(p_w, p_b)$. We leverage this insight to bound a variety of features of $G(\cdot,\cdot)$ in the empirical work to follow.

[8] Suppose, for instance, that $G(\cdot,\cdot)$ is a two type mixture with $\Pr(p_{jw} = 1, p_{jb} = 1) = 1/2$ and $\Pr(p_{jw} = 1/2, p_{jb} = 0) = 1/2$. Then all jobs with zero callbacks are discriminators.

5 Data

We apply our methods to data from three correspondence experiments summarized in Table I. Bertrand and Mullainathan (BM, 2004) applied to 1,112 job openings in Boston and Chicago, submitting four applications to each job. Of the four applications, two were assigned black-sounding names while the remaining two were assigned white-sounding names. The callback rate to applications with black sounding names was 3.1 percentage points lower than to applications with white sounding names.

Table I: Descriptive statistics for resume correspondance studies

		Bertrand & Mullainathan (1)	Nunley et al. (2)	Arceo-Gomez & Campos-Vasquez (3)
Number of jobs		1,112	2,305	802
Applications per job		4	4	8
Treatment/control		Black/white	Black/white	Male/female
Callback rates:	Total	0.079	0.167	0.123
	Treatment	0.063	0.154	0.108
	Control	0.094	0.180	0.138
	Difference	-0.031	-0.026	-0.030
		(0.007)	(0.007)	(0.008)

Notes: This table reports sample characteristics based on data from three resume correspondence experiments. Columns (1) and (2) show statistics from Bertrand and Mullainathan's (2004) and Nunley et al.'s (2015) studies of racial discrimination in the United States. Column (3) reports statistics from Arceo-Gomez & Campos-Vasquez's (2014) study of gender discrimination in Mexico. Standard errors for treatment/control differences, clustered at the job level, are in parentheses.

Nunley et al. (NPRS, 2015) studied racial discrimination in the market for new college graduates by applying to 2,305 listings on an online job board, again sending four resumes per job opening. Unlike BM, the names assigned to the four resumes were sampled without replacement from a pool of eight names, four of which were distinctively black and four of which were distinctively white. This led the fraction of black names sent to each job to vary randomly in increments of 25% from 0% to 100%. The overall callback rate in the NPRS study was more than twice as high as in the BM study, perhaps because the fictitious applicants were more highly educated. On average, black names had a 2.6 percentage point lower callback rate than white names.

Arceo-Gomez and Campos-Vasquez (AGCV, 2014) applied to 802 job openings through an online job portal in a study of race and gender discrimination in Mexico City, Mexico. AGCV sent eight fictitious applications to each job, and the applicants were all recent college graduates. While the AGCV experiment looks at a different context than BM or NPRS, this data set allows us to demonstrate the gains from doubling the number of applications per job opening. To illustrate the identifying power of sending four applications to each protected group, we focus on gender in this experiment, as AGCV used a three-category definition of race.[9] In the AGCV experiment, women were 3.4 percentage points more likely to receive callbacks than men.

6 Are Callbacks Independent Trials?

We begin by considering tests of the Bernoulli trials assumption that undergirds our econometric framework. This assumption would be violated if the likelihood of a callback depends not just on an application's own characteristics but also on the characteristics of other applications sent to the same job. To assess this possibility, we fit linear probability models of the form:

$$Y_{j\ell} = \lambda_0 + X'_{j\ell}\lambda_1 + \bar{X}'_{j\ell}\lambda_2 + \varepsilon_{j\ell}, \tag{7}$$

where $X_{j\ell}$ is a vector of application characteristics and $\bar{X}_{j\ell} = (L-1)^{-1}\sum_{k\neq\ell} X_{jk}$ gives the "leave out" mean of those characteristics among the applications sent to job j excluding application ℓ. While the coefficient vector λ_1 gives the direct effect of application characteristics on callbacks, the coefficient vector λ_2 captures the "peer effect" of other applications to the same job on application ℓ's callback propensity. Assumption 1 restricts these peer effects to be zero ($\lambda_2 = 0$).

For OLS estimates of (7) to identify a causal effect of $\bar{X}_{j\ell}$, we need $\bar{X}_{j\ell}$ to be uncorrelated with any omitted application characteristics $Z_{j\ell}$ that influence callbacks. We therefore focus on the NPRS study which assigned both race and a large number of other application characteristics independently of each other and across applications.[10]

Columns 1 and 2 of Table II report estimates of the parameters in (7) for the NPRS study, with each row showing the coefficients from a separate regression. While applications with distinctively black names are significantly less likely to be called back, we find no significant effect on callback probabilities of changing the racial mix of the other 3 applications to the same job. Across the 12 covariates we consider only one (an indicator for 3+ months of unemployment) finds a significant peer effect at conventional levels, and a joint test fails to reject that all of the leave out mean coefficients are zero ($p = 0.45$). As another composite test, we report the results of a model in which the peer effects are restricted to be proportional to the main effects of the application's own characteristics $X_{j\ell}$. The row titled "predicted callback rate" pools all the application characteristics into an index $X'_{j\ell}\hat{\lambda}_{1(j)}$ where $\hat{\lambda}_{1(j)}$ is the leave out OLS coefficient vector obtained from regressing the

[9] In principle the methods developed here could be extended to a multivariate distribution of callback probabilities for three or more groups. Section 9 explores this possibility by allowing callback rates to differ by resume quality in addition to race or gender.

[10] In contrast, BM assigned application characteristics according to their joint distribution in a training sample, making it likely that the characteristics we study $X_{j\ell}$ are correlated with other omitted characteristics $Z_{j\ell}$ that predict callbacks. The application characteristics were also chosen to yield a good match with the job (see BM p. 996), leading $Z_{j\ell}$ to be correlated with its leave out mean $\bar{Z}_{j\ell}$ and hence with $\bar{X}_{j\ell}$. The AGCV study includes only a small number of randomized resume characteristics that are not predictive of callback outcomes.

callback indicator on application covariates after leaving out all applications to job j. A unit increase in $X'_{j\ell}\hat{\lambda}_{1(j)}$ is associated with roughly half of a callback on average. Though $X'_{j\ell}\hat{\lambda}_{1(j)}$ strongly predicts callbacks, its average value among competing applications $(L-1)^{-1}\sum_{k\neq\ell} X'_{jk}\hat{\lambda}_{1(j)}$ has no statistically discernible impact on callbacks.

Table II: Tests for dependence

Nunley et al. data: resume characteristics		
Variable	Main effect (1)	Leave-out mean (2)
Black	-0.028	-0.019
	(0.010)	(0.027)
Female	0.010	0.009
	(0.010)	(0.027)
High SES	-0.233	-0.674
	(0.174)	(0.522)
GPA	-0.043	-0.153
	(0.066)	(0.198)
Business major	0.008	0.010
	(0.008)	(0.021)
Employment gap	0.011	0.034
	(0.009)	(0.023)
Current unemp.: 3+	0.013	0.005
	(0.012)	(0.032)
6+	-0.008	-0.038
	(0.012)	(0.029)
12+	0.001	0.021
	(0.012)	(0.032)
Past unemp.: 3+	0.029	0.065
	(0.012)	(0.031)
6+	-0.011	-0.016
	(0.012)	(0.033)
12+	-0.004	0.019
	(0.012)	(0.031)
Predicted callback rate	0.476	-0.041
	(0.248)	(0.626)
Joint p-value	0.452	
Sample size	9,220	

Arceo-Gomez & Campos-Vasquez: resume order						
Callbacks	Observations (3)	χ^2 statistic (4)	d.f. (5)	P-value (6)	Exact p-value (7)
Panel A. Four-application sequences					
1	149	2.68	3	0.444	0.592
2	64	8.27	5	0.142	0.249
3	64	3.17	3	0.367	0.513
Panel B. Eight-application sequences					
1	60	10.06	7	0.185	0.319
2	39	27.31	27	0.447	0.516
3	39	61.17	55	0.264	0.287
4	39	67.87	69	0.516	0.595
5	16	40.73	55	0.924	1.000
6	21	29.38	27	0.343	0.390
7	6	8.38	7	0.300	0.539
Panel C. Joint tests					
Independence in all callback strata: χ^2 (247) = 244.9, p = 0.526					
No order effects: χ^2 (7) = 8.3, p = 0.310					
First four minus last four: Difference = 0.011, s.e. = 0.007, p = 0.117					

Notes: This table reports results from tests of the assumption that applications at each job are independent Bernoulli trials with race-specific success probabilities. Columns (1) and (2) show tests based on resume characteristics using data from Nunley et al. (2015). Estimates come from regressions of a callback indicator on a resume characteristic and the mean of this characteristic across other resumes at the same job. The predicted callback rate is the fitted value from a regression of a callback indicator on all resume characteristics, leaving out the reference job. The joint p-value comes from a test of the hypothesis that coefficients on the leave-out mean are zero for all individual characteristics. Standard errors, clustered at the job level, appear in parentheses. Columns (3)-(7) show tests based on resume order in the Arceo-Gomez and Campos-Vasquez (2014) data. These results come from Wald tests of the hypothesis that all callback sequences leading to a particular total number of callbacks are equally likely. Column (3) shows the number of observed sequences in each callback stratum, column (4) shows the Pearson χ^2 goodness of fit statistic, column (5) shows the degrees of freedom for the test, column (6) shows the corresponding p-value, and column (7) shows an exact multinomial goodness of fit p-value obtained by summing probabilities of all sequence configurations that occur with probability less than or equal to the observed configuration under the null. Panel A constructs sequences separately for the first four and last four applications at each job, and panel B uses the full eight application sequence. Panel C shows the results of a joint test of independence across all callback strata, a test that mean callback rates are equal across the eight resume order positions, and a test that the difference in callback rates between the first four and last four resumes equals zero.

A second set of tests for independence exploits data on the specific order in which resumes were sent to jobs in the AGCV experiment (corresponding data were unavailable for BM and NPRS). With independent trials, all callback sequences leading to a particular total number of callbacks t should be equally likely, so each such sequence should constitute a share $\binom{L}{t}^{-1}$ of the sample calling t applications in total. Many plausible forms of dependence would manifest as violations of this condition. If employers stop calling after seeing enough high quality applicants, for example, we should expect to see sequences with runs of callbacks followed by non-callbacks. Likewise, if some employers detect the experiment after receiving several applications, we should see sequences with early callbacks overrepresented and fewer callbacks at later positions in the order.

Columns 3-7 of Table II provide tests of the independence assumption in each callback stratum t of the AGCV data. We form Pearson (1900) χ^2 test statistics equal to quadratic forms in the difference between observed and expected callback sequence frequencies, scaled by the covariance matrix of these differences under the null. Panel A splits the sample of eight applications into two sequences of four at each job in order to increase the expected frequency of each sequence, which may improve the power of the test against certain alternatives. Panel B displays results using the full eight-application sequence. These tests fail to reject the null hypothesis of independence in any callback stratum ($p \geq 0.14$) or across all strata jointly ($p = 0.53$). To focus on alternative hypotheses of particular interest, we also conduct simpler tests for equal callback rates across the eight order positions as well as between the first four and last four positions. These tests again fail to reject independence ($p = 0.31$ and 0.12, respectively). These results indicate that the model of independent Bernoulli trials provides a good approximation to correspondence studies sending eight or fewer applications. We suspect, however, that the quality of this approximation would deteriorate in experiments sending many more applications to each job.

7 Moment and Posterior Estimates

To estimate the identified moments in each experiment we compute shape-constrained GMM (SCGMM) estimates that require the callback frequencies to be rationalizable by a proper discretized probability distribution defined on a 150×150 grid of support points. Imposing shape constraints serves two goals. First, we need the moment estimates to be rationalizable by some $G(\cdot, \cdot) \in \mathscr{G}$ in order to subsequently use them as constraints when estimating bounds via our linear programming method. Second, when the constraints bind, the resulting estimates are typically closer to the truth and more precise (see Chetverikov et al., 2018 for a review). Details of the SCGMM estimation procedure, which involves solving a Quadratic Programming (QP) problem, appear in Appendix C.

Table III uses the shape constrained estimates to summarize key features of the distribution of callback probabilities in each experiment, and reports minimized SCGMM criterion functions (J-statistics) and p-values from bootstrap tests of the shape constraints based on the methods of Chernozhukov et al. (2015). The full set of unconstrained moment estimates appear in Appendix Tables A.I-A.III. Because the shape constraints may make the criterion non-differentiable, we rely on the "numerical bootstrap" procedure of Hong and Li (forthcoming) to construct pointwise valid estimates of standard errors.[11]

Table IV reports LP estimates of the lower bound probability that a given employer is discriminating. In computing both the analytic bounds of Lemma 2 and the sharp bounds of (5), we replace the unknown callback probabilities $\bar{f}$ with estimates $\hat{\bar{f}} = B\hat{\mu}$, where $\hat{\mu}$ is the relevant vector of

[11] Because the asymptotic distribution of the shape constrained estimator will tend to be non-normal (Fang and Santos, 2018), standard errors provide only a heuristic guide to the uncertainty associated with each moment estimate.

Table III: Non-parametric estimates of treatment effect variation in resume correspondence studies

	Bertrand & Mullainathan			Nunley et al.			Arceo-Gomez & Campos-Vasquez		
	p_b	p_w	p_b - p_w	p_b	p_w	p_b - p_w	p_m	p_f	p_m - p_f
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean	0.063	0.094	-0.031	0.153	0.177	-0.023	0.114	0.140	-0.025
	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.005)	(0.009)	(0.009)	(0.008)
Standard deviation	0.152	0.199	0.082	0.290	0.308	0.102	0.231	0.257	0.179
	(0.011)	(0.011)	(0.012)	(0.008)	(0.007)	(0.009)	(0.011)	(0.010)	(0.011)
Correlation with p_w or p_f	0.927	1.000	-0.717	0.944	1.000	-0.336	0.735	1.000	-0.483
	(0.055)	-	(0.089)	(0.018)	-	(0.048)	(0.035)	-	(0.051)
Skewness	-	-	-	3.757	3.648	-4.450	4.067	3.748	-1.403
				(0.074)	(0.087)	(0.405)	(0.140)	(1.161)	(0.385)
Excess kurtosis	-	-	-	-	-	-	8.452	5.756	12.227
							(1.458)	(8.790)	(2.291)
J-statistic:		0.00			23.09			3.33	
P-value:		1.000			0.190			0.790	

Note: This table reports shape-constrained generalized method of moments (SCGMM) estimates of key features of the joint distribution of treatment and control callback rates in three resume correspondence studies. Columns (1)-(3) show estimates for black and white callback rates in Bertrand and Mullainathan (2004), columns (4)-(6) display estimates for black and white callback rates in Nunley et al. (2015), and columns (7)-(9) show estimates for male and female callback rates in Arceo-Gomez and Campos-Vasquez (2014). Standard errors are computed using the numerical bootstrap procedure described by Hong and Li (forthcoming). J-statistics are minimized SCGMM criterion functions. P-values come from bootstrap tests of the hypothesis that the model restrictions are satisfied.

shape-constrained moment estimates produced by our SCGMM procedure. Because the LP algorithm used to solve (5) scales efficiently to large problems, we use a finer discretization with 36 times as many points as the grid used in our earlier SCGMM step.[12]

Bertrand and Mullainathan (2004)

The first rows of columns 1 and 2 of Table III show the mean callback probabilities of white and black applications across jobs. The J-statistic of zero reported in column 2 of Table III indicates that the shape constraints do not bind in the BM data, i.e. that the sample frequencies can be rationalized to numerical precision by a discretized probability distribution. Because the shape constraints do not bind and the BM application design is balanced, the mean callback probabilities match the callback rates reported in Table I perfectly. More interesting are the second moments: there is substantial over-dispersion in callback probabilities, with standard deviations across jobs for each race-specific probability more than double the mean probability. As expected, there is also a strong positive correlation between white and black callback rates, reflecting that some employers simply call back more applications of all types.

Column 3 of Table III reveals substantial heterogeneity in the difference in race specific callback rates $p_{jb}-p_{jw}$ across jobs, with a standard deviation more than twice as large as the mean. The third row shows a strong negative correlation between the discriminatory gap in callback rates $p_{jb}-p_{jw}$ and the white callback probability p_{jw}, suggesting that discrimination tends to be stronger when jobs have higher chances of calling back more white workers. This reflects, in part, a mechanical boundary effect, as an employer with very low callback rates has little opportunity to discriminate. Since the white callback rate in this study is only around 10%, boundary effects are likely to be a quantitatively important phenomenon.

Column 1 of Table IV reports lower bounds on the fraction of jobs engaged in discrimination by the number of total callbacks in the BM experiment. The analytic bounds in Lemma 2 (presented in brackets) imply that at least 38% of the jobs that call back 2 applications are engaged in discrimination, while at least 44% of jobs that call back 3 applications are discriminating. The sharp LP bounds are somewhat tighter than their analytical counterparts, revealing that at least 44% of the jobs calling back two applicants are discriminating. Among jobs that call back three applications, at least half are discriminating on the basis of race. Hence, in this callback stratum, our estimates suggest jobs should not logically be presumed "innocent" of discrimination.

The LP approach also generates informative bounds in callback strata for which analytical bounds are not available. Overall, at least 13% of jobs discriminate on the basis of race. Notably, at least 4% of jobs that call back no applications are engaged in discrimination, while at least 21% of jobs that call back all four applications discriminate on the basis of race. Since neither of

[12]Appendix Table A.IV assesses the sensitivity of our estimates to alternative discretization schemes. The results show that the moment estimates are not sensitive to the number of grid points used in the SCGMM step (as evidenced by the goodness of fit statistic) and that the bounds stabilize with a sufficiently large number of grid points in the linear programming step.

Table IV: Lower bounds on probabilities of discrimination

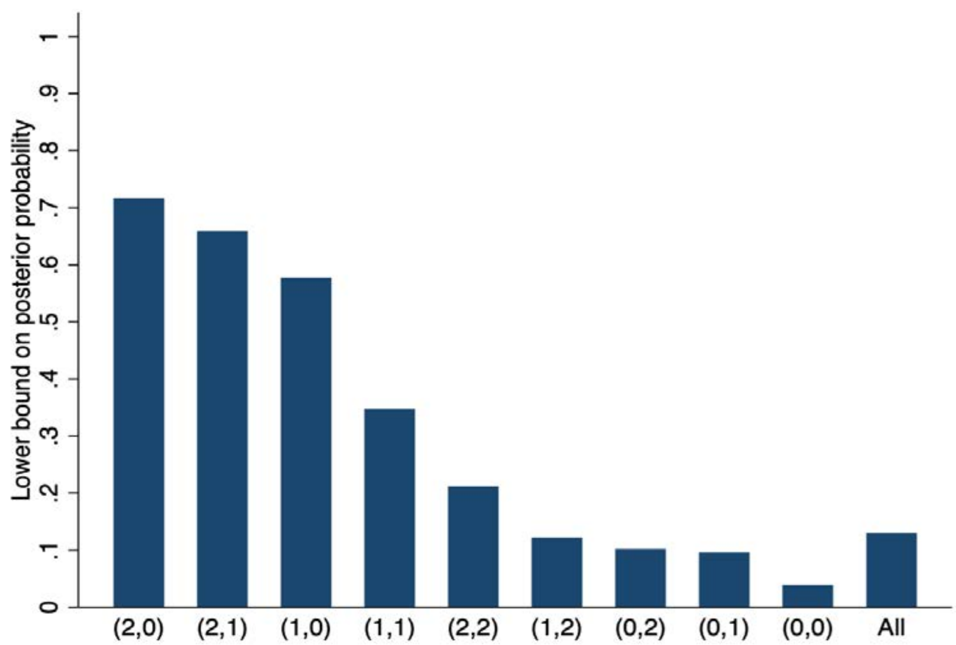
	Bertrand & Mullainathan			Nunley et al.			Arceo-Gomez & Campos-Vasquez		
	$\Pr(p_w \neq p_b)$	$\Pr(p_w < p_b)$	$\Pr(p_b < p_w)$	$\Pr(p_w \neq p_b)$	$\Pr(p_w < p_b)$	$\Pr(p_b < p_w)$	$\Pr(p_f \neq p_m)$	$\Pr(p_f < p_m)$	$\Pr(p_m < p_f)$
Callbacks	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All	0.130	0.000	0.130	0.358	0.154	0.173	0.277	0.089	0.188
0	0.038	0.000	0.038	0.152	0.093	0.048	0.136	0.040	0.095
1	0.424	0.000	0.424	0.672	0.185	0.433	0.895	0.414	0.480
	{0.363}			{0.176}			{0.032}		
2	0.442	0.000	0.442	0.691	0.016	0.675	0.716	0.260	0.456
	{0.379}			{0.282}			{0.540}		
3	0.508	0.000	0.508	0.821	0.067	0.736	0.576	0.047	0.528
	{0.440}			{0.126}			{0.505}		
4	0.212	0.000	0.212	0.421	0.257	0.128	0.503	0.055	0.447
							{0.478}		
5							0.346	0.171	0.175
							{0.313}		
6							0.409	0.212	0.197
							{0.338}		
7							0.486	0.157	0.329
							{0.125}		
8							0.076	0.011	0.065
J-statistic:	29.26	0.00	29.26	62.64	23.46	62.64	369.66	33.88	359.95
P-value:	0.000	1.000	0.000	0.000	0.120	0.000	0.000	0.005	0.000

Notes: This table reports lower bounds on the probability that jobs discriminate based upon race or sex. Bounds are computed via linear programming. Where possible, corresponding analytic bounds based on the formula in Lemma 2 appear in brackets. The first row shows bounds in the population of all jobs, and the remaining rows display bounds conditional on the total number of callbacks. Columns (1)-(3) show results for racial discrimination in the Bertrand and Mullainathan (2004) data, while columns (4)-(6) show results for racial discrimination in the Nunley et al. (2015) data. Columns (1) and (4) display lower bounds on the fraction of jobs with equal callback rates for white and black applicants, columns (2) and (5) report lower bounds on the fraction discriminating against white applicants, and columns (3) and (6) report lower bounds on the fraction discriminating against black applicants. Results for the Nunley et al. (2015) data that condition on the number of callbacks refer to jobs receiving two white and two black applications. Columns (7)-(9) show results for sex discrimination in the Arceo-Gomez and Campos-Vasquez (2014) data. Column (7) reports a lower bound on the fraction of jobs with equal callbacks for men and women, column (8) shows a lower bound on the fraction discriminating against women, and column (9) reports a lower bound on the fraction discriminating against men. J-statistics and p-values come from bootstrap tests of the hypothesis that the lower bound equals zero for all jobs.

these strata exhibit any difference in black-white callback rates, all of the relevant information on discrimination in these strata comes from the total number of callbacks blended with the indirect evidence from the population distribution $G(\cdot,\cdot)$.

Column 2 of Table IV reports LP-based lower bounds on the proportion of jobs with white callback probabilities less than their black callback probabilities. We find a lower bound of exactly zero in each callback stratum, indicating that the callback probabilities can be rationalized without any employers engaging in "reverse discrimination" against whites. Column 3 reports lower bounds on the proportion of jobs with white callback probabilities greater than their black callback probabilities. These lower bound estimates coincide exactly with those reported in column 1. Accordingly, we easily reject the null hypothesis of no discrimination against blacks.

Figure I: Lower bounds on posterior probabilities of discrimination, BM data



Notes: This figure displays lower bounds on the probability that jobs in the Bertrand and Mullainathan (2004) data discriminate based upon race given their callback configurations. Each bar reports a lower bound on the posterior probability of discrimination conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks.

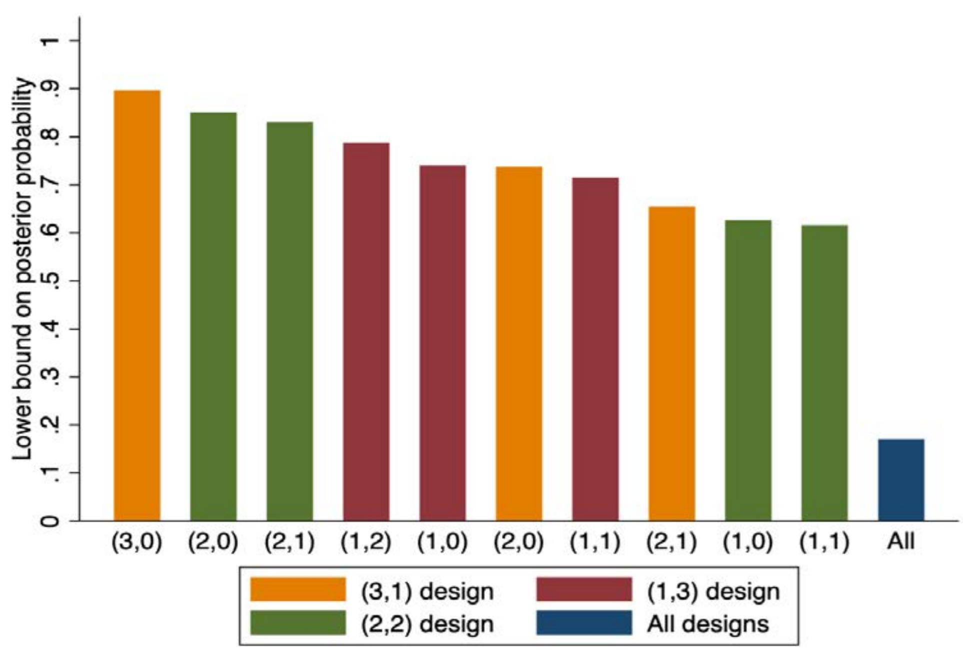
Figure I converts the lower bound estimates in column 3 of Table IV to lower bound posterior probabilities of discrimination. Overall, at least 13% of jobs engage in discrimination. However, at least 72% of jobs that call back two white and no black applications are discriminating, while a job that calls back one white and no black applications has at least a 58% chance of discriminating.

Nunley et al. (2015)

Moment estimates from the NPRS study are reported in columns 4-6 of Table III. Recall that NPRS employed five distinct application designs with $(L_{jw}, L_{jb}) \in \{(4,0),(3,1),(2,2),(1,3),(0,4)\}$. Appendix Table A.II reports design-specific method of moments estimates of all identified moments for the three designs with the largest sample sizes.[13] As expected, the design-specific estimates are generally close to one another and we cannot reject that they are identical. To pool the designs efficiently, we again use an SCGMM estimator that requires the moments be rationalizable by a proper probability distribution $G \in \mathscr{G}$. The minimized SCGMM criterion function provides a measure of the goodness of fit of our model. Applying the bootstrap method of Chernozhukov et al. (2015) yields a p-value of 0.19 for the null hypothesis that the results for all experimental designs are jointly rationalized by a common distribution $G(\cdot,\cdot)$.

Consistent with our findings for the BM data, columns 4-6 of Table III reveal substantial heterogeneity in race-specific callback rates in the NPRS experiment, with standard deviations roughly twice their mean. The imbalanced designs used by NPRS allow us to identify higher moments than the earlier BM study even though the two studies sent the same number of applications per job. While race-specific callback rates are right skewed, racial gaps in callback probabilities $p_{jb} - p_{jw}$ are left-skewed, indicating a long tail of heavy discriminators.

Figure II: Lower bounds on posterior probabilities of discrimination, NPRS data



Notes: This figure displays lower bounds on the probability that jobs in the Nunley et al. (2015) data discriminate against black applicants for the 10 callback configurations with highest posterior bounds. Each bar reports a lower bound on the posterior probability that $p_w > p_b$ conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks. Orange bars correspond to an experimental design with 3 white and 1 black application, green bars correspond to a design with 2 white and 2 black applications, and red bars correspond to a design with 1 white and 3 black applications. The blue bar reports the lower bound on the prior probability of discrimination.

Columns 4-6 of Table IV report estimated lower bounds on the probability of discrimination from the NPRS study for the full population of jobs as well as bounds conditional on total callbacks

[13]The remaining designs were omitted from this analysis due to small sample sizes. Only 22 jobs were in the $(L_{jw} = 0, L_{jb} = 4)$ design while 43 jobs fell in the $(L_{jw} = 4, L_{jb} = 0)$ design.

in a balanced design with $L_{jw} = L_{jb} = 2$. In column 1, our analytic bound formula suggests at least 28% of the jobs calling back two applicants in this design are discriminating – slightly lower than the corresponding estimate in BM. Applying the LP approach tightens the analytic bounds dramatically and provides additional bounds on the prevalence of discrimination among jobs that make no callbacks or that call every application. We estimate that at least 36% of all jobs have different white and black callback probabilities, with that share rising to 69% among employers who call back two applicants in a balanced $(2, 2)$ design.

Some of this discrimination is estimated to be against whites. Column 5 shows that the shape constrained callback probabilities $\hat{\hat{f}}$ imply that at least 15% of employers have white callback probabilities less than their black probabilities. These moments are estimated with error, however, and a bootstrap test of the null hypothesis that all employers have white callback probabilities weakly exceeding their black callback probabilities yields a p-value of 0.12. If we attribute the evidence of reverse discrimination to sampling error, we can take the estimates in column 6 as the relevant lower bounds on discrimination. These results imply that at least 17% of jobs discriminate against black applicants. We decisively reject the null hypothesis that this lower bound is zero.

Figure II converts these lower bound priors into posterior estimates of the share of employers with selected callback configurations engaged in discrimination against black applicants. We estimate that at least 85% of the employers calling back two white and no black applicants in a balanced $(2, 2)$ design are discriminating against blacks. Interestingly, calling back three whites and no blacks in a $(3, 1)$ design is estimated to be even more suspicious, with at least 90% of the employers generating this callback evidence engaged in discrimination against black applicants.

Arceo-Gomez and Campos-Vasquez (2014)

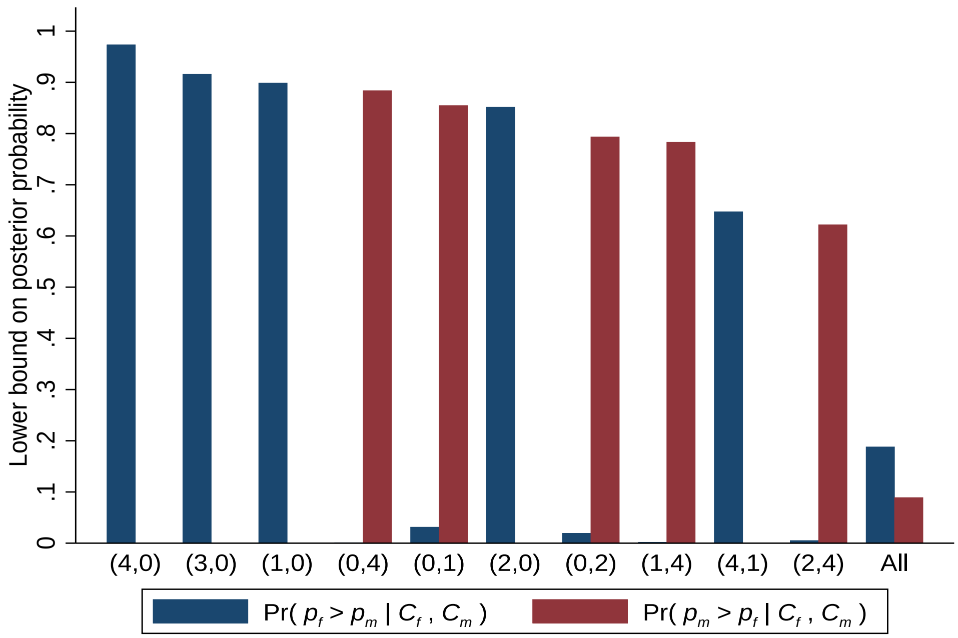
The full set of moment estimates for the AGCV experiment (reported in Appendix Table A.III) reveal that the shape constraints bind strongly in this case, presumably because the design of the AGCV experiment involves many small cells. Despite substantial movement in the moment estimates, the bootstrap p-value from a test of the null hypothesis that the callback frequencies are generated by the model is 0.79, indicating that the raw callback frequencies are rationalizable by a well-behaved underlying joint distribution of callback probabilities.

Columns 7-9 of Table III report key moment estimates from the AGCV data. The behavior of the first two moments is similar to that reported in the prior two experiments, with gender-specific standard deviations roughly twice their mean callback probabilities. However, the greater number of applications used in this design helps enormously with the precision of higher moment estimates.[14] We find strong evidence of left-skew in the distribution of gender gaps in callback probabilities as well as evidence of excess kurtosis in the distribution of gaps. While many jobs discriminate little, there is a thick tail of heavy discriminators.

[14]Though the standard errors reported in Table IV suggest imprecision in our estimates of the higher moments of the female callback rate distribution, this appears to be a consequence of the asymptotic non-normality of the shape-constrained estimator. For example, the numerical bootstrap gives a 90-percent confidence interval of [5.37, 7.49] for the excess kurtosis of p_{jf} while the corresponding standard error equals 8.79.

Columns 7-9 of Table IV report estimated lower bounds on the probability of discrimination in the AGCV experiment. Focusing on the sharp bounds reported in column 7, we find that at least 28% of jobs are engaged in discrimination. Remarkably, this share rises to 90% among jobs calling back one applicant and 72% among jobs calling two. These shares are much higher than the corresponding analytic bounds, showing that cross-stratum restrictions in a design with eight applications are useful for tightening bounds in strata with few callbacks. Evidently, jobs that call back few applicants in the AGCV experiment are very likely to engage in discrimination.

Figure III: Lower bounds on posterior probabilities of discrimination, AGCV data



Notes: This figure displays lower bounds on the probability that jobs in the Arceo-Gomez and Campos-Vasquez (2014) data discriminate based upon sex for the 10 callback configurations with highest posterior bounds. Each bar reports a lower bound on the posterior probability of discrimination conditional on (C_f, C_m), where C_f is the number of female callbacks and C_m is the number of male callbacks. Blue bars report lower bounds on the probability of discriminating against men, and red bars report lower bounds on the probability of discriminating against women.

Some of this discrimination appears to be "reverse" discrimination against women. Column 8 shows that at least 9% of jobs discriminate against women and a bootstrap test of the null hypothesis that this bound equals zero is decisively rejected. An employer that calls back a single application has at least a 41% chance of discriminating against women. Column 9 shows that at least 19% of jobs discriminate against men, and the bootstrap p-value indicates this bound is also statistically distinguishable from zero. The mean difference in callback rates in the ACGV experiment therefore masks gender discrimination operating in both directions. An employer that calls back a single application has at least a 48% chance of discriminating against men.

Figure III plots lower bound posterior probabilities of discrimination against men and women, respectively, for selected callback configurations. At least 97% of the jobs that call back four women and no men are estimated to discriminate against men. But even an employer that calls back a single woman and no men has at least a 90% chance of discriminating against men. Likewise, at least 85% of jobs that call back a single man and no women are estimated to be discriminating against women. That we obtain such strikingly informative posteriors in settings with a single callback demonstrates the tremendous value of indirect evidence in this setting.

8 Experimental Design and Detection Error Tradeoffs

The above analysis demonstrated that it is possible to achieve high posterior certainty that individual jobs are engaged in discrimination when callback rates at those jobs differ dramatically across protected groups. Can such evidence be used to reliably detect a non-trivial share of discriminating jobs? We address this question by studying the tradeoff between Type I and II errors that arises under a simple two-type mixture specification calibrated to match callback rates in the NPRS data given race and other resume characteristics. We then consider how the resulting detection error tradeoffs change as the experimental design is altered to send more applications to each job.

A Mixed Logit Model

We work with a mixed logit model for callbacks of the form:

$$\Pr\left(Y_{j\ell}=1|R_{j\ell},X_{j\ell},\alpha_j,\beta_j\right)=\Lambda\left(\alpha_j-\beta_j 1\left\{R_{j\ell}=b\right\}+X_{j\ell}'\psi\right),$$

where $\Lambda\left(\cdot\right)=\frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the standard logistic CDF, $X_{j\ell}$ is a vector of de-meaned application covariates, and (α_j,β_j) are random coefficients governing the odds of a white callback and discrimination against blacks, respectively. To allow for heterogeneity in white callback rates we assume that $\alpha_j \overset{iid}{\sim} N\left(\alpha_0,\sigma_\alpha^2\right)$. Discrimination is modeled as a two-type (conditional) mixture:

$$\beta_j|\alpha_j=\begin{cases}\beta_0 & \text{w/ prob. } \Lambda\left(\tau_0+\tau_\alpha\alpha_j\right),\\ 0 & \text{w/ prob. } 1-\Lambda\left(\tau_0+\tau_\alpha\alpha_j\right).\end{cases}$$

This specification allows for some fraction of jobs to not discriminate at all, while the remaining jobs depress the odds of calling back blacks relative to whites by roughly β_0%. When $\tau_\alpha\neq 0$, the probability of discrimination depends on α_j, which governs the white callback rate. Note that random assignment of the covariates $X_{j\ell}$ implies they are independent of (α_j,β_j) and therefore excludable from the type probability equation.

Model Estimates

Table V shows the results of fitting the above model to the NPRS experiment by simulated maximum likelihood. Column 1 provides a standard "random effects" logit model with heterogeneity confined to the intercept as in Farber et al. (2016). We find substantial variability across jobs in the overall odds of a callback: a 0.1 standard deviation increase in the intercept α_j is estimated to raise the odds of a callback by 47%. We also find clear evidence of market-wide discrimination: black applications have roughly 46% lower odds of being called back than their white counterparts.

Column 2 allows the race effect β_j to vary across employers, which yields a significant improvement in model fit. The types specification finds that only about 17% of jobs discriminate against blacks – very near the non-parametric lower bound estimate produced earlier by our LP routine

Table V: Mixed logit parameter estimates, NPRS data

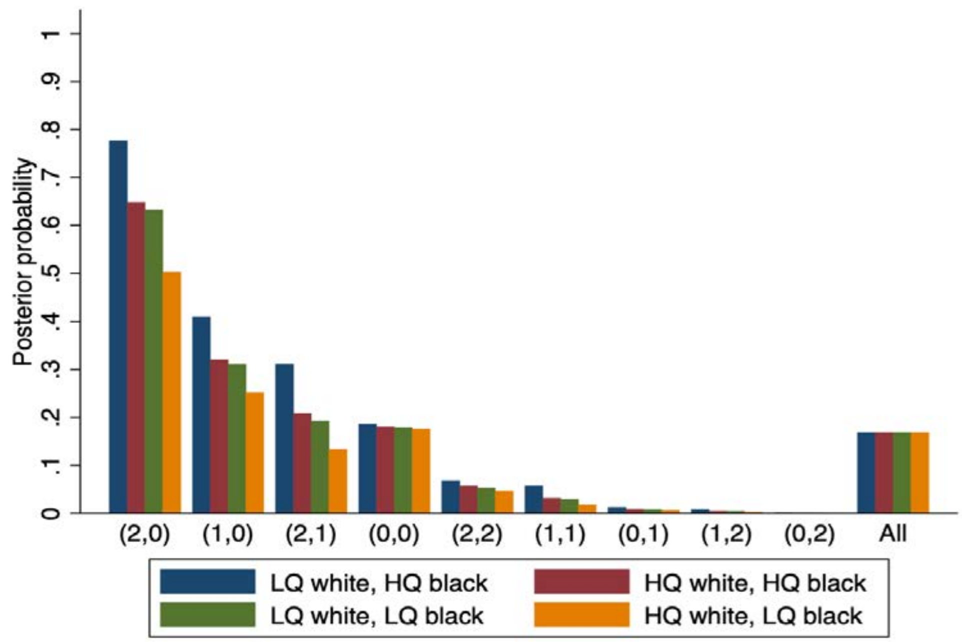
			Types	
		Constant	No selection	Selection
		(1)	(2)	(3)
Distribution of logit(p_w):	α_0	-4.708	-4.931	-4.927
		(0.223)	(0.242)	(0.280)
	σ_α	4.745	4.988	4.983
		(0.223)	(0.249)	(0.294)
Discrimination intensity:	β_0	0.456	4.046	4.053
		(0.108)	(1.563)	(1.576)
Discrimination logit:	τ_0	-	-1.586	-1.556
			(0.416)	(1.098)
	τ_α	-	-	-0.005
				(0.180)
Fraction with $p_w \neq p_b$:		1.000	0.168	0.170
Log-likelihood		-2,792.1	-2,788.2	-2,788.2
Parameters		15	16	17
Sample size		2,305	2,305	2,305

Notes: This table reports simulated maximum likelihood estimates of mixed logit models for callback probabilities in the Nunley et al. (2015) data. Columns (2)-(3) allow for two discrete types of firms, one of which does not discriminate based upon race. All models include resume covariates. Covariates are de-meaned in the estimation sample. Robust standard errors in parentheses.

(see column 6 of Table IV). The degree of discrimination among such jobs is estimated to be severe: the odds of receiving a callback are roughly $\exp(4) - 1 \approx 53$ times higher for white applications than for blacks. Column 3 allows the probability of discrimination to vary with the white callback rate, which yields a negligible improvement in model fit. Surprisingly, α_j and β_j are found to be nearly independent, which implies that the negative correlation between $p_{jb} - p_{jw}$ and p_{jw} reported in Table III is attributable to boundary effects. Again, this model finds roughly 17% of jobs discriminate against blacks. Because we cannot reject the null hypothesis that $\tau_\alpha = 0$, we work with the more parsimonious model in column 2 in the exercises that follow.[15]

[15] Appendix Figure A.I provides a goodness of fit diagnostic for this model, plotting the empirical callback rates in each black / white callback by application design cell against the logit model's predicted callback probability in that cell. The empirical frequencies track the model predictions closely and a naive Pearson χ^2 test fails to reject the null hypothesis that the model rationalizes the cell frequencies up to sampling error.

Figure IV: Mixed logit estimates of posterior discrimination probabilities, NPRS data



Notes: This figure displays mixed logit estimates of the posterior probability that jobs in the Nunley et al. (2015) data discriminate against black workers conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks. Blue bars show posteriors for a design sending two low quality (LQ) white applications and two high quality (HQ) black applications, where low and high quality are defined based on a logit covariate index 1 standard deviation below or above the mean. Red bars show posteriors for a design sending two HQ white and two HQ black applications. Green bars show posteriors for a design sending two LQ white and two LQ black applications. Orange bars show posteriors for a design sending two HQ white and two LQ black applications.

Posteriors

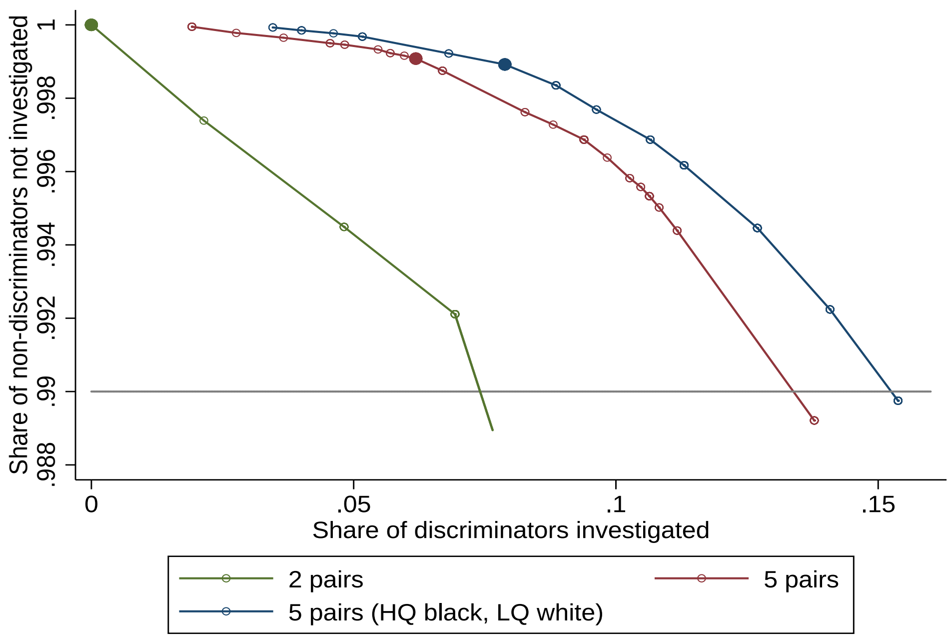
Figure IV reports the distribution of posterior probabilities $\Pr(D_j = 1 | \{Y_{j\ell}, R_{j\ell}, X'_{j\ell}\psi\}_{\ell=1}^{L})$ implied by the parameter estimates reported in column 2 of Table V. To summarize the influence of the covariates, we evaluate the posteriors at two points within each race group, corresponding to the estimated index $X'_{j\ell}\hat{\psi}$ being a standard deviation above or below its empirical mean, which we refer to as "high" and "low" quality applications. By construction, the mean posterior coincides with the estimated fraction of jobs that discriminate. The types model finds that only 17% of jobs are discriminating, yielding a strong prior that the typical job is not violating employment law. Yet calling back only white applicants still justifies a substantial degree of suspicion: 62% of the jobs that call back two whites and no blacks are discriminating.

Imbalances in the covariate mix of applicants can substantially intensify this suspicion. For example, 79% of the jobs that call back two low quality white applications and neither of two high quality black applications are discriminating. Evidently, even in models with a strong presumption of innocence, four applications can provide enough information to cast substantial doubt on whether individual employers are in compliance with employment law. However, it is only under the most extreme callback configurations that we can detect discriminators with reasonable certainty.

Detection Error Tradeoffs

Consider now a hypothetical regulator who forms posteriors taking as prior knowledge the two-type estimates reported in Table V. One may think of the regulator as first learning the parameters of the two-type model from a large experiment and then sending applications to additional vacancies drawn from the same population from which the original study sampled.

Figure V: Detection/error tradeoffs, NPRS data



Notes: This figure displays detection/error tradeoff curves based on models fit to the Nunley et al. (2015) data. Estimates come from decision rules applied to experiments generated from the logit model in column (2) of Table V. The horizontal axis measures the share of discriminating jobs investigated by each decision rule, while the vertical axis measures the share of non-discriminating jobs not investigated. The curves are generated by varying the posterior threshold at which jobs are investigated. The green curve corresponds to an experiment that sends two white and two black applications to each job, and the red curve corresponds to sending five applications of each race. These two curves randomly assign a 2-valued covariate index of resume quality (high or low), defined as +/-1 the empirical standard deviation of this index. The blue curve shows results from sending five low-quality white and five high-quality black applications. Bold points correspond to 80% posterior thresholds.

Figure V displays a rescaling of the Type I and II error rates that arise from investigating all jobs exceeding various posterior thresholds. The horizontal axis gives the share of jobs engaged in discriminating that are investigated. The vertical axis plots the share of non-discriminators that are not investigated. Each point gives the values of these shares corresponding to a particular posterior decision threshold. The bold point corresponds to a posterior threshold of 80%.

In the canonical design with only 4 applications (2 white and 2 black), the 80% posterior threshold yields almost no false accusations. This control over Type I errors comes at the cost of a high Type II error rate – few accusations of any sort are made, leading to a negligible fraction of discriminators detected. Note that conducting a classical hypothesis test (e.g., Fisher's exact test) at the 1% level is equivalent to controlling the fraction of non-discriminators that are not investigated, which is depicted by the horizontal line at 0.99. This rule would yield more investigations but most of these would be erroneous: the equivalent posterior threshold in the 2 pair design is only 33%.

Expanding the design to 5 pairs of applications yields a substantial outward shift in the detection error tradeoff curve. Using a posterior threshold of 80% keeps the fraction of employers erroneously investigated for discrimination below 0.2% while allowing detection of roughly 7.5% of jobs that discriminate. Lowering the posterior threshold further boosts the detection rate above 10% while modestly increasing the Type I error rate. Evidently, ten applications enables accurate detection of a non-trivial fraction of discriminators.

The third line shows the results of an experiment where each job is sent 5 high quality black applications and 5 low quality white applications.[16] Modifying the experimental design in this way yields additional improvements in Type I and II error rates. Using an 80% posterior threshold, the share of non-discriminators investigated remains below 0.2%, while the share of discriminators investigated rises to roughly 10%.

9 Indirect Evidence and Policy

The findings of the previous section indicate that correspondence experiments with as few as 10 applications can reliably detect a substantial share of discriminating employers when the population distribution of callback probabilities is known. We now consider how a Bayesian regulator might go about deciding which jobs to investigate and then assess how partial identification of the callback rate distribution affects this decision rule.[17]

The Regulator's Problem

Suppose the regulator must decide which jobs to investigate based on a vector of direct evidence $\mathcal{E}_j = \{Y_{j\ell}, R_{j\ell}, X_{j\ell}\}_{\ell=1}^{L}$ revealed by a correspondence study. The regulator uses a deterministic decision rule $\delta(\mathcal{E}_j)$ that maps this evidence vector to a binary inquiry decision.[18] Each job has a pair of race specific callback probabilities $\{p_{jw}(x), p_{jb}(x)\}$ that may vary with applicant quality x. We use $H(\cdot)$ to denote the *iid* randomization distribution of $X_{j\ell}$. Consistent with our earlier analysis of the NPRS experiment, we rule out the possibility of discrimination against whites by assuming $\Pr(p_{jw}(x) \geq p_{jb}(x)) = 1$ for all quality levels $x \in \mathcal{X}$.

[16] Of course, the results of such an experiment would be difficult to interpret without an earlier experiment revealing the model parameters, as one would not be able to parse the effects of race from those of quality.

[17] Our analysis is based loosely on the experience of the EEOC, which has the authority to conduct systematic investigations into the discriminatory behavior of particular organizations. Because investigations are costly, the EEOC uses a priority system based on human judgement to decide which complaints to investigate (US Equal Employment Opportunity Commission, 2016). The results below illustrate how direct callback evidence from a correspondence experiment can be blended with indirect evidence to assist or replace informal human judgements.

[18] We confine ourselves to deterministic rules because randomized decision rules violate commonly held horizontal equity principles.

The regulator's loss from applying decision rule $\delta(\mathcal{E}_j) = \delta_j \in \{0, 1\}$ to job j is modeled as:

$$\mathcal{L}_j(\delta_j) = \delta_j \left(\kappa - \Lambda \left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x)) \right] dH(x) \right) \right). \tag{8}$$

One can think of the parameter $\kappa \in (1/2, 1]$ as capturing the cost of conducting an investigation. The term $\Lambda\left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x))\right] dH(x)\right) \in [1/2, 1]$ gives the benefit to the investigation, which is increasing in the average log callback odds advantage of whites over blacks at job j across quality levels. The racial difference in log odds is then mapped back to the unit interval by the logistic CDF to produce the payoff to an investigation. The regulator would like to investigate whenever this payoff exceeds the investigation cost, in which case $\mathcal{L}_j(1)$ is negative.

Because the $\{p_{jw}(x), p_{jb}(x)\}_{x \in \mathcal{X}}$ are not known, the regulator minimizes expected loss (i.e. risk). When the regulator knows the joint distribution of callback probabilities in the population, the risk function can be written:

$$\mathcal{R}_j(G, \delta(\cdot)) = \mathbb{E}[\mathcal{L}_j(\delta_j)] = \mathbb{E}\left[\delta_j(\mathcal{E}_j)\left(\kappa - \Lambda\left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x))\right] dH(x)\right)\right)\right],$$

where $G : [0,1]^{2|\mathcal{X}|} \to [0, 1]$ is the joint distribution of quality-specific callback rates. Because $\{\mathcal{E}_j, p_{jw}(x), p_{jb}(x)\}_{j=1}^{J}$ are *iid* across jobs, we can drop the j subscripts and refer to the risk function as $\mathcal{R}(G, \delta(\cdot))$. Choosing $\delta(\mathcal{E}_j)$ to minimize the risk function pointwise yields the following Lemma, which characterizes the regulator's optimal decision rule in the case where G is known.

Lemma 3 (Optimal Decision Rule)**.** .

$$\delta(\mathcal{E}_j) = 1\left\{\mathbb{E}\left[\Lambda\left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x))\right] dH(x)\right) |\mathcal{E}_j\right] > \kappa\right\} \text{ minimizes } \mathcal{R}(G, \delta).$$

One can think of Lemma 3 as offering an economically motivated standard of *reasonable doubt*: when the posterior expected benefit of an investigation exceeds the investigation cost κ, it is rational to conduct an investigation. Note that in the two-type logit model the difference in log odds at job j equals $\beta_0 D_j$ for all quality levels, so the optimal decision rule amounts to investigating when the posterior probability of discrimination exceeds a cost-based threshold.[19]

Ambiguity

When G is only known to lie in some identified set Θ of distributions, many possible decision rules are consistent with rationality. Among those rules, an important benchmark is the minimax decision rule (Wald, 1945; Savage, 1951; Manski, 2000), which minimizes the maximum risk that may arise from the regulator's decisions. We can define the maximum risk function and the associated

[19]Specifically, in the logit model we have $\delta(\mathcal{E}_j) = 1\{\mathcal{P}(\mathcal{E}_j, G_{logit}) > (\kappa - 1/2)/(\Lambda(\beta_0) - 1/2)\}$.

minimax decision rule respectively as:

$$\mathcal{R}^m(\Theta,\delta) = \sup_{G\in\Theta} \mathcal{R}(G,\delta) \quad \text{and} \quad \delta^{mm} = \arg\inf_{\delta\in\mathscr{D}} \mathcal{R}^m(\Theta,\delta), \tag{9}$$

where $\mathscr{D}$ is the set of deterministic decision rules. Unlike in the case where G is known, a regulator that only knows $G \in \Theta$ cannot consult a single posterior expectation to make the decision of whether to investigate. Rather, the maximum risk of each decision rule must be computed to obtain the minimax decision rule.

Relying on a discretized function space for G simplifies computation of the maximum risk function $\mathcal{R}^m$ consistent with a set of experimental callback probabilities. As explained in Appendix D, when Θ consists of a family of discrete distributions, $\mathcal{R}^m(\Theta,\delta)$ can be computed numerically as the solution to a linear programming problem. The minimax decision rule $\delta^{mm}(\cdot)$ is found by computing $\mathcal{R}^m(\Theta,\delta)$ for each candidate rule $\delta \in \mathscr{D}$ and choosing the rule that yields lowest maximal risk.

Bayes vs. Minimax Decisions

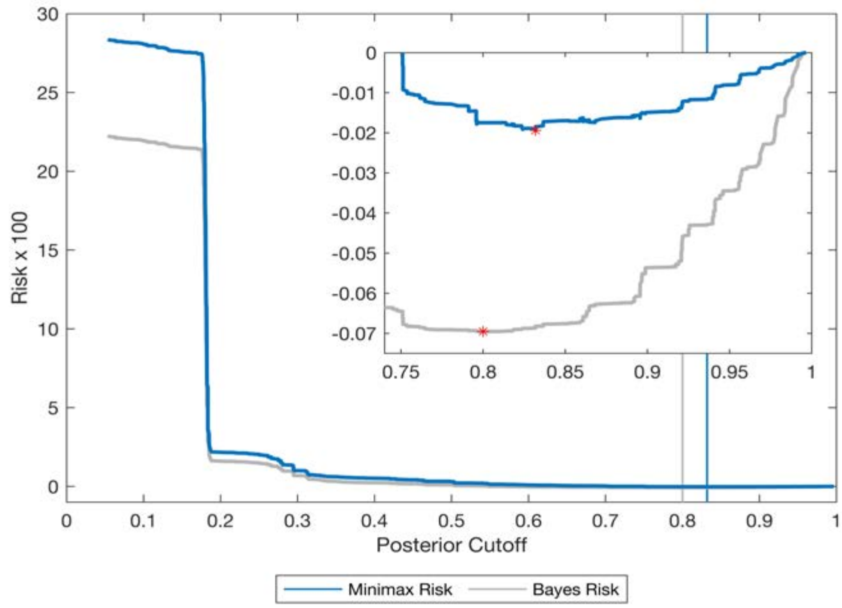
We now compare the decisions made by a Bayesian regulator with a minimax regulator in the hypothetical 5-pair generalization of the NPRS experiment considered in Section 8. As in Figure IV, we assume applications take on only two quality levels (high or low) with equal probability.

We consider a restricted family $\mathscr{D}^\dagger \subset \mathscr{D}$ of decision rules of the form $\delta(\mathcal{E}_j) = 1\{\mathcal{P}(\mathcal{E}_j, G_{logit}) \geq q\}$, where $q \in (0,1)$ is a posterior cutoff and G_{logit} is the logit model reported in column 2 of Table V. Computing the maximal risk for this family of decision rules can be thought of as a way of "second guessing" the risk associated with each logit posterior threshold without debating the logit model's ordering of the underlying evidence configurations.[20] In computing $\mathcal{R}^m(\Theta,\delta)$, we use the logit model predictions of callback probabilities within each of the two quality bins as constraints (see the Appendix for details) and calibrate κ so that, under the logit DGP, an 80% posterior threshold minimizes risk.

Figure VI plots logit (i.e., Bayes) risk and $\mathcal{R}^m(\Theta,\delta)$ against the nominal logit posterior threshold q. As q approaches one, both the maximal and Bayes risks approach zero, as no jobs are investigated in the limit. Conversely as the posterior threshold approaches zero – at which point all jobs are investigated – the maximum risk diverges from the Bayes risk because the least favorable G is one where nearly all jobs are engaged in trivial levels of discrimination that fail to justify the investigation cost. Recall from Table IV, however, that some jobs in the NPRS experiment must be discriminating, which limits the magnitude of this divergence.

[20]With L_w white and L_b black applications there are $2^{(1+L_w)(1+L_b)}$ logically possible decision rules which, in practice, prohibits brute force enumeration when $L_w + L_b > 4$. Restricting attention to logit posterior threshold rules allows us to circumvent this obstacle. In cases where multiple evidence configurations yield the same logit posterior, we consider separate rules that investigate each of these configurations individually.

Figure VI: Bayes and minimax risk, NPRS data



Notes: This figure displays risk functions generated by a hypothetical experiment sending five white and five black applications to jobs in the population studied by Nunley et al. (2015). Resumes are randomly assigned to high or low quality with equal probability, where quality is defined as +/-1 the empirical standard deviation of the logit covariate index. The horizontal axis plots the posterior threshold at which jobs are accused of discrimination. The grey curve displays risk based on the logit data generating process in column (2) of Table V. The cost of an investigation is calibrated so that a Bayesian regulator sets the posterior cutoff at 0.8. The blue curve plots minimax risk calculated by choosing the joint distribution of callback probabilities to maximize risk for each decision rule subject to the moments identified by the Nunley et al. (2015) experiment, restricted to decision rules that order jobs by the logit posterior threshold. Vertical lines indicate risk-minimizing thresholds. The window displays logit and minimax risk at threesholds of 0.75 and above, with risk-minimizing points indicated in red.

While the Bayes risk function is minimized by the decision rule with a posterior threshold of 80%, $\mathcal{R}^m(\Theta, \delta)$ is minimized by a rule with a logit-based threshold of 83%. This higher threshold implies a minimax regulator would investigate fewer jobs than a Bayesian regulator with the same preferences who believes G to be logit. The change in behavior is minimal, however, suggesting that, at least for this specification of preferences, the Bayes decision rule is relatively robust to ambiguity over the nature of the true DGP.[21]

10 Conclusion

Correspondence studies are powerful tools that have been extensively used to detect market-level averages of discriminatory behavior. Revisiting three such studies, we find tremendous heterogeneity across jobs in the degree of discrimination. This heterogeneity presents authorities charged with enforcing anti-discrimination laws with a difficult inferential task. Our analysis suggests that when ensemble evidence is used, sending 10 applications per job enables accurate detection of a

[21]As shown in Appendix Figure A.II, however, had we considered a more aggressive Bayesian benchmark (e.g., a 40% threshold), the minimax and Bayes rules would have departed more substantially, as the maximum risk can greatly exceed the Bayes risk.

non-trivial share of discriminatory employers. This finding opens the possibility that discrimination can be monitored – perhaps in real time – at the employer level.

Our results also provide a number of methodological lessons regarding the design and analysis of correspondence studies, and of experimental ensembles more generally. First, we demonstrate that indirect evidence can serve as a valuable supplement to direct evidence even when heterogeneity distributions are not point identified. Using a few moments of the callback rate distribution in conjunction with only four applications per job, we derived informative lower bounds on the fraction of jobs engaging in illegal discrimination. In the AGCV study, which sent eight applications to each job, we deduced informative lower bound rates of discrimination against men and women separately.

Second, our results highlight that the appropriate use of indirect evidence depends critically on the objectives of the investigator, formalized in our framework by the loss function of a hypothetical regulator. While in point identified settings it is straightforward to characterize the tradeoffs presented by different decision rules, partial identification of heterogeneity distributions tends to undermine identifiability of this tradeoff itself. In our setting acknowledging the ambiguity stemming from partial identification turns out to lead to only slightly more conservative decisions with a realistic loss function. An important topic for future research is the extent to which the policy implications of recent econometric evaluations of teachers, schools, hospitals, and neighborhoods (e.g., Chetty et al., 2014; Angrist et al., 2017; Hull, 2018; Chetty and Hendren, 2018; Chetty et al., 2018) vary with alternative notions of risk.

References

7th Circuit Court of Appeals (2006): "EEOC v. Target Corp." 460 (F. 3d), 946.

Altonji, J. G. and R. M. Blank (1999): "Race and gender in the labor market," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter and D. Card, Elsevier, vol. 3C, chap. 48, 3143–3259.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging lotteries for school value-added: testing and estimation," *Quarterly Journal of Economics*, 132, 871–919.

Arceo-Gomez, E. O. and R. M. Campos-Vasquez (2014): "Race and marriage in the labor market: a discrimination correspondence study in a developing country," *American Economic Review: Papers & Proceedings*, 104, 376–380.

Becker, G. S. (1957): *The Economics of Discrimination*, The University of Chicago Press.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289–300.

Bertrand, M. and E. Duflo (2017): "Field experiments on discrimination," in *Handbook of Field Experiments*, ed. by E. Duflo and A. Banerjee, Elsevier, vol. 1.

Bertrand, M. and S. Mullainathan (2004): "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 94, 991–1013.

Brown, L. D. (2008): "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," *The Annals of Applied Statistics*, 2, 113–152.

Chernozhukov, V., W. K. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *arXiv preprint arXiv:1509.06311*.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018): "The opportunity atlas: mapping the childhood roots of social mobility," Working paper.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the impact of teachers II: teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R. and N. Hendren (2018): "Impacts of neighborhoods on intergenerational mobility II: county-level estimates," *Quarterly Journal of Economics*, 133, 1163–1228.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018): "The econometrics of shape restrictions," *Annual Review of Economics*, 10, 31–63.

Efron, B. (2010): "The future of indirect evidence," *Statistical Science*, 25, 145–157.

——— (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.

——— (2016): "Empirical Bayes deconvolution estimates," *Biometrika*, 103, 1–20.

Efron, B. and C. Morris (1975): "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Fang, Z. and A. Santos (2018): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412.

Farber, H. S., D. Silverman, and T. von Wachter (2016): "Determinants of callbacks to job applications: an audit study," *American Economic Review: Papers & Proceedings*, 106, 314–318.

Fryer, R. G. and S. D. Levitt (2004): "The causes and consequences of distinctively black names," *Quarterly Journal of Economics*, 119, 767–805.

Guryan, J. and K. K. Charles (2013): "Taste-based or statistical discrimination: the economics of discrimination returns to its roots," *The Economic Journal*, 123, F417–F432.

Halevy, Y., D. Persitz, and L. Zrill (2018): "Parametric recoverability of preferences," *Journal of Political Economy*, 126, 1558–1593.

Hong, H. and J. Li (forthcoming): "The numerical bootstrap," *Annals of Statistics*.

Hull, P. D. (2018): "Estimating hospital quality with quasi-experimental data," Working paper.

Levine, D. I., M. W. Toffel, and M. S. Johnson (2012): "Randomized government safety inspections reduce worker injuries with no detectable job loss," *Science*, 336, 907–911.

Manski, C. F. (2000): "Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice," *Journal of Econometrics*, 95, 415–442.

Noubiap, R. F., W. Seidel, et al. (2001): "An algorithm for calculating Γ-minimax decision rules under generalized moment conditions," *The Annals of Statistics*, 29, 1094–1116.

Nunley, J. M., A. Pugh, N. Romero, and R. A. Seals (2015): "Racial discrimination in the labor market for recent college graduates: evidence from a field experiment," *B.E. Journal of Economic Analysis and Policy*, 15, 1093–1125.

Pearson, K. (1900): "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine, Series 5*, 50, 157–175.

Rubin, D. B. (1980): "Randomization analysis of experimental data: the Fisher Randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

Savage, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

Storey, J. D. (2002): "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.

US Equal Employment Opportunity Commission (2016): "Advancing opportunity: a review of the systematic program of the US Equal Employment Opportunity Commission," Technical report.

Wald, A. (1945): "Statistical decision functions which minimize the maximum risk," *Annals of Mathematics*, 46, 265–280.

Appendix A: Proof of Lemma 2

By the law of total probability the fraction of jobs calling c_w white and $t - c_w$ black applications among those calling t total can be written:

$$\bar{f}_t(c_w) = (1 - \bar{\pi}_t)\bar{f}_t^0(c_w) + \bar{\pi}_t \bar{f}_t^1(c_w),$$

where $\bar{f}_t^d(c_w) = \Pr(C_{jw} = c_w | C_{jw} + C_{jb} = t, D_j = d)$ for $d \in \{0, 1\}$. Since $\bar{f}_t^1(c_w) \in [0, 1]$ we have

$$\bar{f}_t(c_w) \geq (1 - \bar{\pi}_t)\bar{f}_t^0(c_w), \;\; \bar{f}_t(c_w) \leq (1 - \bar{\pi}_t)\bar{f}_t^0(c_w) + \bar{\pi}_t,$$

which implies

$$\bar{\pi}_t \geq \max\left\{\tfrac{\bar{f}_t^0(c_w) - \bar{f}_t(c_w)}{\bar{f}_t^0(c_w)}, \tfrac{\bar{f}_t(c_w) - \bar{f}_t^0(c_w)}{1 - \bar{f}_t^0(c_w)}\right\}.$$

Taking the maximum of these lower bounds over $c_w \in \{0, ..., t\}$ yields the bound on $\bar{\pi}_t$ in part i) of Lemma 2.

By Bayes' rule the fraction of discriminators among jobs calling c_w white and $t - c_w$ black applications is given by:

$$\pi(c_w, t - c_w) = 1 - \frac{\bar{f}_t^0(c_w)(1 - \bar{\pi}_t)}{\bar{f}_t(c_w)}.$$

Plugging the bound on $\bar{\pi}_t$ from part i) of the Lemma into this expression gives the bound on $\pi(c_w, t - c_w)$ in part ii).

Appendix B: Discretization of G and Linear Programming Bounds

To compute the solution to the problem in (5), we approximate the CDF $G(p_w, p_b)$ with the discrete distribution

$$G_K(p_w, p_b) = \sum_{k=1}^{K}\sum_{s=1}^{K} \eta_{ks} 1\{p_w \leq \varrho(k, s), p_b \leq \varrho(s, k)\},$$

where the $\{\eta_{ks}\}_{k=1,s=1}^{K,K}$ are probability masses and $\{\varrho(k, s), \varrho(s, k)\}_{k=1,s=1}^{K,K}$ comprise a set of mass point coordinates generated by the function

$$\varrho(k, s) = \underbrace{\frac{\min\{k, s\} - 1}{K}}_{\text{diagonal}} + \underbrace{\frac{\max\{0, k - s\}^2}{K(1 + K - y)}}_{\text{off-diagonal}}.$$

This discretization scheme can be visualized as a two-dimensional grid containing K^2 elements. The diagonal entries on the grid represent jobs where no discrimination is present. The first term above ensures the mass points are equally spaced along the diagonal from $(0, 0)$ to $\left(\frac{K-1}{K}, \frac{K-1}{K}\right)$. The second term spaces off diagonal points quadratically according to their distance from the diagonal in order to accomodate jobs with very low levels of discrimination while economizing on the number of grid points. We use a spacing scheme that places more points near the diagonal because we are particularly interested in the mass exactly on the diagonal. Note that $\lim_{K\to\infty} \varrho(K,s) = 1$, ensuring the grid asymptotically spans the unit square.

With this notation, the constraints in (6) can be written:

$$\bar{f}(c_w, c_b) = \begin{pmatrix} L_w \\ c_w \end{pmatrix} \begin{pmatrix} L_b \\ c_b \end{pmatrix} \sum_{k=1}^{K} \sum_{s=1}^{K} \eta_{ks} \varrho(k,s)^{c_w} (1 - \varrho(k,s))^{L_w - c_w} \varrho(s,k)^{c_b} (1 - \varrho(s,k))^{L_b - c_b}, \tag{10}$$

for $c_w = (1, ..., L_w)$ and $c_b = (0, ..., L_b)$. Hence, our composite discretized optimization problem is to

$$\min_{\{\eta_{ks}\}} 1 - \frac{\begin{pmatrix} L \\ t \end{pmatrix}}{\sum_{(c'_w, c'_b): c'_w + c'_b = t} \bar{f}(c'_w, c'_b)} \sum_{k=1}^{K} \eta_{kk} \varrho(k,k)^t (1 - \varrho(k,k))^{L-t},$$

subject to (10) and

$$\sum_{k=1}^{K} \sum_{s=1}^{K} \eta_{ks} = 1, \quad \eta_{ks} \geq 0,$$

for $k = 1, ..., K$ and $m = 1, ..., K$. We solve this problem numerically using the Gurobi software package. Because setting K too low will tend to yield artificially tight bounds, we set $K = 900$ in all bound computation steps, which yields $(900)^2 = 810,000$ distinct mass points.

Appendix Table A.IV reports linear programming bounds for various choices of K. As expected, the bounds stabilize with a sufficiently large K, and the quadratic spacing described above produces more accurate results than an equally-spaced grid: we obtain similar estimates for a quadratic grid with 300^2 grid points and a rectangular grid with 900^2 points.

Appendix C: Shape Constrained GMM

To accomodate the Nunley et al. (2015) study which employs multiple application designs, we introduce the variable $L_j = (L_{jw}, L_{jb})$ which gives the number of white and black applications sent to job j. Collecting the design-specific callback probabilities $\{\Pr(C_{jw} = c_w, C_{jb} = c_b | L_j = l)\}_{c_w, c_b}$ into the vector $\bar{f}_l$, our model relates these probabilities to moments of the callback distribution via the linear system $\bar{f}_l = B_l \mu$, for B_l a fixed matrix of binomial coefficients. Letting $\bar{f}$ denote the vector formed by "stacking" the $\{\bar{f}_l\}$ across designs in an experiment, we write $\bar{f} = B\mu$. Let η be a $K^2 \times 1$ vector comprised of the probability masses $\{\eta_{ks}\}_{k=1,s=1}^{K,K}$ (see Appendix B). For GMM estimation we set $K = 150$ (larger values yield very similar results). From (3), we can write $\mu = M\eta$ where M is a $dim(\mu) \times K^2$ matrix comprised of entries with typical element $\varrho(k,s)^m \varrho(s,k)^n$. We then have the moment restriction $\bar{f} = BM\eta$.

Let $\tilde{f}$ denote the vector of empirical call back probabilities with typical element:

$$\frac{J^{-1}\sum_{j=1}^{J}1\left\{C_{jw}=c_w,C_{jb}=c_b,L_j=l\right\}}{J^{-1}\sum_{j=1}^{J}1\{L_j=l\}}.$$

Our shape constrained GMM estimator of η can be written as the solution to the following quadratic programming problem:

$$\hat{\eta}=\arg\inf_{\eta}\,(\tilde{f}-BM\eta)'W(\tilde{f}-BM\eta) \tag{11}$$

$$\text{s.t.}\quad \eta\geq 0,\ \mathbf{1}'\eta=1,$$

where W is a fixed weighting matrix. Note that because $G(\cdot,\cdot)$ is not identified, there are many possible solutions $\hat{\eta}$ to this problem, but these solutions will all yield the same values of $BM\hat{\eta}$. Our shape constrained estimate of the moments is $\hat{\mu}=M\hat{\eta}$ while our estimator of the callback probabilities is $\hat{\bar{f}}=BM\hat{\eta}$. We follow a two-step procedure, solving (11) with diagonal weights proportional to the number of jobs used in the application design and then choosing $W=\hat{\Sigma}^{-1}$ where $\hat{\Sigma}=\text{diag}\left(\hat{\bar{f}}^{(1)}\right)-\hat{\bar{f}}^{(1)}\hat{\bar{f}}^{(1)\prime}$ is an estimate of the variance-covariance matrix of the callback frequencies implied by the first step shape-constrained callback probability estimates $\hat{\bar{f}}^{(1)}$.

Hong and Li (forthcoming) standard errors

Standard errors on the moment estimates $\hat{\mu}$ are computed via the numerical bootstrap procedure of Hong and Li (forthcoming) using a step size of $J^{-1/4}$ (we found qualitatively similar results with a step size of $J^{-1/3}$). Our implementation of the numerical bootstrap proceeds as follows: the bootstrap analogue μ^* of $\hat{\mu}$ solves the quadratic programming problem in (11) where $\tilde{f}$ has been replaced by $\left(\tilde{f}+J^{-1/4}f^*\right)$. The bootstrap probabilities f^* have typical element:

$$J^{1/2}\left(\frac{\sum_{j=1}^{J}\omega_j^*1\left\{C_{jw}=c_w,C_{jb}=c_b,L_j=l\right\}}{\sum_{j=1}^{J}\omega_j^*1\{L_j=l\}}-\frac{\sum_{j=1}^{J}1\left\{C_{jw}=c_w,C_{jb}=c_b,L_j=l\right\}}{\sum_{j=1}^{J}1\{L_j=l\}}\right),$$

where $\left\{\omega_j^*\right\}_{j=1}^{J}$ are a set of iid draws from an exponential distribution with mean and variance one. For any function $\phi(\hat{\mu})$ of the moment estimates $\hat{\mu}$ reported, we use as our standard error estimate the standard deviation across bootstrap replications of $J^{-1/4}\left[\phi(\mu^*)-\phi(\hat{\mu})\right]$.

Chernozhukov et al. (2015) goodness of fit test

To formally test whether there exists an η in the K^2 dimensional probability simplex such that $f=BM\eta$ holds, we rely on the procedure of Chernozhukov et al. (2015). Our test statistic (the "J-test") can be written:

$$T_n=\inf_{\eta}\,(\tilde{f}-BM\eta)'\hat{\Sigma}^{-1}(\tilde{f}-BM\eta)$$

$$\text{s.t.}\quad \eta\geq 0,\ \mathbf{1}'\eta=1.$$

Letting $\mathbb{F}^* = f^* - \tilde{f}$ denote the (centered) bootstrap analogue of the callback frequencies $\tilde{f}$ and W^* a corresponding bootstrap weighting matrix, our bootstrap test statistic takes the form:

$$T_n^* = \inf_{\eta,h} (\mathbb{F}^* - BMh)'W^*(\mathbb{F}^* - BMh) \tag{12}$$

$$\text{s.t. } (\tilde{f} - BM\eta)'W(\tilde{f} - BM\eta) = T_n,\ \eta \geq 0,\ \mathbf{1}'\eta = 1,\ h \geq -\eta,\ \mathbf{1}'h = 0.$$

As in the full sample problem, we conduct a two-step GMM procedure in each bootstrap replication, setting $W^* = \left[\text{diag}(BM\eta^{(1)*}) - (BM\eta^{(1)*})(BM\eta^{(1)*})'\right]^{-1}$ where $\eta^{(1)*}$ is a first-step diagonally weighted estimate of the probabilities in the bootstrap sample. The goodness of fit p-value reported is the fraction of bootstrap samples for which $T_n^* > T_n$.

To simplify computation of (12), we re-formulate the problem in two ways. First, we define primary and auxilliary vectors of errors for each moment condition. Letting $\xi_h = \mathbb{F}^* - BMh$ and $\xi_\eta = \tilde{f} - BM\eta$, the problem can be re-posed as:

$$T_n^* = \inf_{\xi_h,\xi_\eta} \ \xi_h' W^* \xi_h,$$

$$\text{s.t. } \xi_\eta' W \xi_\eta = T_n,\ BMh + \xi_h = \mathbb{F}^*,\ BM\eta + \xi_\eta = \tilde{f},\ \mathbf{1}'h = 0,\ \mathbf{1}'\eta = 1,\ h \geq -\eta,\ \eta \geq 0.$$

Now letting $h^+ = h + \eta$, we can further rewrite the problem as:

$$T_n^* = \inf_{\xi_h,\xi_\eta} \ \xi_h' W^* \xi_h,$$

$$\text{s.t. } \xi_\eta' W \xi_\eta = T_n,\ BMh^+ + \xi_h + \xi_\eta = \mathbb{F}^*,\ BM\eta + \xi_\eta = \tilde{f},\ \mathbf{1}'h^+ = 1,\ \mathbf{1}'\eta = 1,\ h^+ \geq 0,\ \eta \geq 0.$$

Note that this final representation replaces a $K^2 \times K^2 + 1$ (inequality) constraint matrix encoding $\xi_h \geq -\xi_\eta$ and $\xi_\eta \geq 0$ with a $2K^2 \times 1$ vector encoding $h^+ \geq 0$ and $\eta \geq 0$. Because this problem still involves a quadratic constraint in ξ_η, we make use of Gurobi's Second Order Cone Programming (SOCP) solver to obtain a solution.

Appendix D: Computing Maximum Risk

We approximate $G\left(p_w^H, p_w^L, p_b^H, p_b^L\right)$ with the discretized distribution

$$G_K\left(p_w^H, p_w^L, p_b^H, p_b^L\right) = \sum_{k=1}^K \sum_{s=1}^K \sum_{k'=1}^K \sum_{s'=1}^K \eta_{ksk's'} 1\left\{p_w^H \leq \varrho(k,s), p_w^L \leq \varrho\left(k',s'\right), p_b^H \leq \varrho(s,k), p_b^L \leq \varrho\left(s',k'\right)\right\},$$

which has K^4 mass points. In practice, we choose $K = 30$, which yields the same number of points as the approximation described in Appendix B.

Generalizing the notation of Appendix C, let the vector $L_j = \left(L_{jw}^H, L_{jw}^L, L_{jb}^H, L_{jb}^L\right)$ record the number of high quality and low quality applications of each race sent to job j and let $C_j =$

$\left(C_{jw}^{H},C_{jw}^{L},C_{jb}^{H},C_{jb}^{L}\right)$ record the corresponding numbers of callbacks. The space of auditing rules we consider is of the form $\delta\left(C_{j},L_{j},q\right)=1\left\{\mathcal{P}\left(C_{j},L_{j},G_{logit}\right)>q\right\}$.With this notation, we can write the risk function

$$\mathcal{R}(q)=\sum_{l\in\mathscr{A}_1}w_l\mathbb{E}\left[\delta(C_j,l,q)\left\{\kappa-\Lambda\left(\sum_{x\in\{H,L\}}\frac{\Lambda^{-1}(p_{wj}^{x})-\Lambda^{-1}(p_{bj}^{x})}{2}\right)\right\}|L_j=l\right],$$

where $\mathscr{A}_1$ is the set of all $2^5=36$ binary quality permutations possible in a design with 5 white and 5 black applications and $w_l=\left(\begin{array}{c}5\\ l_w^H\end{array}\right)\left(\begin{array}{c}5\\ l_b^H\end{array}\right)\left(\frac{1}{2}\right)^{10}$ is the set of weights that arise when quality is assigned at random within race.

To further evaluate the above risk expression we can write:

$$\mathbb{E}\left[\delta(C_j,l,q)\left\{\kappa-\Lambda\left(\sum_{x\in\{H,L\}}\frac{\Lambda^{-1}(p_{wj}^{x})-\Lambda^{-1}(p_{bj}^{x})}{2}\right)\right\}|L_j=l\right]=$$

$$\sum_{c_w^H=0}^{a_{jw}^H}\sum_{c_b^H=0}^{a_{jb}^H}\sum_{c_w^L=0}^{a_{jw}^L}\sum_{c_b^L=0}^{a_{jb}^L}\sum_{k=1}^{K}\sum_{s=1}^{K}\sum_{k'\geq k}^{K}\sum_{s'\geq s}^{K}\delta(c,l,q)\eta_{ksk's'}\left(\begin{array}{c}l_w^H\\ c_w^H\end{array}\right)\left(\begin{array}{c}l_b^H\\ c_b^H\end{array}\right)\left(\begin{array}{c}l_w^L\\ c_w^L\end{array}\right)\left(\begin{array}{c}l_b^L\\ c_b^L\end{array}\right)$$

$$\times\varrho\left(k,s\right)^{c_w^H}\left(1-\varrho\left(k,s\right)\right)^{l_w^H-c_w^H}\varrho\left(s,k\right)^{c_b^H}\left(1-\varrho\left(s,k\right)\right)^{l_b^H-c_b^H}$$

$$\times\varrho\left(k',s'\right)^{c_w^L}\left(1-\varrho\left(k',l=s'\right)\right)^{l_w^L-c_w^L}\varrho\left(s',k'\right)^{c_b^L}\left(1-\varrho\left(s',k'\right)\right)^{l_b^L-c_b^L}$$

$$\times\left\{\kappa-\Lambda\left(\frac{\Lambda^{-1}(\varrho(k,s))-\Lambda^{-1}(\varrho(s,k))}{2}+\frac{\Lambda^{-1}(\varrho(k',s'))-\Lambda^{-1}(\varrho(s',k'))}{2}\right)\right\}.$$

Using this expression, maximal risk can therefore be written as the solution to the following linear programming problem:

$$\mathcal{R}^m(q)\quad=\quad\max_{\{\eta_{ksk's'}\}}\sum_{l\in\mathscr{A}_1}w_l\mathbb{E}\left[\delta(C_j,l,q)\left\{\kappa-\Lambda\left(\sum_{x\in\{H,L\}}\frac{\Lambda^{-1}(p_{wj}^{x})-\Lambda^{-1}(p_{bj}^{x})}{2}\right)\right\}|L_j=l\right]$$

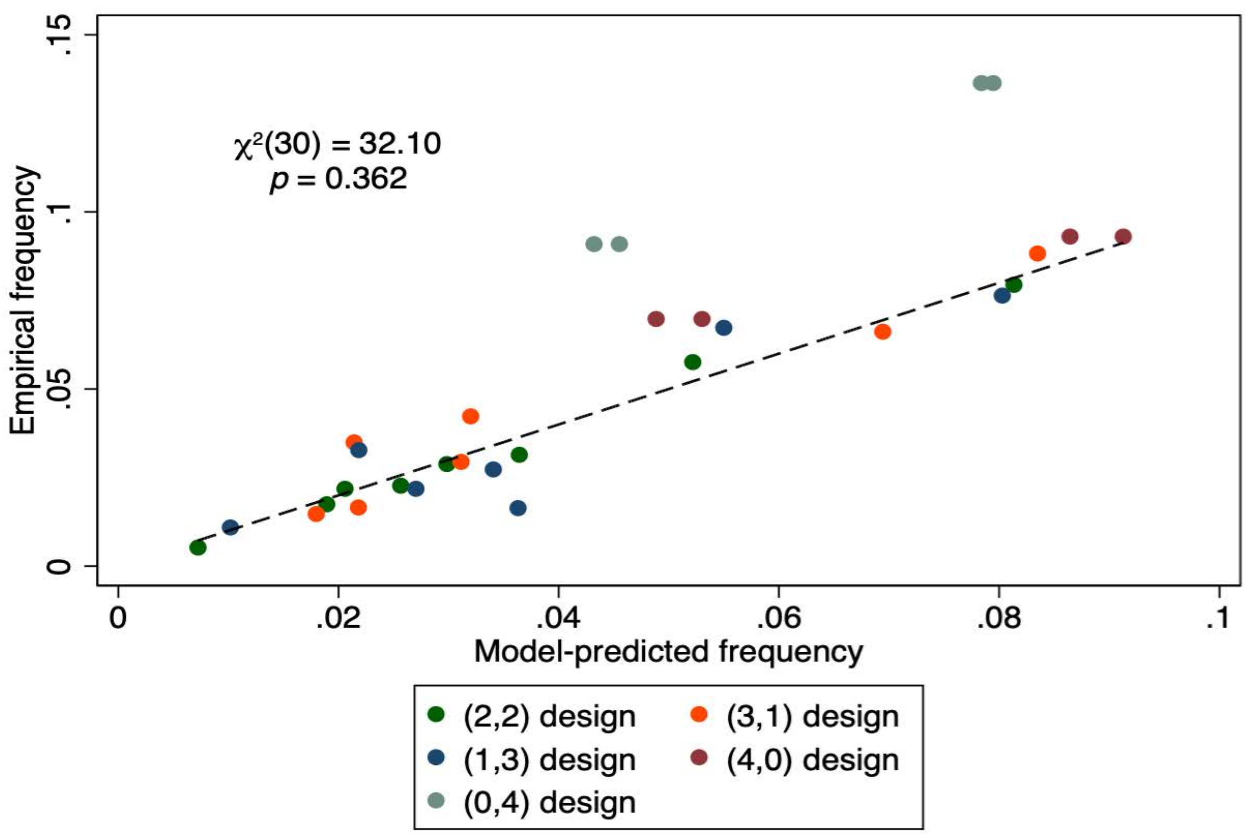
subject to the constraint that the $\eta_{klk'l'}$ are non-negative and sum to one and that the following moment restrictions hold:

$$\begin{aligned}\Pr\left(C_j=c|L_j=l\right)\quad&=\quad\left(\begin{array}{c}l_w^H\\ c_w^H\end{array}\right)\left(\begin{array}{c}l_b^H\\ c_b^H\end{array}\right)\left(\begin{array}{c}l_w^L\\ c_w^L\end{array}\right)\left(\begin{array}{c}l_b^L\\ c_b^L\end{array}\right)\sum_{k=1}^{K}\sum_{s=1}^{K}\sum_{k'=1}^{K}\sum_{s'=1}^{K}\eta_{ksk's'}\\ &\times\quad\varrho\left(k,s\right)^{c_w^H}\left(1-\varrho\left(k,s\right)\right)^{l_w^H-c_w^H}\varrho\left(s,k\right)^{c_b^H}\left(1-\varrho\left(s,k\right)\right)^{l_b^H-c_b^H}\\ &\times\quad\varrho\left(k',s'\right)^{c_w^L}\left(1-\varrho\left(k',s'\right)\right)^{l_w^L-c_w^L}\varrho\left(s',k'\right)^{c_b^L}\left(1-\varrho\left(s',k'\right)\right)^{l_b^L-c_b^L}.\end{aligned}$$

We impose these restrictions for the following set of designs, all of which are present in the Nunley et al. (2015) experiment: $\mathscr{A}_2=\left\{(2,0,2,0),(2,0,0,2),(0,2,2,0),(0,2,0,2)\right\}$.To operationalize these constraints, we replace the unknown cell probabilities $\Pr\left(C_j=c|L_j=l\right)$ for all c and l in

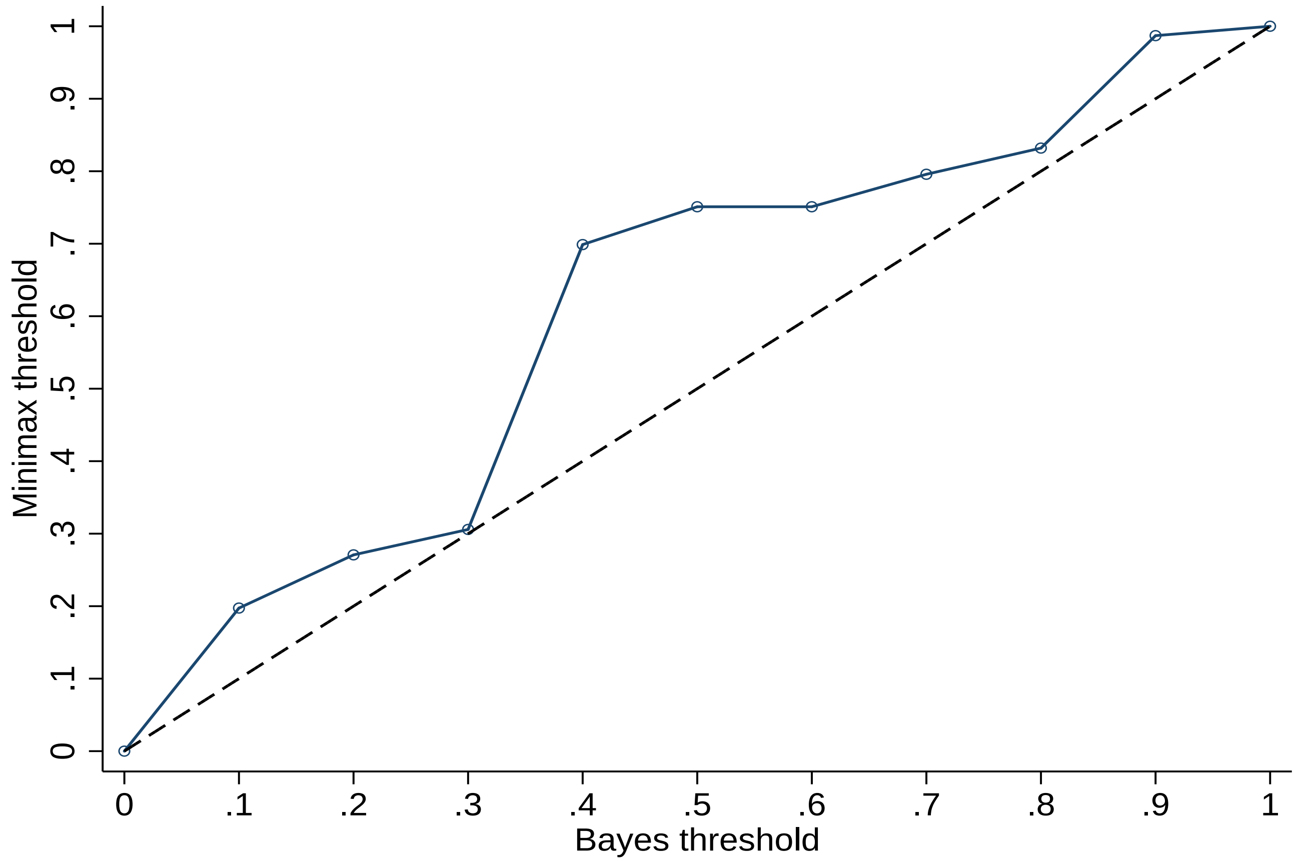
$\mathscr{A}_2$ with their predictions under the logit model reported in column 2 of Table V. Using the logit predictions serves as a form of smoothing that allows us to avoid problems that arise with small cells when considering quality variation due to covariates.

Figure A.I: Mixed logit model fit



Notes: This figure compares mixed logit predicted frequencies for callback events in the Nunley et al. (2015) data with corresponding empirical frequencies. The horizontal axis plots model-predicted probabilities for each possible combination of white and black callback counts (excluding zero total callbacks), separately by experimental design. Model predictions are calculated by simulating the logit model in column (2) of Table X 10,000 times for each job in the Nunley et al. data set. The vertical axis plots the observed frequency of each event. Green dots show frequencies for a design with two white and two black applications, while orange, blue, red, and grey points show frequencies for designs with 3 white and 1 black, 1 white and three black, 4 white and zero black, and 0 white and 4 black applications, respectively. The dashed line is the 45-degree line. The chi-squared statistic and p -value come from a test that all model-predicted and empirical frequencies match, treating the model predictions as fixed.

Figure A.II: Bayes and minimax investigation thresholds



Notes: This figure compares Bayes and minimax decisions for various values of the investigation cost parameter κ. The horizontal axis displays the posterior investigation threshold for a Bayes regulator for each value of κ, and the vertical axis shows the corresponding threshold for a minimax regulator. The dashed line is the 45 degree line.

Table A.I: Moments of callback rate distribution, BM data

Moment	No constraints (1)	Shape constraints (2)
$E[p_w]$	0.094	0.094
	(0.006)	(0.007)
$E[p_b]$	0.063	0.063
	(0.006)	(0.006)
$E[(p_w - E[p_w])^2]$	0.040	0.040
	(0.005)	(0.004)
$E[(p_b - E[p_b])^2]$	0.023	0.023
	(0.004)	(0.003)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.028	0.028
	(0.004)	(0.003)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.015	0.014
	(0.003)	(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.023	0.012
	(0.003)	(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.010	0.010
	(0.003)	(0.002)
	J-statistic:	0.00
	P-value:	1.000
Sample size	1,112	

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Bertrand and Mullainathan (2004) data. Estimates in column (2) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.II: Moments of callback rate distribution, NPRS data

Moment	(2,2) design (1)	(3,1) design (2)	(1,3) design (3)	P-value (4)	Combined estimates (5)
	Design-specific estimates				
$E[p_w]$	0.174	0.199	0.142	0.027	0.177
	(0.010)	(0.025)	(0.015)		(0.007)
$E[p_b]$	0.148	0.149	0.157	0.854	0.153
	(0.010)	(0.015)	(0.013)		(0.007)
$E[(p_w - E[p_w])^2]$	0.089	0.108	-	0.097	0.095
	(0.007)	(0.009)			(0.004)
$E[(p_b - E[p_b])^2]$	0.085	-	0.083	0.857	0.084
	(0.007)		(0.008)		(0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083	0.084	0.080	0.926	0.084
	(0.006)	(0.009)	(0.009)		(0.004)
$E[(p_w - E[p_w])^3]$	-	0.051	-		0.106
		(0.008)			(0.006)
$E[(p_b - E[p_b])^3]$	-	-	0.044		0.092
			(0.007)		(0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044	0.043	-	0.875	0.040
	(0.004)	(0.007)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047	-	0.045	0.819	0.042
	(0.005)		(0.007)		(0.002)
$E[(p_w - E[p_w])^3(p_b - E[p_b])]$	-	0.034	-	-	0.035
		(0.005)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^3]$	-	-	0.037	-	0.037
			(0.006)		(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036	-	-	-	0.038
	(0.004)				(0.002)
				J-statistic:	23.09
				P-value:	0.190
Sample size	1,146	544	550		2,240

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Nunley et al. (2015) data. Columns (1), (2), and (3) show estimates based on jobs that received 2 white and 2 black, 3 white and 1 black, and 1 white and 3 black applications, respectively. Column (4) shows p-values from tests that the moments are the same in each design. Estimates in column (5) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.III: Moments of callback rate distribution, AGCV data

Moment	No constraints (1)	Shape constraints (2)	Moment	No constraints (3)	Shape constraints (4)
$E[p_f]$	0.138 (0.010)	0.140 (0.009)	$E\left[\left(p_f - E[p_f]\right)^4\right]$	0.023 (0.004)	0.025 (0.003)
$E[p_m]$	0.108 (0.009)	0.114 (0.009)	$E[(p_m - E[p_m])^4]$	0.019 (0.004)	0.024 (0.003)
$E\left[\left(p_f - E[p_f]\right)^2\right]$	0.066 (0.009)	0.066 (0.005)	$E\left[\left(p_f - E[p_f]\right)^4 (p_m - E[p_m])\right]$	0.012 (0.003)	0.011 (0.002)
$E[(p_m - E[p_m])^2]$	0.048 (0.005)	0.054 (0.005)	$E\left[\left(p_f - E[p_f]\right)(p_m - E[p_m])^4\right]$	0.013 (0.003)	0.013 (0.002)
$E[(p_f - E[p_f])(p_m - E[p_m])]$	0.043 (0.005)	0.044 (0.004)	$E\left[\left(p_f - E[p_f]\right)^3 (p_m - E[p_m])^2\right]$	0.012 (0.003)	0.011 (0.002)
$E\left[\left(p_f - E[p_f]\right)^3\right]$	0.025 (0.005)	0.063 (0.007)	$E\left[\left(p_f - \mu_f\right)^2 (p_m - E[p_m])^3\right]$	0.012 (0.003)	0.012 (0.002)
$E[(p_m - E[p_m])^3]$	0.031 (0.005)	0.050 (0.007)	$E\left[\left(p_f - E[p_f]\right)^4 (p_m - E[p_m])^2\right]$	0.010 (0.002)	0.009 (0.001)
$E\left[\left(p_f - E[p_f]\right)^2 (p_m - E[p_m])\right]$	0.020 (0.004)	0.017 (0.002)	$E\left[\left(p_f - E[p_f]\right)^2 (p_m - E[p_m])^4\right]$	0.010 (0.002)	0.009 (0.001)
$E\left[\left(p_f - E[p_f]\right)(p_m - E[p_m])^2\right]$	0.022 (0.004)	0.020 (0.002)	$E\left[\left(p_f - E[p_f]\right)^3 (p_m - E[p_m])^3\right]$	0.009 (0.002)	0.009 (0.001)
$E\left[\left(p_f - E[p_f]\right)^3 (p_m - E[p_m])\right]$	0.015 (0.003)	0.015 (0.002)	$E\left[\left(p_f - E[p_f]\right)^4 (p_m - E[p_m])^3\right]$	0.008 (0.002)	0.007 (0.001)
$E\left[\left(p_f - E[p_f]\right)(p_m - E[p_m])^3\right]$	0.016 (0.003)	0.017 (0.002)	$E\left[\left(p_f - E[p_f]\right)^3 (p_m - E[p_m])^4\right]$	0.008 (0.002)	0.008 (0.001)
$E\left[\left(p_f - E[p_f]\right)^2 (p_m - E[p_m])^2\right]$	0.016 (0.003)	0.016 (0.002)	$E\left[\left(p_f - E[p_f]\right)^4 (p_m - E[p_m])^4\right]$	0.009 (0.002)	0.006 (0.001)
		J-statistic:	3.33		
		P-value:	0.790		
		Sample size:	802		

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Arceo-Gomez and Campos-Vasques (2014) data. Estimates in columns (2) and (4) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.IV: Sensitivity of moments and bounds to discretization grid

		Grid spacing	J-statistic (1)	Share discriminating (2)	Share disc., one call (3)	Share disc., two calls (4)	Share disc., three calls (5)
				A. Nunley et al. data			
$K_1 = 50$	$K_2 = 200$	Quadratic	23.13	0.266	0.483	0.549	0.678
$K_1 = 100$	$K_2 = 200$	Quadratic	23.10	0.366	0.692	0.708	0.841
	$K_2 = 400$	Quadratic		0.329	0.611	0.643	0.783
	$K_2 = 600$	Quadratic		0.319	0.588	0.627	0.765
$K_1 = 150$	$K_2 = 300$	Quadratic	23.09	0.401	0.765	0.770	0.880
	$K_2 = 600$	Quadratic		0.368	0.692	0.708	0.835
	$K_2 = 900$	**Quadratic**		**0.358**	**0.672**	**0.691**	**0.821**
		Rectangular		0.410	0.780	0.785	0.887
				B. Arceo-Gomez & Campos-Vasquez data			
$K_1 = 50$	$K_2 = 200$	Quadratic	4.21	0.277	0.882	0.724	0.574
$K_1 = 100$	$K_2 = 200$	Quadratic	3.54	0.288	0.902	0.726	0.595
	$K_2 = 400$	Quadratic		0.277	0.893	0.719	0.578
	$K_2 = 600$	Quadratic		0.276	0.891	0.717	0.573
$K_1 = 150$	$K_2 = 300$	Quadratic	3.33	0.297	0.904	0.725	0.596
	$K_2 = 600$	Quadratic		0.278	0.897	0.718	0.579
	$K_2 = 900$	**Quadratic**		**0.277**	**0.895**	**0.716**	**0.576**
		Rectangular		0.297	0.900	0.715	0.584

Notes: This table explores the sensitivity of our shape-constrained generalized method of moments (SCGMM) and linear programming bounds results to the number of grid points used to approximate the joint distribution of callback probabilities. K1 refers to the number of mass points used in the quadratic programming SCGMM step, while K2 refers to the number of mass points used in the linear programming bounds step. Quadratic grid spacing refers to the scheme described in Appendix A, and rectangular spacing refers to a grid with equally spaced points. Column (1) shows the minimized SCGMM criterion function for each value of K1. Column (2) displays the lower bound on the fraction of discriminating jobs for each combination of K1 and K2. Columns (3)-(5) show corresponding bounds conditional on the total number of callbacks. Panel A displays results for an application design with two white and two black applicants in the Nunley et al. (2015) data, and panel B displays results for the Arceo-Gomez and Campos-Vasquez (2014) data. Bold lines indicate the preferred specification used in the main text.