A Discrimination Report Card

Patrick Kline, UC Berkeley
Evan K. Rose, University of Chicago
Christopher Walters, UC Berkeley

April 8th, 2024

Who discriminates?

- Increasing agreement that wage setting conduct varies systematically across **firms** (Card et al., 2018). What about *recruiting* conduct?
- Large literature uses correspondence studies to measure market-average discrimination against these protected characteristics (Bertrand and Duflo, 2017)
- Little known about discriminatory conduct of specific employers despite widespread interest from the public

Measuring employer-level discrimination

- Recent work uses correspondence experiments combined with empirical Bayes and large-scale inference methods to study discrimination by particular employers
- Kline and Walters (2021): Reanalysis of several correspondence experiments
  - Framework: Correspondence study as ensemble of job-specific micro-experiments, each with its own response probabilities
  - Key findings: Tremendous heterogeneity in discrimination across jobs; possible to detect discrimination at some individual jobs with high confidence
- Kline, Rose, and Walters (2022): Correspondence experiment at 108 large firms
  - Up to 1,000 applications sent to each company
  - Signaled race/gender with distinctive names
  - Key finding 1: Wide variation across firms in bias against Black / female names; top 20% account for $\sim$50% of total
  - Key finding 2: Half of variance across firms explained by two-digit industry

Summarizing firm-level conduct

- Experimental results demonstrate that discrimination is highly concentrated in a small set of employers, but estimate for any given employer may be subject to substantial sampling error
- How should we communicate what we've learned about the biased conduct of firms to a broad audience?
  - Scientific communication generally aided by transparency (Andrews and Shapiro, 2021)
  - But some audiences may find it difficult to interpret complex statistical evidence (Mullainathan, 2002; Mullainathan et al., 2008; Bordalo et al., 2016)
- Scholars and policymakers increasingly construct simple "report cards" summarizing econometric estimates of quality for various institutions: colleges (Chetty et al., 2017), K-12 schools (Bergman et al., 2020; Angrist et al., 2021), teachers (Bergman and Hill, 2018; Pope, 2019), healthcare providers (Brook et al., 2002; Pope, 2009), neighborhoods (Chetty and Hendren, 2018; Chetty et al., 2018)

Today's agenda: discrimination report cards

- An Empirical Bayes report card that grades the discriminatory conduct of firms
- Report card scheme formalizes tradeoff between informativeness and reliability
  - Audience makes pairwise inferences on relative discrimination based on grades
  - Combine EB posterior pairwise ranking probabilities to construct a global partial ordering
  - Asymmetric preferences over correct rankings vs. mistakes $\mapsto$ optimal coarsening with few grades
  - Analogue of False Discovery Rates for summarizing grade reliability
- Time permitting: Survey evidence on beliefs regarding employer discrimination

Related literature

- Audit and correspondence experiments for measuring racial discrimination (Daniel, 1968; Wienk et al., 1979; Heckman and Siegelman, 1993; Heckman, 1998; Bertrand and Mullainathan, 2004; Pager et al., 2009; Nunley et al., 2015; Bertrand and Duflo, 2017; Quillian et al, 2017; Baert, 2018; Gaddis, 2018; Neumark, 2018; Kline, Rose, and Walters, 2022)
- Scientific communication (Savage, 1954; Andrews and Shapiro, 2021; Viviano, Wuthrich, Niehaus, 2021; Korting et al., 2021)
- Limited attention / signal coarsening (Mullainathan, Schwartzstein, and Shleifer, 2008; Pope, 2009; Gilbert et al., 2012; Lacetera, Pope, and Sydnor, 2012; Sejas-Portillo et al., 2020)
- Empirical Bayes inference / selection rules / false discovery rates (Robbins, 1964; Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2002; Armstrong, 2015; Efron, 2016; Armstrong, Kolesár, Plagborg-Møller, 2020; Kline and Walters, 2021; Gu and Koenker, 2023; Chen, 2023)
- Econometrics of ranks (Portnoy, 1982; Berger and Deely, 1988; Laird and Louis, 1989; Sobel, 1993; Mogstad et al., 2020; Andrews et al., 2021; Gu and Koenker, 2022)
- Social choice / vote aggregation (Borda, 1784; Condorcet, 1785; Kemeny, 1959; Smith, 1973; Young and Levenglick, 1978; Young, 1986)

Experimental design

Sampling frame (I/II)

Holding companies split into brands with separate hiring portals (e.g., Berkshire Hathaway into Geico, McLane, Fruit of the Loom, etc.)

Fortune 500

InfoGroup and Burning Glass data merged to measure geographic distribution of establishments and vacancies

123 firms with sufficient expected geographic scope

Hiring platforms investigated to test for feasibility of submitting fictitious applications

108 feasible to audit

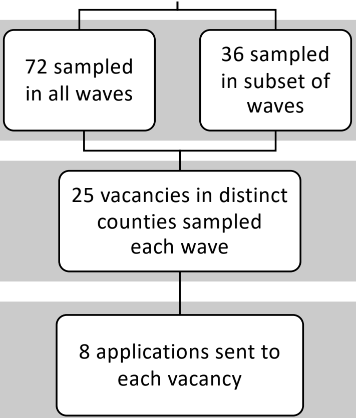
Sampling frame (I/II)

Holding companies split into brands with separate hiring portals (e.g., Berkshire Hathaway into Geico, McLane, Fruit of the Loom, etc.)	Fortune 500
InfoGroup and Burning Glass data merged to measure geographic distribution of establishments and vacancies	123 firms with sufficient expected geographic scope
Hiring platforms investigated to test for feasibility of submitting fictitious applications	108 feasible to audit

Compustat: U.S. employment at 108 sampled firms totaled $\sim$**15M** in 2020

Sampling frame (II/II)

4 not sampled in wave 1 due to COVID interruption; 9 firms dropped before completion due to technological constraints; 19 added in wave 2 or later; 4 posted insufficient jobs to sample in all waves



Job sampled from universe of entry-level vacancies posted on each firm's hiring portal; most recently posted job prioritized

One pair of applications (1 black and 1 white name) sent every 1-2 days; gender (50% male), age (uniform age 20-60), gender identity (5% gender-neutral, 5% same-gender pronouns), and sexual orientation (10% LGBTQ student club, 10% other club) unconditionally randomly assigned

Resume characteristics

Job applications manipulate employer perceptions of several protected characteristics:

- Race & gender: distinctive first names obtained from Bertrand and Mullainathan (2004) + NC data on speeding tickets. Last names from Census
- Age: year of high school graduation

Stratify on race (4B/4W), unconditional random assignment of gender, age, as well as LGBTQ affiliation and gender identity

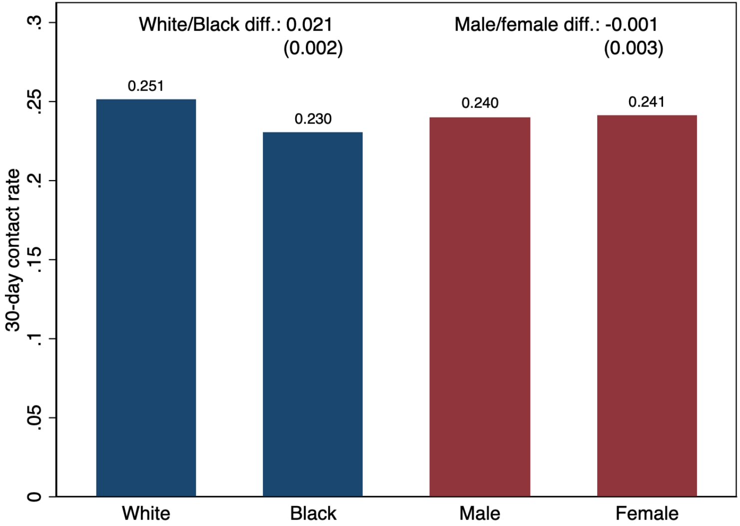
Random assignment of job-appropriate experience, high school, associate degree, resume design, answers to personality tests, etc.

Fully automated sampling of vacancies and submission of apps

Summary stats

	A. All firms			B. Balanced sample		
	White	Black	Difference	White	Black	Difference
Resume characteristics						
Female	0.499	0.499	-0.001	0.500	0.498	0.003
Over 40	0.535	0.535	0.000	0.534	0.533	0.002
LGBTQ club member	0.081	0.082	-0.001	0.079	0.080	-0.001
Academic club	0.040	0.042	-0.002	0.039	0.042	-0.003*
Political club	0.042	0.042	0.001	0.042	0.041	0.001
Gender-neutral pronouns	0.041	0.041	-0.001	0.040	0.040	0.000
Same-gender pronouns	0.043	0.042	0.001	0.042	0.041	0.001
Associate degree	0.476	0.485	-0.009**	0.478	0.485	-0.006*
N applications	41837	41806	83643	32703	32665	65368
N jobs			11114			8667
N firms			108			72
1/2/3/4/5 waves			3/4/15/16/72			

Means: White names favored by 2.1pp, zero average gender difference



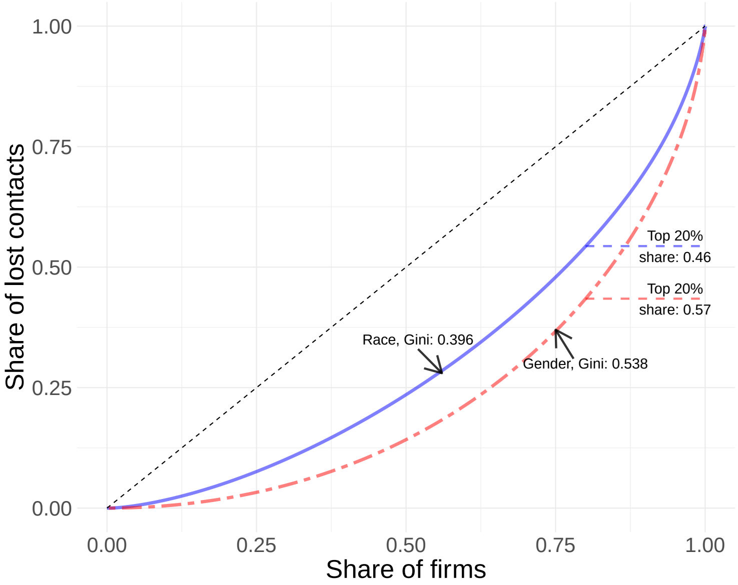
Std. devs.: Substantial heterogeneity across firms for both race and gender

Estimates of firm heterogeneity in race and gender discrimination

	Mean contact gap (1)	Bias-corrected std. dev. of contact gaps (2)
Race (White - Black)	0.021	0.0185
	(0.002)	(0.0031)
Gender (Male - Female)	-0.001	0.0267
	(0.003)	(0.0038)

Estimates from Kline, Rose, and Walters (2022).

Lorenz curves: Top 20% of firms explain ∼50-60% of lost contacts



A Discrimination Report Card

Preliminaries

- n firms, indexed by $i \in \{1, \ldots, n\} \equiv [n]$
- Discrimination at firm i parameterized by $\theta_i \in \mathbb{R}$ (proportional contact gap)
- For each firm observe: $Y_i = (\hat{\theta}_i, s_i)$
- $\{Y_i\}_{i=1}^n$ mutually independent conditional on $\theta = (\theta_1, \ldots, \theta_n)'$
- Large sample approximation

$$\hat{\theta}_i \mid \theta_i, s_i \sim \mathcal{N}(\theta_i, s_i^2)$$

Gambling over contrasts

Suppose smooth $i.i.d.$ prior G over $\{\theta_i\}_{i\in[n]}$ and consider the following risky gamble:

- Observe realizations (y_i, y_j) of (Y_i, Y_j)
- Propose partial ordering $d = (d_i, d_j) \in \{1, 2\}^2$ of θ_i and θ_j
- If ordering correct: payoff $= \lambda \in (0, 1]$
- If ordering incorrect: payoff = -1
- Declare a tie / abstain: payoff = 0

Given posterior $\pi_{ij} = \Pr_G(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j)$, expected utility of choosing d is

$$EU(\pi_{ij}, d) = [\underbrace{\lambda\pi_{ij} - (1 - \pi_{ij})}_{(1+\lambda)\pi_{ij}-1}] \cdot 1\{d_i > d_j\} + [\underbrace{\lambda(1 - \pi_{ij}) - \pi_{ij}}_{(1+\lambda)(1-\pi_{ij})-1}] \cdot 1\{d_i < d_j\}$$

Optimal decision

Maximize EU with posterior threshold rule:

- Set $d_i > d_j$ iff $\pi_{ij} > \frac{1}{1+\lambda}$
- Set $d_i < d_j$ iff $1 - \pi_{ij} > \frac{1}{1+\lambda}$
- Otherwise set $d_i = d_j$

Threshold approaches 1 as $\lambda \to 0$, yielding all ties

No ties when $\lambda = 1$ bc threshold is $1/2$ (and smooth prior)

A scientific reporting interpretation

Consider reporting ranking (d_i, d_j) to audience choosing between firms i and j

Audience receives payoff 1 to choosing correct ranking. Otherwise payoff is 0.

- Audience chooses according to report when ranking is strict.
- If report is a tie, a share $q \in (0,1)$ that are "informed" will make the right choice.
- The remaining share $1-q$ breaks tie correctly with probability $1/2$.

Expected payoff of reporting a tie is $q + (1-q)/2 = (1+q)/2$. Hence, expected utility of a report d is:

$$\frac{1+q}{2} + \frac{1+q}{2} EU\left(\pi_{ij}, d; \frac{1-q}{1+q}\right)$$

$\Rightarrow$ Optimal $\lambda = \frac{1-q}{1+q}$ decreasing in audience sophistication q

Pooling pairs

Now consider all $\binom{n}{2}$ firm pairs. Loss of grades $d = (d_1, \ldots, d_n)' \in [n]^n$ is:

$$L(\theta, d; \lambda) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Bigg[\underbrace{1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}}_{\text{discordant pairs}} - \lambda \Bigg(\underbrace{1\{\theta_i < \theta_j, d_i < d_j\} + 1\{\theta_i > \theta_j, d_i > d_j\}}_{\text{concordant pairs}} \Bigg) \Bigg]$$

Note: when $\lambda = 1$, loss is the negative of Kendall (1938)'s tau coefficient between d and θ, i.e., bubble-sort distance

Quantifying mistakes

Define the *Discordance Proportion* as

$$DP(\theta, d) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \left[1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}\right]$$

$$= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \left|1\{\theta_i > \theta_j\} - 1\{d_i > d_j\}\right| \cdot 1\{d_i \neq d_j\}$$

- *DP* measures frequency of misrankings
- Can limit by coarsening grades / declaring ties

Too much information

Letting $\tau(\theta, d) \in [-1, 1]$ denote Kendall's tau, we can write the loss

$$L(\theta, d; \lambda) = (1 - \lambda)\, DP(\theta, d) - \lambda \tau(\theta, d)$$

- Parameter λ governs trade-off between information content of rankings (τ) and mistake frequency (DP)
- $1 - \lambda$ measures *discordance aversion*
- When $\lambda < 1$, willing to report coarse grades to avoid discordances

Optimal grades

The posterior expected loss of a fixed vector of grades d given data realization y is

$$\begin{aligned}\mathcal{R}(\pi, d; \lambda) &= \mathbb{E}_G[L(\theta, d; \lambda) | Y = y] \\ &= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Big[(1 - \pi_{ij}) 1\{d_i > d_j\} + \pi_{ij} 1\{d_i < d_j\} \\ &\quad - \lambda (1 - \pi_{ij}) 1\{d_i < d_j\} - \lambda \pi_{ij} 1\{d_i > d_j\} \Big]\end{aligned}$$

Bayes optimal grades are

$$d^*(\lambda) = \arg\min_{d \in [n]^n} \mathcal{R}(\pi, d; \lambda)$$

Expected rank correlation and discordance

Recall that loss is a linear combination of DP and τ. Posterior mean loss is:

$$\mathcal{R}(\pi, d; \lambda) = (1-\lambda) DR(\pi, d) - \lambda \bar{\tau}(\pi, d)$$

where

$$DR(\pi, d) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} 1\left\{d_i < d_j\right\} \pi_{ij} + 1\left\{d_i > d_j\right\}\left(1 - \pi_{ij}\right)$$

$$\bar{\tau}(\pi, d) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} 1\{d_i < d_j\}(2\pi_{ij} - 1) + 1\{d_i > d_j\}(1 - 2\pi_{ij})$$

Discordance rates between grades

$\bar{\tau}(\pi, d^*(\lambda))$ is the expected rank correlation of the optimal grades, while $DR(\pi, d^*(\lambda))$ is the expected DP of optimal grades:

The DR between a specific pair of grades g and $g' < g$ is

$$DR_{g,g'}(\lambda) = \frac{\sum_{i=1}^{n} \sum_{j \neq i} 1\left\{d_i^*(\lambda) = g\right\} 1\left\{d_j^*(\lambda) = g'\right\}(1 - \pi_{ij})}{\sum_{i=1}^{n} \sum_{j \neq i} 1\left\{d_i^*(\lambda) = g\right\} 1\left\{d_j^*(\lambda) = g'\right\}}.$$

- $DR_{g,g'}$ analogous to False Discovery Rate of collection of 1-sided contrasts
- DR decomposes into weighted average of the $\{DR_{g,g'}\}$ and $DR_{g,g} = 0$

Condorcet paradox

While objective $\mathcal{R}(\pi, d; \lambda)$ is separable across pairs, logical constraints prevent pairwise optimization via comparing π_{ij} to threshold $(1+\lambda)^{-1}$

Example (Three firms, normal posteriors)

Suppose $\theta_i | Y_i = y_i \sim N(\mu_i, 1)$. Then if posteriors are independent:

$$\pi_{ij} = Pr(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2}}\right)$$

- Let $\lambda = 1/4 \implies (1+\lambda)^{-1} = 0.8$
- Suppose $(\mu_1, \mu_3) = (2, 0)$, so that $\pi_{13} = \Phi(\sqrt{2}) = .92$ and $\pi_{31} = 1 - \pi_{13} = .08$
- Then it is optimal to rank $\theta_1 > \theta_3$.
- But if $\mu_2 \in (0.81, 1.19)$, rank (θ_1, θ_2), (θ_2, θ_3) as ties because $\max\{\pi_{12}, \pi_{23}\} < 0.8$

This is a logical contradiction violating **transitivity**

ILP formulation

Define indicators $d_{ij} = 1\{d_i > d_j\}$ and $e_{ij} = 1\{d_i = d_j\}$. We can rewrite our problem as choosing $\{d_{ij}, e_{ij}\}_{i<j\leq n}$ to minimize

$$\sum_{i=2}^{n}\sum_{j=1}^{i}\left[(1-\pi_{ij})\, d_{ij} + \pi_{ij}\,(1 - e_{ij} - d_{ij}) - \lambda\,(1-\pi_{ij})\,(1 - e_{ij} - d_{ij}) - \lambda\pi_{ij} d_{ij}\right]$$

s.t. to the following transitivity constraints on any triple $(i, j, k) \in [n]^3$:

$$d_{ij} + d_{jk} \leq 1 + d_{ik}, \quad d_{ik} + (1 - d_{jk}) \leq 1 + d_{ij}, \quad e_{ij} + e_{jk} \leq 1 + e_{ik}$$

and $e_{ij} + d_{ij} + d_{ji} = 1$.

Linear objective + linear constraints $\Longrightarrow$ **integer linear programming**

A connection to social choice

When $\lambda = 1$ we seek to minimize

$$\sum_{i=2}^{n} \sum_{j=1}^{i} (2\pi_{ij} - 1)(d_{ji} - d_{ij})$$

If π_{ij} is viewed as the number of votes for $\theta_i > \theta_j$ the constrained minimizer $d^*(1)$ of this objective is the Kemeny - Young voting method (aka Condorcet's rule)

Young (1988) showed that $d^*(1)$ is

- The most likely ranking (aka the maximum likelihood estimator) when all voters have a common probability $> 1/2$ of deciding pairwise contrasts correctly
- The unique ranking rule that is neutral, unanimous, and satisfies reinforcement and independence of remote alternatives details

Condorcet property

Condorcet criterion: if there is a unit i that wins pairwise election against all $j \neq i$, then i will be top ranked.

Theorem (λ-Condorcet Criterion)

Suppose that firm i satisfies $\pi_{ij} > (1+\lambda)^{-1}\ \forall\ j \neq i$. Then $d_i > d_j\ \forall\ j \neq i$.

Moreover, suppose that firm k satisfies $\pi_{ik} > (1+\lambda)^{-1}$ and $\pi_{kj} > (1+\lambda)^{-1}\ \forall\ j \neq i, j \neq k$, then $d_i > d_k > d_j\ \forall\ j \neq i, j \neq k$.

- Equivalent argument yields selection of bottom ranked "losers."
- With $\lambda < 1$, ties emerge. Show in paper that λ-ranking scheme selects notion corresponding to Smith (1973) set.

Empirics: Names

Estimated R^2 of race and sex is 121%!

Table: Summary statistics for first names sample

	Contact rate	# apps	# first names	Wald test of heterogeneity
Male				
Black	0.233	20,927	19	12.6
	(0.003)			[0.82]
White	0.246	20,975	19	15.8
	(0.003)			[0.61]
Female				
Black	0.226	20,879	19	21.2
	(0.003)			[0.24]
White	0.254	20,862	19	19.9
	(0.003)			[0.34]
Estimated contact rate SD				
Total	0.010			
Between race/sex	0.011			

Defining θ

Let N_i give # of apps sent with first name i and C_i give # of contacts within 30 days.

Assuming $C_i | N_i = n \sim Bin(n, p_i)$ we have

$$\mathbb{E}[C_i/N_i] = p_i, \quad \mathbb{V}[C_i/N_i] = p_i(1 - p_i)/N_i$$

Stabilize variance with Bartlett (1936) transform

$$\hat{\theta}_i = \sin^{-1}\sqrt{C_i/N_i}.$$

Why this helps: $\frac{d}{dx}\sin^{-1}\sqrt{x} = \left[2\sqrt{x(1-x)}\right]^{-1}$. Hence, by the Delta method

$$\hat{\theta}_i \mid N_i \sim \mathcal{N}(\theta_i, (4N_i)^{-1}), \text{ where } \theta_i = \sin^{-1}(p_i).$$

Estimating G

Hierarchical model:

$$\hat{\theta}_i | \theta_i \sim \mathcal{N}(\theta_i, (4N_i)^{-1})$$

$$\theta_i | N_i \sim G$$

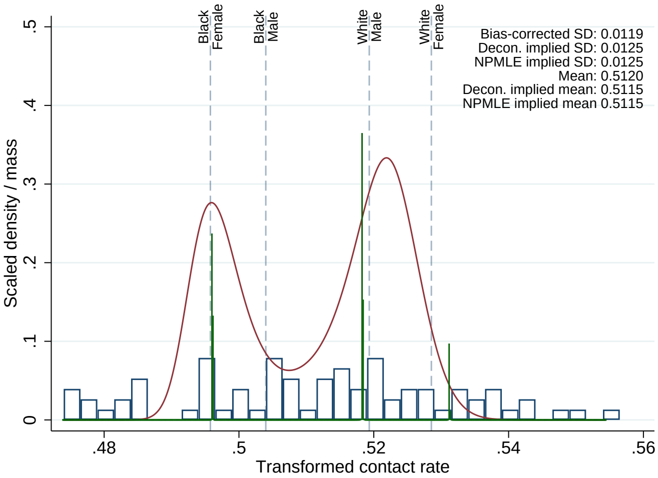
Empirical Bayes: Estimate G via deconvolution, then treat $\hat{G}$ as prior

Two approaches to deconvolution:

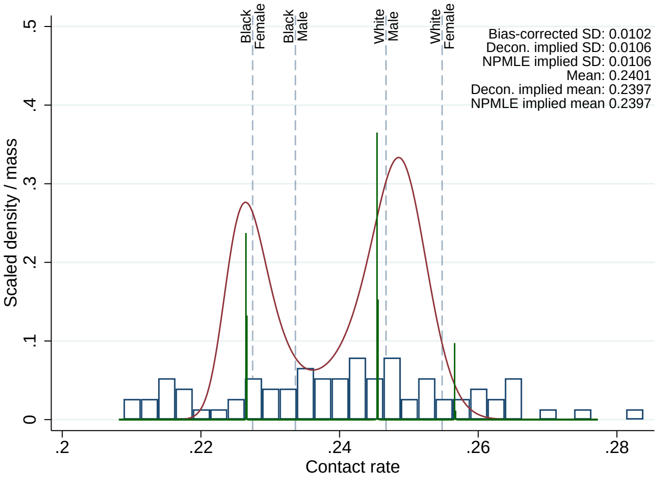
- Efron (2016): model G with exponential family parameterized by fifth-order spline, estimate via penalized MLE
- Koenker and Gu (2017): mass point approximation via NPMLE

True G seems likely to be smooth $\mapsto$ focus on Efron approach, which implies ties are measure zero

Variance-stabilized contact rates ($\sin^{-1}\sqrt{p_i}$)



Contact rates (p_i)



Empirical Bayes posteriors and grades

EB posterior density for θ_i:

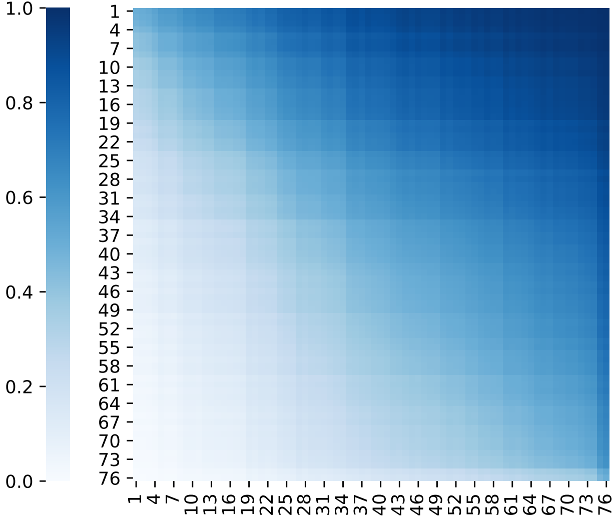
$$\hat{f}(\theta_i|\hat{\theta}_i, s_i) = \frac{\frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - \theta_i}{s_i}\right) d\hat{G}(\theta_i|s_i)}{\int \frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - x}{s_i}\right) d\hat{G}(x|s_i)}$$

Here, std err is $s_i = (4N_i)^{-1/2}$. Pairwise posterior probabilities are:

$$\hat{\pi}_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{x} \hat{f}(x|\hat{\theta}_i, s_i)\hat{f}(y|\hat{\theta}_j, s_j)dydx$$

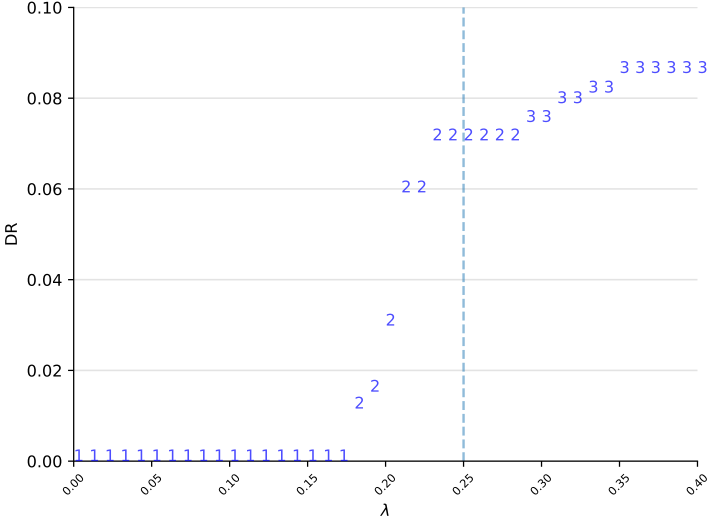
Feed these $\hat{\pi}_{ij}$'s to integer linear programming routine to compute optimal grades for each value of the tuning parameter λ

Posterior contrasts (π_{ij})

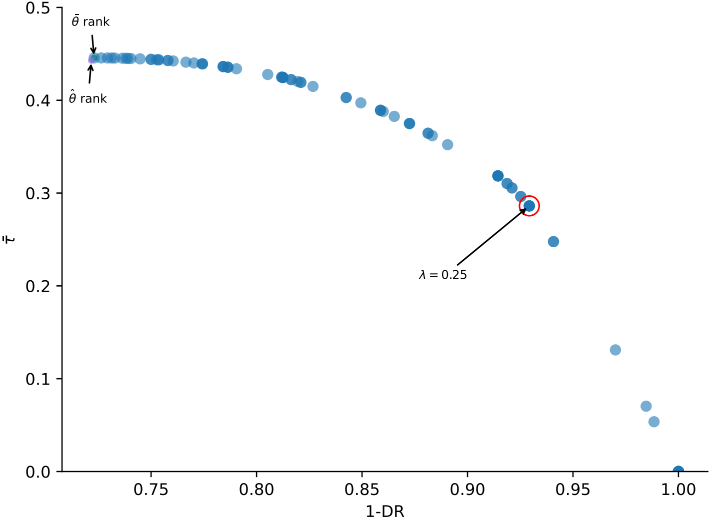


Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

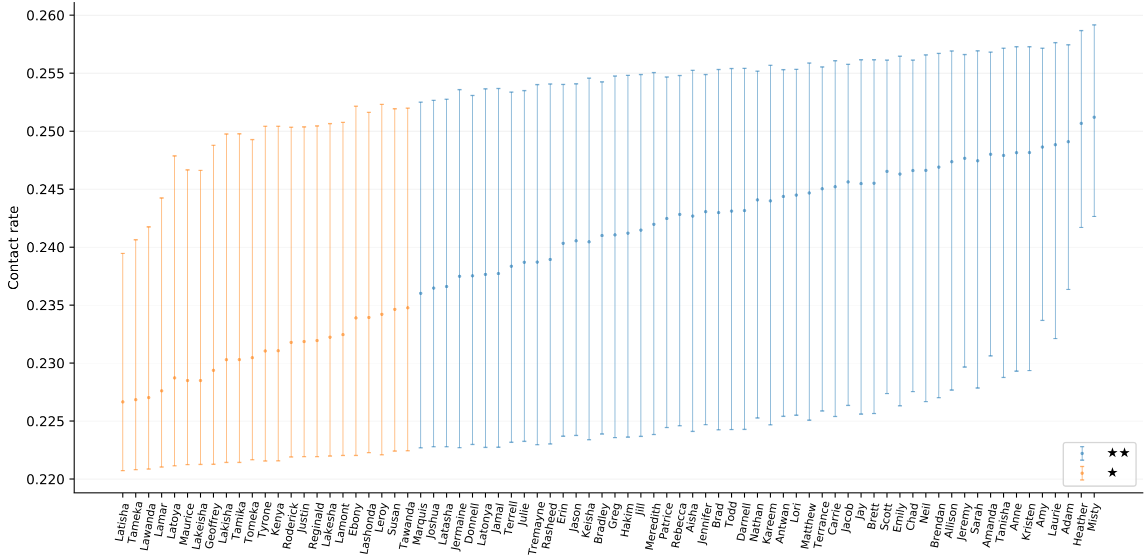
Tune grades to exhibit ∼ 80% posterior confidence threshold



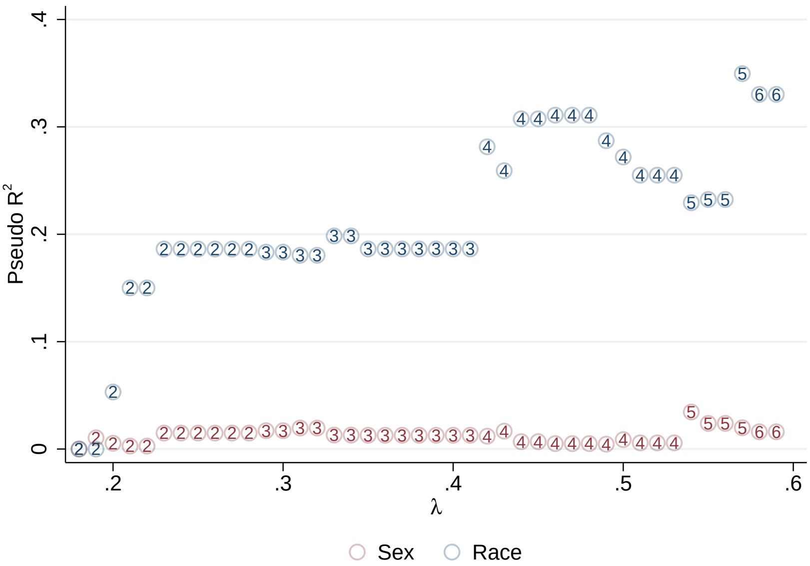
Reporting possibilities



Two grade scheme explains 35% of cross name variance



Grades predict race but not sex



Empirics: Firms

Defining the target parameter

Each firm i has latent race- and gender-specific contact rates $(p_{iw}, p_{ib}, p_{im}, p_{if})$

Focus on proportional contact gaps:

$$\begin{aligned} \textit{Race:} \quad & \theta_i = \ln(p_{iw}) - \ln(p_{ib}) \\ \textit{Gender:} \quad & \theta_i = \ln(p_{im}) - \ln(p_{if}) \end{aligned}$$

Rely on plug-in estimators

$$\hat{\theta}_i = \ln(\hat{p}_{iw}) - \ln(\hat{p}_{ib}),$$

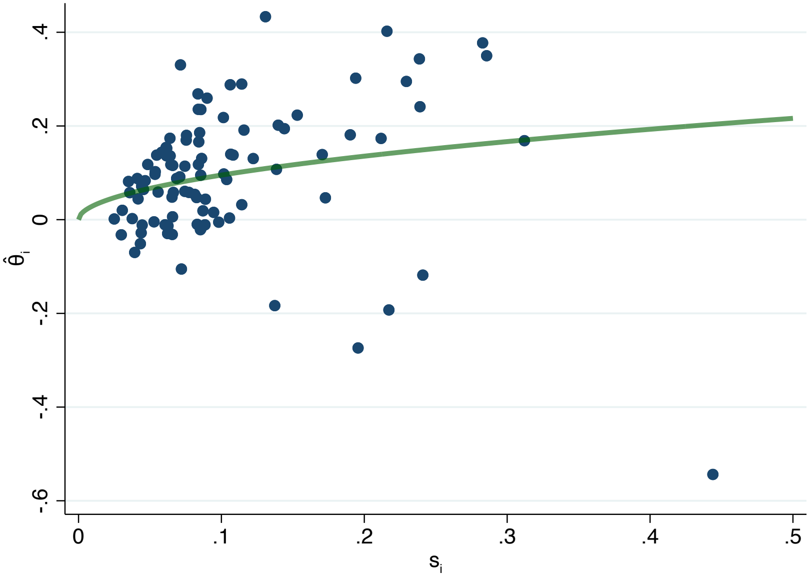
where $(\hat{p}_{ib}, \hat{p}_{iw})$ are sample averages. Standard errors $s_i = \sqrt{\hat{\mathbb{V}}[\hat{\theta}_i]}$ computed via Delta method.

Drop firms with fewer than 40 sampled jobs or callback rates $< 3\%$, leaving $n = 97$

Summary statistics

	Race		Gender	
	White	Black	Male	Female
	(1)	(2)	(3)	(4)
Contact rates	0.256	0.236	0.244	0.248
	(0.004)	(0.003)	(0.004)	(0.004)
Difference	0.020		-0.003	
	(0.002)		(0.003)	
Log difference	0.095		-0.006	
	(0.013)		(0.020)	
# Firms	97			
# Jobs	10,453			
# Apps	78,910			

Race: Standard errors predict point estimates



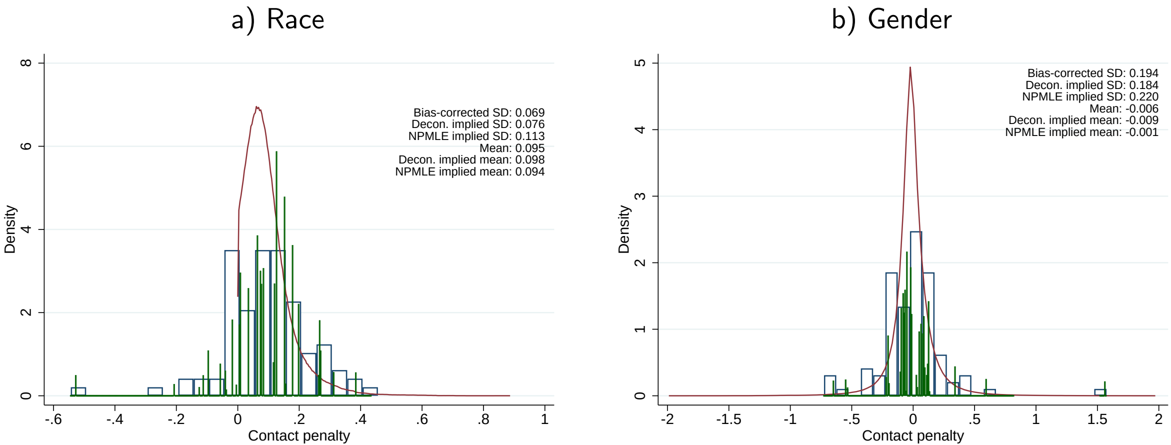
A model of precision-dependence

Work with a model of proportional dependence:

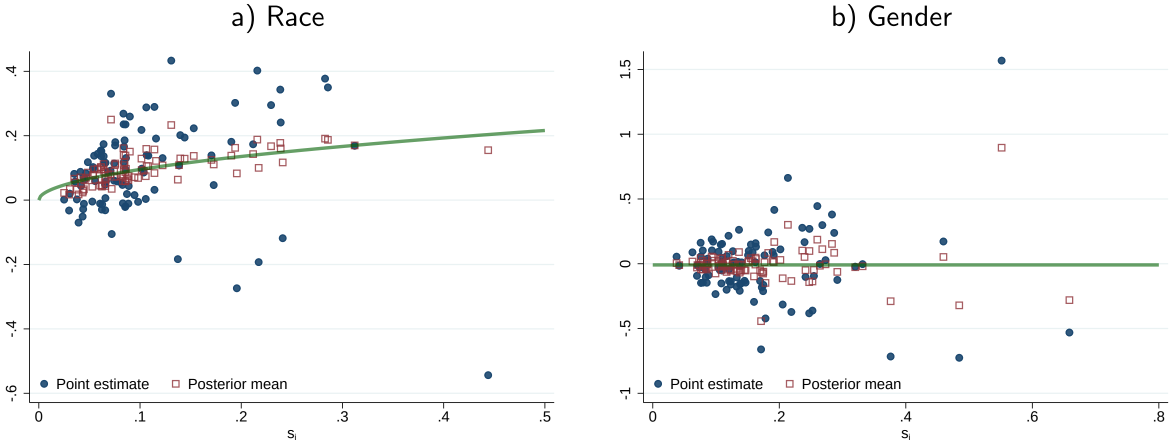
$$\theta_i = \mu + s_i^{\beta} v_i \qquad v_i | s_i \sim G_v$$
$$\hat{\theta}_i = \theta_i + s_i e_i \qquad e_i | s_i, v_i \sim N(0,1)$$

- Estimate μ, β along with $\bar{v} \equiv \mathbb{E}[v_i]$ and $\sigma_v^2 \equiv \mathbb{V}[v_i]$ via GMM details
- Deconvolve standardized residual $\hat{v}_i = (\hat{\theta}_i - \hat{\mu})/s_i^{\hat{\beta}}$ ala Efron (2016) to recover $\hat{G}_v$
- Choose logspline tuning parameter to match GMM estimates of $\bar{v}$ and σ_v^2
- For race, set $\mu = 0$ and assume $G_v(0) = 0$: no firm prefers Black names (test yields $p = 0.94$)

Deconvolution estimates for race and gender



Shrinkage towards firms with similar std errs



Building in industry effects

Allow random effect for industry $\mathrm{k}(i)$:

$$v_i = \underbrace{\eta_{\mathrm{k}(i)}}_{\text{Industry effect}} \times \underbrace{\xi_i}_{\text{Firm Effect}}$$

$$\xi_i \mid s_i, \eta_{\mathrm{k}(i)} \sim G_\xi,$$

$$\eta_k \mid \mathbf{s}_k \sim G_\eta,$$

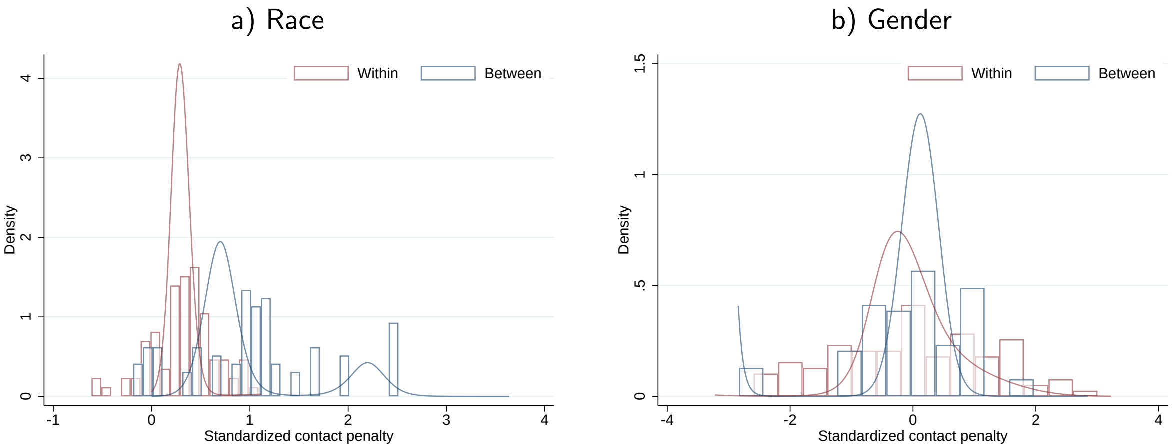
$$\mathbb{E}[\xi_i] = \mu_v, \quad \mathbb{E}[\eta_k] = 1.$$

- Extend Efron (2016)'s deconvolution estimator to hierarchical case, modeling G_ξ and G_η as two fifth-order splines with non-negative support.
- Form posteriors for each θ_i given estimates $\hat{G}_\eta$ and $\hat{G}_\xi$ along with estimates $\{\hat{\theta}_j, s_j\}_{j:\mathrm{k}(j)=\mathrm{k}(i)}$ for all firms in the same industry

GMM estimates: industry R^2 nearly 2/3 for race and 1/2 for gender

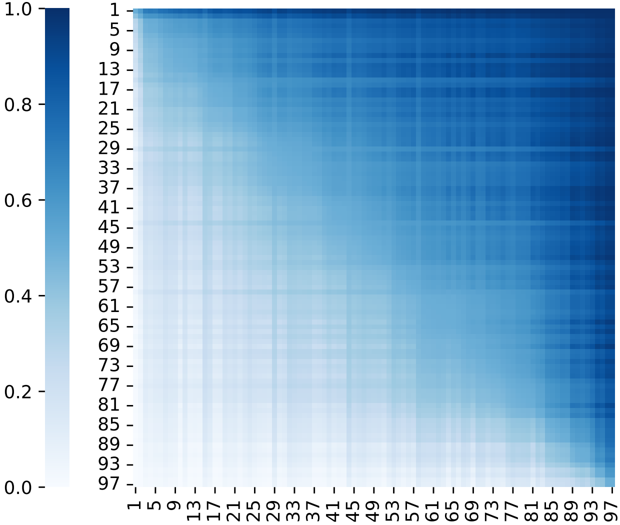
	Race		Gender	
	No industry effects (1)	With industry effects (2)	No industry effects (3)	With industry effects (4)
	a) Model parameters			
β	0.510	0.522	1.255	1.114
	(0.190)	(0.150)	(0.242)	(0.204)
$\bar{v}$	0.308	0.320	0	0
	(0.147)	(0.096)		
μ	0	0	-0.009	0.000
			(0.015)	(0.017)
σ_v	0.207		1.234	
	(0.106)		(0.561)	
σ_η		0.528		0.569
		(0.120)		(0.191)
σ_ξ		0.113		0.645
		(0.054)		(0.213)
J-statistic (d.f.)	0.101	0.087	0.011	1.280
(d. f.)	(1)	(2)	(1)	(2)
	b) Contact penalty distributions			
Mean of θ_i	0.092	0.093	-0.009	0.000
	(0.011)	(0.013)	(0.015)	(0.017)
Std. dev. of θ_i	0.072	0.072	0.180	0.148
	(0.015)	(0.015)	(0.042)	(0.025)
Within share		0.366		0.562
		(0.234)		(0.200)

Significant variation within and between industries



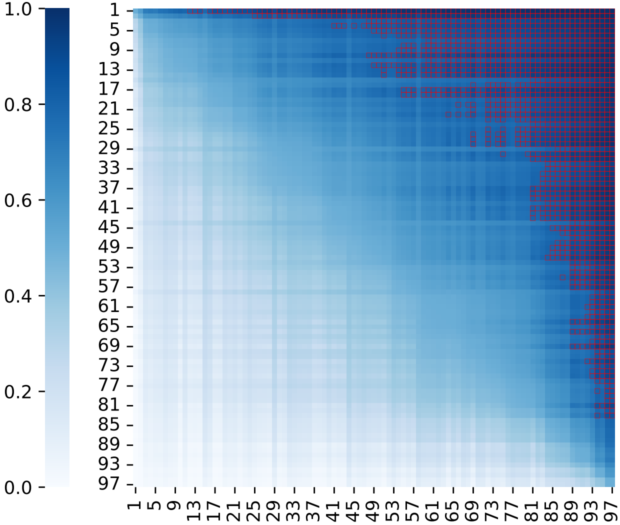
Report Cards: Racial Contact Gaps

Posterior contrasts (π_{ij})



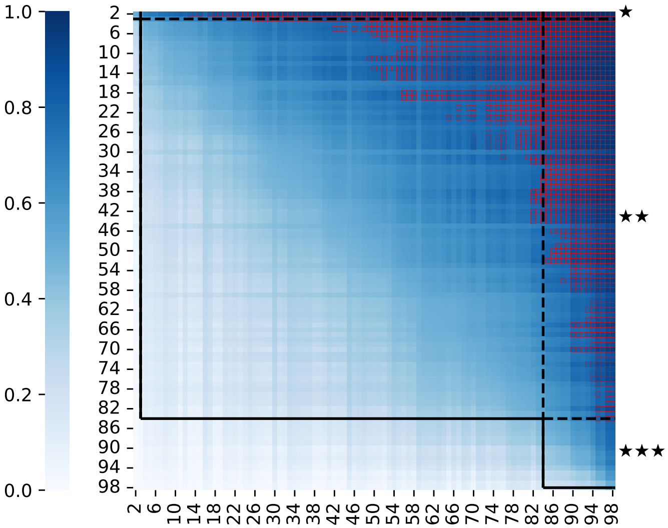
Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

Posterior contrasts (π_{ij})



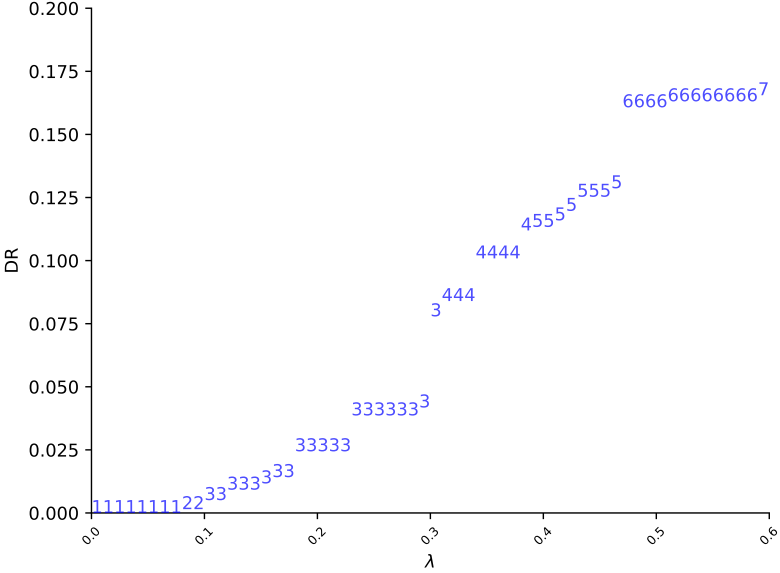
Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

Posterior contrasts (π_{ij})

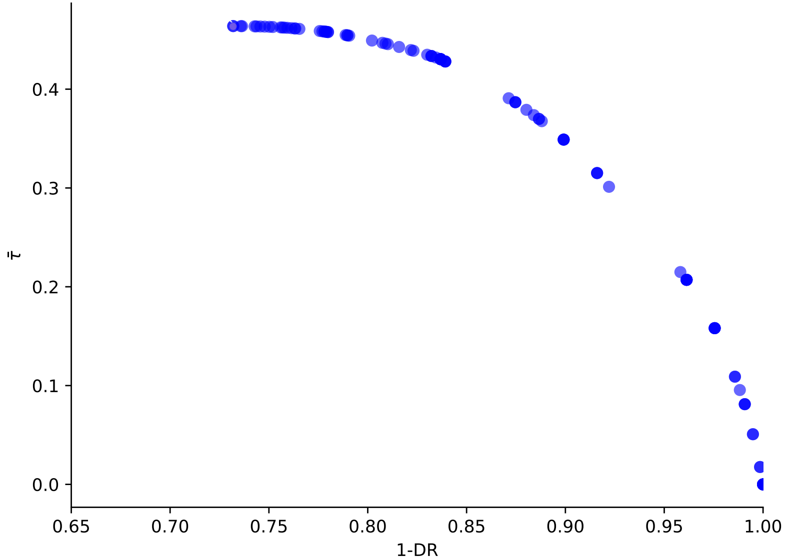


Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

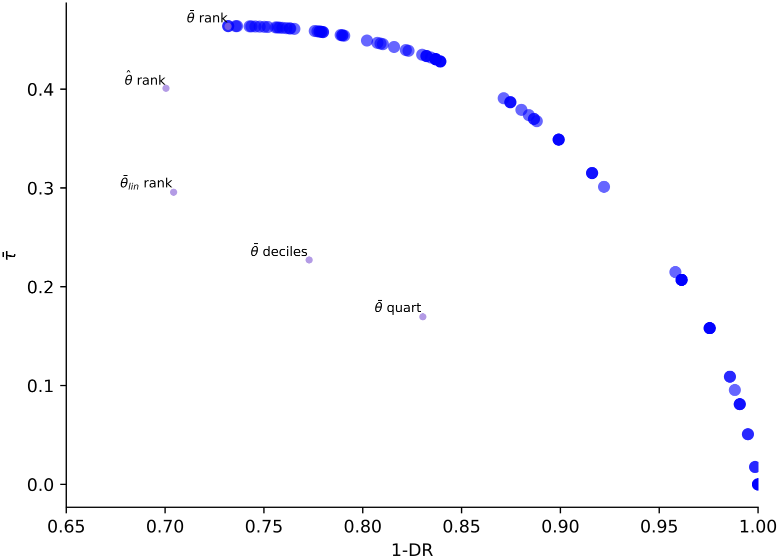
Discordance Rate and # of grades by λ



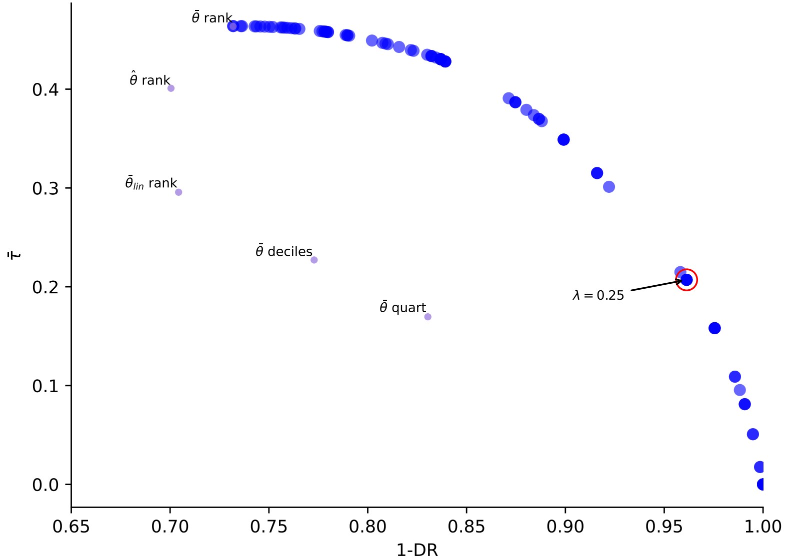
Optimal grades strongly dominate ad-hoc coarsenings



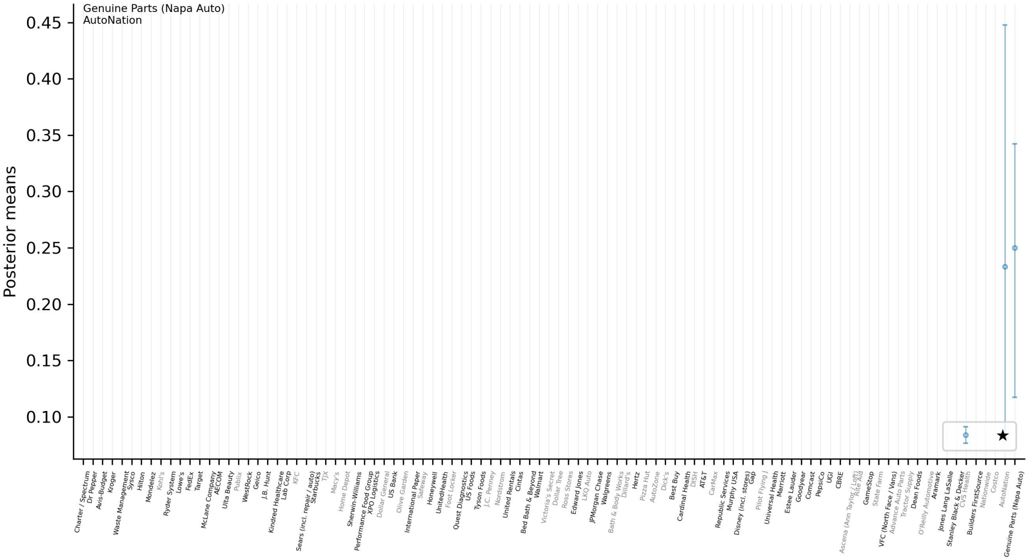
Optimal grades strongly dominate ad-hoc coarsenings



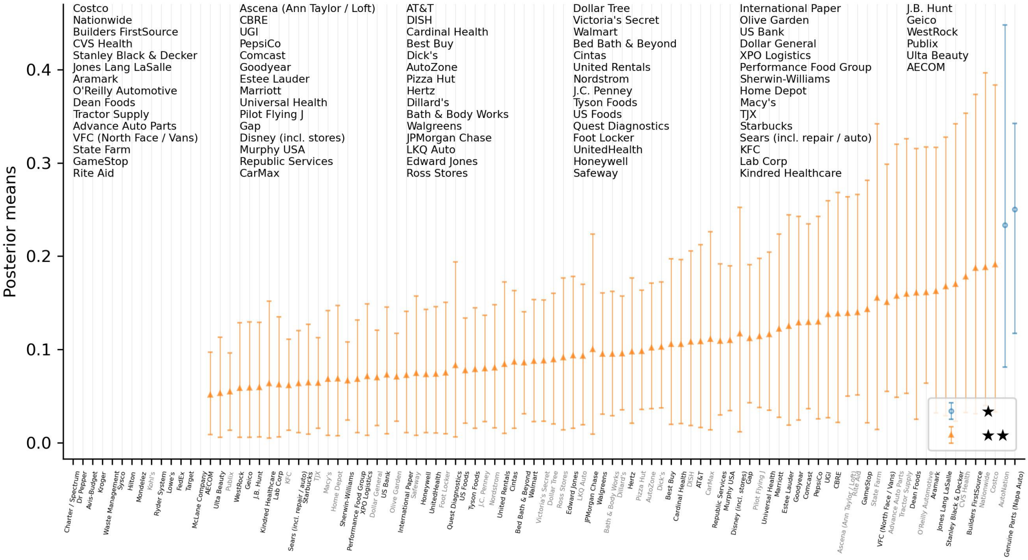
Optimal grades strongly dominate ad-hoc coarsenings



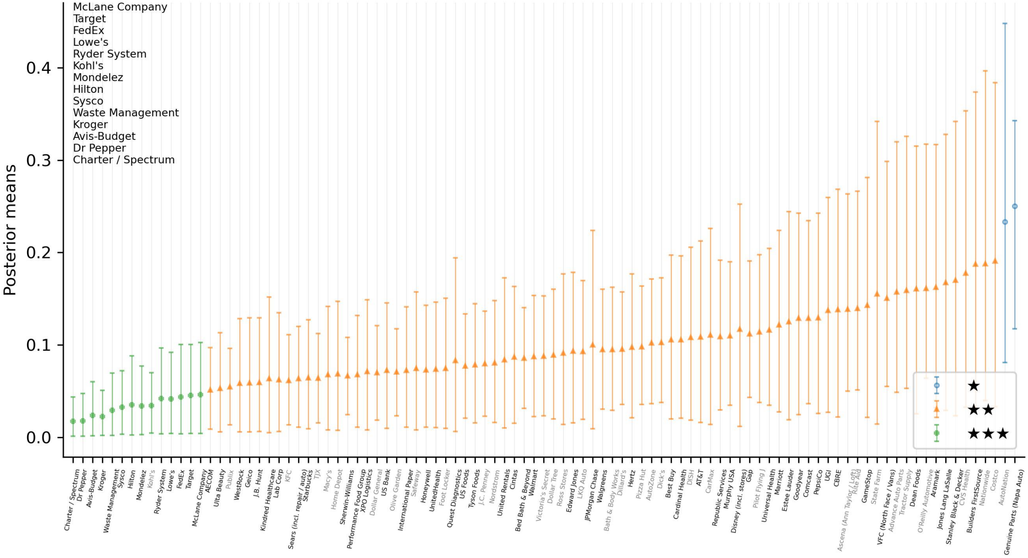
Three total grades, very different conduct estimates, at $\lambda = .25$



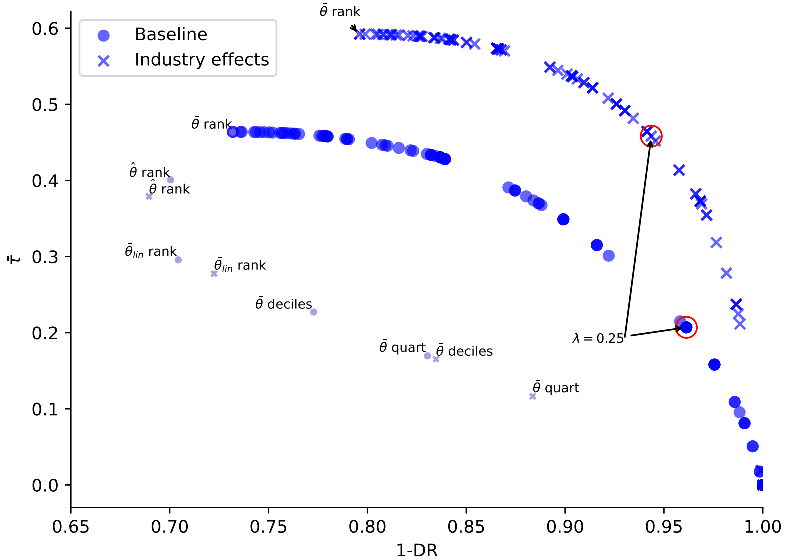
Three total grades, very different conduct estimates, at $\lambda = .25$



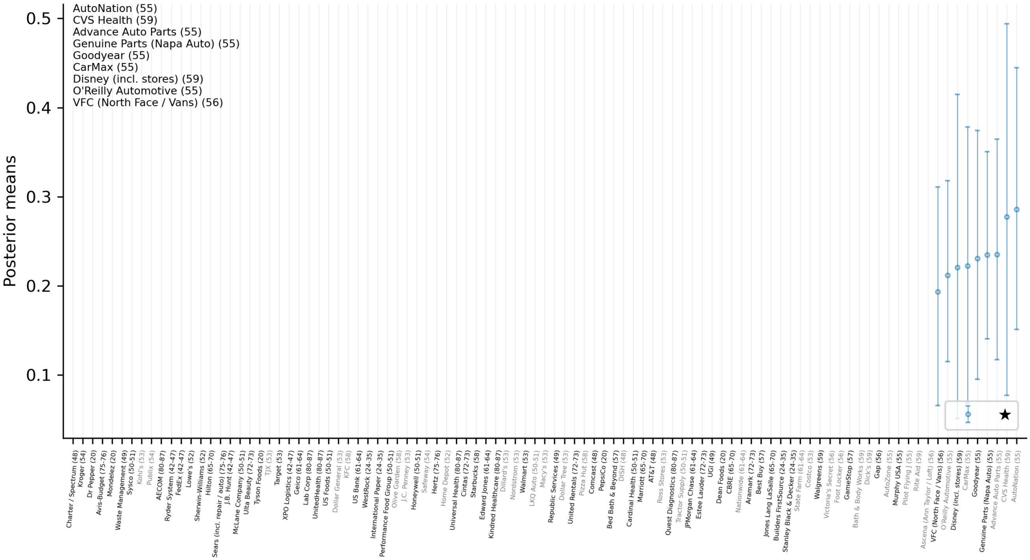
Three total grades, very different conduct estimates, at $\lambda = .25$



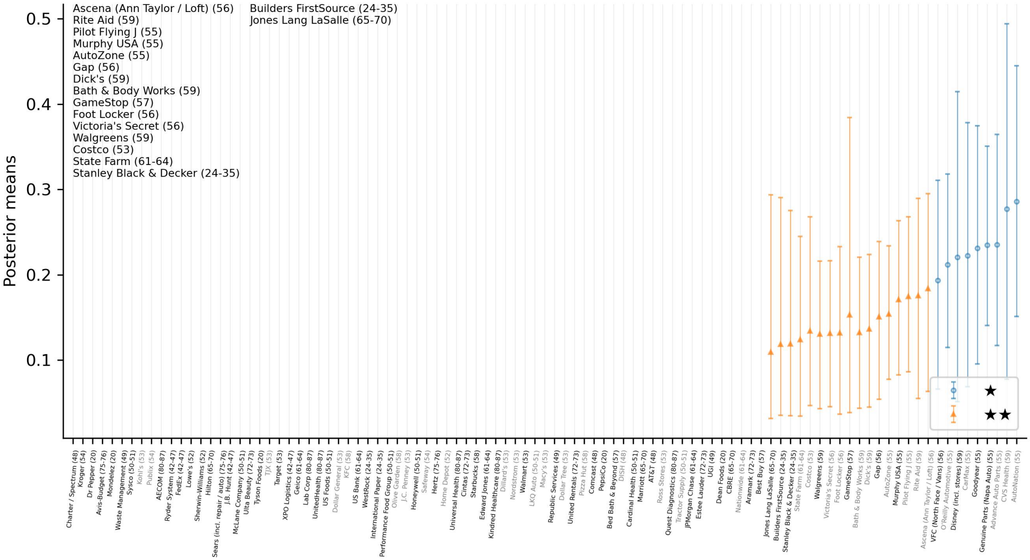
Industry information substantially shifts possibilities frontier



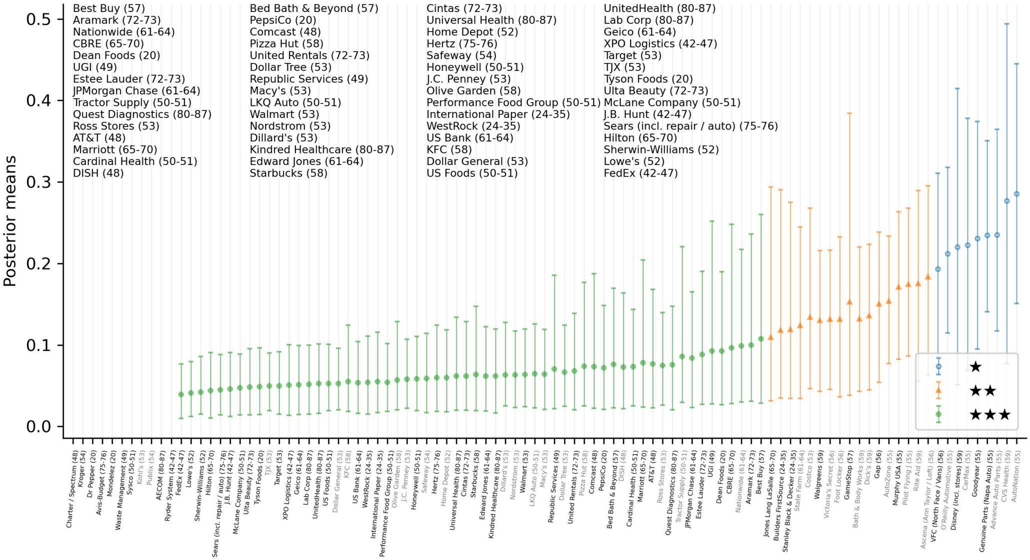
Four total grades at $\lambda = .25$ in industry model



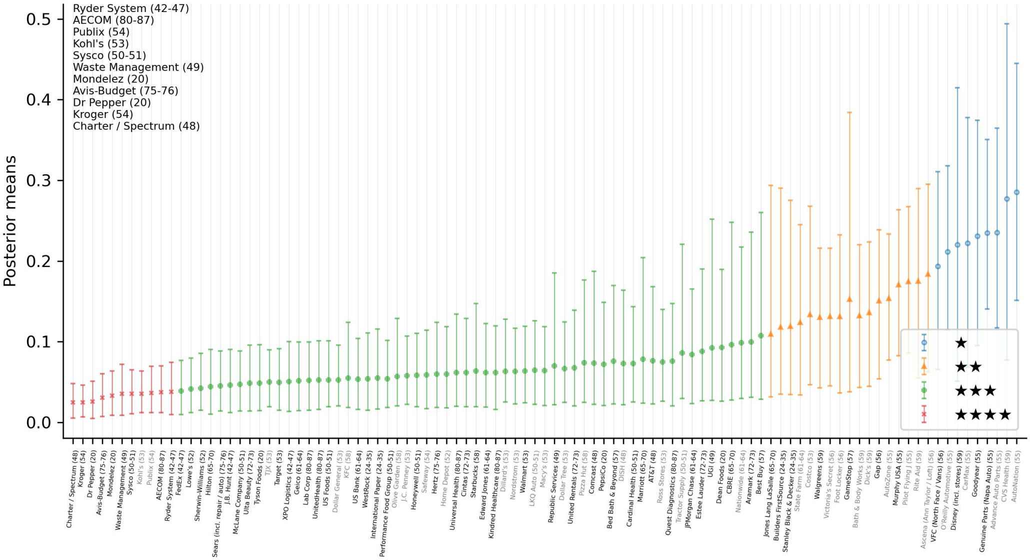
Four total grades at $\lambda = .25$ in industry model



Four total grades at $\lambda = .25$ in industry model



Four total grades at $\lambda = .25$ in industry model



Reliability increasing across non-adjacent grades

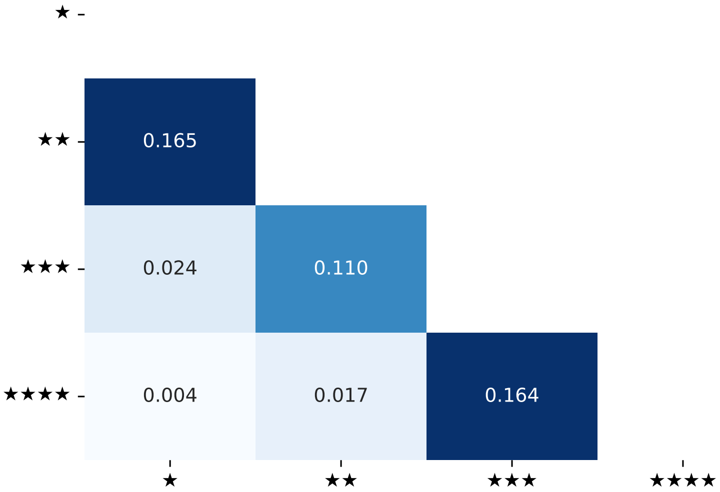
Average posteriors:

0.23

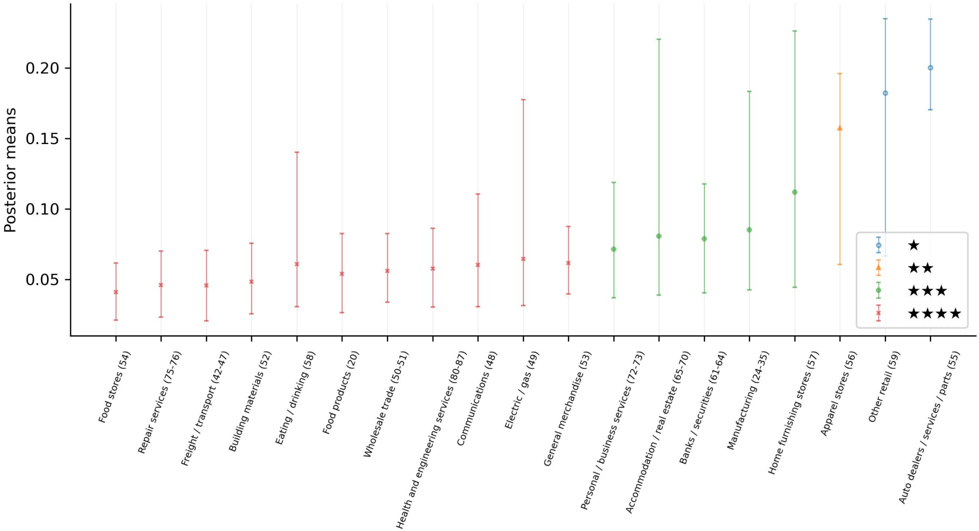
0.14

0.06

0.03



Auto and retail sectors receive lowest grades



Some observations

Two of estimated top 5 discriminators are fed contractors subject to OFCCP oversight

- Fed contractors less biased *on average* but comprise 2/3rds of our sample.
- Top 5 exhibit posteriors means $> 20\%$
- Potential violation of "4/5ths rule" from Uniform Guidelines (1978)

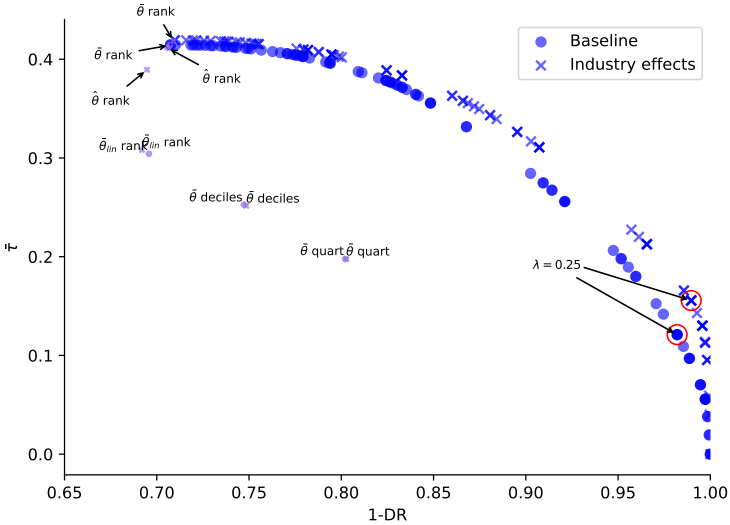
  A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.

Accepting vs failing to reject a null

- Average posterior bias among firms graded as $\star$: 23%
- Average posterior bias among firms graded as $\star\star\star\star$: 3%

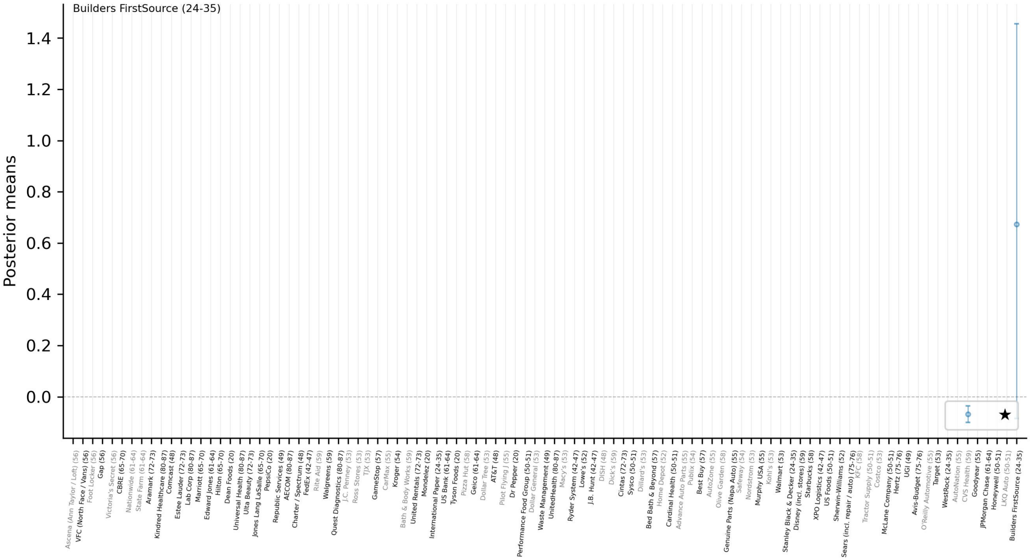
Report Cards: Gender Contact Gaps

Communication tradeoffs for gender

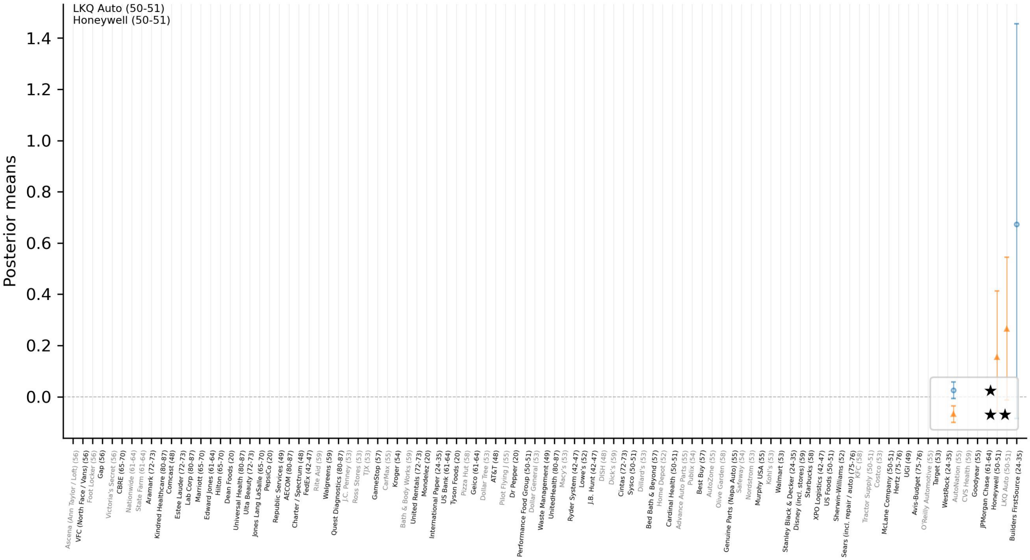


contrasts

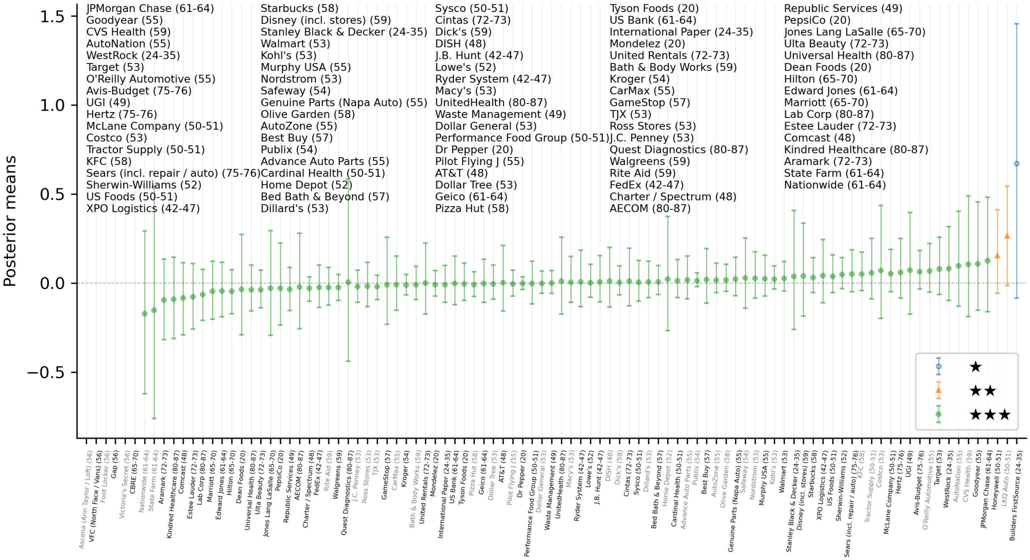
Industry effect gender report card includes 5 grades



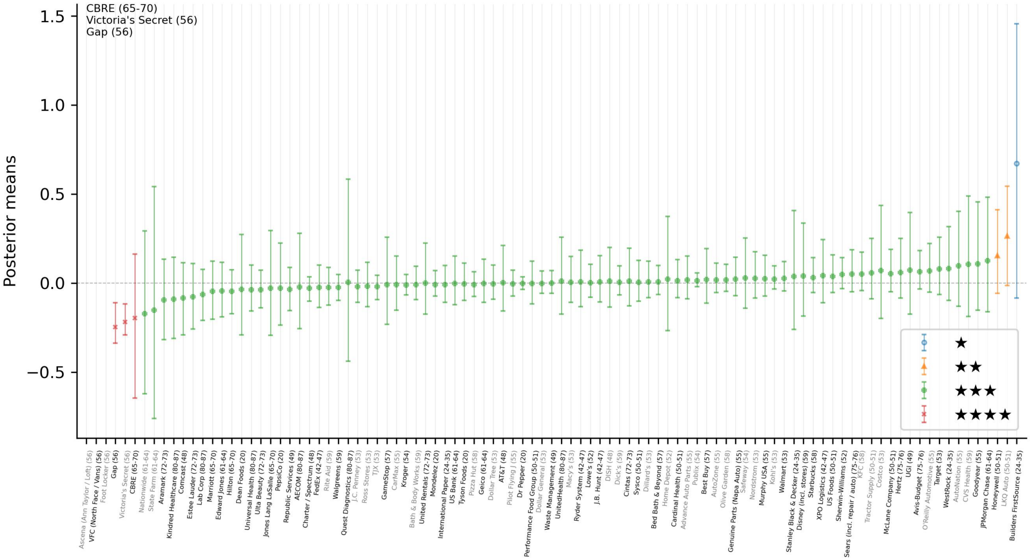
Industry effect gender report card includes 5 grades



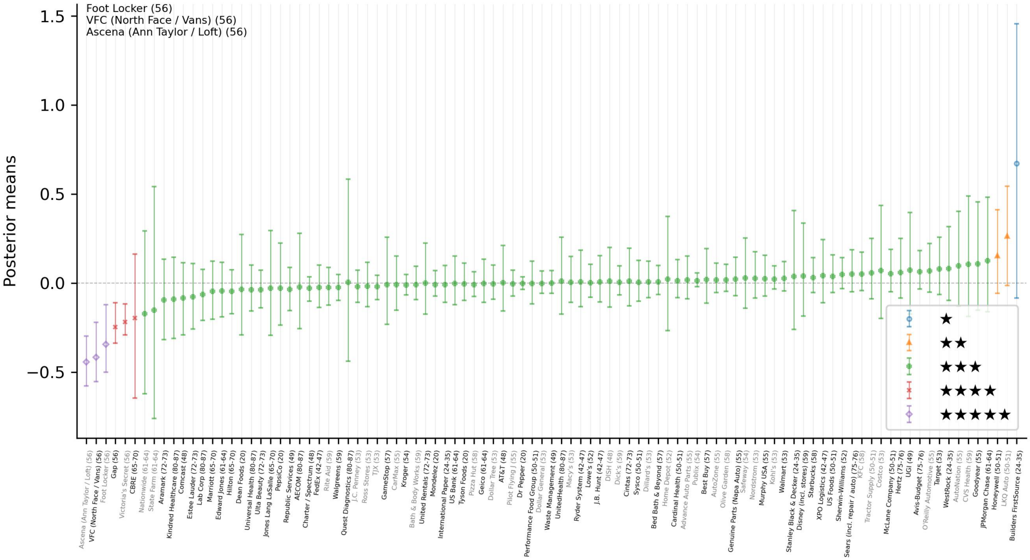
Industry effect gender report card includes 5 grades



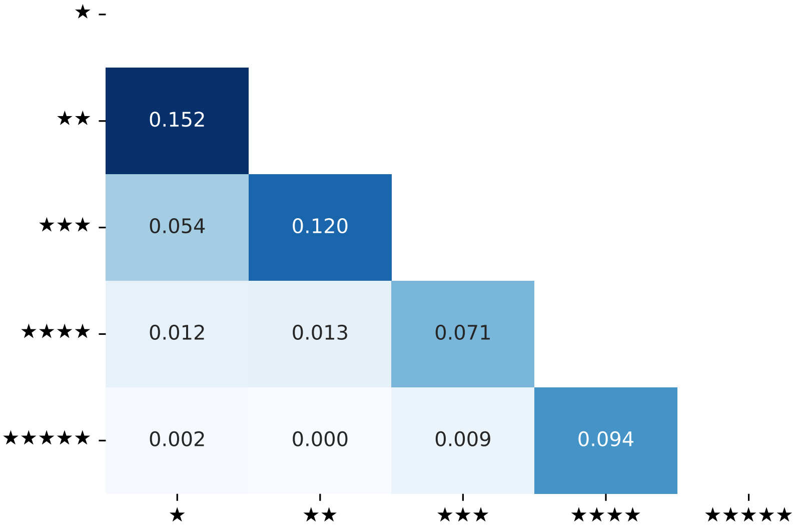
Industry effect gender report card includes 5 grades



Industry effect gender report card includes 5 grades

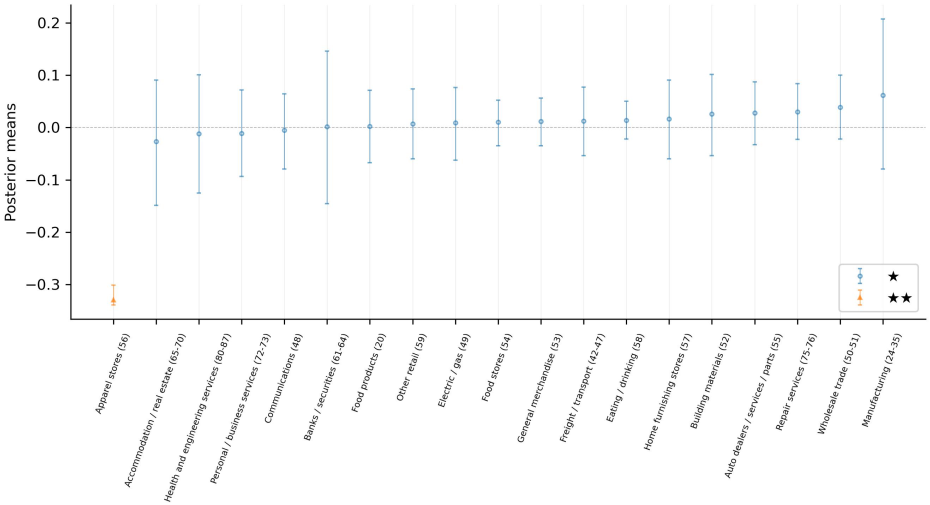


Very confident that firms graded ⋆⋆⋆⋆⋆ prefer women



;

Apparel singled-out at industry level



Recap

New approach to ordinal reporting when concerned about misclassification

- Simple idea: maximize $\bar{\tau} = \mathbb{E}_G[\tau(\theta, d)|Y]$ while limiting DR
- Applicable to many other reporting tasks involving value added or conduct

How much information about discriminatory conduct can be reliably communicated?

- With n grades: $\bar{\tau} = 0.46$, $DR = 0.27$ (or $\bar{\tau} = 0.59$, $DR = 0.20$ w/ industry effects)
- Fixing $\lambda = 0.25$ yields 3 grades, $\bar{\tau} = 0.21$, and $DR = 0.04$ (or 4 grades, $\bar{\tau} = 0.46$, $DR = 0.06$ w/ industry effects)

Ranking package `DRrank` available at `https://github.com/ekrose/drrank`

- Works with any set of posterior probs π_{ij}
- Rapid computation for $n < 500$

DRrank

DRrank is a Python library to implement the Empirical Bayes ranking scheme developed in Kline, Rose, and Walters (2023). This code was originally developed by Hadar Avivi.

Installation:

The package uses the Gurobi optimizer. To use **DRrank** you must first install Gurobi and acquire a license. More guidance is available from Gurobi here. Gurobi offers a variety of free licenses for academic use. For more information, see the following page.

After having successfully set up Gurobipy, install **DRrank** via pip:

```
pip install drrank
```

Usage

1. Load sample data

DRrank grades units based on noisy estimates of a latent attribute. You can construct these estimates however you'd like---all **DRrank** requires is a vector of estimates, $\hat{\theta}_i$, and their associated standard errors, s_i.

To illustrate the package's features, this readme uses the data in *example/name_example.csv*, which contains estimates of name-specific contact rates from the experiment studied in Kline, Rose, and Walters (2023). These contact rates have been adjusted to stabilize their variances using the Bartlett (1936) transformation. Variance-stabilization is useful because the deconvolution procedure used in Step 2 below requires that s_i be independent of θ_i. In cases where variance stabilization is not possible, independence can sometimes be restored by residualizing $\hat{\theta}_i$ against s_i; see Section 5 of Kline, Rose, and Walters (2023) for a detailed example. The transformation used in our names example computes estimates as $\hat{\theta}_i = sin^{-1}\sqrt{\hat{p}_i}$, where $\hat{p}_i$ is share of applications with name i that received a callback. As discussed in the paper, $\hat{\theta}_i$ has asymptotic variance of $(4N_i)^{-1}$, where N_i is the number of applications sent with name i.

Beliefs vs. Experimental Evidence

Perceptions of firm practices

Qualtrics survey (N = 9,189) of beliefs regarding firm recruiting practices

Randomly assigned set of five companies to evaluate

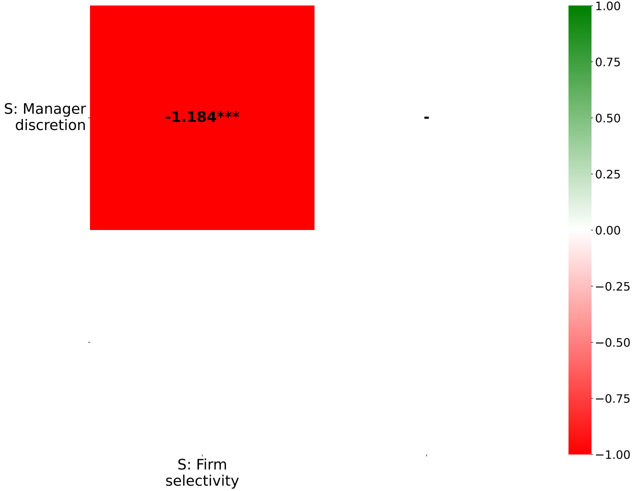
Questions (1-5 scale) all pertain to conduct regarding *entry-level jobs*:

- Please indicate how likely you think it is that each company below would discriminate against (black / female) job-seekers. (**Black / Female discrim.**)
- Please indicate the likelihood that an applicant would be able to successfully pass an interview with each of the following companies (**Firm selectivity**).
- For each company, please indicate how likely you think it is that managers can hire their preferred candidate without input from colleagues or superiors (**Manager discretion**)

Aggregate responses using rank-ordered logit. Firm effects give "wisdom of the crowd."

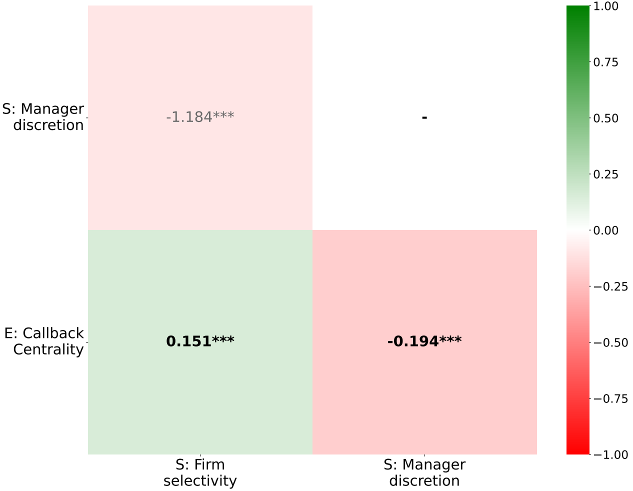
Use std errors to compute bias corrected correlation with experimental contact gaps

Extreme negative correlation btwn perceived discretion and selectivity



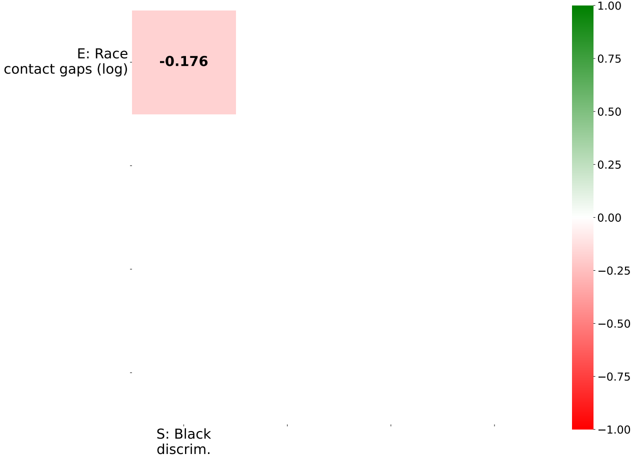
Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Firms believed to exhibit discretion called us from more phone #'s



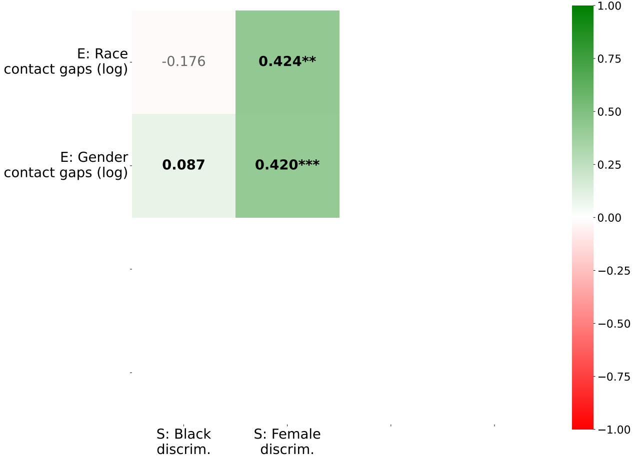
Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Perceived racial discrimination uncorrelated with experimental race gaps



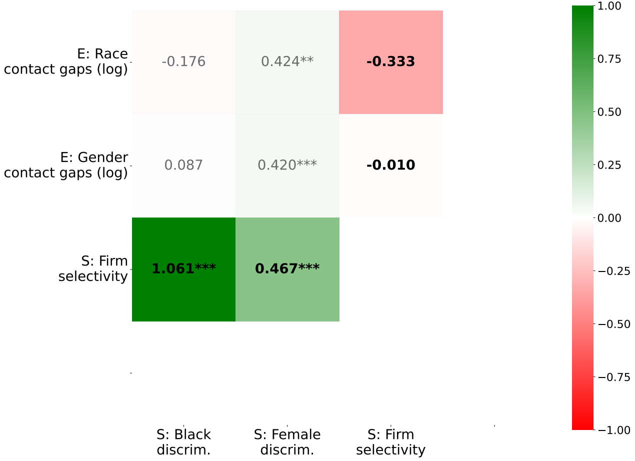
Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Perceived gender discrimination strongly correlated with gender gaps



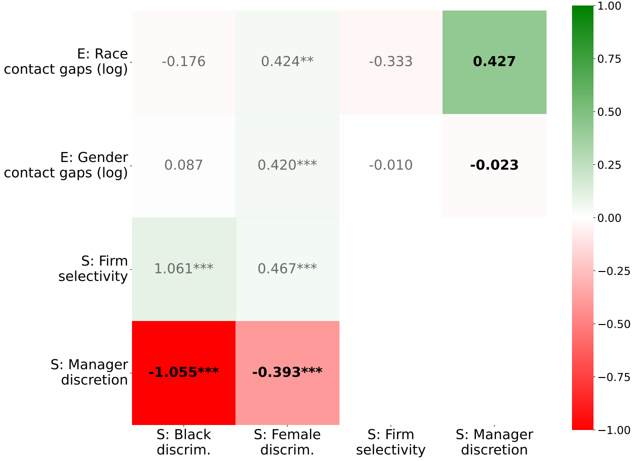
Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Mistaken impression that discrimination pronounced among selective firms



Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Mistaken impression that discretion a negative predictor of bias

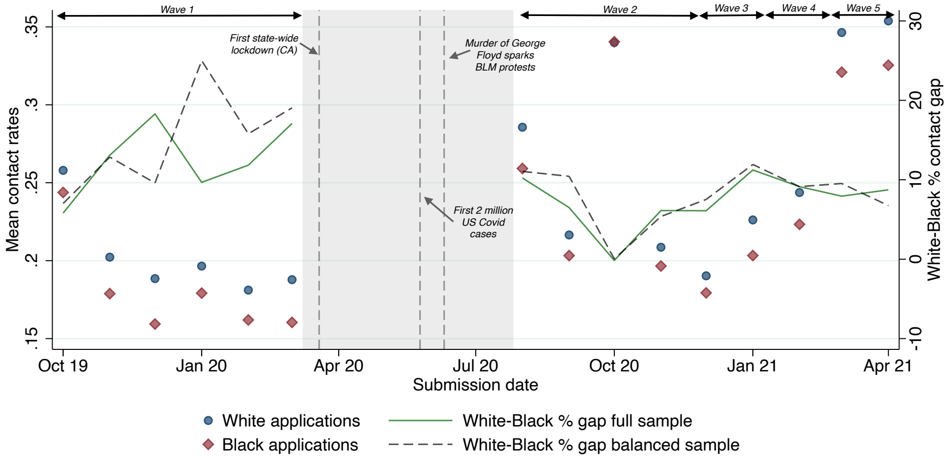


Note: Adjusted Pearson correlation coefficients. E: experimental contact gaps; S: results from the survey.

Taking Stock

- The gender preferences of firms seem to be common knowledge.
- Far less is known about their racial preferences $\Rightarrow$ grades likely to be revelatory.
- Behavioral literature suggests manager discretion a key conduit for bias (Agan et al., 2023). Concordance between perceptions of manager discretion and experimental results corroborates this view.
- Will "sunlight" prove to be the best disinfectant or do firms need guidance about how to reform HR practices?

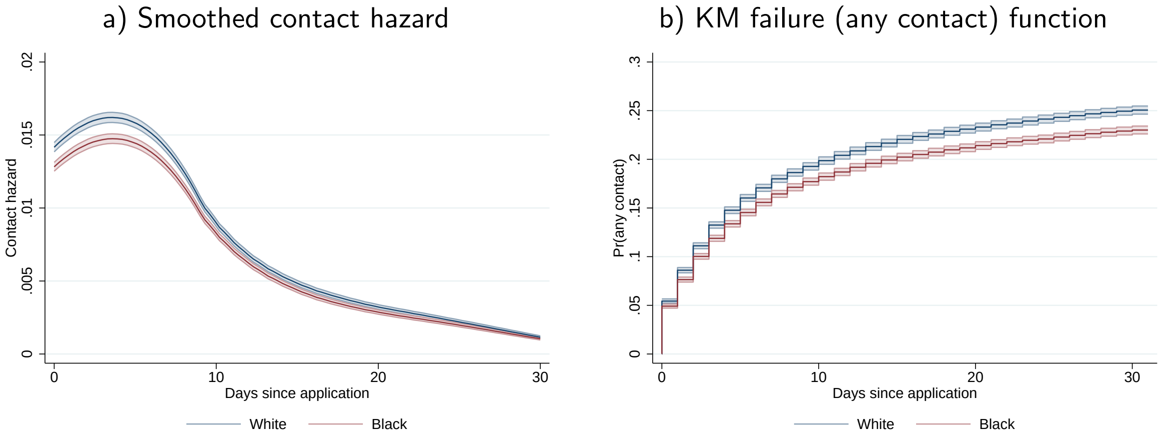
Bonus material



Average Black/white contact gap of 2.1pp, or 9%

- 36% avg. gap reported in meta-analysis of Quillian et al. (2017)
- Level diffs of 3pp in Bertrand and Mullainathan (2004) and 2.6pp in Nunley et al. (2015)
- Discrimination less severe among large firms? (Banerjee et al. 2018)

Contact gap stabilizes by 30 days



Choice properties

Unanimous: Alternative favored in all pairwise comparisons is always chosen

Neutral: Ordering of alternatives does not matter

Reinforcement: Combining data with same preferences does not change ranking of alternatives

Independence of remote alternatives: Relative ordering of adjacent alternatives a_i and a_j depends only on comparisons of a_i and a_j

back

GMM details

Consider the following "studentized" version of $\hat{\theta}_i$:

$$T_i = \frac{\hat{\theta}_i - s_i^{\beta}\mu_v}{\sqrt{s_i^{2\beta}\sigma_v^2 + s_i^2}}.$$

$\mathbb{E}[\hat{\theta}_i|s_i] = E[\theta_i|s_i] = s_i^{\beta}\mu_v \Rightarrow T_i$ should have mean zero

$\mathbb{V}(\hat{\theta}_i|s_i) = s_i^{2\beta}\sigma_v^2 + s_i^2 \Rightarrow T_i$ should have marginal variance one

Combining with independence of v_i and s_i yields moments:

$$\mathbb{E}[T_i] = 0,\ \mathbb{E}[T_i s_i] = 0,\ \mathbb{E}[T_i^2 - 1] = 0,\ \mathbb{E}[(T_i^2 - 1)s_i] = 0. \tag{1}$$

back

GMM details for industry model

$$\mathbb{V}(v_i) = \mathbb{E}[\mathbb{V}(v_i|k)] + \mathbb{V}(\mathbb{E}[v_i|k]) = \mathbb{E}[\eta_k^2]\sigma_\xi^2 + \mathbb{V}(\eta_k \mu_v) = \sigma_\eta^2 \sigma_\xi^2 + \sigma_\xi^2 + \sigma_\eta^2 \mu_v^2$$

Denote the average value of $\hat{v}_i$ in industry k by

$$\bar{v}_k = n_k^{-1} \sum_{i:k(i)=k} \hat{\theta}_i / s_i^\beta = n_k^{-1} \sum_{i:k(i)=k} v_i + n_k^{-1} \sum_{i:k(i)=k} e_i / s_i^\beta,$$

where n_k gives the number of firms in industry k

Variance of $\bar{v}_k$ is $V_k \equiv \left(\sigma_\eta^2 \sigma_\xi^2 / n_k + \sigma_\eta^2 \mu_v^2 + \sigma_\xi^2 / n_k \right) + n_k^{-1} \sum_{i:k(i)=k} s_i^{2(1-\beta)}$

Two more moment conditions:

$$\mathbb{E}\left[(\bar{v}_k - \mu_v)^2 - V_k \right] = 0, \quad \mathbb{E}\left[\left\{ (\bar{v}_k - \mu_v)^2 - V_k \right\} \bar{s}_k \right] = 0.$$

where $\bar{s}_k = n_k^{-1} \sum_{i:k(i)=k} s_i$ denotes the average standard error in industry k back

Extension: weighted loss

Large mistakes more costly. Consider augmented loss function $L^p(\theta, d; \lambda) =$

$$\binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Bigg[\underbrace{1\{\theta_i > \theta_j, d_i < d_j\}(\theta_i - \theta_j)^p + 1\{\theta_i < \theta_j, d_i > d_j\}(\theta_j - \theta_i)^p}_{\text{discordant pairs}} -$$

$$\lambda \Bigg(\underbrace{1\{\theta_i < \theta_j, d_i < d_j\}(\theta_i - \theta_j)^p + 1\{\theta_i > \theta_j, d_i > d_j\}(\theta_j - \theta_i)^p}_{\text{concordant pairs}} \Bigg) \Bigg].$$

The corresponding Bayes risk function takes the linear form

$$\binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \mu_{ji}^{p} d_{ij} + \mu_{ij}^{p}(1 - e_{ij} - d_{ij}) - \lambda \mu_{ji}^{p}(1 - e_{ij} - d_{ij}) - \lambda \mu_{ij}^{p} d_{ij},$$

where $\mu_{ij}^{p} = \mathbb{E}_G[\max\{(\theta_i - \theta_j), 0\}^p \mid Y_i = y_i, Y_j = y_j]$.