# A Discrimination Report Card

Patrick Kline, Evan K. Rose, and Christopher R. Walters*

September 9, 2022

**Abstract**

Understanding which firms discriminate is of great public and academic interest. We use data from a large-scale correspondence experiment to grade the biases against Black names among a collection of large U.S. employers. A novel Empirical Bayes ranking scheme is developed that trades off the informativeness of the assigned grades against the expected frequency of ranking errors. A five grade ranking is found to explain more than one third of the variation in proportional racial contact gaps across firms. Firms awarded the bottom grade exhibit biases against Black names averaging roughly 20%, while those awarded the highest grade exhibit negligible bias. The chances that a randomly selected pair of firms in our analysis sample is misranked by our grading procedure is estimated to be roughly 5%, suggesting the ordinal information conveyed by the grades is highly reliable.

**Preliminary and incomplete; Please do not circulate**

*Kline: UC Berkeley and NBER, pkline@econ.berkeley.edu; Rose: University of Chicago and NBER, ekrose@uchicago.edu; Walters: UC Berkeley and NBER, crwalters@econ.berkeley.edu. We thank Hadar Avivi for outstanding research assistance on this project.

# 1 Introduction

> Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.
>
> *Louis Brandeis*

Though half a century has passed since the enactment of the Civil Right Act, there is ample evidence that American employers still discriminate based upon legally protected characteristics including sex and race (Guryan and Charles, 2013; Bertrand and Duflo, 2017; Quillian et al., 2017; Quillian, Lee and Oliver, 2020). One reason for the stubborn persistence of employer discrimination is the paucity of reliable information regarding the tendencies of specific organizations to engage in such illicit behavior. Job seekers cannot direct their search effort away from biased firms if the identities of these firms are unknown. Corporate executives have little incentive to search for new, more equitable, screening practices if they cannot benchmark their current biases to those of their peers.

This paper details the construction of a discrimination report card designed to communicate objective information regarding the relative biases of a broad collection of Fortune 500 companies. Our analysis relies upon a large-scale correspondence experiment that involved sending as many as one thousand job applications to each of 108 firms. Like past correspondence experiments, we manipulated employer perceptions of applicant race and sex using distinctive names. In Kline, Rose and Walters (2021) we established that experimental applications listing distinctively Black names were roughly seven percent less likely to be contacted than those listing distinctively white names. This penalty displays striking heterogeneity across companies: applying Empirical Bayes (EB) deconvolution methods, we found that the top 20% of discriminating firms were responsible for nearly half of the employer contacts lost as a result of bias against Black names.

While our past work established that discrimination is highly concentrated among a small set of employers, the estimate for any particular employer is subject to substantial sampling error. This statistical noise presents an obstacle to producing reliable estimates for specific firms. Though scientific communication is generally aided by transparency (Andrews and Shapiro, 2021), lay audiences may find it challenging to translate point estimates and standard errors into conclusions of interest. In particular, experimental evidence from psychology suggests audiences often prefer to reason categorically and frequently depart from Bayesian reasoning when drawing inferences (Mullainathan, 2002; Mullainathan, Schwartzstein and Shleifer, 2008; Bordalo et al., 2016), which may generate both over- or under-reactions to statistical evidence.

The present paper aims to limit the scope for misinterpretation of statistical results by developing new methods for assigning ordinal "grades" to firms (or other units) while

maintaining reliability guarantees on the accuracy of the resulting classification scheme. The proposed grading scheme is designed to maximize the amount of information expected to be communicated by our grades subject to a constraint on the expected frequency of firm misrankings engendered by these grades. The information content of our grades is quantified by Kendall's $\tau$ measure of correlation (Kendall, 1938) between our proposed ranking and the true ranking of firms' latent discriminatory behavior. The reliability of these grades is quantified by an analogue of the False Discovery Rate (Benjamini and Hochberg, 1995; Storey, 2002) that we term the Discordance Rate (DR). The DR between two grades gives the expected probability that a firm assigned the higher grade discriminates less than a firm assigned the lower grade. We also develop an extension of the DR that weights mistakes by the cardinal magnitude of the mistake, which captures the idea that large mistakes are more costly than small ones.

We show that the tradeoff between information and reliability posited by our optimization problem can be motivated in terms of a series of lotteries over firm pairs where an analyst guesses which member of each pair exhibits greater bias against Black names. With multiple gambles of this form, the analyst faces logical transitivity constraints requiring the pairwise comparisons to be consistent with a coherent underlying ranking. A tuning parameter $\lambda$ trades off the gains of correctly ranking pairs of firms against the costs of misclassifying them. When $\lambda = 1$, it is optimal to assign every firm a unique grade that maximizes the expected rank correlation with the true latent discrimination levels. Interestingly, these grades turn out to be closely connected to classic proposals for preference aggregation via pairwise elections found in the social choice literature (Borda, 1784; Condorcet, 1785; Young and Levenglick, 1978; Young, 1986) where the posterior probability that one firm is more biased than another serves as a "vote share." When $\lambda < 1$ it is only optimal to rank firm pairs that can confidently be distinguished, potentially yielding very few distinct grades. This coarsening protects against misinterpretation but entails an information cost in terms of reduced correlation with the true ranks.

The grades produced by our procedure yield an easy to interpret classification of firms into groups that facilitate pairwise comparisons of firm conduct. Notably, these grades allow an assessment not only of which firms are discriminating heavily against Black names but also which firms are **not** discriminating against such names. The latter sort of information may prove especially useful in the hunt for "best practices" that lead to inclusive hiring.

Applying non-parametric EB tools to our experiment, we compute optimal grades subject to the restriction that the DR not exceed 5%. In our baseline specification, this restriction yields five grade levels. These grades capture roughly 17% of the between-firm variation in proportional contact penalties and exhibit an expected rank correlation with the true penalties of 0.22. Although our grading system reflects only ordinal considerations, we estimate that firms awarded the lowest grade exhibit biases against Black names

of roughly 20%, while firms awarded the highest grade exhibit negligible biases.

Building on our previous finding that industries vary widely in their treatment of Black names, we develop a hierarchical random effects model that allows us to exploit industry affiliation in assessing the EB risk of grade assignment. Incorporating industry affiliation into the procedure again yields five grades when the DR is limited to 5%. However, these grades explain 36% of the variation in contact penalties across firms and yield a correlation with the latent ranks of 0.34.

These more informative industry-dependent grades constitute our preferred ranking of the companies in our experiment. Even in sectors where discrimination is widespread, some firms are less biased than others. To the extent that these differences are driven by HR practices, an opportunity appears to be present for firms that scored poorly to improve their behavior by imitating the practices of those who scored more highly.

Theoretically, our work extends a burgeoning literature on EB ranking methods. A large empirical literature ranks teachers, schools, hospitals, and neighborhoods using James-Stein style shrinkage rules (e.g., Chetty, Friedman and Rockoff, 2014; Chetty et al., 2018). While these applications typically focus on posterior mean predictions, Laird and Louis (1989) proposed directly computing posterior ranks under a normality assumption on the latent heterogeneity. Posterior ranks may be noisy, however, making them a potentially fraught basis for decisionmaking. Building on an analogy with multiple testing, Gu and Koenker (2020) consider use of non-parametric EB methods to select tail performers subject to FDR constraints. Our proposal generalizes the approach in Gu and Koenker (2020) by accommodating more than two labels and avoids the requirement to treat one of the labels as a null hypothesis. More recent work by Gu and Koenker (2022) considers a ranking of journals based on pairwise citation counts using a penalized Bradley-Terry model (Bradley and Terry, 1952). While our proposed approach shares Gu and Koenker (2022)'s focus on pairwise differences, the method does not require pairwise data on tournaments and offers a more transparent notion of reliability while remaining computationally tractable.

# 2 The experiment

We construct discrimination report cards based on the massive resume correspondence experiment analyzed in Kline, Rose and Walters (2021). The experiment's sampling frame begins with the 2018 list of companies in the Fortune 500. Attention was then restricted to firms with sufficient geographic variation and entry-level job posting for our experiment to be feasible. Over the course of the study, 125 entry-level job vacancies were sampled from each of these employers, with each vacancy corresponding to an establishment in a different U.S. county. Sampling was organized in a series of 5 waves, with a target of 25 jobs sampled for each firm in each wave. The majority of firms (72) were sampled

in all waves; the rest were excluded in some waves due to COVID-19 and technological interruptions. We attempted to send each sampled job four pairs of applications, with each pair consisting of one Black applicant and one white applicant. Some vacancies received fewer than 8 total applications because the job opening closed while applications were still in progress. Our final sample included roughly 84,000 applications.

To signal race and gender, we followed previous correspondence experiments and used distinctive names. Our set of names started with that of Bertrand and Mullainathan (2004), who used 9 unique names for each race and gender group. This list was supplemented with 10 additional names per group from a database of speeding tickets issued in North Carolina between 2006 and 2018.[1] One application within each pair was randomly assigned a distinctively white name while the other was randomly assigned a distinctively Black name. Assignment of sex was not stratified, implying fifty-percent of names are distinctively female, and the rest distinctively male. Each fictitious applicant was randomly assigned a large set of additional characteristics, including educational and previous employment histories.

Our primary outcome is whether an employer attempted to contact the fictitious applicant. Phone numbers and e-mail addresses assigned to the fictitious applicants were monitored to determine when employers reached out for an interview. Contact information was assigned to ensure that no two applicants to the same firm shared an e-mail address or phone number. Our analysis focuses on whether the employer attempted to contact an applicant by any method within 30 days of applying. Full details on the experimental methods and results are available in Kline, Rose and Walters (2021).

# 3 Decision problem

Consider the problem of ranking a collection of $n$ firms, indexed by $i \in \{1, \ldots, n\} \equiv [n]$, according to their values of a scalar measure of discriminatory bias $\theta_i \in \mathbb{R}$. The decision variable $d_i \in [n]$ gives the "grade" assigned to firm $i$. For each firm $i \in [n]$, we have the measurements $Y_i = (\hat{\theta}_i, s_i)$, where $\hat{\theta}_i$ is a consistent estimate of $\theta_i$ and $s_i$ is that estimate's asymptotic standard error. We assume the $\{\hat{\theta}_i\}_{i=1}^n$ are mutually independent and that $\hat{\theta}_i \sim \mathcal{N}(\theta_i, s_i^2)$. The normality assumption can be justified as an asymptotic approximation as each $\hat{\theta}_i$ is based on the contact rate of a large number of underlying applications.

It is convenient to recast the problem of ranking $n$ firms as that of ranking all $\binom{n}{2}$ pairs of firms subject to a set of transitivity constraints. Correctly ranking the bias of a

[1] We classified a name as racially distinctive if more than 90% of individuals with that name are of a particular race, and selected the most common distinctive Black and white names for those born between 1974 and 1979. We assembled distinctive last names from the 2010 U.S. Census, selecting names with high race-specific shares among those that occur at least 10,000 times nationally. See Kline, Rose and Walters (2021) for a complete list of names.

pair of firms yields a *concordance* while ranking the pair incorrectly yields a *discordance*. A pair can also be deemed a tie, which yields neither a discordance nor a concordance.

## 3.1 Gambling over ranks

To build intuition it is helpful to first consider the problem of deciding on the rank of a single pair of firms $i$ and $j$ based upon realizations $(y_i, y_j)$ of independent signals $(Y_i, Y_j)$. Suppose that correctly ranking the pair yields payoff $\lambda \in [0, 1]$ while incorrectly ranking them yields payoff -1. In considering this lottery we view the pair $(\theta_i, \theta_j)$ as *i.i.d.* draws from a prior distribution with CDF $G : \mathbb{R} \to [0, 1]$. The posterior probability that firm $i$ is more biased than firm $j$ will be denoted $\pi_{ij} = \Pr(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j)$. The expected utility of assigning grades $d = (d_1, d_2) \in \{1, 2\}^2$ to these firms can therefore be written

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

The optimal policy is a simple posterior threshold rule:

- Set $(d_i = 2, d_j = 1)$ iff $\pi_{ij} > \frac{1}{1+\lambda}$.
- Set $(d_i = 1, d_j = 2)$ iff $1 - \pi_{ij} > \frac{1}{1+\lambda}$.
- Otherwise, set $d_i = d_j$.

When $\lambda = 1$, it is optimal to follow a maximum a posteriori (MAP) rule, assigning the higher rank to whichever firm has a probability greater than $1/2$ of having the largest value of $\theta$. Assuming $G$ is continuous, ties are measure zero; hence, it is always optimal to rank one member of the firm pair above the other. When $\lambda < 1$ it is better to declare pairs of firms with $\pi_{ij}$ near $1/2$ as tied rather than risk ranking them incorrectly. The quantity $(1 - \lambda)$ can therefore be thought of as measuring discordance aversion.

## 3.2 Compound Loss

Now consider the case where we can gamble on the relative rank of all $\binom{n}{2}$ pairs of firms. Kendall's classic $\tau$ measure of rank correlation can be defined as the share of pairs yielding a concordance minus the share yielding a discordance. The loss function we propose is a generalization of $\tau$ indexed by a scalar $\lambda \in [0, 1]$ that controls the benefit of a concordance relative to the cost of a discordance.

Letting $\theta = (\theta_1, \ldots, \theta_n)'$ denote the vector of latent biases and $d = (d_1, \ldots, d_n)'$ a vector of assigned grades, our loss function can be written:

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

While every discordant pair yields a loss of 1, every concordant pair reduces loss by $\lambda$. When $\lambda = 1$ the loss function equals minus one times Kendall's $\tau$ measure of rank correlation between $d$ and $\theta$, which we will denote $\tau(d, \theta)$. When $\lambda < 1$, ranking mistakes are more costly than forgone concordances, which creates an incentive to deem the pair a tie.

Building on the insight that $\tau(d, \theta) = -L(d, \theta; 1)$, we can also write the loss function:

$$\begin{aligned} L(d, \theta; \lambda) &= (1 - \lambda) \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Big[ 1\{\theta_i < \theta_j, d_i < d_j\} + 1\{\theta_i > \theta_j, d_i > d_j\} \Big] - \tau(d, \theta) \\ &= (1 - \lambda)(1 - DP) - \tau(d, \theta) \end{aligned}$$

where the quantity $DP = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} [1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}]$ is the "Discordance Proportion."

The Discordance Proportion gives the share of firm pairs that are misranked according to their grades. If one views our decision problem as a series of tests of the null hypothesis that each member of a pair exhibits equally discriminatory conduct, then the $DP$ is a directional (sometimes called type III) error rate. That is, we reject equality in favor of an erroneous alternative, yielding a discordance. This representation clarifies that the parameter $\lambda$ trades off the desire to accurately communicate information to the audience by maximizing $\tau(d, \theta)$ against concern about misclassifying the ranks of firms, as reflected by $DP$.

## 3.3 Risk function

Suppose we have an *i.i.d.* prior $G : \mathbb{R} \to [0, 1]$ over the elements of $\theta$ and that we observe a realization $y \in \mathbb{R}^n$ of the random vector $Y$. The posterior probability that firm $i \in [n]$ is more biased than firm $j \neq i$ will again be denoted $\pi_{ij} = \Pr(\theta_i > \theta_j | Y = y)$. Taking

expectations and rearranging, the Bayes risk of assigning grades $d \in [n]^n$ can be written

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

The optimal grades $d^*(\lambda)$ minimize $\mathcal{R}(d;\lambda)$. To simplify this minimization problem, it is convenient to recast the relevant decision variables as the $\binom{n}{2}$ pairwise indicators $d_{ij} = 1\{d_i > d_j\}$ and $e_{ij} = 1\{d_i = d_j\}$. Transitivity requires that for any triple $(i,j,k) \in [n]^3$ the following constraints hold:

$$d_{ij} + d_{jk} \leq 1 + d_{ik}, \tag{3}$$

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

$$e_{ij} + e_{jk} \leq 1 + e_{ik}.$$

Hence, we can rewrite the problem of choosing $d \in [n]^n$ to minimize (2) as that of choosing the binary indicators $\{d_{ij}, e_{ij}\}_{i=2,j=1}^{i=n,j=i}$ to minimize

$$\sum_{i=2}^{n} \sum_{j=1}^{i} (1-\pi_{ij})\, d_{ij} + \pi_{ij}\,(1 - e_{ij} - d_{ij}) - \lambda\,(1-\pi_{ij})\,(1 - e_{ij} - d_{ij}) - \lambda\pi_{ij} d_{ij}, \tag{4}$$

subject to the transitivity constraints in (3) and the logical constraint that $e_{ij} + d_{ij} + d_{ji} = 1$ for all $(i,j) \in [n]^2$. Note that both the objective (4) and the constraints are linear in the control variables. This reformulation therefore yields an integer linear programming problem, the solution to which can be computed with standard optimization packages.

## 3.4 The role of $\lambda$

To develop intuition for the role that $\lambda$ plays in the nature of the solution to our linear programming problem, it is useful to consider the task of ranking a single pair in (4) while ignoring cross-pair constraints. Setting $d_{ij} = 1$ lowers risk whenever

$$\underbrace{1 - \pi_{ij}\,(1+\lambda)}_{d_{ij}=1 \ (i>j)} < \min \left\{ \underbrace{0}_{e_{ij}=1 \ \text{(tie)}}, \underbrace{\pi_{ij}\,(1+\lambda) - \lambda}_{d_{ij}=e_{ij}=0 \ (i<j)} \right\}.$$

Or equivalently, when

$$1 - \pi_{ij} < \min \left\{ \frac{\lambda}{1+\lambda}, \pi_{ij} \right\}.$$

For $\lambda \in [0,1]$, we can rewrite this condition as

$$\pi_{ij} > \frac{1}{1+\lambda}. \tag{5}$$

When $\lambda = 1$, it is optimal to choose $d_{ij} = 1\{\pi_{ij} > 1/2\}$, which can be thought of as a Maximum a Posteriori (MAP) estimate of the pairwise rank. Aside from the (highly improbable) case where $\pi_{ij} = 1/2$, this MAP assignment rule will not involve ties. Consequently, each firm will receive a distinct grade. It is possible to show that these grades will additionally satisfy the transitivity restrictions in (3).

When $\lambda \in (0,1)$, greater posterior certainty is required to conclude that $d_{ij} = 1$ and values of $\pi_{ij}$ near $1/2$ will yield ties. As $\lambda$ approaches zero, fewer distinct grades will be assigned. When $\lambda = 0$, all $n$ firms are assigned the same grade because $\pi_{ij} \leq 1$.

The coarse grades that result when $\lambda < 1$ can generate transitivity violations (i.e., Condorcet cycles) in indifferences that force the constraints in (3) to bind. The following three firm example illustrates the problem.

**Example 1** (Three firms, two-point prior)**.** Suppose $\theta_i \in \{\underline{\theta}, \bar{\theta}\}$ with $\underline{\theta} < \bar{\theta}$ and let $\omega_i = \Pr[\theta_i = \bar{\theta} | Y_i = y_i]$. Then

$$\pi_{ij} = \Pr[\theta_i > \theta_j | Y_i = y_i, Y_j = y_j] = \Pr[\theta_i = \bar{\theta}, \theta_j = \underline{\theta} | Y_i = y_i, Y_j = y_j] = \omega_i(1-\omega_j).$$

Let $\lambda = 1/4$, which implies $(1+\lambda)^{-1} = 0.8$. If $(\omega_1, \omega_3) = (0.9, 0.1)$ then $\pi_{13} = 0.9^2 = 0.81$ and we will rank $\theta_1 > \theta_3$. But if $\omega_2 \in (1/9, 8/9)$, then we must rank $(\theta_1, \theta_2)$ and $(\theta_2, \theta_3)$ as ties because $\max\{\pi_{12}, \pi_{21}, \pi_{23}, \pi_{32}\} < 0.8$. By transitivity, this implies $\theta_1 = \theta_3$ which contradicts our earlier assertion that $\theta_1 > \theta_3$.

## 3.5 Connection to social choice

The literature on ranking methods bears a close connection to problems of social choice. If one views $\pi_{ij}$ as the share of votes for firm $i$ over firm $j$ in a pairwise election then a number of standard preference aggregation schemes suggest themselves. For example, Borda (1784)'s voting method simply ranks each firm $i$ based on the number of elections it has won; i.e., based upon $\sum_{j \neq i} 1\{\pi_{ij} > 1/2\}$.

The Kemeny (1959) distance between the vectors $\theta$ and $d$ can be written

$$K(\theta, d) = \sum_{i=2}^{n} \sum_{j=1}^{i} |1\{\theta_i > \theta_j\} - 1\{\theta_i < \theta_j\} - (1\{d_i > d_j\} - 1\{d_i < d_j\})|$$

Note that

$$\begin{aligned}\mathbb{E}\left[K\left(\theta,d\right)|Y=y\right] &= \sum_{i=2}^{n}\sum_{j=1}^{i}\left(\pi_{ij}+\pi_{ji}\right)1\left\{d_i=d_j\right\}+\left(1+\pi_{ji}-\pi_{ij}\right)1\left\{d_i>d_j\right\}\\ &\quad+\left(1+\pi_{ij}-\pi_{ji}\right)1\left\{d_i<d_j\right\}\\ &\propto \sum_{i=2}^{n}\sum_{j=1}^{i}\left(2\pi_{ij}-1\right)\left(d_{ji}-d_{ij}\right). \qquad (6)\end{aligned}$$

Young and Levenglick (1978) shows that if ties are not possible (i.e., if $\pi_{ij} \neq 1/2$ for all $i \neq j$) that Condorcet (1785)'s voting scheme is equivalent to choosing a ranking $d$ that minimizes (6). Young (1986) shows that this vote aggregation scheme is the unique voting rule that is unanimous, neutral, and satisfies reinforcement and independence of remote alternatives.

Note that $\left(2\pi_{ij}-1\right)\left(d_{ji}-d_{ij}\right)$ is minimized by the MAP thresholding rule $d_{ij}^{MAP} = 1\{\pi_{ij} > 1/2\}$. When $\lambda = 1$ our objective in (4) reduces to (6). Consequently, the most granular version of our grading scheme (which avoids restricting the $DR$) minimizes the expected Kemeny distance between the assigned grades and the true rankings. When $\lambda < 1$ we depart from the Kemeny criterion by calling elections a draw when they are "close."

## 3.6 Extension to weighted loss

Ranking mistakes are likely to be more costly when the magnitude of the mistake is larger. The following family of loss functions weight pairwise concordances and discordances by the $p$'th power of the difference between the cardinal biases of the two firms:

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

The corresponding family of risk functions take the form

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

where $\mu_{ij}^p = \mathbb{E}\left[\max\{(\theta_i-\theta_j),0\}^p \mid Y=y\right]$. Note that $\lim_{p\to 0}\mu_{ij}^p = \pi_{ij}$. Hence, one can think of our baseline risk function in (4) as a limiting case of $\mathcal{R}^p$ as $p$ approaches zero. In what follows, we focus on this limiting case where $p \to 0$ ("binary loss") and the case

where $p = 2$ ("square weighted loss").

## 3.7 Discordance rates

To summarize the reliability of the grades it is useful to compute the expected frequency of discordances between pairs of grades, a quantity we shall refer to as the Discordance Rate (DR). The DR is simply the posterior expected value of the DP:

$$\begin{aligned} DR &= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* < d_j^*\right\} \mathbb{E}[1\left\{\theta_i > \theta_j\right\} \mid Y = y] + 1\left\{d_i^* > d_j^*\right\} \mathbb{E}[1\left\{\theta_i < \theta_j\right\} \mid Y = y] \\ &= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* < d_j^*\right\} \pi_{ij} + 1\left\{d_i^* > d_j^*\right\} (1 - \pi_{ij}). \end{aligned}$$

As the second line reveals, this quantity is trivial to compute, since it depends only on the optimized decisions $\{d_i^*\}_{i=1}^n$ and the posterior probabilities $\{\pi_{ij}\}_{i \neq j}$.

We can also define the conditional Discordance Rate between a specific pair of grades $g$ and $g' < g$ as

$$\begin{aligned} DR_{g,g'} &= \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} \mathbb{E}\left[1\left\{\theta_i < \theta_j\right\} | Y = y\right]}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\}} \\ &= \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} (1 - \pi_{ij})}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\}}. \end{aligned}$$

The denominator of each pairwise rate can be thought of as giving the number of rejections of the null hypothesis that a pair of firms exhibits equal discriminatory conduct in favor of the alternative that the firm assigned to group $g'$ is more biased than the firm assigned to group $g$. Hence, $DR_{g,g'}$ is an analogue of the directional false discovery rate (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2005), giving the expected share of pairs with differing grades that are misranked.

Because the conditional DRs are symmetric ($DR_{g,g'} = DR_{g',g}$), we report them as a lower triangular matrix. Note that the $DR$ is a weighted average of the conditional DRs with positive weight put on the "on-diagonal" terms $DR_{g,g}$, which are necessarily zero. When working with $p$-weighted loss a corresponding weighted version of the conditional discordance rate can be employed:

$$\begin{aligned} DR^p_{g,g'} &= \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} \mathbb{E}\left[\max\{(\theta_i - \theta_j), 0\}^p \mid Y = y\right]}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} \mathbb{E}\left[(\theta_i - \theta_j)^p \mid Y = y\right]} \\ &= \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} (1 - \mu^p_{ij})}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i^* = g\right\} 1\left\{d_j^* = g'\right\} m^p_{ij}}, \end{aligned}$$

where $m_{ij}^p = \mathbb{E}_G\left[(\theta_i - \theta_j)^p \mid Y = y\right]$. Note that the $p$-weighted discordance rate nests the corresponding unweighted rate as $DR_{g,g'}^0 = DR_{g,g'}$. For any $p > 0$, $DR_{g,g'}^0$ is guaranteed to lie in the unit interval.

If we view the grades $\{d_i^*\}_{i=1}^n$ as exchangeable random variables and take $DR_{g,g'}$ as an assessment of the conditional misranking probability $\Pr(\theta_i > \theta_j \mid d_i^* = g, d_j^* = g')$ we can convert the conditional DR's into Bayes Factors expressing the informativeness of assigned grade pairs about the relative ranks of firms. By Bayes' rule,

$$\begin{aligned} BF_{g,g'} &= \frac{\Pr\left(d_i^* = g, d_j^* = g' \mid \theta_i \leq \theta_j\right)}{\Pr\left(d_i^* = g, d_j^* = g' \mid \theta_i > \theta_j\right)} \\ &= \frac{\Pr\left(\theta_i \leq \theta_j \mid d_i^* = g, d_j^* = g'\right)\Pr\left(\theta_i \leq \theta_j\right)}{\Pr\left(\theta_i > \theta_j \mid d_i^* = g, d_j^* = g'\right)\Pr\left(\theta_i > \theta_j\right)} \\ &= \frac{1 - DR_{g,g'}}{DR_{g,g'}}, \end{aligned}$$

where we have used the fact that under a smooth *i.i.d.* prior $\Pr\left(\theta_i > \theta_j\right) = \Pr\left(\theta_i \leq \theta_j\right)$. The Bayes Factor $BF_{g,g'}$ conveys the odds that a randomly selected firm assigned grade $g$ is more biased than a randomly selected firm assigned grade $g' < g$. Corresponding $p$-weighted Bayes Factors can be formed in the same manner using $DR_{g,g'}^p$.

# 4 Ranking Names

As an introductory illustration, we now rank the employer contact propensities of the names used in our correspondence experiment. The experiment utilized 76 first names, which were split equally between the nominal categories of: Black male, Black female, white male, and white female. The same names were sent to each firm.

In our past work (Kline, Rose and Walters, 2021) we were unable to reject the null hypothesis that the first names randomly assigned to our applications exhibited equal employer contact probabilities conditional on their nominal race and sex. In the following exercise, we treat race and sex as "missing labels" and examine whether the grades assigned to first names by our ranking algorithm effectively recover these labels from data on firm contacts alone.

## 4.1 Estimating $G$

Abusing notation somewhat, we let $i$ in this section refer to a first name and denote the number of applications with name $i$ sent in the experiment by $N_i$. The number of employer contacts received within 30 days by those applications is denoted by $C_i$. If the contacts are viewed as independent Bernoulli trials with name specific contact probabilities $p_i$ then the contact rate $C_i/N_i$ of name $i$ will exhibit a mean of $p_i$ and variance of $p_i(1-p_i)/N_i$.

This dependence of the variance on the contact probability complicates ranking exercises as the names that deserve the highest grades will also tend to exhibit the most noise.

To stabilize the variance, we rank names according to a Bartlett (1936) transformation of their contact rates

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

The logic of this transform follows from the observation that $\frac{d}{dx}\sin^{-1}\sqrt{x} = \left[2\sqrt{x(1-x)}\right]^{-1}$. Consequently, the Delta method implies $\hat{\theta}_i$ has asymptotic distribution $\mathcal{N}(\theta_i, (4N_i)^{-1})$, where $\theta_i = \sin^{-1}\sqrt{p_i}$.

To estimate the distribution $G$ from which $\theta_i$ was drawn we use the non-parametric maximum likelihood (NPMLE) estimator developed by Koenker and Mizera (2014) and provided by the 'GLVmix' command in the R package REBayes (Koenker and Gu, 2017). The NPMLE estimates a discrete approximation to the joint distribution of $\theta_i$ and $(4N_i)^{-1}$, which in turn yields a discrete estimate $\hat{G}$ of the marginal distribution $G$ of $\theta_i$.

A plot of the $\hat{G}$ produced by the NPMLE is provided in Figure 1. The bars correspond to histograms of $\hat{\theta}$ while the yellow spikes represent the probability mass function of the $\theta_i$. This discrete distribution does an excellent job matching the mean value of the $\hat{\theta}_i$ and its bias corrected variance, which we compute as its sample variance minus its average squared standard error $n^{-1}\sum_{i=1}^n s_i^2 = n^{-1}\sum_{i=1}^n (4N_i)^{-1}$.

For comparison, we also plot the $G$ estimate produced by Efron (2016)'s log-spline estimator, which models $G$ as a penalized spline in the exponential family. We have chosen the spline order and penalization parameter to match unbiased estimates of $G$'s first two moments. Despite being continuous, the bimodal shape of the log-spline estimate is remarkably consistent with that of the NPMLE.

The lower panel of Figure 1 converts these $G$ estimates back into probability points via the inverse transform $\theta \mapsto \sin(\theta)^2$. The NPMLE finds two large mass points: one at $p = 0.226$ and another at $p = 0.244$. The roughly 2 percentage point gap between these points turns out to almost exactly match the Black white contact gap in the experiment of 2.1 percentage points. Likewise, the distance between the modes of the log-spline estimate is roughly 2.1 percentage points. The NPMLE also finds a smaller mass point at $p = 0.260$.

In what follows, we rely on the NPMLE estimate $\hat{G}$ of $G$, which maps most naturally to the missing labels problem. The top left panel of Figure 2 depicts the posterior contrast probabilities $\pi_{ij} = \Pr(\theta_i > \theta_j \mid \hat{\theta}_i, \hat{\theta}_j, N_i, N_j)$; see the Appendix for details on how these posteriors were computed. Names have been ordered according to their Condorcet rank (i.e., their grade when $\lambda = 1$). Name pairs with adjacent ranks exhibit $\pi_{ij}$'s near 1/2, indicating that we have little confidence in their relative order. Reassuringly, names with vastly different ranks exhibit $\pi_{ij}$'s near 0 or 1, indicating that the experimental data are

informative about the relative ordering of these pairs.

## 4.2 Reporting possibilities

The top right panel of Figure 2 depicts the Discordance Rate that arises from minimizing $\mathcal{R}(d;\lambda)$ – that is, from solving (4) subject to (3) – for different choices of $\lambda$. The number used to demarcate each solution corresponds to the number of distinct grades the integer linear programming procedure produces for that choice of $\lambda$. A sharp "elbow" emerges around $\lambda = 0.18$, above which the $DR$ grows rapidly.

The upper right panel depicts the trade off between grade reliability $1 - DR$ and informativeness $\bar{\tau}(d) = \mathbb{E}[\tau(d,\theta)]$ associated with our choice of $\lambda$. The data are potentially quite informative about name rankings: as $\lambda$ approaches 1, $\bar{\tau}(d)$ exceeds 0.5. However, the reliability of such a report would be fairly low: achieving $\bar{\tau}(d) = 0.5$ requires a Discordance Rate of roughly 0.25. To limit the chances of misclassification, we set $\lambda$ to the largest value that ensures $DR \leq 0.05$. From the top right panel of Figure 2, this choice yields two grades.

## 4.3 Grades and demographics

Figure 3 lists the first names according to their Condorcet ranking, along with the posterior mean and mode of each name's contact probability $p_i = \sin(\theta_i)^2$. Each name's color reflects its assigned grade. Variation in $N_i$ generates non-monotonicity of the posterior means in the Condorcet ranking.[2] To understand this phenomenon, consider two names $i$ and $j$ with $\hat{\theta}_i > \hat{\theta}_j$ but $N_i \ll N_j$. Because name $i$ has a much larger standard error than name $j$, its posterior will be shrunk more heavily towards the grand mean, which may lead the posterior mean of name $i$ to be lower than that of name $j$. However, the fact that $\hat{\theta}_i > \hat{\theta}_j$ suggests that name $i$ has a higher contact probability than name $j$, implying $\pi_{ij} > 1/2$. The posterior mode, by contrast, is less influenced by the spread of the distribution and hence by variation in $N_i$.

The coarsened grades that emerge from our procedure align closely with our race labels. 38 of the 64 highly ranked names are distinctively white, while none of the 12 low ranked names are white. By contrast, female names are only slightly under-represented among top graded names. Hence, the 2 group ranking recovers the missing race label with limited error and is not predictive of sex labels.

It is natural to wonder if a three or four grade solution would be more predictive of sex. Appendix Figure A1 reports the pseudo-$R^2$ and Area under the Curve (AUC) from

[2]The variation in $N_i$ is primarily attributable to the fact that a subset of our first and last name pairs were taken from the study of Bertrand and Mullainathan (2004), while the remaining name pairs were drawn from North Carolina data on speeding tickets and Census data. The number of last names considered differed across the two data sources, leading to imbalances in the average number of last names (and hence applications) per first name.

a series of logistic regressions of the name's sex on grade indicators for different choices of $\lambda$. Note that if we were to set $\lambda = 1$, this regression would necessarily predict sex perfectly, as every name would receive its own dummy indicator. Even with four grades, however, we find a pseudo-$R^2$ for sex of only 0.03. By contrast, a corresponding logistic regression of race on assigned grades yields a pseudo-$R^2$ above 0.2.

Figure 4 shows the grades that result from minimizing expected square-weighted loss $\mathcal{R}^2(d, \lambda)$, which weights large mistakes in pairwise rank more heavily than small mistakes. Remarkably, under this weighting scheme, both the posterior means and modes of the names are monotone in the $n$ assigned grades that result when setting $\lambda = 1$. Evidently, variance heterogeneity makes small but not large mistakes likely. Moreover, the procedure now chooses a 5 grade solution when $\lambda$ is chosen to yield a square-weighted Discordance Rate of 5%.

Once again, the grade categories are better predictors of race than sex. As shown in Figure A2, a logistic regression of sex on these five grade categories yields a pseudo-$R^2$ of 0.09, while a corresponding race regression yields a pseudo-$R^2$ of 0.2. That said, it is interesting that the top ranked names Heather and Misty are distinctively white and female while the bottom ranked names Lawanda, Tameka, and Latisha are distinctively Black and female. This within gender discrepancy is consistent with the race by sex interaction in contact gaps previously noted in Kline, Rose and Walters (2021).

We conclude that our grades are fairly strong predictors of a name's race but not its sex. Given that the overall gender gap in contact rates is indistinguishable from statistical noise in our experiment, the failure to predict gender is not particularly surprising. The ability to predict race for a wide range of choices of $\lambda$, however, suggests that our grading scheme can be effective at detecting latent group structure even when the number of units being ranked is relatively modest.

# 5 Ranking firms

We turn now to ranking firms in their relative treatment of Black versus white names. The behavior of each firm $i$ in our experiment is characterized by the race-specific contact probabilities $(p_{iw}, p_{ib})$. These probabilities are "population" quantities giving the hypothetical 30 day contact rates that would arise if we were to sample an infinite number of job vacancies from firm $i$ and send each job four pairs of applications with distinctively Black and white names. The sample contact rates $(\hat{p}_{iw}, \hat{p}_{ib})$ provide unbiased estimates of these contact probabilities.

To remove the potential influence of firm heterogeneity in baseline contact rates on our analysis, we focus on the following proportional measure of bias against Black names

at firm $i$:

$$\theta_i = 1 - p_{ib}/p_{iw}.$$

Our estimator of $\theta_i$ will be $\hat{\theta}_i = 1 - \hat{p}_{ib}/\hat{p}_{iw}$. Because the number of applications sent to each firm is large, we employ the Delta method to construct a standard error $s_i$ for each $\hat{\theta}_i$ based on the job clustered standard errors of the sample contact rates.

In what follows, we exclude the four firms in the original experiment with callbacks at fewer than five jobs, since estimated contact ratios for these firms may be unreliable. Summary statistics for the remaining estimation sample of 104 firms are provided in Table 2.

## 5.1 Estimating $G$

To estimate the population distribution of contact penalties, we rely on Efron (2016)'s log-spline estimator. The log-spline estimator is motivated by a hierarchical model for contact penalty estimates with heteroscedastic but independent measurement error:

$$\hat{\theta}_i \mid \theta_i, s_i \sim \mathcal{N}(\theta_i, s_i^2), \quad \theta_i \mid s_i \sim G, \text{ for all } i \in [n].$$

The distribution $G$ is assumed to belong to an exponential family flexibly parameterized by a natural spline. The spline parameters are estimated by penalized maximum likelihood. As before, we choose the penalization parameter and the spline order to match the mean and bias corrected variance of $\theta_i$, the latter of which is obtained by subtracting from the sample variance of $\hat{\theta}_i$ its average squared standard error $\frac{1}{n}\sum_{i=1}^{n} s_i^2$.

Figure 5 plots the log-spline estimate $\hat{G}$ overlaid against the histogram of $\hat{\theta}_i$. Reflecting the noise in the $\{\hat{\theta}_i\}_{i=1}^n$, $\hat{G}$ is less dispersed than the histogram of penalty estimates. Unlike with our earlier analysis of names, the density $\hat{G}$ is unimodal, and appears nearly symmetric. Though we have not imposed non-negativity on the distribution $\hat{G}$ in estimation, discrimination against white names is estimated to be exceedingly rare. In accord with this finding, an application of the high dimensional inequality testing procedure of Bai, Santos and Shaikh (2021) to the joint restriction that all firms weakly prefer white to Black names (i.e., the null that $\theta_i \geq 0 \ \forall i \in [n]$) yields a p-value of 0.94.

As a robustness check, we also compute NPMLE estimates using the GLVmix procedure developed by Koenker and Gu (2017), which estimates a bivariate discrete distribution for $(\theta_i, s_i)$. The resulting marginal distribution of $\theta_i$, depicted in Figure 5, exhibits many mass points and is also unimodal. Though the NPMLE yields moment estimates that depart somewhat from unbiased estimators of the same quantities, the two estimates of $G$ appear quite comparable in their overall shape. Since discrete distributions where firms exhibit exactly equivalent conduct strike us as implausible, we rely upon the log-

spline estimates in what follows. The continuity of the log-spline $\hat{G}$ is inherited by the EB posterior distribution of $\theta_i|Y$, which has the added benefit of enabling computation of posterior credible intervals for each $\theta_i$.

## 5.2 Industry effects

In Kline, Rose and Walters (2021) we found large differences in the magnitude of contact gaps across 2-digit NAICS industries. Many of these industries have only 2 or 3 firms, precluding a fixed effects approach. We therefore work with a hierarchical random effects model taking the form:

$$\hat{\theta}_i = \exp(\eta_{k(i)}) + \xi_i + s_i\epsilon_i,$$

where the function $k : \{1, \ldots, n\} \to \{1, \ldots, K\}$ returns a firm's industry, and $s = (s_1, \ldots, s_n)'$ denotes the vector of firm standard errors. The industry effect $\eta_{k(i)}$ captures correlation in discriminatory conduct among firms in the same industry, while the firm effect $\xi_i$ captures departures from the industry average. The exponential mapping of the industry effect guarantees that all industries discriminate against Black names on average – a shape restriction we found fits the data better than allowing $\eta_k$ to have support below zero. However, the firm specific deviations from these industry averages have unbounded support, which allows for the possibility of discrimination against white names by particular firms. The remaining error $\epsilon_i$ represents the sampling error in our firm specific estimates.

To complete the model, we assume

$$\begin{bmatrix} \eta_j \\ \xi_i \\ \varepsilon_i \end{bmatrix} \mid s \overset{i.i.d.}{\sim} \mathcal{N}\left( \begin{bmatrix} \mu_\eta \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\eta^2 & 0 & 0 \\ 0 & \sigma_\xi^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \text{ for all } (i,k) \in [n] \times [K]. \tag{7}$$

Treating the first step standard errors as known allows us to separate the magnitude $\sigma_\xi^2$ of within industry variability in discrimination from the variation attributable to the noise component $\epsilon_i$. While the assumption of multivariate normality is restrictive, we show in the Appendix that our parametric model closely replicates the observed marginal and within industry distributions of $\theta_i$ recovered by non-parametric methods. We also provide a test of the presumed independence of $\theta_i$ from $s_i$.

Maximum simulated likelihood estimates of model (7) appear in Table 3. Column (1) reports estimates of the three parameters $\mu_\eta$, $\sigma_\eta$, and $\sigma_\xi$, while column (2) shows the standard deviations of the between- and within-industry variance components along with the overall mean and standard deviation of $\theta_i$. Consistent with the results of Kline, Rose and Walters (2021), the estimates indicate substantial variation in discrimination across industries: $(\frac{0.047}{0.059})^2 \times 100 = 63$ percent of the variance in $\theta_i$ is explained by the industry

component $\exp(\eta_{k(i)})$, while 37 percent is explained by the within-industry deviation $\xi_i$. The mean of $\theta_i$ implied by the hierarchical model is 0.073, exactly matching the mean proportional gap shown in the first row of Table 2, while the implied standard deviation of $\theta_i$ is slightly smaller than the square root of the unbiased variance estimate shown in Figure 5 (0.059 vs. 0.077). The close correspondence between these estimates suggests that the parametric model in (7) fits key identified features of the distribution of $\theta_i$ reasonably well.

## 5.3 Reporting possibilities

Figure 6 plots the pairwise posterior ranking probabilities $1 - \pi_{ij}$ with firms ordered by their rank under $\lambda = 1$. Panel (a) shows results from our baseline specification with the log-spline estimate of the marginal mixing distribution as prior, while panel (b) reports results based on the parametric log-normal/normal model with industry effects. Note that, in this depiction, the firm assigned rank 0 is the one deemed most discriminatory. Consequently, most other firms are more likely than not to have lower values of $\theta$. Firms of middling rank, on the other hand, are more difficult to distinguish from others.

Including industry effects tightens the posteriors, which leads the $\pi_{ij}$'s to become more dispersed around 1/2. This phenomenon is apparent in the more distinct ridge along the diagonal of the matrix in panel (b) compared to the baseline specification in panel (a).

Figure 7 displays only the pairwise probabilities that satisfy the naive thresholding rule $1 - \pi_{ij} > \frac{1}{1+\lambda}$ when $\lambda$ has been set to 0.36. The resulting frontier implies numerous transitivity violations. For example, in panel (a), firm 7 cannot be distinguished from firm 4 or firm 28, suggesting they are all "tied." However, firm 28 is clearly distinguishable from firm 4, yielding a contradiction. Super-imposed on the figure we show a frontier corresponding to the 5 grades that solve 4 subject to 3 while constraining the Discordance Rate to 5%. One can think of these frontiers as a transitivity constrained version of the thresholding rule.

Figure A4 plots the number of distinct grades that result from minimizing $\mathcal{R}(d;\lambda)$ along with the Discordance Rate of those grades as a function of the tuning parameter $\lambda$. As expected, the number of grades increases with $\lambda$ as does the $DR$. In the absence of industry effects, setting $\lambda = 0.36$ yields five groups and an unconditional DR of roughly 5%. For a fixed $\lambda$, introducing industry effects tends to increase the number of grades because the posteriors become more accurate, but this also tends to increase the DR.

Figure A3 illustrates the empirical tradeoff between the information content of our grades, quantified by $\bar{\tau}(d) = \mathbb{E}[\tau(d,\theta)|Y]$, and their reliability, as quantified by the Discordance Rate. Without industry effects, setting $\lambda = 1$ yields $\bar{\tau}(d) = 0.42$ and a Discordance Rate of 0.29. Including industry effects increases $\bar{\tau}(d)$ to 0.56 and lowers the DR to 0.22. Holding the DR at .05, we are able to achieve $\bar{\tau}(d) = 0.25$ in the baseline model, while

including industry effects raises the achievable $\bar{\tau}(d)$ to nearly 0.4.

# 6 Firm Grades

Figure 10 plots a posterior mean estimate of each firm's bias $\theta_i$ along with 95% credible intervals, which are constructed by connecting the posterior 2.5th percentile of $\theta_i$ to the posterior 97.5th percentile. The firms are ordered by their Condorcet rank (i.e., their grades under $\lambda = 1$). In this draft we have replaced each firm's name with the rank of its bias estimate $\hat{\theta}_i$, which allows us to compare firm orderings across loss functions. After the paper has undergone peer review, we will release the underlying firm names.

While the Condorcet ranks of the posterior means are highly correlated with the ranks of the bias estimates, heteroscedasticity makes the correlation less than perfect. For example, the firm with the 3rd largest point estimate of bias (AKA firm #3) has a Condorcet rank of 2 and the second largest posterior mean, while the firm with the 7th largest bias point estimate has a Condorcet rank of 1 and the largest posterior mean. The posterior credible interval for firm #3 is wider than that of firm #7, reflecting its larger standard error. This larger standard error leads the EB posterior mean to exhibit "shrinkage" towards the center of the distribution. Reflecting this phenomenon, firms #1 and #2, which are not ranked in the top 10, exhibit unusually wide credible intervals.

The shading of each interval reflects the grade assigned by our ranking procedure when setting $\lambda$ to the highest value consistent with a Discordance Rate less than or equal to 5%. The five resulting grades are labeled in descending order. Most firms receive the middle grade of $\star\star\star$, which reflects both the noise in our estimates and the shape of the estimated distribution $\hat{G}$. By contrast, only two firms out of 104 are assigned the grade of $\star$, suggesting they are the heaviest discriminators against Black names. A single firm is assigned the score of $\star\star\star\star\star$ suggesting that they are the least biased against Black names of the lot. A depiction of how grades vary with firm-specific standard errors and bias estimates is provided in Appendix Figure A3, which reveals that large standard errors usually result in a middling grade.

Though we have used stars to represent the firm ranks, it is important to remember that these grades were designed to convey ordinal rather than cardinal information. There is nothing in our integer linear programming problem that guarantees a grade of $\star$ implies a particularly egregious level of discrimination. Conversely, there is nothing that guarantees firms assigned a grade of $\star\star\star\star\star$ exhibit no bias against Black (or white) names.

As it turns out, however, the grades assigned by our procedure yield groups of firms with large cardinal differences in contact gaps. The firm assigned the grade of $\star\star\star\star\star$ has a posterior mean value of $\theta_i$ very close to zero, while the two top-graded firms exhibit posterior means clustered near a 20% penalty for Black names.

Although a legal precedent for audit studies has yet to be established, a commonly applied standard in discrimination cases is the so called "four fifths rule," described in the Uniform Guidelines on Employee Selection (Commission, 1978) which state that

> 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.

Judged by the posterior credible intervals in Figure 10, the two firms assigned $\star$ grades appear more likely than not to be violating this standard.

Rankings derived from the model with industry random effects are displayed in Figure 11. Conditioning on industry increases the predictive power of the model, which gives rise to changes in the number of firms in each group. The number of firms assigned grade $\star$ rises from 2 to 8. This group of firms is overwhelmingly comprised of companies in the Auto Sector and Services industry (SIC code 55) . Likewise, the number of firms assigned grade $\star\star$ rises from 4 to 7. The composition of the bottom grades is less affected and the same firm continues to be the sole recipient of the grade $\star\star\star\star\star$.

To understand what share of the total variance in discrimination is captured by these grades, it is useful to compute the between-grade variance in $\theta_i$. We do so by computing an estimate using grade-average posterior means, as detailed in Appendix Section 7. The variance estimate is weighted by the number of firms per grade, so that the ratio of between-grade to total variance can be given an $R^2$ interpretation. The estimated between-grade standard deviation in contact penalties is 0.032 for the five grades reported in Figure 10, implying an $R^2$ of roughly 17%. The five grades computed from the model with industry effects and reported in Figure 11 are substantially more informative, explaining 36% of the total variance.

## 6.1 Grades under square-weighted loss

Figure 12 displays rankings derived from minimizing $\mathcal{R}^2(d;\lambda)$, which is motivated by square weighted loss, without conditioning on industry effects. The parameter $\lambda$ is chosen to ensure the square-weighted $DR$ is roughly 5%. As before, firms are ordered by their ranking under $\lambda = 1$, which aligns closely with the Condorcet ranks reported in Figure 10. Relative to the binary loss solution, square weighting leads to an additional grade, and a more even distribution of firms across grades.

Using industry effects to evaluate the square-weighted risk $\mathcal{R}^2(d;\lambda)$ generates the rankings portrayed in Figure 13. Including industry effects has a dramatic effect on the number of grades assigned, leading to 19 distinct grade categories. This proliferation of grades arises because the square-weighted risk function is more tolerant of small mistakes,

most of which capture firm specific deviations from industry averages. Consequently, the procedure is able to more grade categories than we have industries, while constraining the square-weighted $DR$ to lie below 5%.

As in the binary loss case, these grades explain a meaningful share of total variance in $\theta_i$. The estimated between-grade standard deviation in contact penalties is 0.042 for the six grades reported in Figure 12, implying an $R^2$ of roughly 30%. The grades computed from the model with industry effects and reported in Figure 13 explain 33% of the total variance.

## 6.2 Misclassification and Bayes Factors

Figure 14 assesses the reliability of grade classifications across different loss functions by reporting the lower-triangular matrix of estimated between-grade DRs in our baseline model that omits industry effects. Panel a), for example, shows that 17% of the firm comparisons across grades $\star$ and $\star\star$ are expected to be misordered, implying the bias against Black names of a firm graded $\star\star$ is in fact higher than that of a firm graded $\star$. The Discordance Rate naturally declines when comparing non-adjacent grades. The expected share of misordered comparisons across $\star$ and $\star\star\star$, for example, is below 5%. Adjacent grades have $DR$'s between 15 and 23%, implying Bayes Factors between 3.3 and 5.7. Discordance rates for non-adjacent grades are much lower, in most cases falling below .04. These rates suggest firms in non-adjacent grades are highly likely to exhibit different behavior, with Bayes Factors exceeding 24.

Panel b) of Figure 14 reports square-weighted DR estimates for the grades that minimize $\mathcal{R}^2(d;\lambda)$. Under square weighted loss, we find weighted Discordance Rates between adjacent grades ranging from 9 to 18% implying Bayes factors ranging from 4.6 to 10.1. Weighted Discordance Rates between non-adjacent grades are uniformly smaller than .03. Our finding that the square weighted DRs in panel (b) fall below those in panel (a), reflects that most of the ranking mistakes likely to arise are small in magnitude, meaning that if one firm is erroneously listed as more biased than another, the two firms likely exhibit similar levels of bias.

Figure 15 summarizes the reliability of the grades obtained when conditioning on industry effects. In panel a), which considers the case of binary loss, Discordance Rates between adjacent grades range from 6% to 23%. DR's between non-adjacent grades are negligible. Panel b) of Figure 15 reports the matrix of Discordance Rates between the 19 grades obtained under square weighted loss. $DR$'s between adjacent grades range from 0.07 to 0.39. Even between non-adjacent grades, many $DR$'s are substantial, ranging as high as 29%. However, firms separated by 3 or more grades have substantially smaller misclassification probabilities.

# 7 Conclusion

We have proposed a new computationally efficient method for grading units based upon noisy measurements and used it to grade the conduct of firms in a large-scale correspondence experiment. In practice, we find that a 5 grade model exploiting industry affiliation explains more than one third of the variation across firms in their proportional contact biases against Black names while controlling the DR to roughly 5%. The most heavily discriminating firms are estimated to exhibit biases against Black names of roughly 20%, which is potentially a violation of the fourth-fifths rule stipulated by the Uniform Guidelines on Employee Selection.

The information conveyed in this report card is of potentially great value to the public. Regulatory agencies such as the Equal Employment Opportunity Commission (EEOC) and the Office of Federal Contract Compliance (OFCCP) have broad discretion to launch investigations into possible violations of equal employment opportunity laws, especially by federal contractors. Some of the firms receiving low grades turn out to be federal contractors, suggesting this information may be help in targeting future compliance efforts. The evidence we have provided may also prove useful to unemployed workers seeking to direct their search strategies towards less biased firms and to employees and executives of these companies seeking to benchmark their hiring conduct to peer organizations.

# References


**Andrews, Isaiah, and Jesse M Shapiro.** 2021. "A model of scientific communication." *Econometrica*, 89(5): 2117–2142.

**Bai, Yuehao, Andres Santos, and Azeem M Shaikh.** 2021. "A Two-Step Method for Testing Many Moment Inequalities." *Journal of Business & Economic Statistics*, 1–33.

**Bartlett, MSo.** 1936. "The square root transformation in analysis of variance." *Supplement to the Journal of the Royal Statistical Society*, 3(1): 68–78.

**Benjamini, Yoav, and Daniel Yekutieli.** 2005. "False discovery rate–adjusted multiple confidence intervals for selected parameters." *Journal of the American Statistical Association*, 100(469): 71–81.

**Benjamini, Yoav, and Yosef Hochberg.** 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

**Bertrand, Marianne, and Esther Duflo.** 2017. "Field experiments on discrimination." In *Handbook of Field Experiments*. Vol. 1, , ed. Esther Duflo and Abhijit Banerjee. Elsevier.

**Bertrand, Marianne, and Sendhil Mullainathan.** 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review*, 94(4): 991–1013.

**Borda, JC de.** 1784. "Mémoire sur les élections au scrutin." *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*.

**Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer.** 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753–1794.

**Bradley, Ralph Allan, and Milton E Terry.** 1952. "Rank analysis of incomplete block designs: I. The method of paired comparisons." *Biometrika*, 39(3/4): 324–345.

**Chetty, Raj, John N Friedman, and Jonah E Rockoff.** 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American economic review*, 104(9): 2633–79.

**Chetty, Raj, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter.** 2018. "The opportunity atlas: Mapping the childhood roots of social mobility." National Bureau of Economic Research.

**Commission, Equal Employment Opportunity.** 1978. "Uniform guidelines on employee selection procedures." *Federal Register*, 43(166): 38290–38315.

**Condorcet, Marquis de.** 1785. "Essay on the Application of Analysis to the Probability of Majority Decisions." *Paris: Imprimerie Royale*.

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

**Gu, Jiaying, and Roger Koenker.** 2020. "Invidious Comparisons: Ranking and Selection as Compound Decisions." *arXiv preprint arXiv:2012.12550*.

**Gu, Jiaying, and Roger Koenker.** 2022. "Ranking and Selection from Pairwise Comparisons: Empirical Bayes Methods for Citation Analysis." Vol. 112, 624–29.

**Guryan, Jonathan, and Kerwin Kofi Charles.** 2013. "Taste-based or statistical discrimination: the economics of discrimination returns to its roots." *The Economic Journal*, 123(572): F417–F432.

**Kemeny, John G.** 1959. "Mathematics without numbers." *Daedalus*, 88(4): 577–591.

**Kendall, Maurice G.** 1938. "A new measure of rank correlation." *Biometrika*, 30(1/2): 81–93.

**Kline, Patrick M, Evan K Rose, and Christopher R Walters.** 2021. "Systemic discrimination among large US employers." National Bureau of Economic Research.

**Koenker, Roger, and Ivan Mizera.** 2014. "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109(506): 674–685.

**Koenker, Roger, and Jiaying Gu.** 2017. "REBayes: an R package for empirical Bayes mixture methods." *Journal of Statistical Software*, 82: 1–26.

**Laird, Nan M, and Thomas A Louis.** 1989. "Empirical Bayes ranking methods." *Journal of Educational Statistics*, 14(1): 29–46.

**Mullainathan, Sendhil.** 2002. "Thinking through categories." Working Paper, Harvard University.

**Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer.** 2008. "Coarse thinking and persuasion." *The Quarterly journal of economics*, 123(2): 577–619.

**Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen.** 2017. "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time." *Proceedings of the National Academy of Sciences*, 114(41): 10870–10875.

**Quillian, Lincoln, John J Lee, and Mariana Oliver.** 2020. "Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback." *Social Forces*, 99(2): 732–759.

**Storey, John D.** 2002. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 479–498.

**Young, H Peyton.** 1986. "Optimal ranking and choice from pairwise comparisons." *Information pooling and group decision making*, 113–122.

**Young, H Peyton, and Arthur Levenglick.** 1978. "A consistent extension of Condorcet's election principle." *SIAM Journal on applied Mathematics*, 35(2): 285–300.

# Appendix: Computing posterior quantities

The posterior density (or PMF in the case of a discrete prior) for $\theta_i$ given $Y_i$ and prior $G$ can be written as:

$$f_G(x|Y_i) = \frac{\mathcal{L}(\hat{\theta}_i|\theta_i = x, s_i)dG(x)}{\int \mathcal{L}(\hat{\theta}_i|\theta_i = u, s_i)dG(u)}$$
$$\mathcal{L}(\hat{\theta}_i|\theta_i = x, s_i) = \frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - x}{s_i}\right).$$

Taking $\hat{G}$ as an estimate of $G$, posterior means are computed as $\int x f_{\hat{G}}(x|Y_i)dx$, while the lower and upper limits of 95% credible intervals are given by the 2.5th and 97.5th percentiles of the posterior cumulative distribution $F_{\hat{G}}(x|Y_i) = \int_{-\infty}^{x} f_{\hat{G}}(u|Y_i)du$.

We also use $\hat{G}$ to compute the matrix of pairwise posterior ranking probabilities $\pi_{ij}$. Bayes rule implies that:

$$\Pr(\theta_i > \theta_j|Y_i, Y_j) = \frac{\mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i > \theta_j, s_i, s_j)}{\mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i > \theta_j, s_i, s_j) + \mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i \leq \theta_j, s_i, s_j)}.$$

The simplicity of this expression stems from the fact that the prior probability that any firm out ranks another, $\Pr(\theta_i > \theta_j|s_i, s_j)$, is simply $\frac{1}{2}$ when, as we have assumed, *theta* is independent of $s$. The likelihoods that comprise the right-hand side of this expression can be written:

$$\mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i > \theta_j, s_i, s_j) = \frac{\int_{-\infty}^{\infty}\int_{y}^{\infty} \mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i = x, \theta_j = y, s_i, s_j)dG(x)dG(y)}{\int_{-\infty}^{\infty}\int_{y}^{\infty} dG(x)dG(y)}$$
$$\mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i = x, \theta_j = y, s_i, s_j) = \frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - x}{s_i}\right)\frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j - y}{s_j}\right).$$

Posterior differences in firm biases $\mu_{ij}^p = \mathbb{E}\left[\max\{(\theta_i - \theta_j), 0\}^p \mid Y_i, Y_j\right]$ are computed analogously. Specifically:

$$\mathbb{E}\left[\max\{(\theta_i - \theta_j), 0\}^p \mid Y_i, Y_j\right] = \int_{-\infty}^{\infty}\int_{y}^{\infty} (x - y)^p \rho(x, y|Y_i, Y_j)dxdy$$

where $\rho(x, y|Y_i, Y_j)$ is the posterior distribution of $\theta_i, \theta_j$ given $Y_i, Y_j$, which can be written:

$$\rho(x, y|Y_i, Y_j) = \frac{\mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i = x, \theta_j = y, s_i, s_j)dG(x)dG(y)}{\int\int \mathcal{L}(\hat{\theta}_i, \hat{\theta}_j|\theta_i = x, \theta_j = y, s_i, s_j)dG(x)dG(y)}.$$

All posterior expressions are evaluated for each firm and pair of firms by numerically integrating against $\hat{G}$. We then solve the linear programming problem to minimize either binary or weighted risk using Gurobi.

**Between grade variance**

Letting $M$ denote the total number of grades, the (firm-weighted) between grade variance of $\theta_i$ can be written

$$\sum_{g=1}^{M} \omega_g \bar{\theta}_g^2 - \left( \sum_{g=1}^{M} \omega_g \bar{\theta}_g \right)^2 = \sum_{g=1}^{M} \omega_g \left(1 - \omega_g\right) \bar{\theta}_g^2 - \sum_{g=1}^{M} \sum_{g' \neq g} \omega_g \omega_{g'} \bar{\theta}_g \bar{\theta}_{g'},$$

where $\bar{\theta}_g = \frac{\sum_{i=1}^{n} D_{ig} \theta_i}{\sum_{i=1}^{n} D_{ig}}$, $D_{ig} = 1\{d_i^* = g\}$ is an indicator for being assigned grade $g \in [M]$, and $\omega_g = n^{-1} \sum_{i=1}^{n} D_{ig}$ gives the share of firms assigned grade $g$.

We compute a Bayes unbiased estimate of each $\bar{\theta}_g$ by simply averaging the firm specific posterior firm means $\mathbb{E}\left[\theta_i | Y\right]$ within grade. The posterior mean estimate of each $\bar{\theta}_g^2$ is slightly harder to compute because

$$\bar{\theta}_g^2 = \frac{\sum_{i=1}^{n} D_{ig} \theta_i^2}{\sum_{i=1}^{n} D_{ig}} + \frac{\sum_{i=1}^{n} \sum_{i' \neq i} D_{ig} D_{i'g} \theta_i \theta_{i'}}{\sum_{i=1}^{n} \sum_{i' \neq i} D_{ig} D_{i'g}}.$$

Our posterior mean estimate of this quantity is computed analogously as

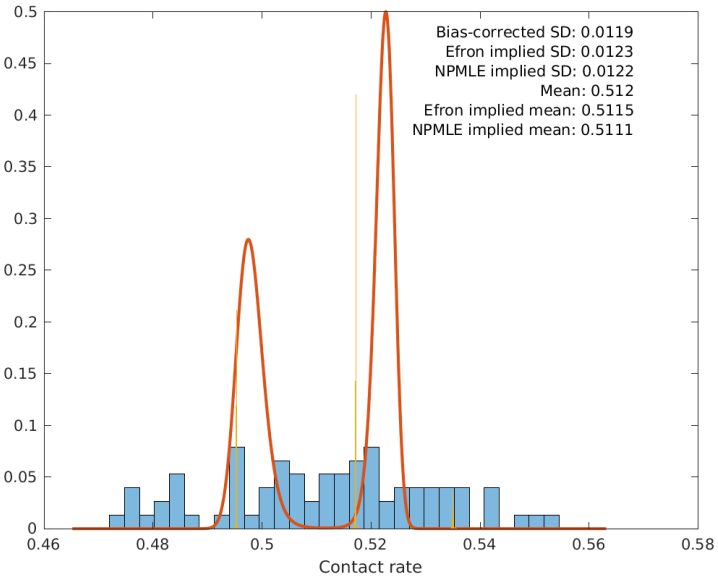
$$\mathbb{E}\left[\bar{\theta}_g^2 | Y\right] = \frac{\sum_{i=1}^{n} D_{ig} \mathbb{E}\left[\theta_i^2 | Y\right]}{\sum_{i=1}^{n} D_{ig}} + \frac{\sum_{i=1}^{n} \sum_{i' \neq i} D_{ig} D_{i'g} \mathbb{E}\left[\theta_i | Y\right] \mathbb{E}\left[\theta_{i'} | Y\right]}{\sum_{i=1}^{n} \sum_{i' \neq i} D_{ig} D_{i'g}},$$

where each $\mathbb{E}\left[\theta_i | Y\right]$ and $\mathbb{E}\left[\theta_i^2 | Y\right]$ is evaluated numerically using the estimated $\hat{G}$.

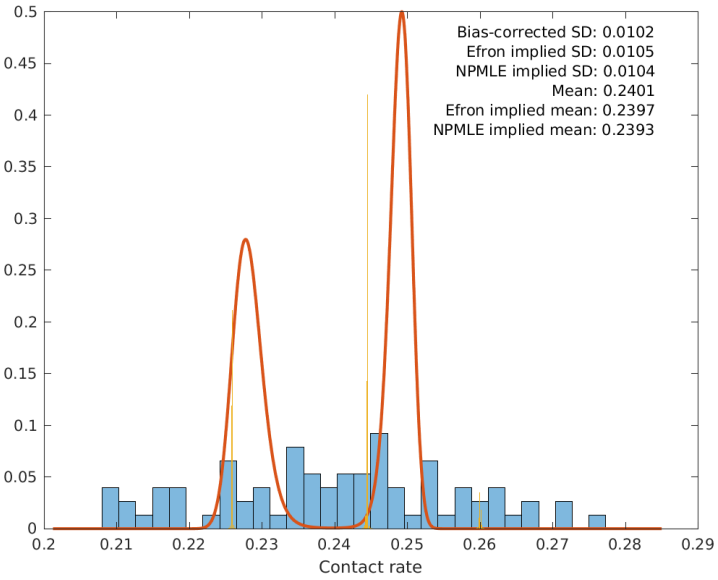
# Figures

Figure 1: Deconvolutions of name-specific contact rate estimates

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

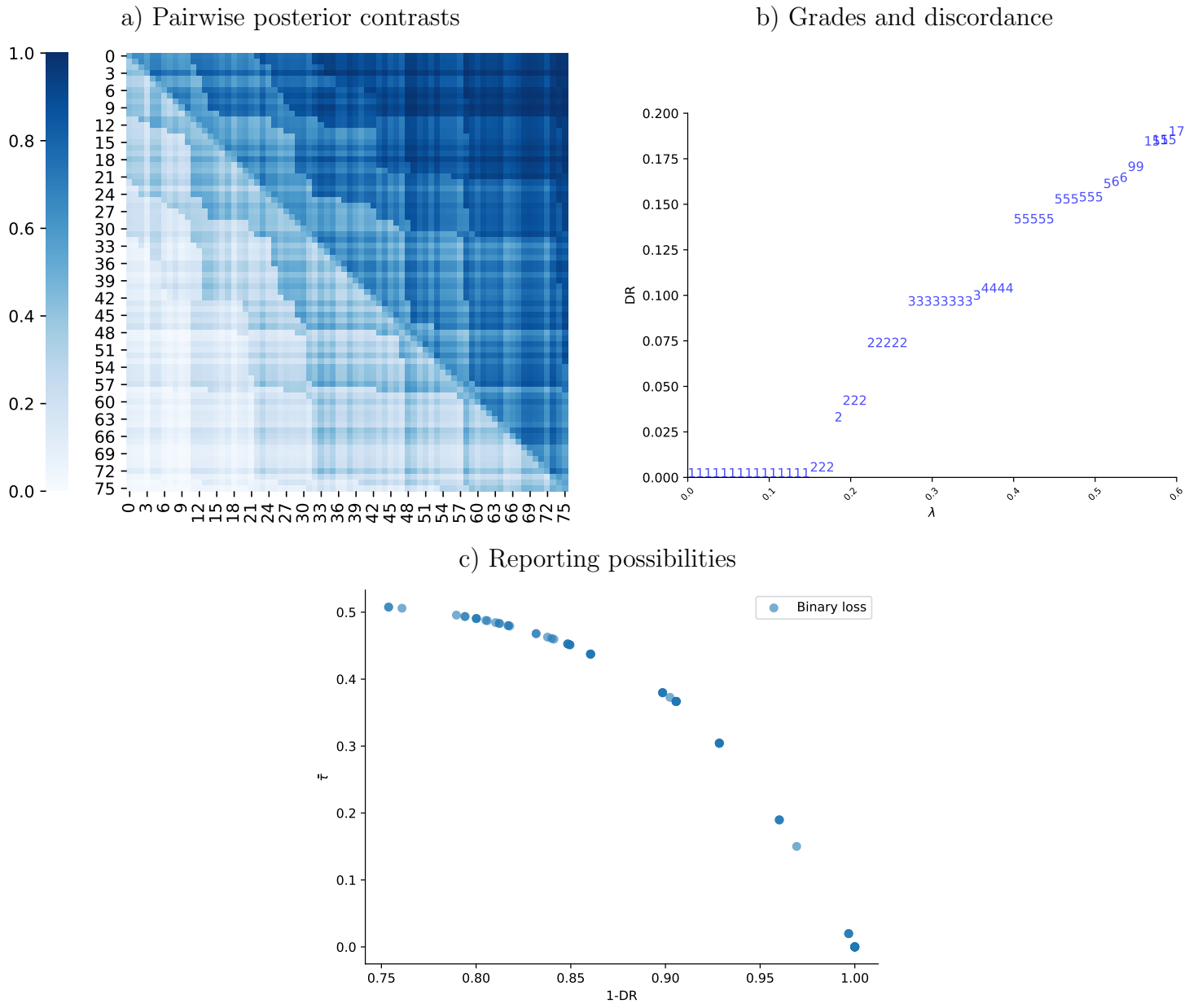



b) Contact rates ($p_i$)



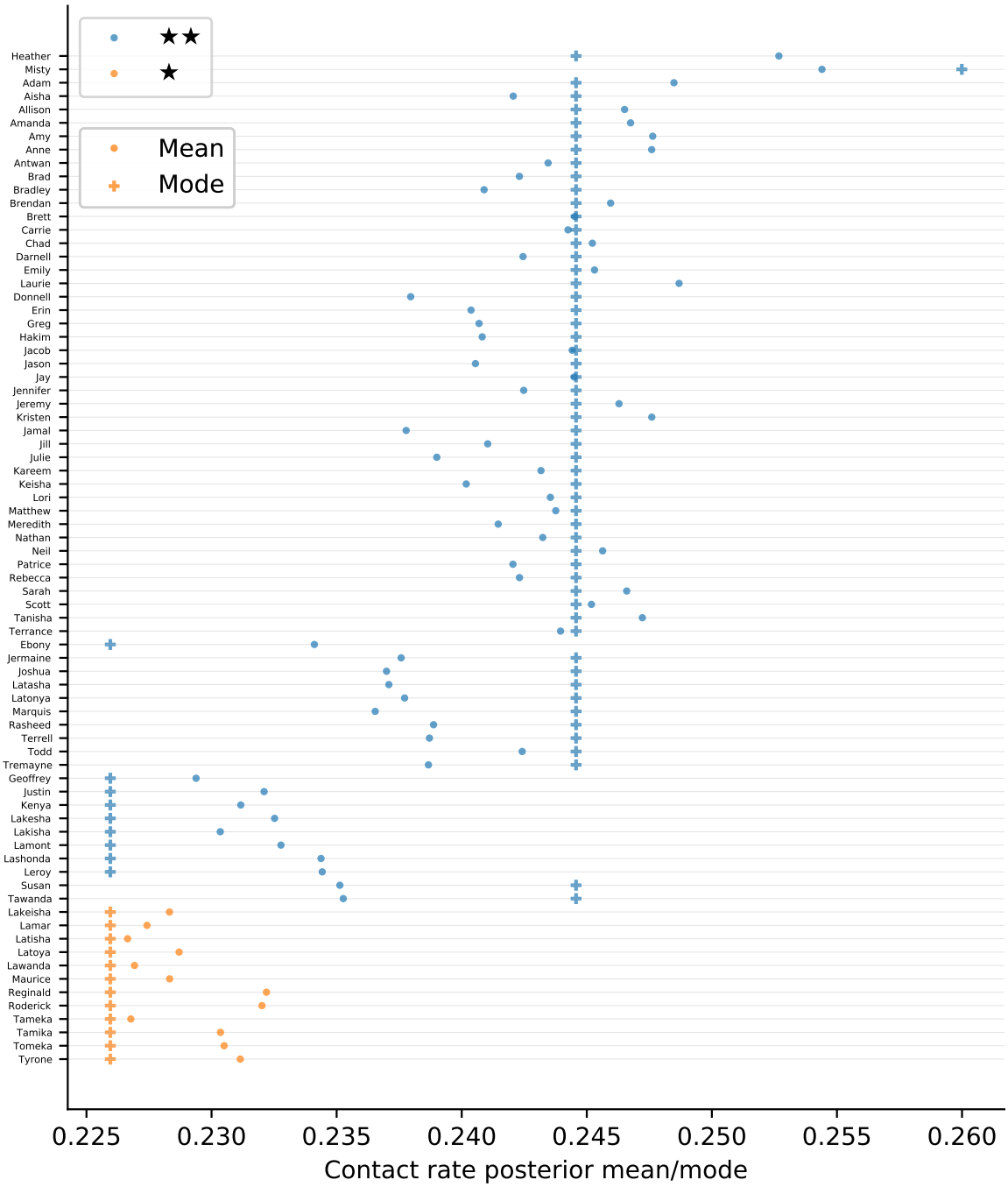

*Notes*: This figure presents non-parametric estimates of the distribution name-specific contact rates. Panel a deconvolves transformed contact rates $\hat{\theta}_i = \sin^{-1}\left(\sqrt{\hat{p}_i}\right)$, where $\hat{p}_i$ is the contact rate for applications sent with first name $i$. The blue histogram show the distribution of estimated transformed contact rate. The red line shows the population contact rate distribution estimated by applying a variation on Efron (2016)'s Empirical Bayes deconvolution procedure that allows for heteroscedastic measurement error and re-scaled to be comparable to the histogram. The tuning parameters in this deconvolution step are chosen so that the resulting distribution matches the corresponding mean and bias-corrected variance estimate as closely as possible, resulting in a fourth-order spline and zero penalization. The gold mass points plot the distribution of population contact rate from a NPMLE estimate using the procedure from Koenker and Gu (2017). Panel b plots the same set of distributions after transforming back into contact rates $p_i$.

Figure 2: Name ranking exercises



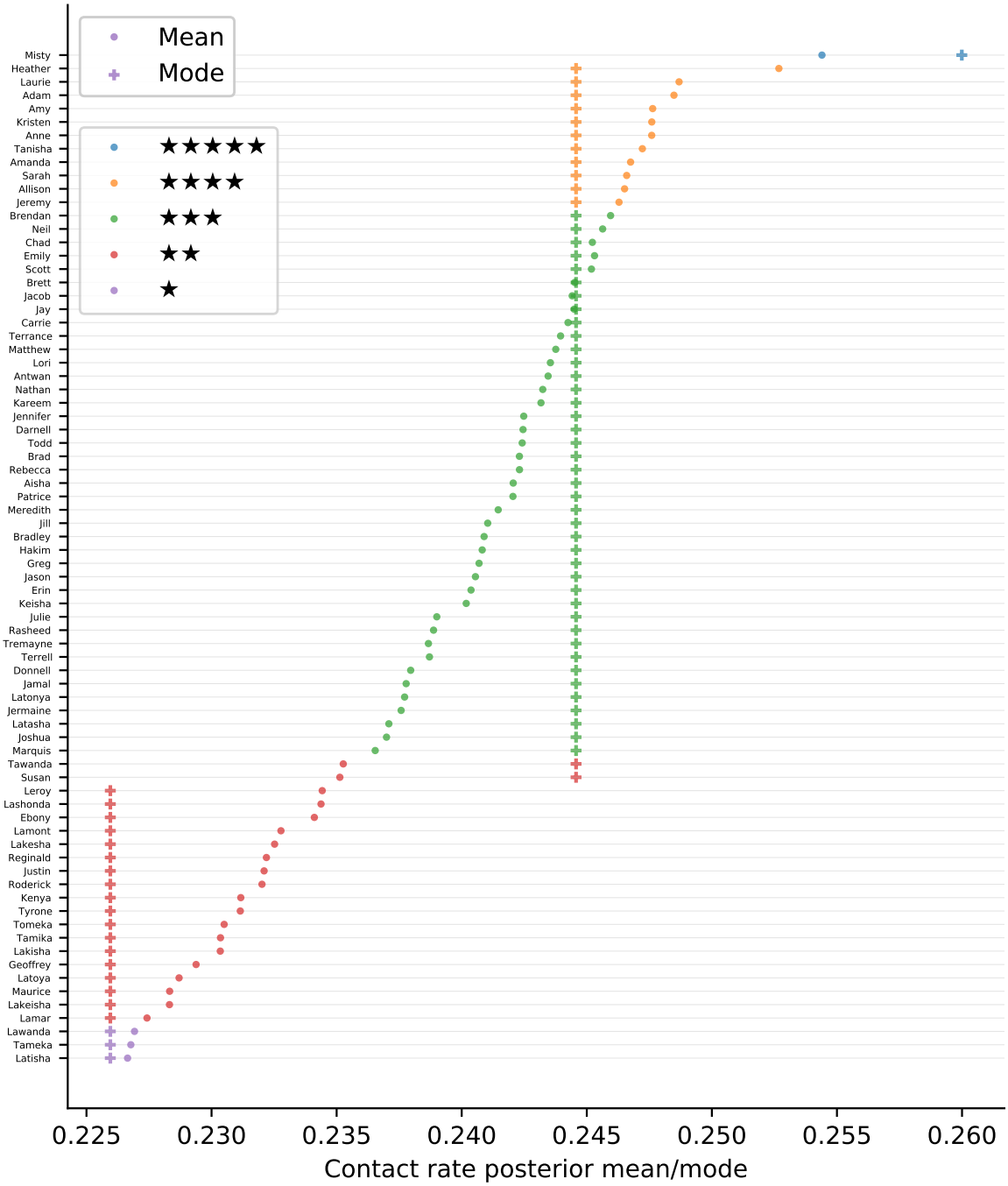

*Notes*: This figure presents several results from the ranking procedure for names. Panel a shows pairwise posterior ordering probabilities for all names. Posteriors are computed using the NPMLE estimate plotted in Figure 1 as the prior. Names are ordered by their rank under $\lambda = 1$. Panel b shows estimated Discordance Rates (DR) for an intermediate range of $\lambda$. DR is the expected Discordance Proportion (DP), or share of name pairs whose latent relative rank is the opposite of their grade rank. Panel c shows the expectation of Kendall's $\tau$ between $\theta_i$ and name grades against Discordance Rates (DR) for a range of grades indexed by $\lambda$.

Figure 3: Posterior means/modes and grades for names



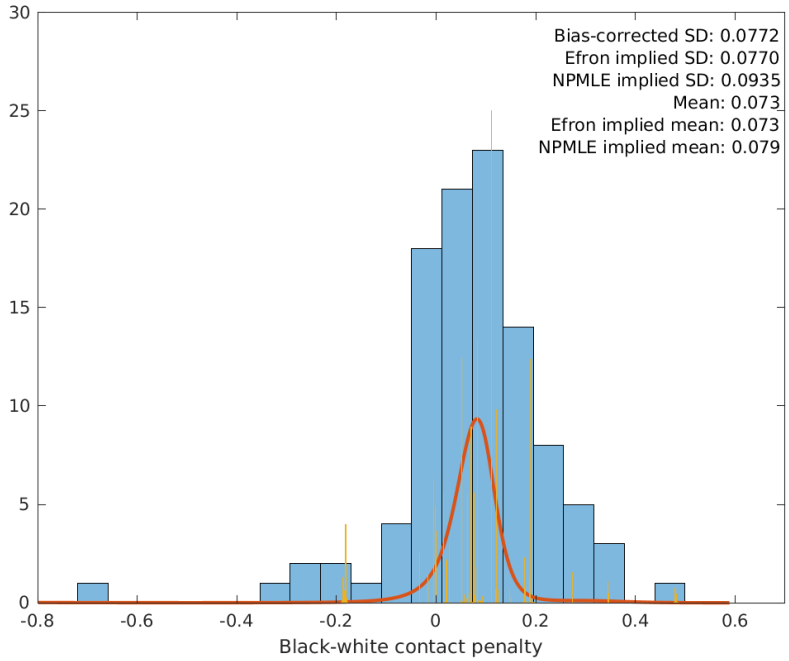

*Notes*: This figure shows posterior mean and mode contact rates and assigned grades for names. Results are shown for the largest value of $\lambda$ ensures DR is $\leq 0.05$. Names are ordered by their rank under $\lambda = 1$, when each name is assigned its own grade.

Figure 4: Posterior means/modes and grades for names under square-weighted loss




*Notes*: This figure shows posterior mean and mode contact rates and assigned grades for names. Results are shown for the largest value of $\lambda$ ensures DR is $\leq 0.05$. Names are ordered by their rank under $\lambda = 1$, when each name is assigned its own grade.

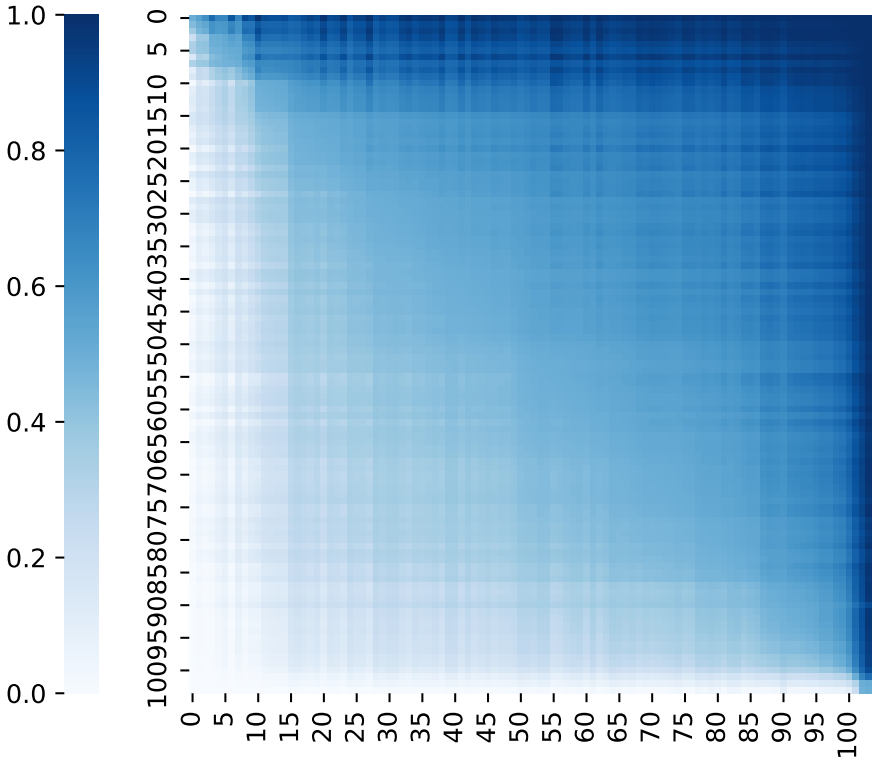
Figure 5: Deconvolution estimates of firm-level contact penalty distribution



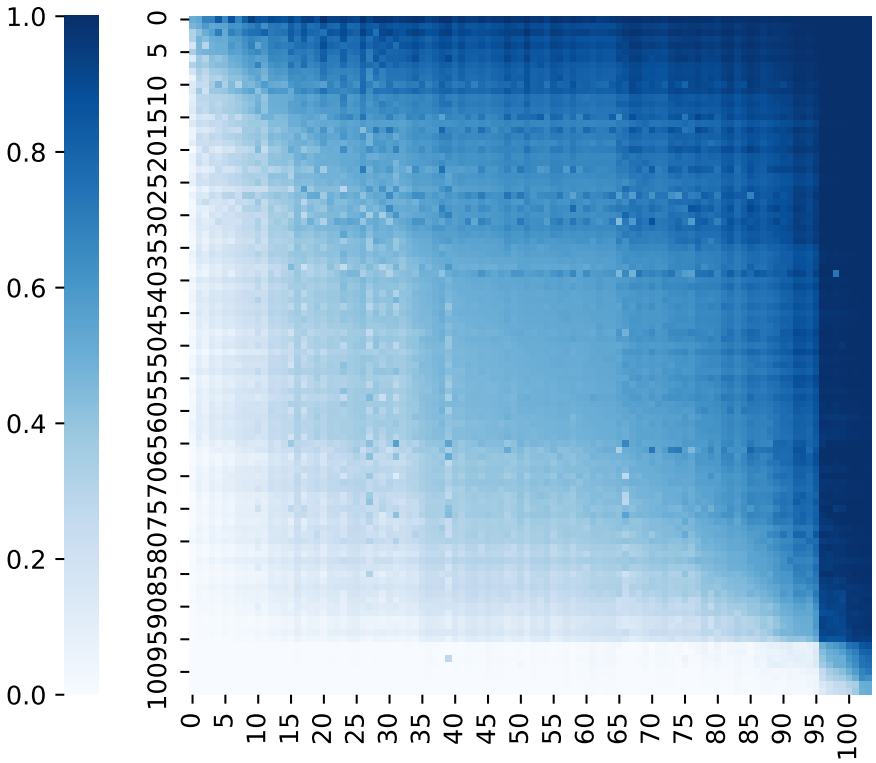

*Notes*: This figure presents non-parametric estimates of the distribution of firm-specific contact penalties. The blue histogram show the distribution of estimated firm proportional contact penalties. The red line shows the population contact penalty distribution estimated by applying a variation on Efron (2016)'s Empirical Bayes deconvolution procedure that allows for heteroscedastic measurement error. The tuning parameters in this deconvolution step are chosen so that the resulting distribution matches the corresponding mean and bias-corrected variance estimate as closely as possible, resulting in a fifth-order spline and penalization parameter of 0.0179. The gold mass points plot the marginal distribution of $\theta_i$ from a NPMLE estimate that allows for unrestricted correlation between $\theta_i$ and $s_i$ using the procedure from Koenker and Gu (2017).

Figure 6: Posterior contrasts

(a) Baseline



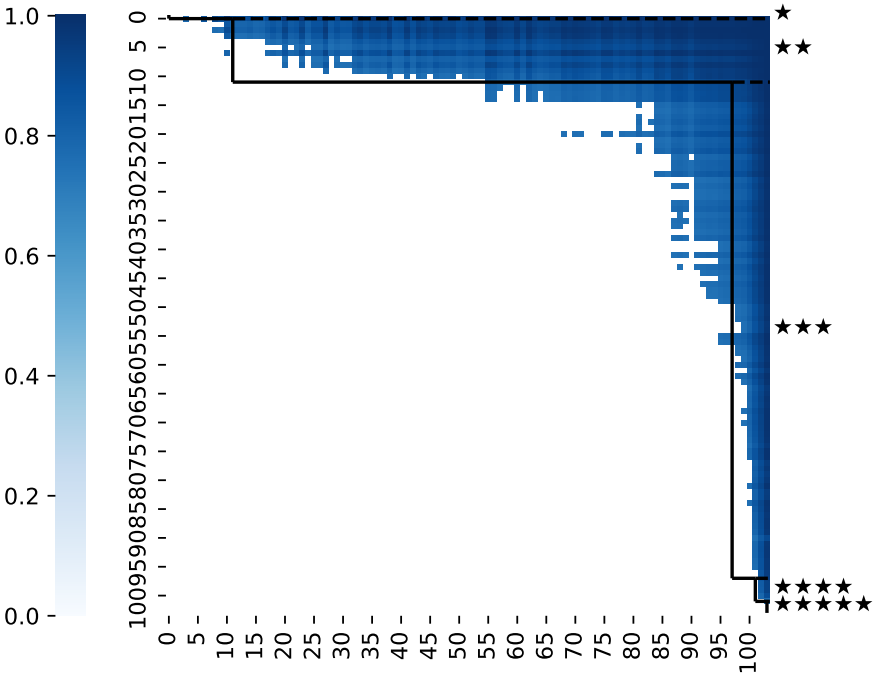

(b) Industry effects



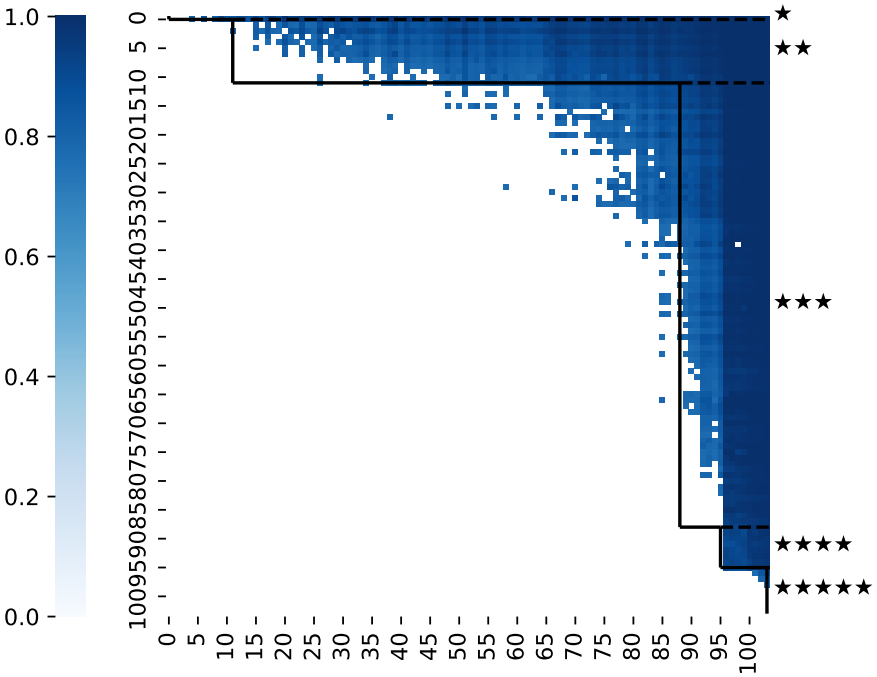

*Notes*: This figure plots pairwise posterior ordering probabilities for all firms in the study. Posteriors are computed using the smooth distribution plotted in Figure 5 as the prior in panel a, and using the equivalent distribution model with industry effects described in Section 5.2 in panel b. Firms ordered by rank under $\lambda = 1$. The figure plots $\pi_{ij}$ so that firm zero would be in the most discriminatory grade if all firms were assigned their own.

Figure 7: Optimal pairwise rankings and global orderings

(a) Baseline

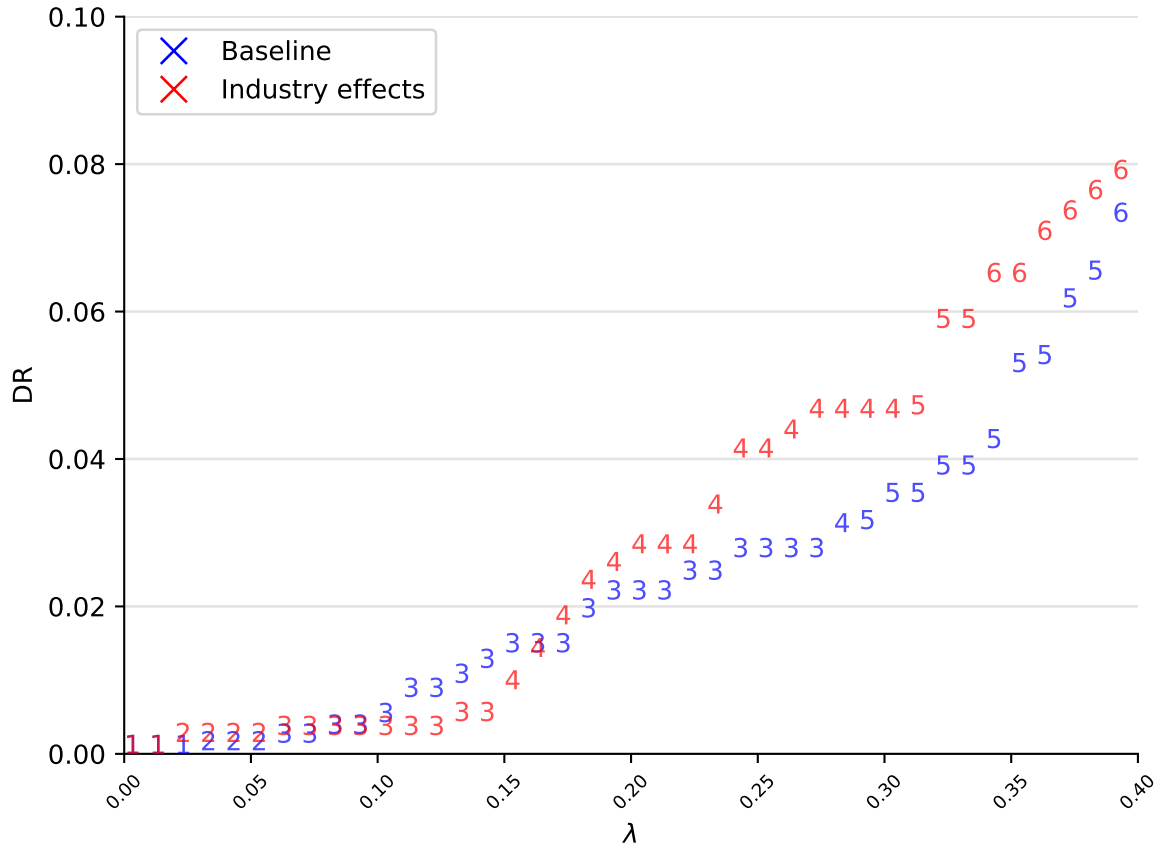



(b) Industry effects



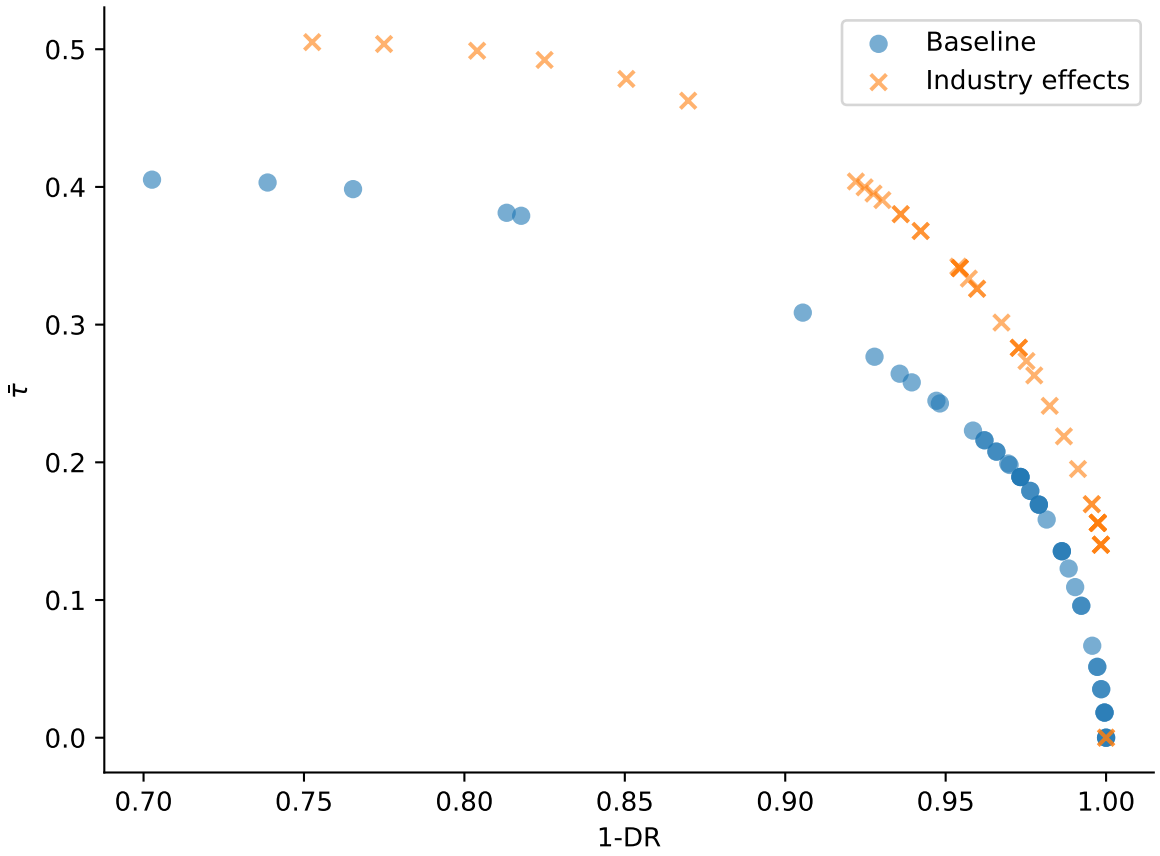

*Notes*: This figure plots the posterior ordering probabilities from Figure 6 for firm pairs where $\pi_{ij} > 1/(1+\lambda)$, indicating the pairwise expected utility maximizing decision would rank firm in the column below the firm in the row for $\lambda = 0.34$ in panel a and $\lambda = 0.31$ in panel b. This value of $\lambda$ maximizes $\tau$ while constraining discordance rates to fall below 0.05 f. The black lines denote the boundaries of optimal grades for this $\lambda$ for the firms in the rows. Panel b repeats the same exercise, but uses the industry random effect model to compute posteriors.

Figure 8: Grades and discordance as a function of $\lambda$



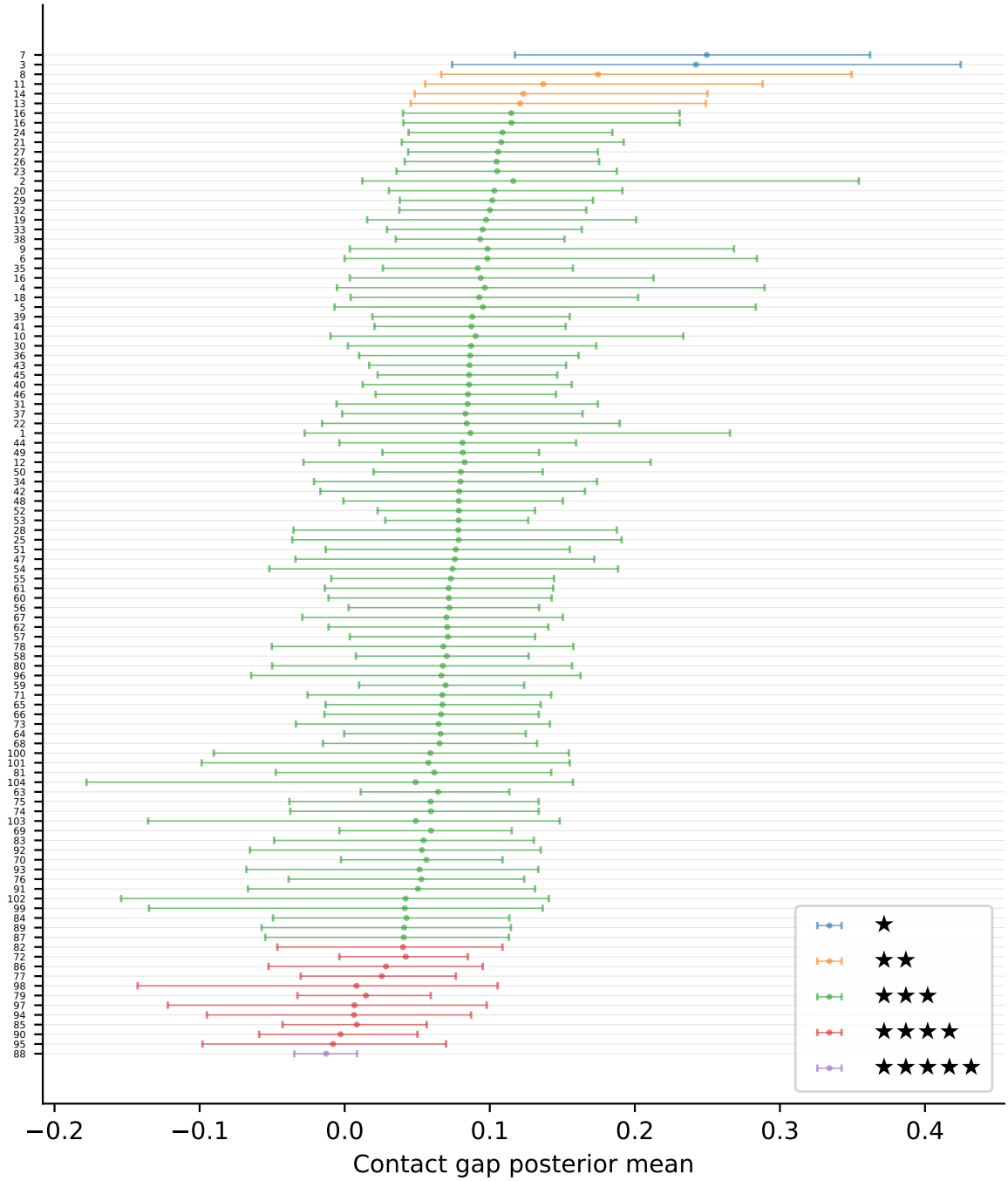

*Notes*: This figure shows estimated Discordance Rates (DR) for an intermediate range of $\lambda$. DR is the expected Discordance Proportion (DP), or share of firm pairs whose latent relative rank is the opposite of their grade rank. For small $\lambda$, which implies ranking mistakes are relative costly, optimal grades consist of just one or two grades. Larger $\lambda$ yield more grades and less weight on potential ranking mistakes, even when the number of grades remains the same. The results in Figure 10 correspond to $\lambda = 0.34$, the value that maximizes $\tau$ while constraining discordance rates to fall below 0.05.

Figure 9: Reporting possibilities



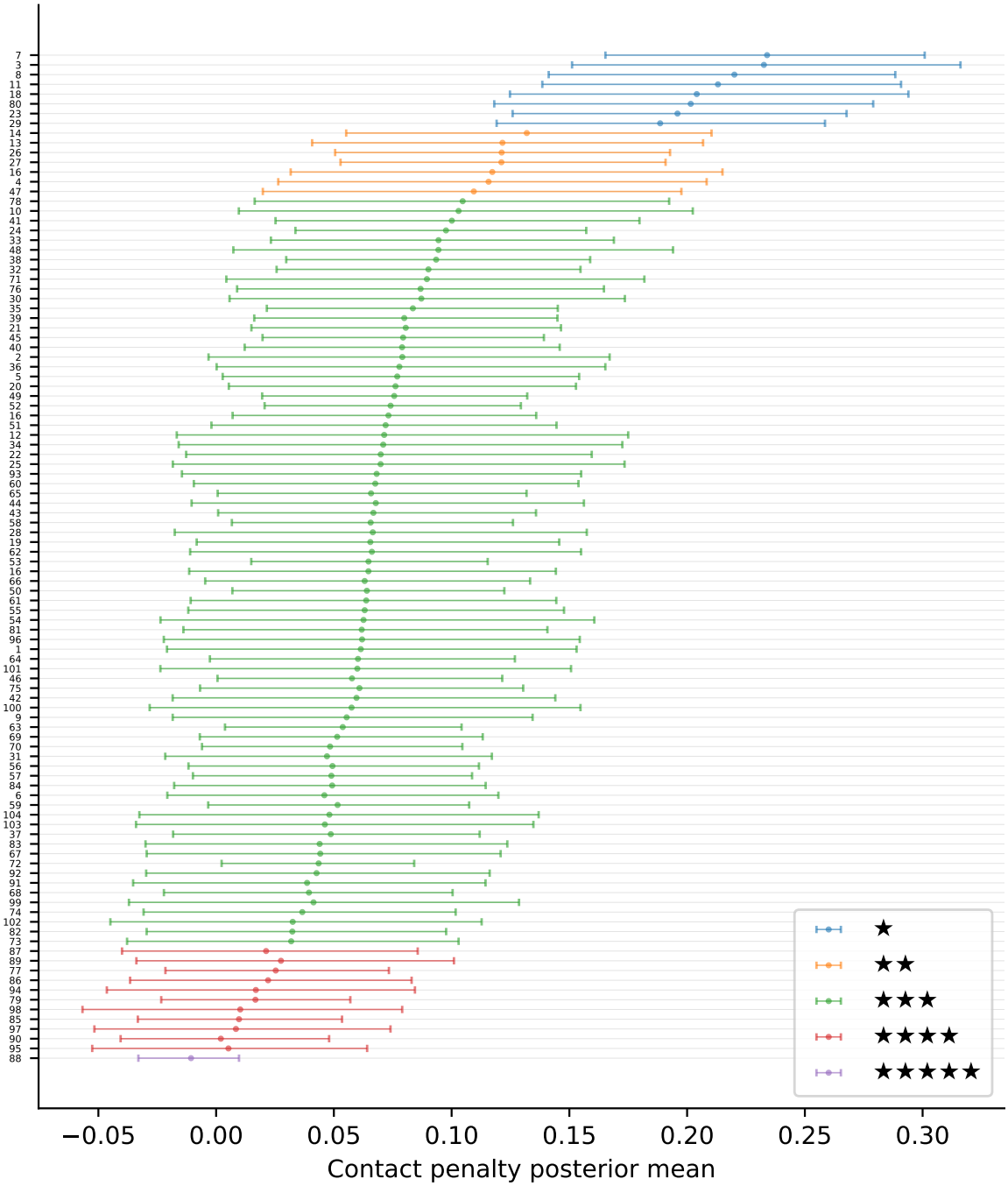

*Notes*: This figure shows the expectation of Kendall's $\tau$ between $\theta$ and our grades against Discordance Rates (DR) for a range of grades indexed by $\lambda$. DR is the expected Discordance Proportion (DP), or share of firm pairs whose latent relative rank is the opposite of their grade rank.

Figure 10: Binary loss: Posterior means and grades



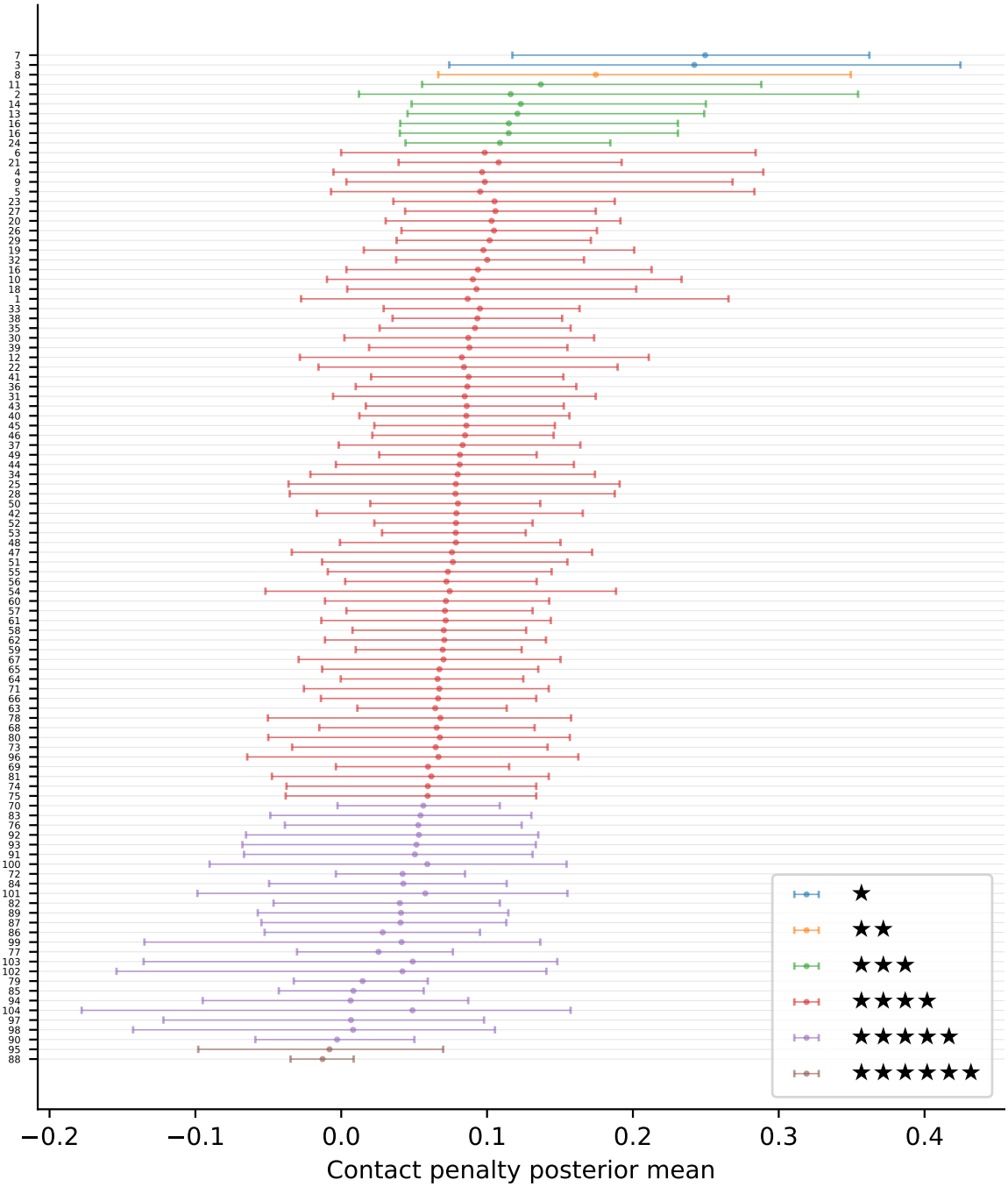

*Notes*: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades. $\lambda$ is chosen to maximize $\tau$ while constraining discordance rates to be below 0.05, which occurs at a value of 0.34. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade, and labeled by their raw contact penalty rank, with #1 showing the largest bias towards white applicants.

Figure 11: Binary loss with industry effects: Posterior means and grades



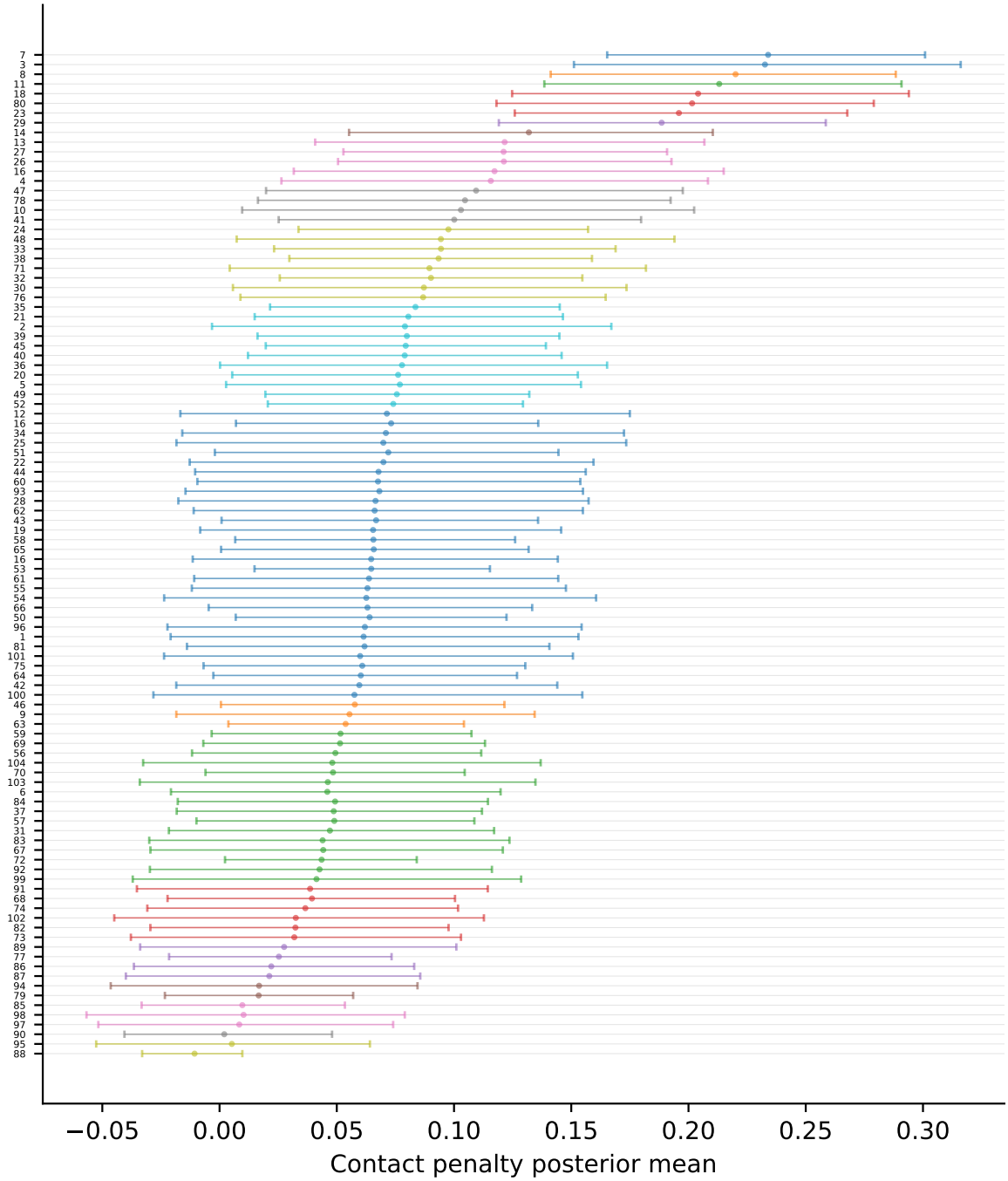

*Notes*: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades from industry random effect model. $\lambda$ is chosen to maximize $\tau$ while constraining discordance rates to be below 0.05, which occurs at a value of 0.31. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade, and labeled by their raw contact penalty rank, with #1 showing the largest bias towards white applicants.

Figure 12: Square-weighted loss: Posterior means and grades




*Notes*: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades. $\lambda$ is chosen to maximize weighted $\tau$ while constraining weighted discordance rates to be below 0.05, which occurs at a value of 0.23. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade, and labeled by the raw proportional contact gap rank, with #1 showing the largest gap in favor of white applicants.

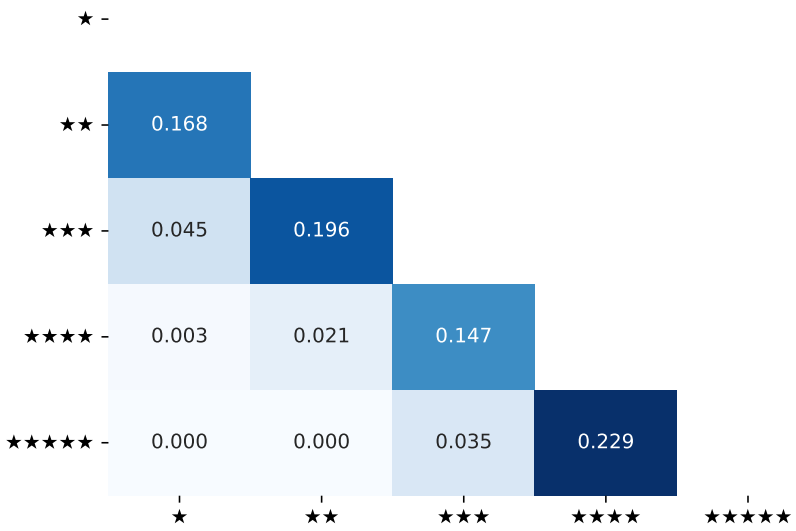
Figure 13: Square-weighted loss with industry effects: Posterior means and grades



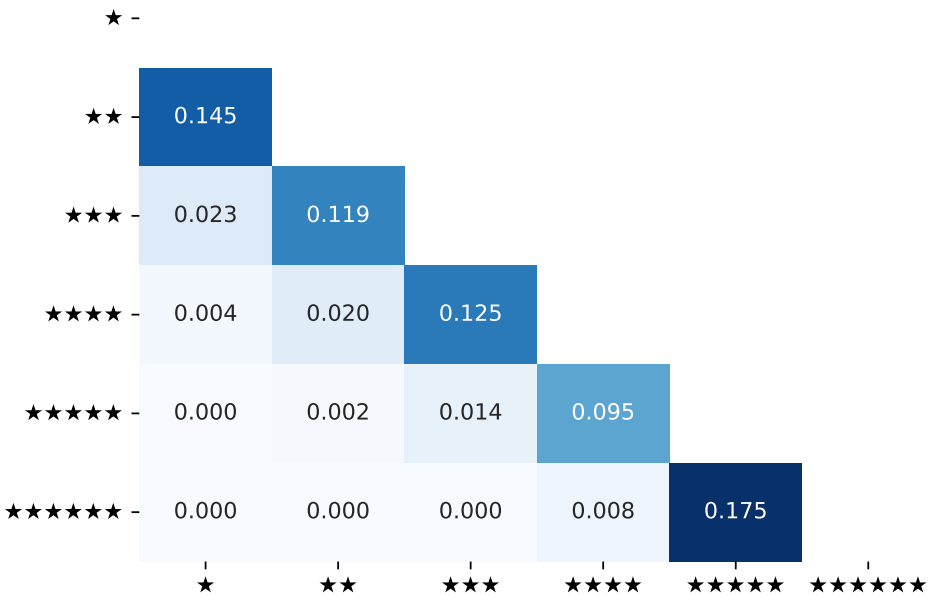

*Notes*: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades from the location shift model. $\lambda$ is chosen to maximize weighted $\tau$ while constraining weighted discordance rates to be below 0.05, which occurs at a value of 0.39. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade, and labeled by the raw proportional contact gap rank, with #1 showing the largest gap in favor of white applicants.

Figure 14: DR across weighted and unweighted loss

a) Binary



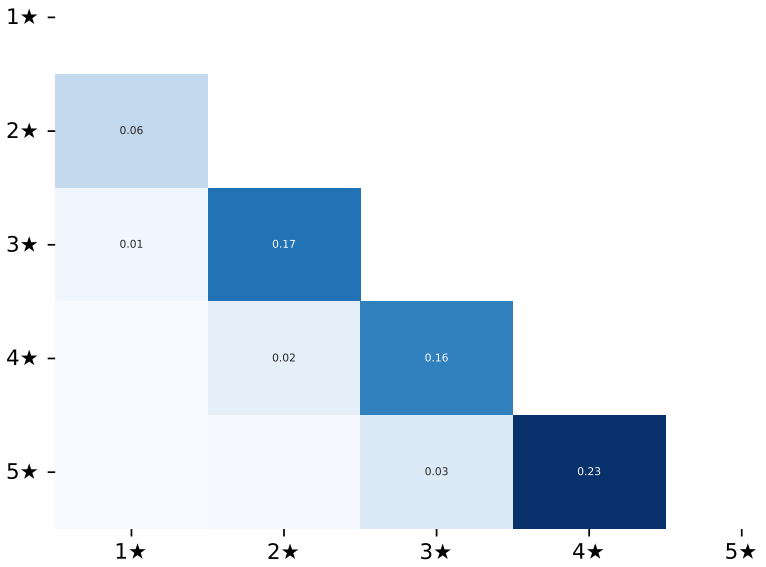

b) Square weighted



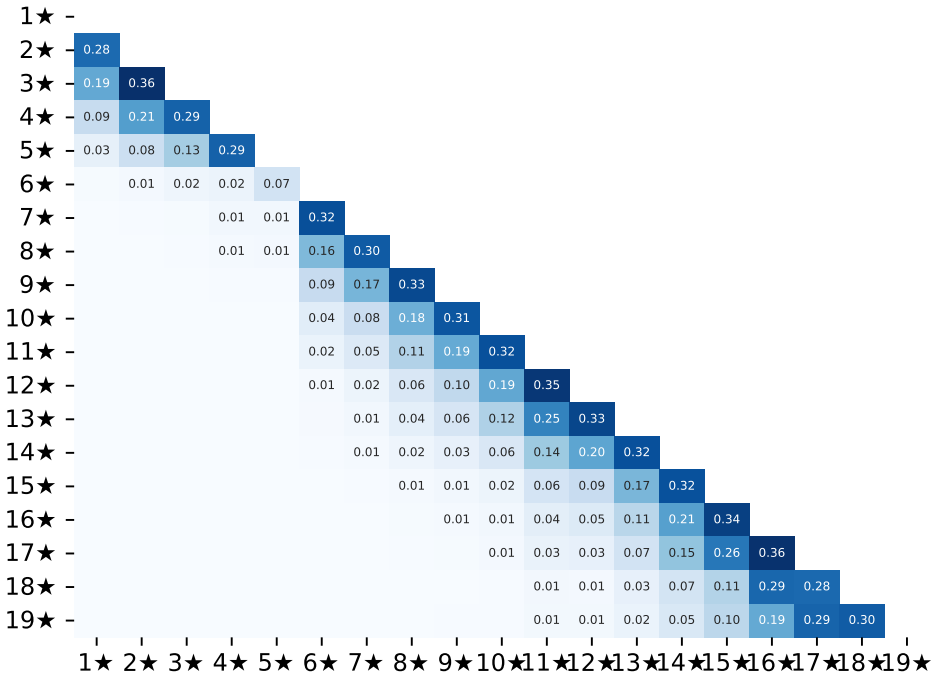

*Notes*: This figure shows mean Discordance Rates (DR) across grade pairs for unweighted and square weighted loss functions. In panel a, $DR_{g,g'}$ is the expected share of pairwise comparisons between firms in $g$ and $g'$ where the grade rank differs from the latent firm rank. The upper-left most estimate for example, indicates a 24% chance that a random firm in grade ★'s true rank is below a randomly chosen firm's in grade ★★'s. Panel b presents the square-weighted analogue, which weights ranking mistakes by the expected square distance between firm pairs. In both panels, DR decays quickly when comparing non-adjacent grades.

Figure 15: DR across weighted and unweighted loss for industry effects model

a) Binary




b) Square weighted




*Notes*: This figure shows mean Discordance Rates (DR) across grade pairs for unweighted and square weighted loss functions. In panel a, $DR_{g,g'}$ is the expected share of pairwise comparisons between firms in $g$ and $g'$ where the grade rank differs from the latent firm rank. Panel b presents the square-weighted analogue, which weights ranking mistakes by the expected square distance between firm pairs. Only cells where DR is above 0.01 are annotated. In both panels, DR decays quickly when comparing non-adjacent grades.

# Tables

Table 1: Summary statistics for first names sample

| | Contact rate | # apps | # first names |
|---|---|---|---|
| Male | | | |
| Black | 0.233 | 20,927 | 19 |
| | (0.003) | | |
| White | 0.246 | 20,975 | 19 |
| | (0.003) | | |
| Female | | | |
| Black | 0.226 | 20,879 | 19 |
| | (0.003) | | |
| White | 0.254 | 20,862 | 19 |
| | (0.003) | | |

*Notes:* This table presents summary statistics the sample of applications used in the analysis of first names. The table presents the mean 30-day contact rate, total number of applications sent, and number of unique first names used for each race and sex combination. Contact rates are re-weighted to balance the distribution of names across experimental waves. Although Black and white names were sent in pairs during the experiment, the total number of applications across race groups is not identical because some jobs closed before both applications could be sent. The gender of the name was unconditionally assigned. See Appendix Table A1 for a list of firstnames used and their sources.

Table 2: Summary statistics for firm sample

| | # Firms | # Jobs | # Apps | Contact rates and gaps | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | White | Black | Difference | Prop. bias | Mean SE |
| All | 104 | 10,723 | 80,956 | 0.248 (0.016) | 0.230 (0.015) | 0.018 (0.003) | 0.073 (0.015) | 0.101 |
| 2-digit SIC industry (code) | | | | | | | | |
| Food products (20) | 2 | 137 | 1,058 | 0.265 (0.170) | 0.281 (0.159) | -0.016 (0.011) | -0.149 (0.137) | 0.125 |
| Apparel manufacturing (23) | 2 | 198 | 1,530 | 0.206 (0.031) | 0.178 (0.038) | 0.028 (0.007) | 0.143 (0.057) | 0.071 |
| Other manufacturing (24) | 4 | 375 | 2,898 | 0.119 (0.038) | 0.106 (0.038) | 0.013 (0.002) | 0.141 (0.035) | 0.184 |
| Freight / transport (42) | 4 | 452 | 3,282 | 0.191 (0.016) | 0.201 (0.021) | -0.010 (0.006) | -0.047 (0.033) | 0.076 |
| Communications (48) | 2 | 170 | 1,110 | 0.263 (0.134) | 0.248 (0.132) | 0.015 (0.001) | 0.073 (0.032) | 0.079 |
| Electric / gas (49) | 3 | 315 | 2,408 | 0.265 (0.108) | 0.249 (0.114) | 0.016 (0.006) | 0.103 (0.056) | 0.070 |
| Wholesale durable (50) | 2 | 152 | 1,140 | 0.195 (0.035) | 0.176 (0.027) | 0.018 (0.008) | 0.089 (0.027) | 0.073 |
| Wholesale nondurable (51) | 11 | 1,104 | 8,126 | 0.304 (0.069) | 0.289 (0.073) | 0.015 (0.006) | 0.082 (0.030) | 0.071 |
| Building materials (52) | 3 | 376 | 2,712 | 0.298 (0.126) | 0.287 (0.117) | 0.011 (0.013) | 0.013 (0.034) | 0.060 |
| General merchandise (53) | 12 | 1,371 | 10,418 | 0.321 (0.049) | 0.294 (0.045) | 0.028 (0.006) | 0.096 (0.024) | 0.070 |
| Food stores (54) | 5 | 529 | 4,012 | 0.453 (0.090) | 0.428 (0.086) | 0.025 (0.012) | 0.054 (0.035) | 0.053 |
| Auto dealers / services (55) | 8 | 887 | 6,910 | 0.256 (0.041) | 0.203 (0.034) | 0.053 (0.011) | 0.205 (0.038) | 0.084 |
| Apparel stores (56) | 4 | 398 | 3,084 | 0.238 (0.076) | 0.204 (0.065) | 0.034 (0.015) | 0.148 (0.046) | 0.095 |
| Furnishing stores (57) | 4 | 477 | 3,668 | 0.277 (0.038) | 0.253 (0.033) | 0.024 (0.012) | 0.084 (0.042) | 0.075 |
| Eating/drinking (58) | 5 | 524 | 4,174 | 0.386 (0.028) | 0.361 (0.029) | 0.025 (0.009) | 0.066 (0.022) | 0.060 |
| Other retail (59) | 7 | 812 | 6,240 | 0.206 (0.057) | 0.180 (0.049) | 0.027 (0.010) | 0.132 (0.042) | 0.112 |
| Banks / credit (60) | 4 | 496 | 3,686 | 0.099 (0.035) | 0.095 (0.031) | 0.003 (0.006) | 0.007 (0.075) | 0.217 |
| Insurance / real estate (63) | 7 | 462 | 3,408 | 0.121 (0.053) | 0.114 (0.051) | 0.007 (0.007) | 0.063 (0.077) | 0.176 |
| Accommodation (70) | 2 | 241 | 1,842 | 0.202 (0.038) | 0.193 (0.059) | 0.009 (0.021) | 0.064 (0.118) | 0.074 |
| Business services (73) | 3 | 373 | 2,748 | 0.218 (0.120) | 0.212 (0.126) | 0.007 (0.006) | 0.098 (0.081) | 0.093 |
| Auto repair services (75) | 3 | 332 | 2,516 | 0.286 (0.022) | 0.280 (0.028) | 0.006 (0.012) | 0.024 (0.040) | 0.045 |
| Health services (80) | 5 | 420 | 3,038 | 0.142 (0.049) | 0.152 (0.038) | -0.010 (0.021) | -0.217 (0.150) | 0.204 |
| Engineering services (87) | 2 | 122 | 948 | 0.071 (0.051) | 0.064 (0.054) | 0.008 (0.002) | 0.272 (0.228) | 0.208 |

*Notes:* This table presents summary statistics the sample of applications used in the analysis of firm contact penalties names. "White" and "Black" refer to average firm-level contact rates for white and Black applications. Difference is the average difference. Prop. bias is the primary "proportional bias" measure used in the analysis, or $1 - \hat{p}_{ib}/\hat{p}_{iw}$. Mean SE is the average standard error of firm-level bias estimate.

Table 3: Hierarchical random effects estimates

| | Parameter estimates (1) | | Random effect distributions (2) |
|---|---|---|---|
| $\mu_\eta$ | -2.919 | $SD(\exp(\eta_{k(i)}))$ | 0.047 |
| | (0.171) | | (0.013) |
| $\sigma_\eta$ | 0.635 | $SD(\xi_i)$ | 0.035 |
| | (0.147) | | (0.014) |
| $\sigma_\xi$ | 0.035 | $E[\theta_i]$ | 0.073 |
| | (0.014) | | (0.009) |
| Parameters | 3 | $SD(\theta_i)$ | 0.059 |
| Log-likelihood | 85.95 | | (0.010) |

*Notes:* This table reports maximum simulated likelihood estimates of a hierarchical random effects model treating each firm's proportional contact gap $\theta_i$ as the sum of a log-normally distributed industry component $\eta_{k(i)}$ and a normally distributed within-industry component $\xi_i$. The simulated likelihood is computed using 1,000 draws of each random effect. Column (1) reports estimates of the parameters of the distributions of $\eta_k$ and $\xi_i$. Column (2) reports implied standard deviations of the between- and within-industry components along with the mean and standard deviation of $\theta_i$, with standard errors computed by the delta method.

# Appendix A Additional Figures and Tables

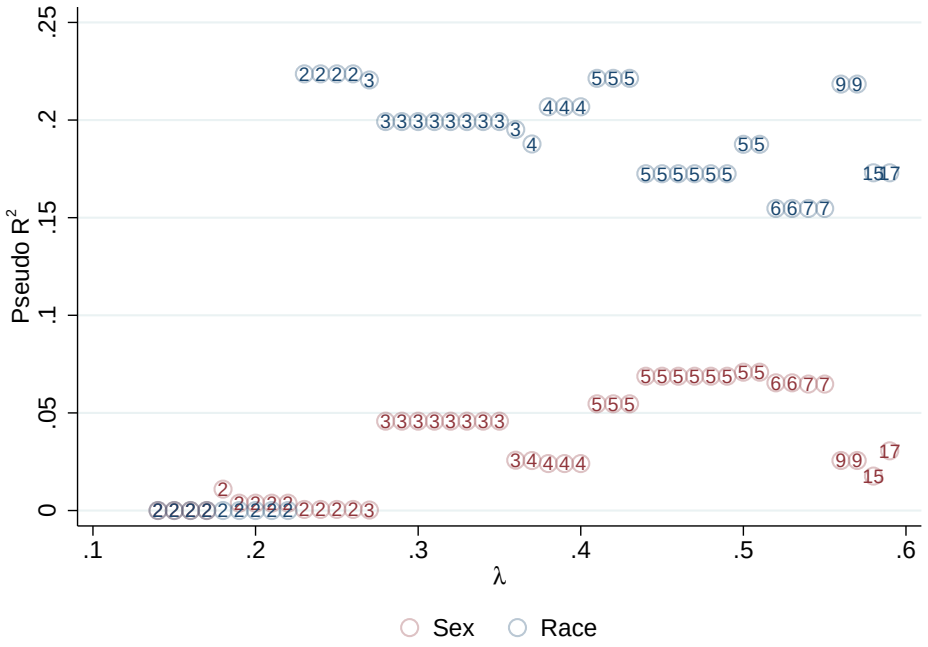
Table A1: First names assigned by race and gender

| | Black male | | White male | | Black female | | White female | |
|---|---|---|---|---|---|---|---|---|
| | Name | Source | Name | Source | Name | Source | Name | Source |
| 1 | Antwan | NC | Adam | NC | Aisha | Both | Allison | BM |
| 2 | Darnell | BM | Brad | Both | Ebony | Both | Amanda | NC |
| 3 | Donnell | NC | Bradley | NC | Keisha | BM | Amy | NC |
| 4 | Hakim | BM | Brendan | Both | Kenya | BM | Anne | BM |
| 5 | Jamal | Both | Brett | BM | Lakeisha | NC | Carrie | BM |
| 6 | Jermaine | Both | Chad | NC | Lakesha | NC | Emily | Both |
| 7 | Kareem | Both | Geoffrey | BM | Lakisha | Both | Erin | NC |
| 8 | Lamar | NC | Greg | BM | Lashonda | NC | Heather | NC |
| 9 | Lamont | NC | Jacob | NC | Latasha | NC | Jennifer | NC |
| 10 | Leroy | BM | Jason | NC | Latisha | NC | Jill | Both |
| 11 | Marquis | NC | Jay | BM | Latonya | Both | Julie | NC |
| 12 | Maurice | NC | Jeremy | NC | Latoya | Both | Kristen | Both |
| 13 | Rasheed | BM | Joshua | NC | Lawanda | NC | Laurie | BM |
| 14 | Reginald | NC | Justin | NC | Patrice | NC | Lori | NC |
| 15 | Roderick | NC | Matthew | Both | Tameka | NC | Meredith | BM |
| 16 | Terrance | NC | Nathan | NC | Tamika | Both | Misty | NC |
| 17 | Terrell | NC | Neil | BM | Tanisha | BM | Rebecca | NC |
| 18 | Tremayne | BM | Scott | NC | Tawanda | NC | Sarah | Both |
| 19 | Tyrone | Both | Todd | BM | Tomeka | NC | Susan | NC |

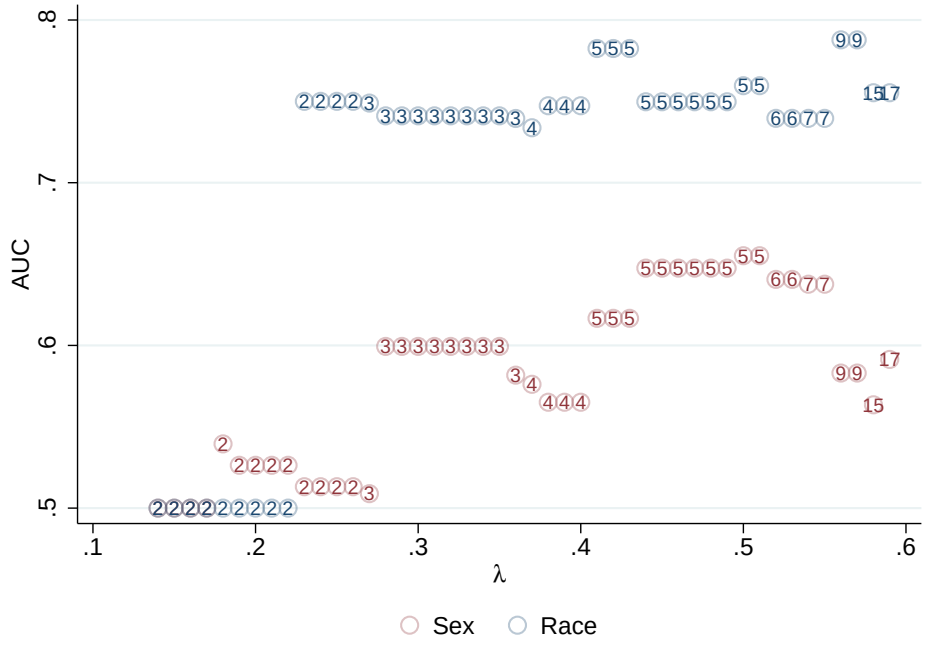
*Notes:* This table lists the first names assigned by race and gender and their sources. "BM" indicates that the name appeared in original set of nine names used for each group in Bertrand and Mullainathan (2004). "NC" indicates the name was drawn from data on North Carolina speeding infractions and arrests. "Both" indicates the name appeared in both sources. Names from N.C. speeding tickets were selected from the most common names where at least 90% of individuals are reported to belong to the relevant race and gender group.

Figure A1: Predictive power of grades name for race and sex labels

a) Pseudo $R^2$



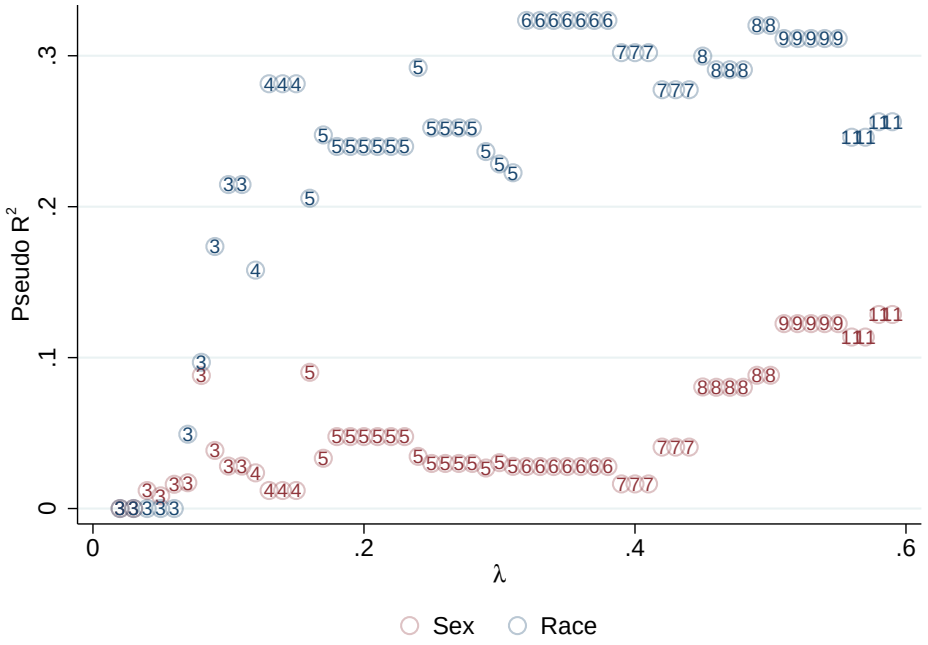

b) Area under the curve



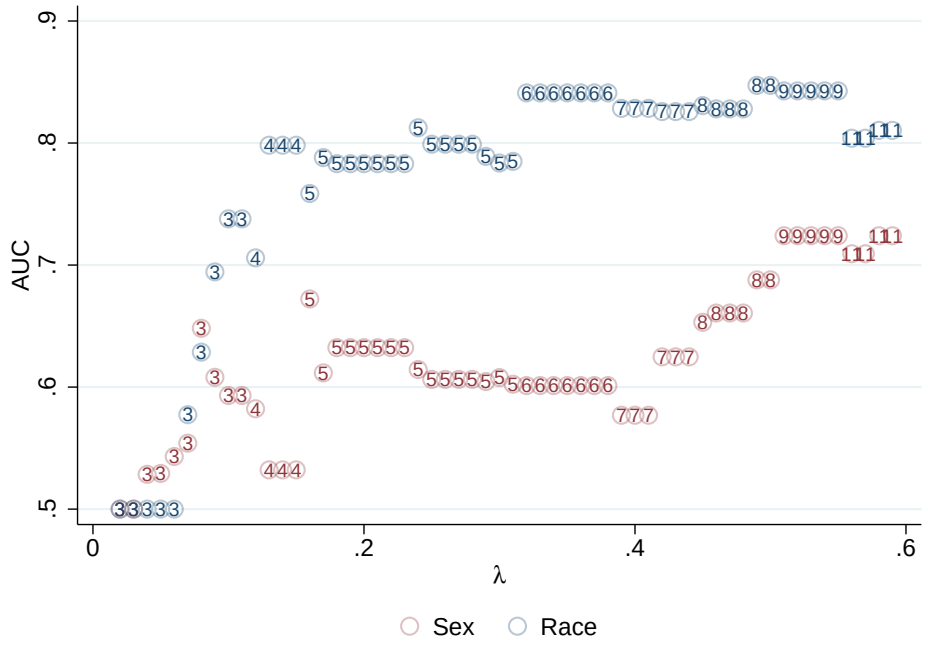

*Notes*: This figure plots the psuedo-$R^2$ (panel a) and AUC (panel b) for a series of logistic regressions using an indicator for the race or sex of the name as the outcome and dummies for assigned grades as the explanatory variables for an intermediate range of $\lambda$. The number shown indicates the number of grades assigned.

Figure A2: Predictive power of square-weighted name grades for race and sex labels

a) Pseudo $R^2$

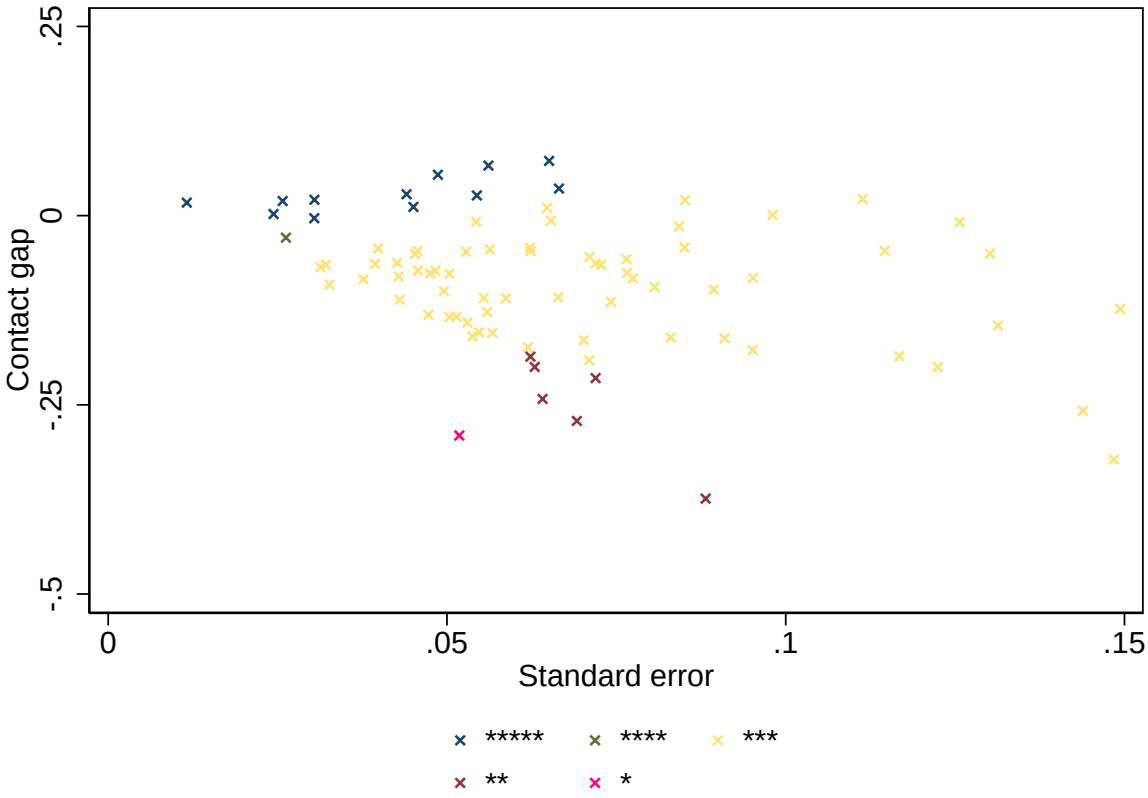



b) Area under the curve



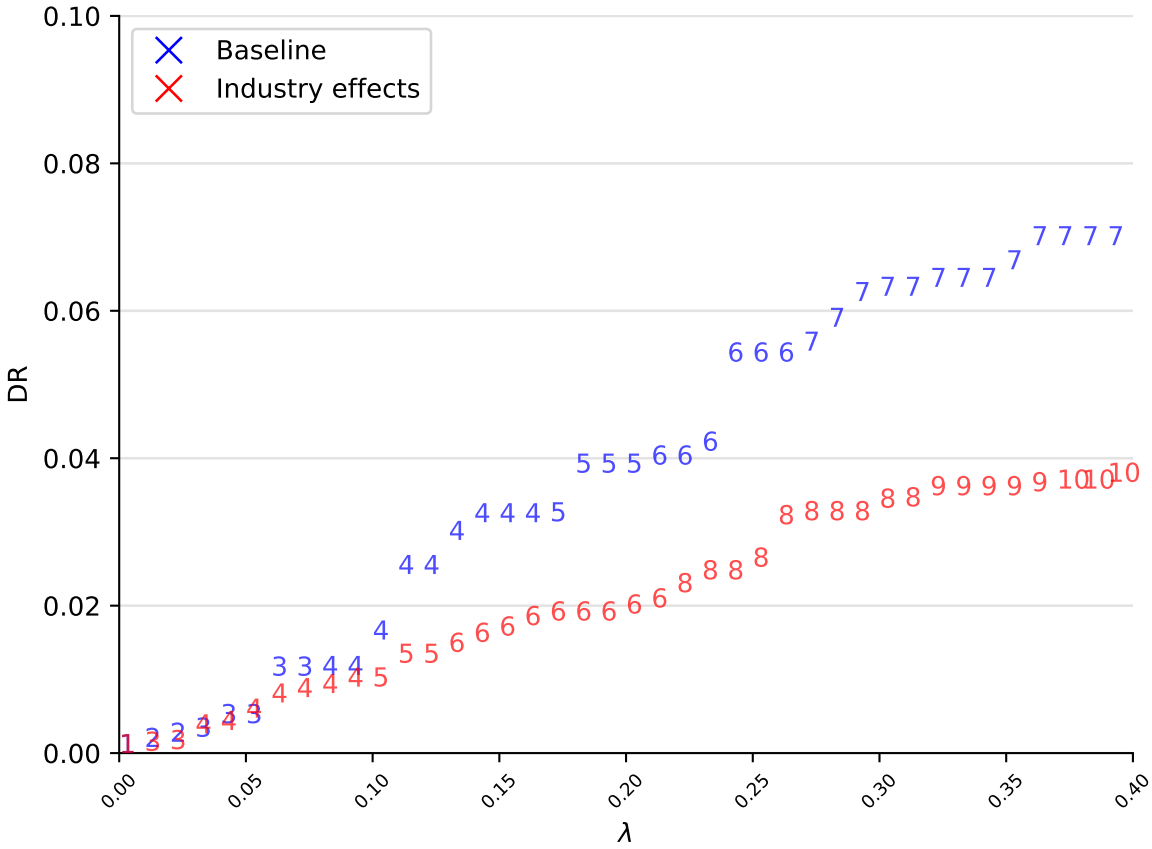

*Notes*: This figure plots the psuedo-$R^2$ (panel a) and AUC (panel b) for a series of logistic regressions using an indicator for the race or sex of the name as the outcome and dummies for assigned grades under square-weighted loss as the explanatory variables for an intermediate range of $\lambda$. The number shown indicates the number of grades assigned.

Figure A3: Contact penalties, standard errors, and report card grades




*Notes*: This figure plots the estimated proportional contact penalty for a Black name at each firm against the standard error of the contact penalty estimate. The shape and color of each point indicate the grade assigned to the firm using the same specification as Figure 10, which sets $\lambda = 0.36$.

Figure A4: Square weighted loss: Grades and discordance as a function of $\lambda$




*Notes*: This figure shows estimated weighted Discordance Rates (DR) for an intermediate range of $\lambda$. Weighted DR is the expected weighted Discordance Proportion (DP), or share of firm pairs whose latent relative rank is the opposite of their grade rank, weighted by the expected squared gap in $\theta$ for the firm pair. For small $\lambda$, which implies ranking mistakes are relative costly, optimal grades consist of just one or two grades. Larger $\lambda$ yield more grades and less weight on potential ranking mistakes, even when the number of grades remains the same.