A Discrimination Report Card

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

June 23, 2023

Abstract

We develop an Empirical Bayes grading scheme that balances the informativeness of the assigned grades against the expected frequency of ranking errors. Applying the method to a massive correspondence experiment, we grade the racial biases of 97 U.S. employers. A four-grade ranking limits the chances that a randomly selected pair of firms is mis-ranked to 5% while explaining nearly half of the variation in firms' racial contact gaps. The grades are presented alongside measures of uncertainty about each firm's contact gap in an accessible rubric that is easily adapted to other settings where ranks and levels are of simultaneous interest.

Keywords: Discrimination, Empirical Bayes, Ranking, Correspondence Experiment

JEL Codes: J71, C11, C13

---

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 may have little incentive to search for more equitable recruiting practices if they cannot benchmark their organizations' biases to those of their peers.

This paper constructs a *discrimination report card* that summarizes objective information regarding the relative biases of a broad collection of Fortune 500 companies. Our analysis leverages a massive correspondence experiment sending up to 1,000 job applications to each of 108 firms. In a previous analysis of this experiment (Kline, Rose and Walters, 2022), we established that applications listing distinctively Black names received seven percent fewer employer contacts than those listing distinctively white names. This contact 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 to racial discrimination.

While this earlier work established that discrimination is highly concentrated among a small subset of employers, the experimental estimate for any particular firm is subject to significant sampling uncertainty. This statistical noise presents an obstacle to reliable estimation of discrimination by specific firms. Though scientific communication is generally aided by transparency (Andrews and Shapiro, 2021), lay audiences can find it challenging to translate point estimates and standard errors into conclusions of interest. Experimental evidence from psychology suggests audiences frequently depart from Bayesian reasoning, which may generate over- or under-reactions to statistical evidence (Mullainathan, 2002; Mullainathan, Schwartzstein and Shleifer, 2008; Bordalo et al., 2016). These considerations suggest that a low-dimensional, easily-digestible summary of the experimental findings is important for effectively communicating the results to a broad audience. Likely for similar reasons, scholars, policymakers, and private businesses increasingly report simple "report cards" summarizing estimates of quality for various types of institutions, includ-

ing colleges (Chetty et al., 2017), K-12 schools (Bergman, Chan and Kapor, 2020; Angrist et al., 2021), teachers (Bergman and Hill, 2018; Pope, 2019), healthcare providers (Brook et al., 2002; Pope, 2009; Kolstad, 2013), and neighborhoods (Chetty and Hendren, 2018b; Chetty et al., 2018a).

This paper develops new methods for grading firms (or other units) based upon noisy measures of their conduct while maintaining statistical guarantees on the reliability of the resulting grades. The information content of our classification scheme is quantified by Kendall's τ measure of correlation (Kendall, 1938) between our proposed ranking and the true ranking of firms' latent discriminatory behavior. The reliability of the report card 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 worse grade discriminates less than a firm assigned the better grade. We also develop an extension of DR that weights mistakes by their cardinal magnitudes, which captures the idea that large mistakes are more costly than small ones.

We show that the tradeoff between these notions of information and reliability arises naturally from a series of lotteries over firm pairs where an analyst guesses which member of each pair exhibits worse conduct. When facing multiple gambles of this form, the analyst faces an optimization problem subject to logical transitivity constraints requiring all pairwise comparisons to be consistent with a coherent underlying ranking. A parameter λ 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. These maximally-informative 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), with the posterior probability that one firm is more biased than another serving as a "vote share." When $\lambda < 1$ it is only optimal to strictly rank firm pairs that can be distinguished with sufficiently high posterior probability, potentially yielding ties and therefore a low number of distinct "grades." These coarse grades protect against misinterpretation at the cost of losing information, thereby reducing correlation with the true ranks.

The grades generated by our procedure yield a simple classification of firms into groups that facilitates pairwise comparisons of firm conduct. The grades allow an assessment not only of which firms are discriminating most, but also which firms are discriminating *least*. The latter sort of information may prove especially useful in the hunt for best practices that lead to inclusive hiring. For example, our past work documented smaller contact gaps at firms that contacted applicants from fewer phone numbers, a proxy for centralized human resources (HR) practices. Berson, Laouenan and Valat (2020) provide corroborating evidence that contact gaps are lower among French firms with centralized

recruiting practices, while Challe et al. (2022) report encouraging experimental evidence on the potential for HR reforms to foster inclusive hiring at large firms. Our report card results suggest directions for work comparing these and other policies across firms graded as high and low performers.

To supplement the ordinal information conveyed by our grades, we report Empirical Bayes posterior means and credible intervals summarizing the *level* of discriminatory conduct at each employer. This ordinal and cardinal information is visualized in a single report card that provides a powerful and easily-accessible rubric for assessing absolute and relative discriminatory conduct. Although this rubric was designed to communicate information about discrimination, we expect the methods proposed here to prove useful in other settings where noisy measures of institutional performance and conduct have been studied, including assessments of school value added (Angrist et al., 2017), hospital quality (Chandra et al., 2016), health insurance plans (Abaluck et al., 2021), and regional intergenerational mobility (Chetty and Hendren, 2018a). Routines for implementing our ranking procedure are available online at `https://github.com/ekrose/drrank`.

As an introductory illustration of the method, we rank the contact rates of the first names used in our correspondence experiment. A non-parametric deconvolution suggests that name-specific contact rates cluster around two distinct values capturing mean contact rates for distinctively white and Black names. Weighing the loss from incorrectly ordering a pair of names four times as heavily as the gain from correctly ordering them, our ranking procedure stratifies the names into two groups with distinct grades. These grades are shown to be strongly predictive of a name's nominal race but not its sex. Allowing additional grades has little impact on these correlations, suggesting that our ranking procedure is suitable for recovering "missing labels" with a low-dimensional structure.

Proceeding to our primary application of ranking firm biases against Black applicants, we compute optimal grades subject to the same preferences over correct and incorrect rankings used for first name pairs. In a single pairwise gamble, these preferences (which correspond to a particular choice of λ) require at least 80% posterior confidence to justify a strict ordering of firms. In our baseline specification, applying this choice of λ to generate a transitive ordering over all firms yields three unique grade levels. These grades capture roughly 19% of the between-firm variation in proportional contact penalties and yield an expected rank correlation with the true penalties of 0.21. Although our grading system reflects only ordinal considerations, we estimate that the average white/Black gap in contact rates among firms awarded the worst grade is 24%, while the gap among firms awarded the best grade is only 3%.

Our earlier work found that industry explains roughly half of the variation in racial discrimination levels across firms (Kline, Rose and Walters, 2022). Motivated by this finding, we extend our procedure to build industry information into the report card grades. This extension is achieved by augmenting Efron (2016)'s log-spline deconvolution

approach to flexibly estimate separate distributions of discrimination within and between industries. Consistent with our past work, we find that industry affiliation accounts for roughly half of the cross-firm variation in proportional contact penalties. Incorporating industry affiliation into the ranking procedure with the same choice of λ yields four grades. These improved grades explain 47% of the variation in contact penalties across firms and yield a correlation with the latent ranks of 0.32, while limiting the expected share of firm pairs that are mis-ranked to 5.2%.

These more informative industry-dependent grades constitute our preferred ranking of the companies in our experiment. Firms assigned the worst grade in this ranking contact white applicants 22% more often than Black applicants, similar to the lowest category in the ranking without industry effects. However, 5 firms receive this label in the model with industry effects compared to only 2 in the baseline model, an indication of the extra information conveyed by industry. Similar to the specification without industry, firms receiving the best grade in the industry effects model (11 firms) exhibit very small racial biases. To the extent that these differences are driven by HR practices or other firm policies, there may be opportunities for the substantial set of firms that scored poorly to improve their behavior by imitating the practices of those who scored more highly.

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., 2018b). Portnoy (1982) established conditions under which ranking based on such rules maximizes the probability of a correct ordering, while Laird and Louis (1989) proposed directly computing posterior mean ranks under a normality assumption on the latent heterogeneity. Both sorts of ranks may be noisy, however, leading to a proliferation of ranking mistakes when the number of units grows large. A recent econometrics literature confirms that this problem can become severe in practice and proposes approaches to testing hypotheses regarding either ranks themselves or the levels of highly-ranked units (Andrews, Kitagawa and McCloskey, 2019; Mogstad et al., 2020).

Building on the analogy with multiple testing, Gu and Koenker (2020) consider the use of non-parametric EB methods to select tail performers subject to constraints on the False Discovery Rate, which limits the number of ordering mistakes expected when selecting top performers. Our proposal generalizes the approach in Gu and Koenker (2020) by accommodating more than two grades and avoids the requirement to treat one of the grades 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 allows users to trade off transparent notions of the information content and reliability of the resulting grades.

After undergoing peer review, we plan to release the names of the firms in this study 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 violations by federal contractors. Many of the firms receiving poor grades turn out to be federal contractors, suggesting this information may be of help in targeting future compliance efforts. However, compliance efforts are inevitably long and costly and many firms remain out of compliance even after having been fined (Maxwell et al., 2013).

As the introductory quote by Brandeis suggests, shining some statistical light on the problem of discrimination may have a more immediately salutary effect than regulatory enforcement efforts. Little scientific information about the discriminatory conduct of particular firms is available to the public. The most powerful "disinfectant" may well be the decentralized reactions of employees, customers, and leaders of these organizations to the provision of such information.

2   The experiment

We construct discrimination report cards based on the resume correspondence experiment analyzed in Kline, Rose and Walters (2022). The experiment's sampling frame began 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. This restriction was intended to ensure nation-wide coverage of each firm's recruitment conduct and to minimize the chances that multiple sampled job vacancies were being managed by the same individual.

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. 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. Distinctive last names were taken from the 2010 U.S. Census. We selected names with high race-specific shares among those that occur at least 10,000 times nationally.

One application within each pair was randomly assigned a distinctively white name while the other was randomly assigned a distinctively Black name. Fifty-percent of names were distinctively female and the rest distinctively male, but assignment of sex was not stratified. Each fictitious applicant was independently 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. Further details on the experimental design are available in Kline, Rose and Walters (2022).

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 discrimination $\theta_i \in \mathbb{R}$. The decision variable $d_i \in [n]$ gives the *grade* assigned to firm i. Larger values of d_i indicate a firm is more biased. Hence, when $d_i > d_j$ for two firms i and j, we say that firm i received a "worse" grade than firm j.

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 θ_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 by 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 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 reversing their true rank yields payoff -1. We can also declare the comparison a draw by assigning the firms

equal ranks, which amounts to abstaining from the gamble and yields certain payoff 0.

In considering this lottery, we view the pair (θ_i, θ_j) of firm types as *i.i.d.* draws from a continuous prior distribution $G : \mathbb{R} \to [0, 1]$. The posterior probability that firm i is more biased than firm j will be denoted $\pi_{ij} = \Pr(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j)$. Because G is continuous, ties are measure zero and $\pi_{ij} = 1 - \pi_{ji}$.

The expected utility of assigning grades $d = (d_1, d_2) \in \{1, 2\}^2$ to these firms can therefore be written

$$
\begin{aligned}
EU(\pi_{ij}, d) &= [\lambda \pi_{ij} - \pi_{ji}] \cdot 1\{d_i > d_j\} + [\lambda \pi_{ji} - \pi_{ij}] \cdot 1\{d_i < d_j\} \\
&= [(1+\lambda)\pi_{ij} - 1] \cdot 1\{d_i > d_j\} + [(1+\lambda)(1 - \pi_{ji}) - 1] \cdot 1\{d_i < d_j\}.
\end{aligned}
$$

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 $\pi_{ji} > \frac{1}{1+\lambda}$.

- 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 greater probability of having the largest value of θ. But when $\lambda < 1$, it is better to assign pairs of firms with π_{ij} near $1/2$ equal grades 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 (1938)'s classic τ 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 τ indexed by a scalar $\lambda \in [0, 1]$ that controls the benefit of a concordance relative to the cost of a discordance.

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

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

8

While every discordant pair yields a loss of 1, every concordant pair reduces loss by λ. When $\lambda = 1$ the loss function equals minus one times Kendall's τ measure of rank correlation between d and θ, 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:

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

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

where the quantity $DP(d, \theta) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} [1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}]$ is the *Discordance Proportion*. The Discordance Proportion gives the share of firm pairs that are misranked according to their grades. Interpreting our decision problem as a series of tests of the null hypothesis that each member of a pair discriminates equally, the Discordance Proportion may be seen as 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 λ 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 $DP(d, \theta)$.

3.3 Risk function

While one would ideally like to choose grades d that balance the rank correlation $\tau(d, \theta)$ against the Discordance Proportion $DP(d, \theta)$, these quantities are not directly observed. Suppose, however, that we have a continuous *i.i.d.* prior $G : \mathbb{R} \to [0, 1]$ over the elements of θ and that we observe a realization y of the random vector $Y = (Y_1, \ldots, Y_n)'$. The posterior probability that firm $i \in [n]$ is more biased than firm $j \neq i$ will again be denoted by $\pi_{ij} = \Pr(\theta_i > \theta_j | Y = y)$.

The posterior expectation under G of both the unknown $\tau(d, \theta)$ and $DP(d, \theta)$ can be expressed in terms of the pairwise probabilities π_{ij}. The posterior expected rank correlation $\bar{\tau}(d) = \mathbb{E}[\tau(d, \theta) | Y = y]$ is given by

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

Likewise, the posterior expected value of $DP(d, \theta)$, a quantity we term the *Discordance*

Rate (DR), is

$$DR(d) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\{d_i < d_j\}\pi_{ij} + 1\{d_i > d_j\}\pi_{ji}. \tag{2}$$

Consequently, the posterior expected loss (i.e., the Bayes risk) of assigning grades $d \in [n]^n$ can be written:

$$\begin{aligned}
\mathcal{R}(d; \lambda) &= \mathbb{E}[L(d, \theta; \lambda) \mid Y = y] \\
&= (1 - \lambda)DR(d) - \lambda\bar{\tau}(d) \\
&= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \pi_{ji}1\{d_i > d_j\} + \pi_{ij}\left(1 - 1\{d_i = d_j\} - 1\{d_i > d_j\}\right) - \\
&\quad \lambda\pi_{ji}\left(1 - 1\{d_i = d_j\} - 1\{d_i > d_j\}\right) - \lambda\pi_{ij}1\{d_i > d_j\}.
\end{aligned} \tag{3}$$

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 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:

$$d_{ij} + d_{jk} \leq 1 + d_{ik}, \tag{4}$$

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

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

Hence, we can rewrite the problem of choosing $d \in [n]^n$ to minimize (3) as that of choosing the binary indicators $\{d_{ij}, e_{ij}\}_{i=2, j=1}^{i=n, j=i}$ to minimize

$$\sum_{i=2}^{n} \sum_{j=1}^{i} \pi_{ji}d_{ij} + \pi_{ij}\left(1 - e_{ij} - d_{ij}\right) - \lambda\pi_{ji}\left(1 - e_{ij} - d_{ij}\right) - \lambda\pi_{ij}d_{ij}, \tag{5}$$

subject to the transitivity constraints in (4) and the logical constraint that $e_{ij} + d_{ij} + d_{ji} = 1$ for all $(i, j) \in [n]^2$. Note that both the objective (5) 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 λ

To develop intuition for the role that λ 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 (5)

while ignoring cross-pair constraints. Setting $d_{ij} = 1$ minimizes risk whenever

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

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

$$\pi_{ij} > (1 + \lambda)^{-1}, \tag{6}$$

which is equivalent to the rule derived in section 3.1. Hence, with $\lambda = 1$, it is optimal to choose $d_{ij} = 1\{\pi_{ij} > 1/2\}$, which can be thought of as a MAP estimate of the pairwise rank. When $\lambda \in (0, 1)$, greater posterior certainty is required to conclude that $d_{ij} = 1$ and values of π_{ij} near $1/2$ will yield ties even though G is continuous. As λ 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 from applying the pairwise thresholding rule in (6) when $\lambda < 1$ can generate a form of Condorcet cycle in *indifferences* that violates the transitivity constraints in (4) even if they would be satisfied under $\lambda = 1$. The following three firm example illustrates the problem.

Example 1 (Three firms, normal posteriors). Suppose $n = 3$ and $\theta_i | Y_i = y_i \sim N(\omega_i, 1)$. If the $\{\theta_i\}_{i=1}^3$ are mutually independent, then:

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

Let $\lambda = 1/4$, which implies $(1+\lambda)^{-1} = 0.8$. If $(\omega_1, \omega_3) = (2, 0)$, so that $\pi_{13} = \Phi(\sqrt{2}) = .92$ and $\pi_{31} = 1 - \pi_{13} = .08$, then it is optimal to rank $\theta_1 > \theta_3$. But if $\omega_2 \in (0.81, 1.19)$, it is optimal to rank both (θ_1, θ_2) and (θ_2, θ_3) as ties because $\max\{\pi_{12}, \pi_{23}\} < 0.8$. By transitivity, this implies $\theta_1 = \theta_3$ which contradicts our earlier assertion that $\theta_1 > \theta_3$. $\square$

Note that if we had set $\lambda = 1$ in the above example transitivity would have been satisfied because the posteriors themselves are transitive in the sense that for any triple (i, j, k) of firms, $\pi_{ij} > \pi_{ji}$ and $\pi_{jk} > \pi_{kj}$ imply $\pi_{ik} > \pi_{ki}$. This transitivity derives from the scalar index structure of the posteriors, revealed by the fact that

$$\pi_{ij} > \pi_{ji} \iff \omega_i > \omega_j.$$

Sobel (1993) establishes the transitivity of posteriors in a broader exponential family subject to a corresponding index restriction. In general, however, when the observations are heteroscedastic, such index representations are not available and transitivity is not

assured. When transitivity fails, the constraints in (4) will bind and multiple units may receive the same grade even when $\lambda = 1$.

Finally, note that coarse grades need not be a consequence of transitivity violations. If $\omega_2 \in (-1.19, 0.81)$ in the preceding example, it is optimal to rank $\theta_1 > \theta_3$, $\theta_1 > \theta_2$, and $\theta_2 = \theta_3$. Thus pairwise thresholding yields two grades and no transitivity violations. Whether the transitivity constraints bind therefore depends on the structure of the pairwise posterior contrasts.

3.5 Connections to social choice

The literature on ranking methods bears a close connection to problems of social choice. If we re-interpret π_{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.[1] 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 $\sum_{j \neq i} 1\{\pi_{ij} > 1/2\}$. If (as we have assumed) G is continuous, then the Borda measure is equivalent to the posterior mean rank, a quantity studied by Laird and Louis (1989).

The ranking procedure devised in section 3.3 turns out to be closely tied to Condorcet (1785)'s voting scheme. To develop this connection, it is useful to define the Kemeny (1959) distance between the vectors θ and d, which can be written

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

Integrating out θ yields

$$\mathbb{E}[K(\theta, d) | Y = y] = \sum_{i=2}^{n} \sum_{j=1}^{i} (\pi_{ij} + \pi_{ji}) 1\{d_i = d_j\} + (1 + \pi_{ji} - \pi_{ij}) 1\{d_i > d_j\}$$
$$+ (1 + \pi_{ij} - \pi_{ji}) 1\{d_i < d_j\}.$$

When ties are not possible (i.e., when $\pi_{ij} = 1 - \pi_{ji}$ for all $i \neq j$) we obtain the simplification

$$\mathbb{E}[K(\theta, d) | Y = y] \propto \sum_{i=2}^{n} \sum_{j=1}^{i} (2\pi_{ij} - 1)(d_{ji} - d_{ij}). \tag{7}$$

Young and Levenglick (1978) show that Condorcet (1785)'s voting scheme is equivalent to choosing a ranking d that minimizes (7). Young (1986) establishes that this vote aggregation scheme is the unique rule that is unanimous, neutral, and satisfies reinforcement

---

[1]In developing this analogy, we temporarily depart from the convention that $d_i > d_j$ implies firm i has been assigned a "worse" grade than firm j, referring instead to firms with high d_i as highly ranked.

and independence of remote alternatives. It can also be viewed as a type of maximum likelihood estimator giving "the ranking of all candidates that is most likely to be correct" (Young, 1986, p.114).

Note that $(2\pi_{ij} - 1)(d_{ji} - d_{ij})$ is minimized by the MAP thresholding rule $d_{ij}^{MAP} = 1\{\pi_{ij} > 1/2\}$. When $\lambda = 1$, our objective in (5) reduces to (7). Consequently, the most granular version of our grading scheme 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. Here, a close election is one where $\pi_{ij} < (1 + \lambda)^{-1}$.

Condorcet rankings obey the famous Condorcet winner criterion: a unit that wins all pairwise elections between candidates (that is, satisfies $\pi_{ij} > 1/2 \ \forall j \neq i$) will be ranked first. The following proposition reveals that when $\lambda < 1$ our grades fulfill a modified version of the Condorcet winner criterion.

Proposition 1 (λ-Condorcet Criterion). Suppose that firm i satisfies $\pi_{ij} > (1+\lambda)^{-1} \ \forall \ j \neq i$. Then $d_i > d_j \ \forall \ j \neq i$. Moreover, suppose that firm k satisfies $\pi_{ik} > (1 + \lambda)^{-1}$ and $\pi_{kj} > (1 + \lambda)^{-1} \ \forall \ j \neq i, j \neq k$. Then $d_i > d_k > d_j \ \forall \ j \neq i, j \neq k$.

We leave the short proof for Appendix A. By symmetry of the objective in (5), the firm assigned the lowest grade by our method must achieve the highest grade when the sign of the estimand being ranked is reversed. Hence, Proposition 1 also implies that any Condorcet *loser* – i.e., any candidate firm i with $\pi_{ji} > (1 + \lambda)^{-1}$ for all $j \neq i$ – must be assigned the lowest grade.

Another well-known property of Condorcet rankings is that when no Condorcet winner exists, the top ranked candidate must be a member of the Smith (1973) set: the smallest non-empty subset of candidates such that every candidate in the subset is majority-preferred over every candidate not in the subset. The following proposition establishes a corresponding property of our grades in the case where $\lambda < 1$.

Proposition 2 (λ-Smith criterion). Let $\mathcal{S}$ denote a collection of firms exhibiting the following dominance property: $\pi_{ij} > (1 + \lambda)^{-1} \ \forall i \in \mathcal{S}, j \notin \mathcal{S}$. Then the top graded firms must be a member of $\mathcal{S}$.

The proof is again left for the appendix. Symmetrically, Proposition 2 implies the firm assigned the lowest grade must be a member of the Smith loser set of candidates that are majority non-preferred to all others. Finally, we note that when $\lambda < 1$ and no ordering is possible within the Smith set, all firms in the set will receive equal grades.

Proposition 3 (Unordered λ-Smith candidates are tied). Let $\mathcal{S}$ denote a collection of firms exhibiting the following dominance property: $\pi_{ij} > (1 + \lambda)^{-1} \ \forall i \in \mathcal{S}, j \notin \mathcal{S}$. Moreover, suppose $\pi_{ij} < (1 + \lambda)^{-1} \ \forall (i, j) \in \mathcal{S}$. Then all firms in $\mathcal{S}$ receive the highest grade.

As with the preceding propositions, the proof appears in Appendix A.

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:

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

A loss function corresponding to the $(p = 2, \lambda = 1)$ case was previously considered by Sobel (1990). The corresponding family of risk functions take the form

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

where $\mu_{ij}^p = \mathbb{E}\left[\max\{(\theta_i - \theta_j), 0\}^p \mid Y = y\right]$. Note that $\lim_{p \to 0} \mu_{ij}^p = \pi_{ij}$. Hence, one can think of our baseline risk function in (5) 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") but also report results analyzing the case where $p = 2$ ("square-weighted loss") in Appendix D.

3.7 Discordance rates

To summarize the reliability of the grades it is useful to report the Discordance Rate $DR(d^*)$, which gives the posterior expected frequency of discordances between pairs of grades. From (2), this quantity is trivial to compute, as 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

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

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 discriminate equally 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 overall 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:

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

where $m_{ij}^p = \mathbb{E}_G\left[(\theta_i - \theta_j)^p \mid Y = y\right]$. 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 $\Pr\left(\theta_i > \theta_j\right) = 1 - \Pr\left(\theta_i \leq \theta_j\right)$. 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 of the methods developed in the previous section, 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.

Table 1 lists the mean contact rates of names in each of these categories, along with the number of applications. Distinctively white and female names were called back most often in the experiment, followed by white male names, then Black male names, with Black female names being called back least often. Though the same names were intended to be sent to each firm, the COVID-19 epidemic and other disruptions led to minor imbalances reflected in the sample counts. In our past work (Kline, Rose and Walters, 2022) 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, which suggests these imbalances had little effect on mean contact rates by race or sex. For completeness, Table 1 reports for each demographic group the p-value of a Wald test of the null hypothesis that the name specific contact rates are equal. The smallest such p-value is 0.24.

To get a sense of what share of the variation in name contact probabilities is likely explained by the names' putative race and sex labels, we compute a bias-corrected estimate of the between demographic group variance using the formula $\frac{G-1}{G}\left(S^2 - \overline{s^2}\right)$ where $G = 4$ is the number of demographic groups, S^2 is the sample variance across demographic groups of the point estimates reported in Table 1, and $\overline{s^2}$ is the average squared standard error across those groups. This calculation yields an unbiased estimate of the between demographic group variance of $(0.011)^2$. Applying a corresponding calculation to the name specific means and standard errors yields a bias-corrected estimate of the variability in contact probabilities across all 76 first names of $(0.010)^2$. The finding that our between demographic group variance estimate exceeds our between name variance estimate suggests that the nominal race and sex of a first name explains nearly all of the variation across first names in contact probabilities; that is, we expect that a regression of the latent name specific contact probabilities on race and sex interactions would yield an R^2 very close to one.

In principle, even if race and sex perfectly predict employer treatment of names, the causal factors generating this association could be other features of names that correlate strongly with race and sex. A candidate factor that has attracted substantial attention from social scientists is employer stereotypes about the likely productivity of individuals with different names (Fryer Jr and Levitt, 2004; Gaddis, 2017). This hypothesis was evaluated by Bertrand and Mullainathan (2004), who found that the average socioeconomic status of the first names considered in their experiment (as proxied by average maternal education) varied widely within race but were insignificantly related to contact rates.[2] Because roughly half (40 out of 76) of our first names coincide with those studied by

---

[2]Recent work by Crabtree et al. (2022) directly elicits perceptions of educational attainment and income by first name on a variety of online platforms. Remarkably, they find that perceptions of social class exhibit variability across racially distinctive first names in the same race category comparable in magnitude to the variability found between race categories (see their Figure 4).

Bertrand and Mullainathan (2004), our finding of insignificant contact probability differences within race and gender casts further doubt on the view that employer responses are driven primarily by features of names other than their likely race or sex.

The finding that race and sex provide an accurate low dimensional summary of the 76 name specific contact probabilities suggests it is possible to build a highly informative ranking of the names involving just a few grades. Below, we investigate this conjecture in two ways. First, we examine how the expected Kendall's τ produced by our procedure scales with the number of grades assigned. Second, we treat each name's nominal race and sex as "missing labels" and study the extent to which the coarse grades assigned to first names by our ranking algorithm can recover these labels from data on firms' sample contact rates.

4.1   Estimating G

Abusing notation somewhat, 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 has mean p_i and variance $p_i(1-p_i)/N_i$. This dependence of the variance on the contact probability complicates ranking exercises, as contact rates for names that deserve the best and worst grades – that is, those with p_i near one or zero – will also be estimated with the least 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 θ_i was drawn we first apply the non-parametric maximum likelihood (NPMLE) estimator developed by Koenker and Mizera (2014) as implemented by the 'GLmix' command in the R package REBayes (Koenker and Gu, 2017). The NPMLE estimates a discrete approximation to the distribution of θ_i assuming that $\hat{\theta}_i \mid \theta_i, N_i \sim \mathcal{N}(\theta_i, (4N_i)^{-1})$. Supporting the maintained independence of θ_i from N_i, a regression of $\hat{\theta}_i$ on $\ln N_i$ yields a statistically insignificant relationship (p=0.17).[3]

---

[3]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 plot of the estimated marginal distribution $\hat{G}$ of θ_i produced by the NPMLE is provided in Figure 1. The bars correspond to histograms of $\hat{\theta}_i$ while the yellow spikes represent the estimated probability mass function $d\hat{G}$ of the θ_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 $n^{-1} \sum_{i=1}^n s_i^2 = n^{-1} \sum_{i=1}^n (4N_i)^{-1}$.

Figure 1 also plots the estimated density of θ_i produced by Efron (2016)'s log-spline estimator, which models G as a 5th order spline in the exponential family. Estimation of the spline parameters is conducted via penalized maximum likelihood, where N_i is treated as independent of θ_i. The penalization parameter has been chosen by GMM to match unbiased estimates of G's first two moments as closely as possible. Despite being continuous, the bimodal shape of the log-spline estimate is remarkably consistent with that of the NPMLE. For reference, the sample mean values of $\hat{\theta}_i$ for each nominal race and sex category are portrayed on the Figure as vertical lines. Despite not using race in the estimation procedure, the two modes of the log-spline estimate fall near the race specific means as do the modes of the NPMLE estimate.

The lower panel of Figure 1 converts these 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 1.8 percentage point gap between these mass points is very near 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 third mass point at $p = 0.260$, which lies just below the estimated average contact rate for distinctively white female names.

As noted earlier, the discrete $\hat{G}$ produced by the NPMLE is a data dependent approximation to the true G. Even if differences in the treatment of names are driven primarily by employer perceptions of their race and sex, it seems unlikely that the true G is literally characterized by a few mass points, as small differences across names in their perceived race should generate corresponding contact rate differences. In fact, although three discrete modes are visible from the Figure, the NPMLE solution involves dozens of atoms with positive mass, most of which are imperceptible. In what follows, we rely on the log-spline estimate $\hat{G}$ of G which, as in the theoretical analysis of Section 3, implies that ties are measure zero.

4.2   Reporting possibilities

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 $\lambda = 1$). To ease interpretation, we have labeled the name with the highest ranked contact probability

1 and that with the lowest ranked contact probability 76. Name pairs with adjacent ranks exhibit π_{ij}'s near 1/2, indicating that we have little confidence in their relative order. Reassuringly, names with vastly different ranks exhibit π_{ij}'s near 0 or 1, indicating that the experimental data are informative about the relative ordering of these pairs.

The top right panel of Figure 2 depicts the Discordance Rate that arises from minimizing $\mathcal{R}(d; \lambda)$ – that is, from solving (5) subject to (4) – for different choices of λ. The number used to demarcate each solution corresponds to the number of distinct grades the integer linear programming procedure produces for that choice of λ. A sharp "elbow" emerges around $\lambda = 0.18$, above which the DR grows rapidly.

The bottom panel depicts the trade off between grade reliability $1 - DR$ and informativeness $\bar{\tau}$ associated with our choice of λ. The data are potentially quite informative about name rankings: as λ approaches 1, the expected rank correlation $\bar{\tau}$ approaches 0.44. However, the reliability of such a report would be fairly low, yielding a Discordance Rate of 0.28. For comparison, we also show the results of naively ranking based on $\hat{\theta}$ or the EB posterior mean $\bar{\theta}_i = \mathbb{E}[\theta_i | Y = y]$. Remarkably, both naive approaches yield ranks with $\bar{\tau}$ and DR very similar to those produced by our procedure when $\lambda = 1$.

To improve the reliability of the grades, we set $\lambda = .25$, implying via equation (6) that, in the absence of transitivity considerations, we would require 80% posterior certainty of each pairwise ranking decision that is not an abstention. This choice yields two grades that exhibit a remarkably high level of informativeness ($\bar{\tau} = 0.29$) and an acceptable level of reliability ($DR = .07$). For comparison, lowering the implicit posterior threshold to 70% by setting $\lambda = 0.41$ would yield three grades and increase their informativeness by 11% (to $\bar{\tau} = 0.32$) at the expense of a 21% increase in DR. Conversely, requiring $\lambda < 0.18$ would generate only one grade, yielding both $\bar{\tau}$ and DR of zero by construction.

4.3   Grades and demographics

Figure 3 lists the first names according to their Condorcet ranking, along with the posterior mean of each name's contact probability $p_i = \sin(\theta_i)^2$. In addition to the posterior means, which are depicted as dots, we report posterior credible intervals connecting the 2.5th percentile of each name's posterior distribution of contact probabilities to the 97.5th percentile of its posterior distribution. One should expect that approximately 72 (i.e., 95%) of these 76 intervals contain their name's true latent contact rate.[4] While the credible intervals tend to be fairly short—spanning between two and three percentage points in most cases—there is clearly enough uncertainty about each name's contact probability to significantly complicate the task of ranking them.

The Condorcet ranks can be thought of as a one dimensional encoding of the $\binom{76}{2} =$

---

[4]The asymmetry of the credible intervals reflects both that the estimated mixing distribution $\hat{G}$ of θ_i is asymmetric and that we have fed the interval limits through the nonlinear transformation $\theta \mapsto \sin(\theta)^2$.

2, 850 pairwise contrasts depicted in the upper left panel of Figure 2. When the posteriors are transitive, as they are in this case, this reduction amounts to the most likely ordering of the name contact probabilities. Variation in N_i across names, and hence the precision with which contact rates are measured, could in principle generate substantial non-monotonicity of the posterior mean in the Condorcet rank. In practice, however, names' Condorcet rankings are very nearly monotone in their posterior means.

The Condorcet ranks are extremely correlated with race. Of the top 38 ranked first names, only 8 are distinctively Black. Though the three top ranked names "Misty," "Heather," and "Laurie" are all distinctively female, the presumptive sex of a name turns out to be only weakly related to its Condorcet rank: 19 of the top 38 names are distinctively male. Hence, the Condorcet ranks manage to recover the race labels from contact rates with very little error but serve as unreliable proxies of a name's sex.

By construction, the Condorcet ranks exhibit the strongest expected rank correlation with the latent θ ranks. The coarse ranks that emerge when $\lambda < 1$ sacrifice rank correlation in exchange for the prospects of incurring fewer mistakes. Each name's color reflects its assigned grade (i.e., its coarse rank). A depiction of how the grades vary with name-specific contact rates and their standard errors is provided in Appendix Figure E1. As expected, names with higher contact rates tend to earn the better grade $\star\star$. However, heteroscedasticity in the estimates prevents the grades from being characterized by a single cutoff contact rate.

Though we saw earlier that the expected rank correlation of our grades with the true latent ranks is 0.29, it is also of interest to know how much p_i varies across grades. As described in Appendix B, we can use our EB posteriors to compute an estimate of the variance of p_i across grades. Though our procedure assigns only two grades to the names, we estimate that the (name-weighted) between grade standard deviation in contact probabilities is 0.006. Since the marginal standard deviation of p_i is roughly 0.010, a regression of the latent p_i on our grades should yield an R^2 of 35%.

The coarse grades that emerge from our procedure continue to align closely with our race labels: 35 of the 53 names (66%) in the top grade are distinctively white, while just 3 of the 23 names (13%) in the second grade are white. Notably, the top two names are also female; however, they do not appear in their own grade. Hence, a two-group ranking recovers the missing race label with limited error and, consistent with our findings in Table 1, suggests that white female names are particularly favored.

It is natural to wonder if a solution with more grades would be more predictive of sex. Appendix Figure E2 reports the pseudo-R^2 and Area under the Curve (AUC) from a series of logistic regressions of the name's sex on grade indicators for different choices of λ. 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. However, the four-grade solution with the smallest value of λ yields a pseudo-R^2 for sex of 0.012. With five grades we

find a pseudo-R^2 for sex of 0.034. By contrast, a corresponding logistic regression of race on assigned grades yields pseudo-R^2s for four- and five-grade solutions of 0.28 and 0.23, respectively.

Appendix Figure E4 shows the grades that result from minimizing expected square-weighted loss $\mathcal{R}^2(d, \lambda)$, which weights large mistakes in pairwise rankings more heavily than small mistakes. As above, we set $\lambda = 0.25$, which now yields five distinct grades. Once again, the grade categories are better predictors of race than sex. As shown in Appendix Figure E5, a logistic regression of sex on these five grade categories yields a pseudo-R^2 of roughly 0.05, while a corresponding race regression yields a pseudo-R^2 of 0.24. Appendix D provides further details on the square-weighted rankings.

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 λ, 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 conduct of each firm i in our experiment is characterized by the race-specific contact probabilities (p_{iw}, p_{ib}). These probabilities represent the hypothetical 30 day contact rates that would arise for applications with distinctively white and Black names, respectively, if we were to sample an infinite number of job vacancies from firm i and send each job four pairs of applications. The sample contact rates $(\hat{p}_{iw}, \hat{p}_{ib})$ provide unbiased estimates of these contact probabilities.

To mitigate the potential influence of firm heterogeneity in baseline contact rates on our measure of discrimination, we focus on the following proportional measure of bias against Black names at firm i:

$$\theta_i = \ln(p_{iw}) - \ln(p_{ib}).$$

Our estimator of θ_i will be the plug-in analog $\hat{\theta}_i = \ln(\hat{p}_{iw}) - \ln(\hat{p}_{ib})$. 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 sampling covariance matrix of the sample contact rates. Although $\hat{\theta}_i$ is not fully variance-stabilized, the log transform removes any direct dependence of the variance on θ_i itself.[5]

---

[5]Specifically, a second-order Taylor expansion of $\hat{p}_{iw}/\hat{p}_{ib} = \exp\left(\hat{\theta}_i\right)$ around the point (p_{iw}, p_{ib}) yields

In what follows, we exclude the eleven firms in the original experiment with callback rates below 3% or fewer than 40 total sampled jobs, since the estimated contact ratios for these firms may be unreliable. Summary statistics for the remaining estimation sample of 97 firms are provided in Table 2. The unweighted average value of $\hat{\theta}_i$ across these 97 firms is 0.095, implying the typical firm in our sample exhibits a bias against Black names of roughly 10%. Detailed point estimates and uncertainty measures for all 97 firms used in our analysis are provided in Appendix E2.

Twenty-one of the 97 estimated contact gaps are negative, indicating a preference for distinctively Black names. The firm-specific estimates are noisy, however, possessing an average standard error of 0.104. To test whether all firms in fact weakly prefer white to Black names (i.e., the joint null that $\theta_i \geq 0 \ \forall i \in [n]$) we apply the high dimensional inequality testing procedure of Bai, Santos and Shaikh (2021). This procedure yields a p-value of 0.94, suggesting the observed negative point estimates are likely attributable to chance.

Although the asymptotic variance of $\hat{\theta}_i$ does not mechanically depend on θ_i, it is possible for θ_i and s_i to be correlated. The top panel of Appendix Figure E6 plots $\hat{\theta}_i$ against s_i, revealing that firms with more precise estimates tend to show less bias against Black names. The Spearman correlation between between $\hat{\theta}_i$ and s_i is 0.36 ($p < 0.001$).

5.1 A model of precision dependence

In light of the above findings, we assume that each θ_i is non-negative and may depend (statistically) on its standard error s_i. A simple model satisfying these criteria is:

$$\theta_i = \exp\left(\beta \ln s_i + \ln v_i\right) = s_i^\beta v_i, \quad v_i \mid s_i \sim G_v \text{ for all } i \in [n]. \tag{8}$$

The parameter β governs how the conditional distribution of bias varies with the standard error s_i. The latent variable v_i captures heterogeneity in discriminatory conduct among firms with similar standard errors, and follows a distribution G_v with strictly positive support. When β is positive, both the mean and variance of θ_i increase monotonically with s_i. To complete the model, we link our estimates $\hat{\theta}_i$ to θ_i as follows:

$$\hat{\theta}_i = \theta_i + s_i e_i, \quad e_i \mid s_i, v_i \sim \mathcal{N}(0,1) \text{ for all } i \in [n],$$

where $s_i e_i$ is the noise in $\hat{\theta}_i$ attributable to the fact that a finite number of jobs were sampled from firm i.

To evaluate the plausibility of this model, we scrutinize some of the moment conditions

---

the approximation $\mathbb{V}\left[\hat{p}_{iw}/\hat{p}_{ib}\right] \approx \theta_i^2 \left\{ \frac{\mathbb{V}[\hat{p}_{iw}]}{p_{iw}^2} + \frac{\mathbb{V}[\hat{p}_{ib}]}{p_{ib}^2} - 2\frac{\mathbb{C}[\hat{p}_{iw},\hat{p}_{ib}]}{p_{iw}p_{ib}} \right\}$. Consequently, the Delta method implies that $\mathbb{V}\left[\hat{\theta}_i\right] \approx \frac{\mathbb{V}[\hat{p}_{iw}]}{p_{iw}^2} + \frac{\mathbb{V}[\hat{p}_{ib}]}{p_{ib}^2} - 2\frac{\mathbb{C}[\hat{p}_{iw},\hat{p}_{ib}]}{p_{iw}p_{ib}}$.

it implies. Letting $\mathbb{E}[v_i|s_i] = \mu_v > 0$ and $\mathbb{V}(v_i|s_i) = \sigma_v^2 > 0$, consider the following "studentized" version of $\hat{\theta}_i$:

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

Our model implies that T_i has mean zero and marginal variance one. Moreover, T_i should be independent of s_i. These restrictions imply the following four moment conditions:

$$\mathbb{E}[T_i] = 0, \ \mathbb{E}[T_i s_i] = 0, \ \mathbb{E}[T_i^2 - 1] = 0, \ \mathbb{E}[(T_i^2 - 1)s_i] = 0. \tag{9}$$

Imposing these moment conditions via two-step efficient GMM yields the parameter estimates reported in Table 3. The minimized value of the GMM criterion function suggests the model's over-identifying restrictions – which test the joint requirement that the T_i have mean zero and constant variance across all values of s_i – are satisfied ($p = 0.97$). The GMM estimate of β is $\hat{\beta} \approx 1/2$, indicating that θ_i is roughly proportional to $\sqrt{s_i}$. The large estimated value of σ_v reveals that discriminatory conduct varies substantially among firms with similar standard errors.

The top panel of Appendix Figure E6 superimposes the estimated conditional expectation function $\hat{\mathbb{E}}[\theta_i|s_i] = s_i^{\hat{\beta}}\hat{\mu}_v$ on the scatterplot of $\hat{\theta}_i$ against s_i. Consistent with the J-test from GMM estimation, the estimated conditional mean fits the cloud of points closely. The bottom panel of Appendix Figure E6 plots values of the estimated residual $\hat{T}_i = \dfrac{\hat{\theta}_i - s_i^{\hat{\beta}}\hat{\mu}_v}{\sqrt{s_i^{2\hat{\beta}}\hat{\sigma}_v^2 + s_i^2}}$ against s_i. Consistent with our model, $\hat{T}_i$ exhibits roughly constant variance and a mean near zero throughout the observed range of s_i.

5.2 Estimating G

To estimate the population distribution of contact penalties, we deconvolve the residual $\hat{v}_i = \hat{\theta}_i/s_i^{\hat{\beta}}$ which, by the Delta method, obeys

$$\hat{v}_i \mid v_i, s_i \sim \mathcal{N}\left(v_i, s_i^{2(1-\beta)}\right), \ \ v_i \mid s_i \sim G_v, \ \text{for all } i \in [n].$$

Relying again on a variant of Efron (2016)'s log-spline estimator, we parametrize G_v as a fifth order natural spline with strictly positive support. The spline parameters are estimated by penalized maximum likelihood with the penalty term chosen to minimize the distance to our earlier GMM estimates $(\hat{\mu}_v, \hat{\sigma}_v^2)$ of the first two moments of v_i. We then integrate over the empirical distribution of s_i to convert the estimated $\hat{G}_v$ into an estimate $\hat{G}_\theta$ of the distribution of contact gaps θ_i .

Figure 4 plots the log-spline estimate $\hat{G}_\theta$ overlaid against the histogram of contact gap estimates. $\hat{G}_\theta$ is less dispersed than the histogram, reflecting the noise in the estimates

$\{\hat{\theta}_i\}_{i=1}^n$. Unlike with our earlier analysis of names, the density $\hat{G}_\theta$ is unimodal but highly skewed. While most firms exhibit little bias against Black names, some exhibit large biases of 20-40%. By construction, no firms are estimated to discriminate against white names.

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 $(\theta_i, N_i s_i^2)$ under the assumption that θ_i is independent of N_i. The resulting marginal distribution of θ_i exhibits many mass points and is also unimodal. The NPMLE estimate of the variance of the θ_i's departs somewhat from both the log-spline estimate and a simple bias-corrected variance estimator $n^{-1} \sum_i [(\hat{\theta}_i - \bar{\theta})^2 - s_i^2]$. However, the NPMLE and log-spline estimates appear comparable in their overall shape. Since a discrete distribution with exact ties seems implausible, we rely again on the log-spline estimates in what follows. The EB posterior distribution of $\theta_i | Y_i$ inherits the continuity of the log-spline deconvolution estimate of the prior distribution, which has the added benefit of simplifying computation of posterior credible intervals for each θ_i.

5.3 Industry effects

In Kline, Rose and Walters (2022) 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 to incorporating industry affiliation into the model. We therefore employ a hierarchical random effects specification of v_i taking the form:

$$v_i = \eta_{k(i)} \xi_i,$$

$$\xi_i \mid s_i, \eta_{k(i)} \sim G_\xi, \quad i \in \{1, ..., n\},$$

$$\eta_k \mid \mathbf{s}_k \sim G_\eta, \quad k \in \{1, ..., K\},$$

where the function $k : \{1, \ldots, n\} \to \{1, \ldots, K\}$ returns a firm's industry and $\mathbf{s}_k$ is the vector of standard errors for all firms with $k(i) = k$. The industry effect $\eta_{k(i)}$ captures correlation in discriminatory conduct among firms in the same industry, while the firm effect ξ_i captures departures from the industry average. As a normalization we assume $\mathbb{E}[\eta_k] = 1$, which implies $\mathbb{E}[\xi_i] = \mu_v$.

The marginal variance of v_i in this model is $\sigma_v^2 = \sigma_\eta^2 \sigma_\xi^2 + \sigma_\eta^2 \mu_v^2 + \sigma_\xi^2$, where σ_ξ^2 gives the variance of ξ_i and σ_η^2 the variance of η_k. To separately identify the between and within industry variance components, we add two new moment conditions to the set listed in (9). Denote the average value of $\hat{v}_i$ in industry k by

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

where n_k gives the number of firms in industry k. The variance of $\bar{v}_k$ in this model can be shown to be $V_k \equiv \left(\sigma_\eta^2 \sigma_\xi^2/n_k + \sigma_\eta^2 \mu_v^2 + \sigma_\xi^2/n_k\right) + n_k^{-1} \sum_{i:k(i)=k} s_i^{2(1-\beta)}$. Letting $\bar{s}_k = n_k^{-1} \sum_{i:k(i)=k} s_i$ denote the average standard error in industry k, our two new moment conditions can be written

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

The first condition simply equates the empirical squared deviations of the $\bar{v}_k$ around the model implied mean to the model implied variance. The second condition prohibits heteroscedasticity with respect to $\bar{s}_k$.

GMM estimates of the parameters of this hierarchical model are reported in the second column of Table 3. The model's over-identifying restrictions again appear to be satisfied ($p = 0.95$). While the variance σ_η^2 of the industry component is estimated to be nearly 10 times as large as the variance σ_ξ^2 of the firm specific component, the multiplicative influence of these components on v_i implies that roughly half of the marginal variance in v_i stems from within industry variation.[6]

To identify the marginal distribution of θ_i, we assume that both G_η and G_ξ lie in the exponential family parameterized by a 5th order spline. Generalizing Efron (2016)'s log-spline estimator to the hierarchical case, these distributions are estimated by penalized maximum likelihood (see Appendix C for details). The two penalty parameters in this likelihood function are chosen so that the resulting distributions match GMM estimates of the between-industry and total variances of θ_i.

Estimates of G_ξ and G_η are displayed in the top panel of Figure 5. Table 4 reports moments of the within- and between-industry distributions implied by the log-spline estimates as well as moments of the overall contact ratio $\theta_i = s_i^\beta \eta_{k(i)} \xi_i$. Standard errors for the moments are computed via the Delta method, treating the fifth-order splines as correctly specified models for the log density functions. The mean contact gap, between-industry standard deviation, and total standard deviation reported in Table 4 closely match the corresponding GMM estimates of these parameters in Table 3.

As can be seen in Figure 5, the industry component η_k is more variable than the firm component ξ_i and exhibits positive skew and excess kurtosis, reflecting that some industries feature particularly heavy discrimination against Black names. Recall however that the location of the industry effect distribution is not informative as we have normalized $\mathbb{E}[\eta_k] = 1$. The bottom panel of Figure 5 shows that the implied distribution of θ_i is similar to the estimate from the model without industry effects in Figure 4, with a peak at small contact penalties and a long right tail. As expected, the deconvolved distribution is more compressed than the empirical distribution of estimated contact gaps.

---

[6]By the law of total variance $\mathbb{V}[v_i] = \mathbb{V}[\mathbb{E}[v_i|k(i)]] + \mathbb{E}[\mathbb{V}[v_i|k(i)]]$. Plugging in $v_i = \eta_{k(i)} \xi_i$ reveals that the within share $\mathbb{V}[\mathbb{E}[v_i|k(i)]]/\mathbb{V}[v_i]$ evaluates to $(\sigma_\eta^2 + 1)\sigma_\xi^2/\sigma_v^2$.

5.4 Reporting possibilities

Figure 6 plots the pairwise posterior ranking probabilities π_{ij} with firms ordered by their rank under $\lambda = 1$. Following our earlier convention with the names, these ranks range from 1 (the largest contact penalty) to 97 (the smallest contact penalty). 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 hierarchical log-spline model with industry effects. Because the firm assigned rank 1 is deemed most discriminatory, many other firms are more likely than not to have lower values of θ. Firms of middling rank, on the other hand, are more difficult to distinguish from others. Including industry effects tightens the posteriors, which leads the π_{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 $\pi_{ij} > (1+\lambda)^{-1}$ when λ has been set to 0.25. The resulting frontier implies numerous transitivity violations. For example, in panel (a), firm #9 cannot be distinguished from firm #4 or firm #49, suggesting each of these pairs in isolation would be labeled a tie. However, firm #49 is clearly distinguishable from firm #4, yielding a contradiction. Super-imposed on the figure we show a frontier corresponding to the three grades that solve (5) subject to (4) when $\lambda = 0.25$. These frontiers can be viewed as a transitivity-constrained version of the thresholding rule.

Figure 8 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 parameter λ. As expected, the number of grades tends to increase with λ as does the DR. In the absence of industry effects, setting $\lambda = 0.25$ yields three groups and an unconditional DR of roughly 3.9%. Introducing industry effects yields four groups and decreases the DR to 5.2%.

Figure 9 illustrates the empirical tradeoff between the information content of our grades, quantified by the expected rank correlation $\bar{\tau}$, and their reliability, as quantified by the Discordance Rate. Without industry effects, setting $\lambda = 1$ yields $\bar{\tau} = 0.46$ and a Discordance Rate of 0.27. Including industry effects increases the $\bar{\tau}$ of the Condorcet ranks to 0.51 and lowers their DR to 0.24. In contrast, ranking naively on $\hat{\theta}_i$ yields both a higher Discordance Rate and lower $\bar{\tau}$ than the Condorcet ranks, indicating such an approach is both less informative and less reliable. Interestingly, ranking based upon the EB posterior means yields a $\bar{\tau}$ and DR essentially equivalent to the Condorcet ranks.[7]

To improve the reliability of the Condorcet ranks, we set $\lambda = 0.25$. In the absence of transitivity violations, this choice of λ requires a posterior threshold of at least 80% to make pairwise ranking decisions. As depicted in Figure 7, however, numerous transitivity

---

[7]While ranking based upon posterior means is known to possess certain optimality properties when G is normal and the normal noise is homoscedastic (Portnoy, 1982), our environment features both heteroscedasticity and a decidedly non-normal mixing distribution $\hat{G}$.

violations emerge in this example. Resolving these violations raises the required posterior certainty above 80% in most instances, yielding a Discordance Rate of only 3.9% in the baseline specification without industry effects and 5.2% in the hierarchical specification with industry effects. Fortunately, the resulting grades remain highly informative: $\bar{\tau}$ is 0.21 in our baseline specification and 0.32 when industry effects are included.

6   Discrimination report cards

The report cards generated by our grading procedure provide a concise, low-dimensional summary of differences in discrimination across firms. This can be seen in Figure 10, which displays the report card results for the baseline specification without industry effects. In addition to the report card grades, the Figure plots a posterior mean estimate of each firm's bias θ_i along with 95% credible intervals, which are constructed by connecting the posterior 2.5th percentile of θ_i to the posterior 97.5th percentile. The lower limit of each credible interval is positive as a result of our support restriction ruling out bias against white applicants. The firms are ordered by their Condorcet ranks (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. Firms that are federal contractors, and hence subject to higher regulatory standards regarding equal opportunity laws, have been listed in black, while those that are not contractors are listed in gray. After the paper has undergone peer review, we intend to label the report card with firms' actual names.

Setting $\lambda = 0.25$ generates a report card with three grades, represented in Figure 10 by a number of $\star$'s between one (the worst grade) and three (the best). The shading of credible intervals reflects the grade assigned to each firm. Most firms receive the middle grade of $\star\star$, which reflects both the noise in our estimates and the shape of the estimated distribution $\hat{G}$. By contrast, only two firms out of 97 are assigned the grade of $\star$, suggesting they are the heaviest discriminators against Black names. Fourteen firms are assigned the score of $\star\star\star$, which indicates that this group is the least-biased against Black applicants.

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 sixth largest point estimate of bias (AKA firm #6) has a Condorcet rank of 1 and the largest posterior mean, while the firm with the largest bias point estimate (AKA firm #1) has a Condorcet rank of 2 and the second largest posterior mean. This rank reversal reflects that firm #6 has a larger standard error than firm #1, which leads the EB posterior mean to apply more shrinkage of the point estimate towards the center of the distribution.

Appendix Figure E7 depicts the relationship between report card grades and firm-

27

specific bias estimates and standard errors. Firms assigned the best grade of $\star\star\star$ tend to have both small contact gap estimates and standard errors, while firms assigned the grade $\star\star$ range widely in their standard errors but have modest contact gap estimates falling uniformly below 0.2. Firms assigned the worst grade of $\star$ exhibit very large contact gap estimates and widely varying standard errors. Appendix Figure E13 depicts the grade assignments that result from different choices of λ.

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. One of us (Kline, 2023) has recently cautioned against focusing excessively on rankings without also considering absolute standards of conduct. 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 $\star\star\star$ exhibit no bias against Black names.

As it turns out, however, the grades assigned by our procedure yield groups of firms with large cardinal differences in contact gaps. The firms assigned the grade of $\star\star\star$ have an average posterior mean value of θ_i of 0.03, while the two firms assigned the worst grade exhibit posterior means indicating a 24% penalty against Black names on average. Notably, both of these heavily discriminating firms are federal contractors.

Our past work (Kline, Rose and Walters, 2022) found that federal contractors, who are subject to monitoring by OFCCP for compliance with equal employment laws, tend to be substantially less biased against Black names on average, which is consistent with a variety of other evidence on the causal effects of affirmative action provisions on hiring behavior (e.g., McCrary, 2007; Kurtulus, 2016; Miller, 2017). Indeed, an early audit study of federal contractors by Newman (1978) found evidence of a systematic preference for Black over white applicants among such firms. It is somewhat surprising then that the Condorcet ranks indicate that the four most heavily discriminating firms are all federal contractors. This finding is, to some extent, a reflection of the fact that the vast majority of the firms in our sample of large employers are contractors (63 of 97). The mean Condorcet rank of federal contractors is 54 (with rank 1 showing the most bias against Black applicants) while the mean Condorcet rank of non-contractors is 42.

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.

Our estimates suggest the contact rates for fictitious applicants in our experiment may

have violated this standard.

6.1 Incorporating industry effects

As a result of the substantial variation in contact penalties across industries, a report card that incorporates industry information is substantially more informative. This can be seen in Figure 11, which displays discrimination report card results based on a model with industry effects. Adding industry information while maintaining the preference parameter λ at 0.25 yields a report card with four grades rather than three. The number of firms assigned the worst grade of $\star$ increases from two to five, while nine firms are now assigned the second-worst grade $\star\star$. Eleven firms are assigned the best grade $\star\star\star\star$. Appendix Figure E14 depicts the grade assignments that result from different choices of λ.

The average value of the posterior mean $\bar{\theta}_i$ among the firms assigned the grade $\star$ is 0.22, indicating an expected bias against Black names in this group of 22%. In contrast, the average value of $\bar{\theta}_i$ among the eleven firms assigned grade $\star\star\star\star$ is 0.03, suggesting a negligible effect of race on callback outcomes in this group. This finding indicates that many large firms are nearly unbiased, an important possibility result for companies seeking to improve the fairness of their recruiting process.

The small number of grades generated by our report card procedure explain a substantial portion of the total variation in discrimination across employers, especially when we incorporate industry. To summarize the explanatory power of the grades, we again utilize the grade-average posterior means as detailed in the Appendix. The variance estimate is weighted by the number of firms per grade, so that the ratio of between-grade to total variance has an R^2 interpretation. The estimated between-grade standard deviation in contact penalties is 0.034 for the three grades reported in Figure 10, implying an R^2 of roughly 19%. Adding industry boosts the R^2 to 47%. In other words, the four categories displayed in Figure 11 explain nearly half of the variance in discrimination across the 97 companies in our experiment.

Our ranking procedure allows us to grade the conduct of entire industries in addition to individual firms. Figure 12 plots posterior estimates $\mathbb{E}[\eta_k | Y = y] n_k^{-1} \sum_{i:k(i)=k} s_i^{\hat{\beta}}$ of industry mean contact penalties. The most biased industry is estimated to be SIC 55, "Auto dealers / services," with a posterior mean contact penalty of 25%. In an industry grading scheme with $\lambda = 0.25$ (which yields three total grades), SIC 55 is assigned its own unique grade of $\star$, indicating that this industry can be distinguished as the most biased in the experiment with high confidence. A group of six industries receive the best grade of $\star\star\star$, all of which have posterior mean contact gaps of roughly 5%. The role of common industry-level practices in generating the stark differences between these low- and high-performers is an interesting topic for further inquiry.

6.2   Grades under square-weighted loss

Appendix D details the firm rankings derived from minimizing our square-weighted notion of risk $\mathcal{R}^2(d; \lambda)$. Under this grading scheme, misrankings involving small differences in discriminatory conduct between firms yield negligible losses, which makes a more aggressive partitioning of the firms optimal. Choosing $\lambda = 0.25$ yields a six grade ranking without industry effects and an eight grade ranking when industry affiliation is taken into account. Because more grades are employed under square-weighting than under our default scheme, the variance in contact penalties explained by the grades is somewhat larger than for the grades reported in Figures 10 and 11. As explained in the Appendix, the square-weighted grades nevertheless maintain tight control over the expected probability of mistakes involving misrankings of firms with large absolute differences in discriminatory conduct.

6.3   Misclassification and Bayes Factors

In addition to conveying substantial information about discrimination, our report card system limits the ranking errors generated by the resulting classification. Figure 13 assesses the reliability of report card grades 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 11% of the firm comparisons across grades $\star$ and $\star\star$ are expected to be misordered. The Discordance Rate naturally declines when comparing non-adjacent grades. The expected share of misordered comparisons across grades $\star$ and $\star\star\star$ is below 1%. Adjacent grades have DR's between 11 and 14%, implying Bayes Factors between 7 and 6, respectively. The discordance rate for non-adjacent grades of 0.8% implies a Bayes Factor exceeding 11.

Our preferred report card with industry effects limits misrankings between firms selected as high- and low-performers. Panel (b) of Figure 13 summarizes the reliability of the grades obtained when conditioning on industry effects. Discordance Rates between adjacent grades range from 15% to 17%. DRs for grades separated by two categories are less than 4%, and the DR between the worst grade ($\star$) and the best grade ($\star\star\star\star$) is 0.4%. Combined with the results from Section 6.1, this implies that a comparison of the best- and worst-performers in Figure 11 isolates firms with large differences in discriminatory conduct while yielding a negligible chance of misclassification.

7   Conclusion

We have proposed a new Empirical Bayes method for ranking units based upon noisy measurements and used it to grade the discriminatory conduct of firms towards distinctively Black names in a large-scale correspondence experiment. The experiment is shown

to contain a wealth of information about the relative conduct of firms: our most granular (Condorcet) grades taking into account industry affiliation yield an expected correlation with the true firm ranks of 0.51. These grades are noisy, however, resulting in (expected) mistakes in nearly one quarter of the $\binom{97}{2} = 4,656$ possible pairwise firm comparisons.

A generalization of the Condorcet scheme based on a desired 80% posterior certainty threshold for pairwise contrasts yields a report card with only four grades. These coarse grades turn out to be substantially more reliable than the Condorcet ranks, lowering the chances that a randomly sampled pair of firms is misordered to 5.2%. The four grades are also highly informative, offering an expected correlation with the true firm ranks of 0.32. In addition to conveying information about the ranking of firm conduct, the grades capture important differences in conduct levels. Firms assigned the worst grade favor white applicants over Black applicants by 23% while those assigned the best grade favor white applicants by only 3%.

The finding of negligible contact gaps in a large group of firms provides a possibility result for employers seeking to improve the fairness of their hiring processes. Recent research points towards centralization of hiring processes as a possible means of dampening bias in large organizations (Berson, Laouenan and Valat, 2020; Challe et al., 2022), a conjecture that aligns with findings in behavioral economics that snap judgments by individuals are especially susceptible to bias (e.g., Agan et al., 2023). Further corroboration of this view comes from Miller (2017)'s finding that temporary exposure to the heightened scrutiny over HR practices accompanying federal contractor status has persistent effects on the composition of firm hires.

We plan to eventually release the information in this report card to the public along with the identities of the companies in the experiment. Our hope is that this information will prompt corporate leaders to rethink whether their inclusive hiring goals are being met. Much work remains to establish which sorts of reforms to organizational practices can improve the fairness and efficiency of corporate recruiting efforts. Releasing these data for use by other researchers will hopefully accelerate the pace of research into strategies for subduing hiring discrimination.

References

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc. 2021. "Mortality Effects and Choice Across Private Health Insurance Plans." *The Quarterly Journal of Economics*, 136(3): 1557–1610.

Agan, Amanda Y, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias." National Bureau of Economic Research.

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

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey. 2019. "Inference on Winners." *NBER Working Paper*, , (w25456).

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging lotteries for school value-Added: Testing and estimation." *The Quarterly Journal of Economics*, 132(2): 871–919.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2021. "Race and the mismeasure of school quality." *NBER Working Paper*, , (w29608).

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, MS. 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.

Bergman, Peter, and Matthew J. Hill. 2018. "The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers." *Economics of Education Review*, 66: 104–113.

Bergman, Peter, Eric W. Chan, and Adam Kapor. 2020. "Housing search frictions: evidence from detailed search data and a field experiment." *NBER Working Paper*, , (w227209).

Berson, Clémence, Morgane Laouenan, and Emmanuel Valat. 2020. "Outsourcing recruitment as a solution to prevent discrimination: A correspondence study." *Labour Economics*, 64: 101838.

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.

Brook, Robert H., Elizabeth A. McGlynn, Paul G. Shekelle, Martin Marshall, Sheila Leatherman, John L. Adams, Jennifer Hicks, and David J. Klein. 2002. *Report Cards for Health Care: Is Anyone Checking Them?* Santa Monica, CA:RAND Corporation.

Challe, Laetitia, Sylvain Chareyron, Yannick L'horty, and Pascale Petit. 2022. "The effect of pro diversity actions on discrimination in the recruitment of large companies: a field experiment." *TEPP working paper # 2022-18*.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson. 2016. "Health care exceptionalism? Performance and allocation in the US health care sector." *American Economic Review*, 106(8): 2110–44.

Chetty, Raj, and Nathaniel Hendren. 2018*a*. "The impacts of neighborhoods on intergenerational mobility II: County-level estimates." *The Quarterly Journal of Economics*, 133(3): 1163–1228.

Chetty, Raj, and Nathaniel Hendren. 2018*b*. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*." *The Quarterly Journal of Economics*, 133(3): 1163–1228.

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, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. "Mobility report cards: the role of colleges in intergenerational mobility." *NBER Working Paper*, , (w23618).

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie Jones, and Sonya Porter. 2018*a*. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." *NBER Working Paper*, , (w25147).

Chetty, Raj, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. 2018*b*. "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.*

Crabtree, Charles, S Michael Gaddis, John B Holbein, and Edvard Nergård Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." *Sociological Science*, 9: 454–472.

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

Fryer Jr, Roland G, and Steven D Levitt. 2004. "The causes and consequences of distinctively black names." *The Quarterly Journal of Economics*, 119(3): 767–805.

Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies." *Sociological Science*, 4(19): 469–489.

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. 2023. "A COMMENT ON:"Invidious Comparisons: Ranking and Selection as Compound Decisions" by Jiaying Gu and Roger Koenker." *Econometrica*, 91(1): 47–52.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic discrimination among large US employers." *The Quarterly Journal of Economics*, 137(4): 1963–2036.

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.

Kolstad, Jonathan T. 2013. "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards." *American Economic Review*, 103(7): 2875–2910.

Kurtulus, Fidan Ana. 2016. "The impact of affirmative action on the employment of minorities and women: a longitudinal analysis using three decades of EEO-1 filings." *Journal of Policy Analysis and Management*, 35(1): 34–66.

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

Maxwell, Nan, Aravind Moorthy, Caroline Massad Francis, Dylan Ellis, et al. 2013. "Using Administrative Data to Address Federal Contractor Violations of Equal Employment Opportunity Laws." Mathematica Policy Research.

McCrary, Justin. 2007. "The effect of court-ordered hiring quotas on the composition and quality of police." *American Economic Review*, 97(1): 318–353.

Miller, Conrad. 2017. "The persistent effect of temporary affirmative action." *American Economic Journal: Applied Economics*, 9(3): 152–90.

Mogstad, Magne, Joseph Romano, Azeem Shaikh, and Daniel Wilhelm. 2020. "Inference for Ranks with Applications to Mobility across Neighborhoods and Academic Achievement across Countries." *NBER Working Paper*, , (w26883).

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.

Newman, Jerry M. 1978. "Discrimination in recruitment: An empirical analysis." *ILR Review*, 32(1): 15–23.

Pope, Devin G. 2009. "Reacting to rankings: Evidence from "America's Best Hospitals"." *Journal of Health Economics*, 28(6): 1154–1165.

Pope, Nolan G. 2019. "The effect of teacher ratings on teacher performance." *Journal of Public Economics*, 172: 84–110.

Portnoy, Stephen. 1982. "Maximizing the probability of correctly ordering random variables using linear predictors." *Journal of Multivariate Analysis*, 12(2): 256–269.

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.

Smith, John H. 1973. "Aggregation of preferences with variable electorate." *Econometrica: Journal of the Econometric Society*, 1027–1041.

Sobel, Marc J. 1990. "Complete ranking procedures with appropriate loss functions." *Communications in Statistics-Theory and Methods*, 19(12): 4525–4544.

Sobel, Marc J. 1993. "Bayes and empirical Bayes procedures for comparing parameters." *Journal of the American Statistical Association*, 88(422): 687–693.

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.

Figures

Figure 1: Deconvolutions of name-specific contact rate estimates

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

Bias-corrected SD: 0.0119
Decon. implied SD: 0.0125
NPMLE implied SD: 0.0125
Mean: 0.5120
Decon. implied mean: 0.5115
NPMLE implied mean 0.5115

b) Contact rates (p_i)

Bias-corrected SD: 0.0102
Decon. implied SD: 0.0106
NPMLE implied SD: 0.0106
Mean: 0.2401
Decon. implied mean: 0.2397
NPMLE implied mean 0.2397

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 hollow blue histogram shows the distribution of estimated variance-stabilized contact rates. The red line shows a deconvolution estimate of the population contact rate distribution. The deconvolution procedure parameterizes the log-density with a fifth-order spline, and the parameters are estimated by penalized maximum likelihood, with penalization parameter chosen to match the mean and bias-corrected variance estimate as closely as possible. The dark green mass points plot the distribution of population contact rates estimated by non-parametric maximum likelihood (NPMLE). The vertical dashed lines plot mean contact rates for each race and gender group of names. Panel (b) converts the estimated distributions of variance-stabilized contact rates into distributions of contact rates p_i.

Figure 2: Name ranking exercises

a) Pairwise posterior contrasts

b) Grades and discordance

c) Reporting possibilities

Notes: This figure summarizes the results from grading contact rates for names. Panel (a) shows pairwise posterior ordering probabilities for all names. Posteriors are computed using the log-spline estimate plotted in Figure 1 as the prior. Names are ordered by their rank under $\lambda = 1$. Shading indicates the posterior probability that the contact rate for the name on the vertical axis exceeds the contact rate for the name on the horizontal axis. Panel (b) shows estimated Discordance Rates (DR) for an intermediate range of λ. Panel (c) plots the expectation of Kendall's τ rank correlation between true contact rates and grades against Discordance Rates (DR) for a range of grades indexed by λ. The red circle highlights the DR and expected τ corresponding to $\lambda = 0.25$. "$\hat{\theta}$ rank" refers to ranks based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon Empirical Bayes posterior means.

Figure 3: Posterior means and grades of first names

Notes: This figure shows posterior mean contact rates, 95% credible intervals, and assigned grades for names. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Names are ordered by their rank under $\lambda = 1$, when each name is assigned its own grade.

Figure 4: Deconvolution estimates of firm-level contact penalty distribution

Bias-corrected SD: 0.069
Decon. implied SD: 0.077
NPMLE implied SD: 0.112
Mean: 0.095
Decon. implied mean: 0.097
NPMLE implied mean: 0.094

Notes: This figure presents non-parametric estimates of the distribution of firm-specific contact penalties. The blue histogram shows the distribution of estimated proportional contact penalties. The red line shows a log-spline deconvolution estimate of the population contact penalty distribution. The dark green mass points plot a non-parametric maximum likelihood (NPMLE) estimate of the population contact penalty distribution. The bias-corrected standard deviation estimate is computed by subtracting the average squared standard error from the sample variance of estimated contact penalties, then taking the square root.

Figure 5: Deconvolution estimates of between- and within-industry contact penalty distributions

a) Standardized contact penalty distribution

b) Marginal contact penalty distribution

Bias-corrected SD: 0.069
Decon. implied SD: 0.072
Mean: 0.095
Decon. implied mean: 0.086

Notes: This figure presents estimates of the hierarchical random effects model described in Section 5.3. Panel (a) presents the estimated within- and between- industry distributions of standardized contact penalties. The blue histogram shows the distribution of $\hat{\eta}_k$, computed as the industry mean of $\hat{v}_i/\hat{\mu}_v$, where $\hat{v}_i = \hat{\theta}_i/s_i^{\hat{\beta}}$ and $\hat{\mu}_v$ and $\hat{\beta}$ are GMM estimates from Table 3. The red histogram displays the distribution of $\hat{\xi}_i = \hat{v}_i/\hat{\eta}_{k(i)}$. Blue and red densities show corresponding log-spline deconvolution estimates of the distributions of η_k and ξ_i. The deconvolution procedure parameterizes the log-density of each distribution with a fifth-order spline, and the parameters are estimated by penalized maximum likelihood, with penalization parameter chosen to match the GMM mean and variance estimates from Table 3 as closely as possible. Panel (b) shows the marginal distribution of contact penalties θ_i implied by the estimates from panel (a) along with a histogram of estimated contact penalties $\hat{\theta}_i$. See Appendix C for complete details.

Figure 6: Posterior contrasts

a) Baseline

b) Industry effects

Notes: This figure plots pairwise posterior ordering probabilities for firm-specific contact penalties. Posteriors are computed using the log-spline estimates from Figure 4 as the prior distributions and Firms are ordered by their ranks under $\lambda = 1$. The rank implying the largest θ_i is denoted by 1. Shading indicates the posterior probability that the contact penalty for the firm on the vertical axis exceeds the contact penalty for the firm on the horizontal axis.

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 optimal decision would rank the firm on the horizontal axis below the firm on the vertical axis. Both panels use $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. The black lines denote the boundaries of optimal grades for this λ 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 λ

Notes: This figure shows estimated Discordance Rates (DR) as a function of λ. The number on each point indicates the number of unique grades in the underlying grading scheme. The vertical dashed line shows results for the benchmark case of $\lambda = 0.25$.

Figure 9: Reporting possibilities

Notes: This figure shows the expectation of Kendall's τ rank correlation between θ and assigned grades (labeled $\bar{\tau}$) against Discordance Rates (DR) for a range of grades indexed by λ. Red circles highlight the DR and $\bar{\tau}$ corresponding to $\lambda = 0.25$. "$\hat{\theta}$ rank" refers to ranks based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon Empirical Bayes posterior means.

Figure 10: Posterior means and grades of firms

Notes: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. 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. Firms labeled with black text are federal contractors, whereas firms in gray are not.

Figure 11: Posterior means and grades of firms (Industry effects model)

Notes: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades from the industry random effect model. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. 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. Firms labeled with black text are federal contractors, whereas firms in gray are not.

Figure 12: Posterior means and grades of industries

Notes: This figure shows posterior means, 95% credible intervals, and assigned grades for industry mean proportional contact penalties. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Each industry is labeled by its name and two-digit SIC code.

Figure 13: DR in baseline and industry effects model

a) Baseline

b) Industry effects

Notes: This figure shows mean Discordance Rates (DR) across grade pairs for the baseline model and the model with industry effects. In both panels, $\text{DR}_{g,g'}$ is the expected share of pairwise comparisons between firms in grades g and g' where the grade rank differs from the latent firm rank. The upper-left most estimate in panel a, for example, indicates a 11% chance that a random firm in grade ★'s true rank is below a randomly chosen firm's in grade ★★'s. 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	Wald test of heterogeneity
Male				
Black	0.233	20,927	19	12.6
	(0.003)			[0.82]
White	0.246	20,975	19	15.8
	(0.003)			[0.61]
Female				
Black	0.226	20,879	19	21.2
	(0.003)			[0.24]
White	0.254	20,862	19	19.9
	(0.003)			[0.34]
Estimated contact rate SD				
Total	0.010			
Between race/sex	0.011			

Notes: This table presents summary statistics for 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 assigned to each application was unconditionally randomized. The final column reports Wald tests for equality of contact probabilities across the first names in each demographic group. Under the null hypothesis of equal contact probabilities, each test statistic is distributed $\chi^2(18)$. Corresponding *p*-values are reported in brackets. The estimated contact rate SD is a bias-corrected estimate of the standard deviation of name-specific contact rates, computed by subtracting the average squared standard error from the sample variance of contact rate estimates then taking the square root. The between race/sex standard deviation is a corresponding bias-corrected estimate of the variation in mean contact rates across race and sex groups. See Appendix Table E1 for a list of first names used in the analysis.

Table 2: Summary statistics for firm sample

	# Firms	# Jobs	# Apps	Contact rates and gaps				
				White	Black	Difference	Log dif	Mean SE
All	97	10,453	78,910	0.256 (0.004)	0.236 (0.003)	0.020 (0.002)	0.095 (0.013)	0.104
2-digit SIC industry (code)								
Food products (20)	1	100	788	0.435 (0.041)	0.440 (0.040)	-0.005 (0.019)	-0.011 (0.045)	0.045
Apparel manufacturing (23)	2	200	1,538	0.205 (0.026)	0.175 (0.023)	0.031 (0.012)	0.177 (0.062)	0.088
Other manufacturing (24)	4	375	2,904	0.119 (0.012)	0.104 (0.012)	0.015 (0.009)	0.179 (0.116)	0.211
Freight / transport (42)	4	458	3,300	0.194 (0.017)	0.197 (0.017)	-0.003 (0.008)	-0.014 (0.039)	0.076
Communications (48)	2	175	1,124	0.273 (0.036)	0.225 (0.033)	0.048 (0.021)	0.163 (0.086)	0.120
Electric / gas (49)	3	320	2,419	0.261 (0.021)	0.247 (0.020)	0.014 (0.010)	0.120 (0.060)	0.094
Wholesale durable (50)	2	152	1,143	0.194 (0.028)	0.177 (0.027)	0.017 (0.011)	0.088 (0.057)	0.081
Wholesale nondurable (51)	11	1,117	8,194	0.299 (0.011)	0.288 (0.011)	0.011 (0.007)	0.092 (0.032)	0.091
Building materials (52)	3	377	2,755	0.297 (0.019)	0.285 (0.019)	0.012 (0.008)	0.024 (0.039)	0.062
General merchandise (53)	12	1,380	10,440	0.320 (0.010)	0.292 (0.010)	0.028 (0.006)	0.108 (0.030)	0.083
Food stores (54)	5	530	4,030	0.451 (0.018)	0.425 (0.018)	0.026 (0.010)	0.063 (0.029)	0.058
Auto dealers / services (55)	8	891	6,930	0.257 (0.012)	0.204 (0.011)	0.053 (0.008)	0.237 (0.040)	0.107
Apparel stores (56)	4	400	3,093	0.237 (0.017)	0.202 (0.016)	0.035 (0.010)	0.173 (0.067)	0.117
Furnishing stores (57)	4	482	3,679	0.286 (0.018)	0.251 (0.017)	0.035 (0.011)	0.131 (0.045)	0.086
Eating/drinking (58)	4	500	4,000	0.368 (0.018)	0.337 (0.018)	0.032 (0.009)	0.086 (0.027)	0.053
Other retail (59)	7	816	6,281	0.206 (0.011)	0.182 (0.011)	0.024 (0.007)	0.133 (0.060)	0.138
Banks / credit (60)	2	252	1,947	0.119 (0.015)	0.121 (0.016)	-0.002 (0.010)	-0.073 (0.116)	0.150
Securities brokers (62)	1	125	965	0.122 (0.021)	0.111 (0.019)	0.011 (0.012)	0.098 (0.102)	0.102
Insurance / real estate (63)	5	398	2,907	0.142 (0.015)	0.142 (0.016)	0.000 (0.010)	0.015 (0.108)	0.203
Accommodation (70)	2	243	1,850	0.200 (0.022)	0.199 (0.023)	0.001 (0.012)	0.043 (0.068)	0.094
Business services (73)	3	375	2,812	0.214 (0.017)	0.212 (0.017)	0.003 (0.007)	0.101 (0.076)	0.113
Auto / repair services (75)	3	340	2,551	0.285 (0.022)	0.275 (0.022)	0.010 (0.010)	0.046 (0.037)	0.062
Health services (80)	4	400	2,886	0.150 (0.015)	0.144 (0.015)	0.006 (0.008)	-0.071 (0.067)	0.127
Engineering services (87)	1	47	374	0.122 (0.047)	0.117 (0.046)	0.005 (0.005)	0.044 (0.042)	0.042

Notes: This table presents summary statistics the sample of applications used in the analysis of firm contact penalties. "White" and "Black" refer to average firm-level contact rates for white and Black applications. Difference is the average difference. Log dif is the industry average of the primary contact penalty measure used in the analysis: $\ln(\hat{p}_{iw}) - \ln(\hat{p}_{ib})$. Mean SE is the average standard error of firm-level log difference estimate. Standard errors in parentheses.

Table 3: GMM estimates of contact penalty parameters

	No industry effects (1)	With industry effects (2)
β	0.510 (0.190)	0.517 (0.121)
μ_v	0.313 (0.074)	0.292 (0.074)
σ_v	0.207 (0.106)	
σ_η		0.452 (0.171)
σ_ξ		0.144 (0.066)
Within share		0.556
J-statistic (d.f.)	0.101 (1)	0.111 (2)

Notes: This table reports generalized method of moments (GMM) estimates of the parameters of the model $\theta_i = s_i^\beta v_i$, with $\mathbb{E}[v_i] = \mu_v$ and $\mathbb{V}[v_i] = \sigma_v^2$. Column (2) allows an industry component of the form $v_i = \eta_{k(i)}\xi_i$, where $k(i)$ is the industry of firm i and $\mathbb{E}[\eta_k] = 1$. Estimates come from two-step optimally-weighted GMM with an identity weighting matrix in the first step. The variance matrix in column (2) is clustered by industry. The within share is $\frac{\mathbb{E}[\mathbb{V}[v_i|k(i)]]}{\mathbb{V}[v_i]} = \frac{(\sigma_\eta^2+1)\sigma_\xi^2}{\sigma_\eta^2\sigma_\xi^2+\sigma_\eta^2\mu_v^2+\sigma_\xi^2}$.

Table 4: Moments of random effect distributions

	Industry effect (η_k) (1)	Firm effect (ξ_i) (2)	Contact penalty (θ_i) (3)
Mean	1.000	0.292	0.086
	-	(0.040)	(0.012)
Std. dev.	0.495	0.138	0.072
	(0.192)	(0.049)	(0.019)
Skewness	1.856	2.293	3.545
	(1.824)	(4.580)	(2.778)
Excess kurtosis	7.735	30.067	43.653
	(6.740)	(52.585)	(70.467)

Notes: This table reports estimated moments of the distributions of industry and firm effects along with moments of the marginal distribution of contact penalties. Results are derived from hierarchical log-spline deconvolution estimates, with spline parameters estimated by penalized maximum likelihood. Standard errors are computed by the Delta method, with the variance matrix for the spline parameters computed from the negative inverse Hessian of the log likelihood function.

Appendix

(for online publication only)

Appendix A   Proofs of propositions

This section provides proofs of the propositions discussed in Section 3.5, which are re-stated here for completeness.

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

Proof. First, we establish that no firm can be tied with firm i. Suppose $\exists \ j \ s.t. \ d_j = d_i = d$. Let $\tilde{d} = \inf\{\{d' \in d^*(\lambda) \ s.t. \ d' > d\} \cup \{\infty\}\}$. Then changing firm i's grade to a value in $(d, \tilde{d})$ yields strictly lower loss, because $\sum_{j \neq i \ s.t. \ d_j = d} \pi_{ji} - \lambda\pi_{ij} < 0$, and comparisons between i and all other firms $j \ s.t. \ d_j \neq d$ are unaffected.

Now suppose $\exists \ d \in d^*(\lambda) \ s.t. \ d > d_i$. Let $d' = \inf\{d \in d^*(\lambda) \ s.t. \ d > d_i\}$. Then $\forall j \ s.t. \ d_j = d'$, the risk of re-assigning $d_i = d' + \epsilon < \inf\{\{d \in d^*(\lambda) \ s.t. \ d > d'\} \cup \{\infty\}\}$ is strictly lower because $\sum_{j \neq i \ s.t. \ d_j = d'} \pi_{ji} - \lambda\pi_{ij} < 0 < \sum_{j \neq i \ s.t. \ d_j = d'} \pi_{ij} - \lambda\pi_{ji}$, and comparisons between i and all other firms $j \ s.t. \ d_j \neq d'$ are unaffected. Since the same argument applies to firm k removing firm i from set of firms under consideration, the proof of the second part of the claim is identical. $\square$

Proposition 2 (λ-Smith criterion)**.** Let $\mathcal{S}$ denote a collection of firms exhibiting the following dominance property: $\pi_{ij} > (1 + \lambda)^{-1} \ \forall i \in \mathcal{S}, j \notin \mathcal{S}$. Then the top graded firms must be a member of $\mathcal{S}$.

Proof. First, note that if $\mathcal{S}$ is a singleton, then Proposition 1 applies directly. Otherwise, let $\tilde{d} = \sup\{d_i \ s.t. \ i \in \mathcal{S}\}$ and let $\bar{\mathcal{S}}$ denote the set $\{i \in \mathcal{S} \ s.t. \ d_i = \tilde{d}\}$. Suppose $\exists \ j \notin \mathcal{S} \ s.t. \ d_j > \tilde{d}$. Let $d' = \inf\{d \in d^*(\lambda) \ s.t. \ d > \tilde{d}\}$ and $\underline{\mathcal{S}}$ denote the set $\{j \notin \mathcal{S} \ s.t. \ d_j = d'\}$. Then swapping grades such that all firms in $\bar{\mathcal{S}}$ receive grade d' and all firms in $\underline{\mathcal{S}}$ receive grade $\tilde{d}$ must decrease risk, because $\sum_{i \in \bar{\mathcal{S}}} \sum_{j \in \underline{\mathcal{S}}} \pi_{ji} - \lambda\pi_{ij} < 0 < \sum_{i \in \bar{\mathcal{S}}} \sum_{j \in \underline{\mathcal{S}}} \pi_{ij} - \lambda\pi_{ji}$, comparisons between all firms within $\bar{\mathcal{S}}$ and $\underline{\mathcal{S}}$ are unaffected, and comparisons between all firms $k \notin \{\bar{\mathcal{S}} \cup \underline{\mathcal{S}}\}$ are unaffected. Thus no firm $j \notin \mathcal{S}$ may be ranked above the top graded member of $\mathcal{S}$. $\square$

Proposition 3 (Unordered λ-Smith candidates are tied)**.** Let $\mathcal{S}$ denote a collection of firms exhibiting the following dominance property: $\pi_{ij} > (1 + \lambda)^{-1} \ \forall i \in \mathcal{S}, j \notin \mathcal{S}$. Moreover, suppose $\pi_{ij} < (1 + \lambda)^{-1} \ \forall (i, j) \in \mathcal{S}$. Then all firms in $\mathcal{S}$ receive the highest grade.

Proof. First, we show that all firms $j \notin \mathcal{S}$ must be ranked below every member of $\mathcal{S}$. Suppose not. Let $d' = \inf\{d_j \ s.t. \ j \notin \mathcal{S}, \exists \ i \in \mathcal{S} \ s.t. \ d_j > d_i\}$, $\underline{\mathcal{S}} = \{j \notin \mathcal{S} \ s.t. \ d_j = d'\}$, $\tilde{d} = \sup\{d_i \ s.t. \ i \in \mathcal{S}, d_i < d'\}$, $\bar{\mathcal{S}} = \{i \in \mathcal{S} \ s.t. \ d_i = \tilde{d}\}$. Then setting grades so that all firms in $\underline{\mathcal{S}}$ receive a grade $m \in (d', \tilde{d})$ and all firms in $\bar{\mathcal{S}}$ receive grade d' must decrease

risk because $\sum_{i \in \bar{\mathcal{S}}} \sum_{j \in \underline{\mathcal{S}}} \pi_{ji} - \lambda \pi_{ij} < 0 < \sum_{i \in \bar{\mathcal{S}}} \sum_{j \in \underline{\mathcal{S}}} \pi_{ij} - \lambda \pi_{ji}$, implying it is optimal to rank all firms in $\bar{\mathcal{S}}$ above those in $\underline{\mathcal{S}}$. Moreover, $\sum_{i \in \bar{\mathcal{S}}} \sum_{j \in \mathcal{S} \ s.t. \ d_j = d'} \pi_{ji} - \lambda \pi_{ij} > 0$ implies that it is optimal to tie firms in $\bar{\mathcal{S}}$ with firms in $\mathcal{S}$ that already have grade d', while $\sum_{j \in \underline{\mathcal{S}}} \sum_{i \in \mathcal{S} \ s.t. \ d_i = d'} \pi_{ji} - \lambda \pi_{ij} < 0$ implies it is optimal to rank any firms in $\mathcal{S}$ that already have grade d' above those firms $\notin \mathcal{S}$ reassigned to grade m, and $\sum_{i \in \bar{\mathcal{S}}} \sum_{j \notin \mathcal{S} \ s.t. \ d_j = \tilde{d}} \pi_{ji} - \lambda \pi_{ij} < 0$ implies it is optimal to rank firms in $\bar{\mathcal{S}}$ above any firms $\notin \mathcal{S}$ that currently have grade $\tilde{d}$. Comparisons to all firms with grades higher than d' are unaffected, as well as to any firms with grades below $\tilde{d}$. The top grades thus consist exclusively of firms in $\mathcal{S}$. To see that they also must be tied, note that because $\pi_{ji} - \lambda \pi_{ij} > 0 \ \forall (i, j) \in \mathcal{S}$, collapsing any two adjacent grades for firms in $\mathcal{S}$ must decrease risk. $\qquad \square$

Appendix B Computing posteriors

This appendix details computation of posterior distributions for the firm contact gap analysis of Section 5. Computation of posteriors for the name contact rate analysis of Section 4 is a special case of this framework setting the dependence parameter β to zero and the standard error s_i for name i to $(4N_i)^{-1}$. Under the model in (8), the posterior density for $v_i = \theta_i/s_i^\beta$ is given by

$$f_v(x|Y_i; G_v, \beta) = \frac{\mathcal{L}(\hat\theta_i|v_i = x, s_i; \beta)dG_v(x)}{\int \mathcal{L}(\hat\theta_i|v_i = u, s_i; \beta)dG_v(u)},$$

$$\mathcal{L}(\hat\theta_i|v_i = x, s_i; \beta) = \frac{1}{s_i^{1-\beta}}\phi\left(\frac{(\hat\theta_i/s_i^\beta) - x}{s_i^{1-\beta}}\right).$$

Taking $\hat{G}_v$ as a deconvolution estimate of G_v and $\hat\beta$ as a GMM estimate of β, posterior means for θ_i are computed as $s_i^{\hat\beta} \times \int x f_v(x|Y_i; \hat{G}_v, \hat\beta)dx$, 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 $F_\theta(t|Y_i; \hat{G}_v, \hat\beta) = \int_{-\infty}^{t/s_i^{\hat\beta}} f_v(x|Y_i; \hat{G}_v, \hat\beta)dx$.

We also use $\hat{G}_v$ and $\hat\beta$ to compute the matrix of pairwise posterior ranking probabilities π_{ij}. We have:

$$\begin{aligned}
\pi_{ij} &= \Pr(\theta_i > \theta_j|Y_i, Y_j; G_v, \beta) \\
&= \Pr((s_i/s_j)^\beta v_i > v_j|Y_i, Y_j; G_v, \beta) \\
&= \int_{-\infty}^\infty \int_{-\infty}^{(s_i/s_j)^\beta x} f_v(x|Y_i; G_v, \beta)f_v(u|Y_j; G_v, \beta)dudx.
\end{aligned}$$

Posterior moments of 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:

$$\mathbb{E}\left[\max\{(\theta_i - \theta_j), 0\}^p \mid Y_i, Y_j\right] = \int_{-\infty}^\infty \int_{-\infty}^{(s_i/s_j)^\beta x} (s_i^\beta x - s_j^\beta u)^p f_v(x|Y_i; G_v, \beta)f_v(u|Y_j; G_v, \beta)dudx.$$

We plug $\hat{G}_v$ and $\hat\beta$ into these formulas to construct empirical Bayes posteriors by numerical integration, then solve the linear programming problem to minimize either binary or weighted risk using Gurobi.

B.1 Industry effects

Posteriors for the hierarchical industry effects model of Section 5.3 condition on the data for all firms in an industry. Let $\mathbf{Y}_k$ denote the $2n_k \times 1$ vector of estimates $\hat\theta_i$ and standard errors s_i for all firms in industry k, and let $\boldsymbol{\xi}_k$ denote the $n_k \times 1$ vector of within-industry deviations ξ_i for all firms in this industry. The joint posterior density for $\boldsymbol{\xi}_k$ and the

industry effect η_k at the point where $\eta_k = x$ and $\xi_k = \mathbf{z} = (z_1, ..., z_{n_k})'$ is given by:

$$f_{\eta,\boldsymbol{\xi}}(x, \mathbf{z}|\mathbf{Y}_k; G_\eta, G_\xi, \beta) = \frac{\left[\prod_{i:k(i)=k} \mathcal{L}(\hat{\theta}_i|v_i = x \times z_i, s_i; \beta) dG_\xi(z_i)\right] dG_\eta(x)}{\int_u \int_{\mathbf{t}} \left[\prod_{i:k(i)=k} \mathcal{L}(\hat{\theta}_i|v_i = u \times t_i, s_i; \beta) dG_\xi(t_i)\right] dG_\eta(u)}.$$

We form Empirical Bayes joint posteriors given by $f_{\eta,\boldsymbol{\xi}}(x, \mathbf{z}|\mathbf{Y}_k; \hat{G}_\eta, \hat{G}_\xi, \hat{\beta})$, where $\hat{\beta}$ is the GMM estimate of β from column (2) of Table 3, and $\hat{G}_\xi$ and $\hat{G}_\eta$ are hierarchical deconvolution estimates from panel (a) of Figure 5. We then integrate over these joint posteriors by simulation to compute posterior means and quantiles for each random effect along with pairwise posterior probabilities π_{ij} for the model with industry effects.

B.2 Between grade variance

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

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

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 $w_g = n^{-1} \sum_{i=1}^n D_{ig}$ 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 $\mathbb{E}[\theta_i|Y]$ within grade. The posterior mean estimate of each $\bar{\theta}_g^2$ is slightly harder to compute because

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

$$= (nw_g)^{-2} \left\{\sum_{i=1}^n D_{ig} \theta_i^2 + \sum_{i=1}^n \sum_{i' \neq i} D_{ig} D_{i'g} \theta_i \theta_{i'}\right\}.$$

Our posterior mean estimate of this quantity is computed analogously as

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

$$= (nw_g)^{-2} \left\{\sum_{i=1}^n D_{ig} \mathbb{E}\left[\theta_i^2|Y\right] + \left(\sum_{i=1}^n D_{ig} \mathbb{E}\left[\theta_i|Y\right]\right)^2 - \sum_{i=1}^n D_{ig} \mathbb{E}\left[\theta_i|Y\right]^2\right\},$$

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

Appendix C   Hierarchical log-spline estimator

Efron's (2016) empirical Bayes log-spline deconvolution approach uses a flexible exponential family mixing distribution model with density parameterized by a flexible B-th order spline. The spline parameters are then estimated by penalized maximum likelihood. We adapt this approach to estimate the within- and between-industry distributions for the hierarchical model of Section 5.3. The between-industry distribution G_η is approximated with a discrete probability mass function defined on a set of M_η support points $\{\bar{\eta}_1, ..., \bar{\eta}_{M_\eta}\}$. The mass at the m-th support point $\bar{\eta}_m$ is given by

$$g_{\eta,m}(\alpha_\eta) = \exp\left(q'_{\eta,m}\alpha_\eta - \log\left(\sum_{\ell=1}^{M_\eta} \exp(q'_{\eta,\ell}\alpha_\eta)\right)\right),$$

where $q_{\eta,m}$ is a $B \times 1$ vector of values of natural spline basis functions for point m (as detailed in Efron 2016) and α_η is a $B \times 1$ vector of coefficients. Similarly, we approximate the within-industry distribution G_ξ with a discrete distribution defined on support $\{\bar{\xi}_1, ..., \bar{\xi}_{M_\xi}\}$, with mass function

$$g_{\xi,m}(\alpha_\xi) = \exp\left(q'_{\xi,m}\alpha_\xi - \log\left(\sum_{\ell=1}^{M_\xi} \exp(q'_{\xi,\ell}\alpha_\xi)\right)\right)$$

for $B \times 1$ spline basis and coefficient vectors $q_{\xi,m}$ and α_ξ, respectively.

With this specification of the mixing distributions, the joint likelihood contribution for firms in industry k is given by:

$$\mathcal{L}\left(\hat{\boldsymbol{\theta}}_k|\boldsymbol{s}_k; \alpha_\eta, \alpha_\xi\right) = \sum_{\ell=1}^{M_\eta} g_{\eta,\ell}(\alpha_\eta) \left\{ \prod_{i:k(i)=k} \left[\sum_{m=1}^{M_\xi} g_{\xi,m}(\alpha_\xi) \frac{1}{s_i^{1-\beta}} \phi\left(\frac{(\hat{\theta}_i/s_i^\beta) - \bar{\eta}_\ell\bar{\xi}_m}{s_i^{1-\beta}}\right) \right] \right\},$$

where $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{s}_k$ are vectors collecting the $\hat{\theta}_i$ and s_i for firms with $k(i) = k$. Following Efron (2016), we estimate the parameters α_η and α_ξ by penalized maximum likelihood. Our approach extends the Efron (2016) estimator to add a separate penalty for the within- and between-industry spline coefficients. Specifically, the parameter estimates are computed as:

$$(\hat{\alpha}_\eta, \hat{\alpha}_\xi) = \arg\max_{(\alpha_\eta, \alpha_\xi)} \sum_{k=1}^{K} \log \mathcal{L}\left(\hat{\boldsymbol{\theta}}_k|\boldsymbol{s}_k; \alpha_\eta, \alpha_\xi\right) - c_\eta\sqrt{\alpha'_\eta\alpha_\eta} - c_\xi\sqrt{\alpha'_\xi\alpha_\xi}.$$

We set the spline order equal to $B = 5$. In models with industry effects the number of support points is set equal to $M_\eta = M_\xi = 200$, with points equally spaced on the supports of η_k and ξ_i. Models without industry effects use $M_\xi = 1,000$ and $M_\eta = 1$

with $\bar{\eta}_1 = 1$, so that η_k has a degenerate distribution at unity. The lower limits of the supports are set to zero. The upper limits of the support for each component is set equal to the maximum of the empirical distribution of corresponding estimates or five standard deviations above the GMM-estimated mean, whichever is larger. Since the scales of the two mixing distributions are not separately identified we impose the constraint $\sum_m g_{\eta,m}(\alpha_m)\bar{\eta}_m = 1$. The penalty terms c_η and c_ξ are calibrated so that mean contact ratio and variances of the within- and between-industry components come as close as possible to matching GMM estimates of these same quantities, as measured by the quadratic distance between log-spline and GMM estimates (scaled by the inverse variance matrix of the GMM estimates). The model without industry effects chooses c_ξ to minimize the quadratic difference between model-implied and GMM estimates of the mean and total variance of contact gaps. In practice all parameters match well, as can be seen by comparing Tables 3 and 4.

Appendix D   Results under square weighted loss

In this Appendix we discuss the grades that result from minimizing the square-weighted risk function $\mathcal{R}^2(d; \lambda)$ introduced in Section 3.6. Mirroring the formulation in (3), square-weighted risk can be written as a linear combination of a square-weighted Discordance Rate and a square-weighted rank correlation:

$$\mathcal{R}^2(d; \lambda) = (1 - \lambda)DR^2(d) - \lambda\bar{\tau}^2(d). \tag{10}$$

The square-weighted rank correlation coefficient

$$\bar{\tau}^2(d) = \frac{1}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} m_{ij}^2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i < d_j\right\}(\mu_{ij}^2 - \mu_{ji}^2) + 1\left\{d_i > d_j\right\}(\mu_{ji}^2 - \mu_{ij}^2)$$

provides a measure of the information conveyed by a set of grades. The square-weighted Discordance Rate

$$DR^2(d) = \frac{1}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} m_{ij}^2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 1\left\{d_i < d_j\right\}\mu_{ij}^2 + 1\left\{d_i > d_j\right\}\mu_{ji}^2$$

is an average of the $DR_{g,g'}^2$ introduced in Section 3.7 that summarizes the reliability of a set of grades. In words, $DR^2(d)$ gives the chance that a randomly selected pair of firms, sampled with weights proportional to the square of the difference in their latent contact gaps, is misranked.

We use the decomposition given by (10) in the applications below to construct a reporting possibilities frontier contrasting $DR^2(d^*(\lambda))$ with $\bar{\tau}^2(d^*(\lambda))$ for different choices of λ. In both applications, square-weighting shifts out the reporting possibilities frontier relative to the frontier implied by binary (i.e. $p = 0$) loss, which is intuitive as one should expect less loss from ranking exercises when small mistakes are inconsequential. As a result, switching from binary to square-weighted loss tends to yield more grades when λ is held constant.

D.1   Square-weighted results for first names

Panel (a) of Figure E3 depicts the number of grades that emerge from minimizing square-weighted risk as a function of the preference parameter λ. At our preferred value of $\lambda = 0.25$, five grades emerge with a square weighted discordance rate of 5.4%. Panel (b) of Figure E3 shows that square-weighting shifts out the reporting possibilities frontier. As in our binary baseline, naively ranking either the point estimates $\hat{\theta}_i$ or the posterior means $\bar{\theta}_i$ according to their empirical distribution yields grades with information content and reliability very close to those delivered by the square-weighted grades with $\lambda = 1$.

The five grades that result from choosing $\lambda = 0.25$ under square-weighting are shown in Figure E4. Interestingly, under this weighting scheme, the posterior means of the names are more nearly monotone in the n assigned grades that result when setting $\lambda = 1$ than was found under binary loss. Two distinctively white female names "Misty" and "Heather" earn the top grade, while three out of the four names assigned the bottom grade "Lawanda", "Tameka", and "Latisha" are distinctively Black and female. As shown in Figure E5, however, a logistic regression of sex on these five grade categories yields a pseudo-R^2 of roughly 0.05, while a corresponding race regression yields a pseudo-R^2 of 0.24 revealing that these grades are much stronger predictors of race than sex.

D.2 Square-weighted results for firms

Figure E8 shows the number of grades that emerge from different choices of λ. Regardless of whether industry effects are included, square weighting substantially increases the number of grades assigned relative to the solutions under binary loss. At $\lambda = 0.25$ square weighting yields six grades without industry effects and eight grades when industry effects are included. While the number of grades increases, their reliability remains high, with the six grade ranking exhibiting a square-weighted discordance rate of 4.7% and the eight grade ranking a corresponding discordance rate of 3.1%. Evidently, it is possible to assign many grades without exposing oneself to many large mistakes.

Figure E9 displays the reporting possibilities under binary and square weighting. Square-weighting the losses pushes out the reporting possibilities frontier relative to binary ranking. Naively ranking based upon the posterior mean $\bar{\theta}_i$ again yields results very similar to the grades generated by setting $\lambda = 1$. However ranking based upon the naive (i.e., unshrunk) point estimates $\hat{\theta}_i$ yields a combination of very high square-weighted Discordance Rates and a modest square-weighted rank correlation that is dominated by our square-weighted grades when $\lambda = 0.25$.

Figure E10 displays the firm rankings derived from minimizing square-weighted risk without conditioning on industry effects. 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 three additional grades, and a more even distribution of firms across grades. Using industry effects to evaluate the square-weighted risk generates the eight grade ranking portrayed in Figure E11. Notably, a single firm with the highest posterior mean contact penalty earns the worst grade.

As in the binary loss case, these grades explain a meaningful share of total variance in θ_i. The estimated between-grade standard deviation in contact penalties is 0.047 for the six grades reported in Figure E10, implying an R^2 of 37%. The grades computed from the model with industry effects and reported in Figure E11 explain 56% of the total variance. Figures E15 and E16 depict the grades assigned under different choices of λ for

the square-weighting schemes excluding and including industry effects respectively.

Figure E16 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 10 to 15% implying Bayes factors ranging from 8 to 6. Weighted Discordance Rates between non-adjacent grades are uniformly smaller than 3%. Our finding that the square-weighted DRs in Figure E16 fall below those in Figure E16, 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.

Appendix E   Additional Figures and Tables

Table E1: 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.

Table E2: Detailed results by firm

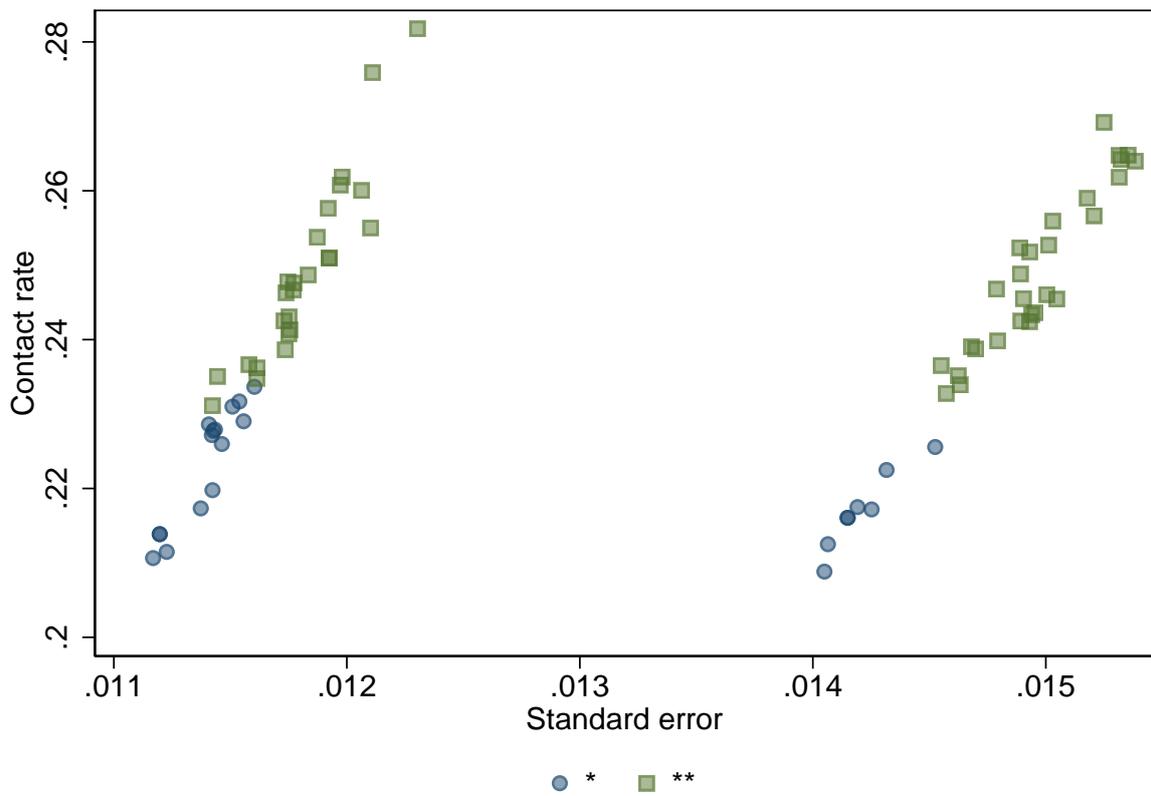| Firm (SIC)	θ_i	Baseline model				Industry effects model			
		Post. mean	Post. CI	Grd	Cond. rank	Post. mean	Post. CI	Grd	Cond. rank
6 (55)	0.33 (0.07)	0.25	[0.12, 0.34]	1	1	0.27	[0.16, 0.38]	1	2
1 (55)	0.43 (0.13)	0.23	[0.08, 0.45]	1	2	0.33	[0.16, 0.5]	1	1
3 (53)	0.38 (0.28)	0.19	[0.03, 0.38]	2	3	0.17	[0.03, 0.32]	2	8
2 (63)	0.4 (0.22)	0.19	[0.04, 0.4]	2	4	0.13	[0.02, 0.27]	3	17
4 (24)	0.35 (0.29)	0.19	[0.03, 0.37]	2	5	0.15	[0.02, 0.32]	2	12
5 (59)	0.34 (0.24)	0.18	[0.03, 0.35]	2	6	0.17	[0.03, 0.34]	2	9
27 (24)	0.17 (0.31)	0.17	[0.02, 0.34]	2	7	0.15	[0.02, 0.32]	2	14
8 (56)	0.3 (0.23)	0.17	[0.03, 0.33]	2	8	0.16	[0.03, 0.36]	2	11
7 (73)	0.3 (0.19)	0.16	[0.03, 0.32]	2	9	0.10	[0.01, 0.24]	3	29
11 (55)	0.27 (0.08)	0.16	[0.06, 0.32]	2	10	0.24	[0.12, 0.36]	1	5
13 (51)	0.24 (0.24)	0.16	[0.03, 0.32]	2	11	0.12	[0.02, 0.25]	3	19
10 (51)	0.29 (0.11)	0.16	[0.05, 0.33]	2	12	0.10	[0.02, 0.19]	3	31
9 (55)	0.29 (0.11)	0.16	[0.05, 0.32]	2	13	0.27	[0.11, 0.42]	1	3
12 (23)	0.26 (0.09)	0.15	[0.06, 0.3]	2	14	0.13	[0.03, 0.29]	3	18
97 (63)	-0.54 (0.44)	0.16	[0.01, 0.34]	2	15	0.14	[0.01, 0.32]	3	15
25 (59)	0.17 (0.21)	0.14	[0.02, 0.28]	2	16	0.15	[0.02, 0.3]	2	13
14 (59)	0.24 (0.08)	0.14	[0.05, 0.27]	2	17	0.12	[0.04, 0.21]	3	20
15 (56)	0.24 (0.09)	0.14	[0.05, 0.26]	2	18	0.11	[0.03, 0.24]	3	21
22 (63)	0.18 (0.19)	0.14	[0.02, 0.27]	2	19	0.11	[0.01, 0.23]	3	26
16 (49)	0.22 (0.15)	0.14	[0.03, 0.26]	2	20	0.11	[0.02, 0.23]	3	24
18 (51)	0.2 (0.14)	0.13	[0.03, 0.24]	2	21	0.09	[0.02, 0.19]	3	30
17 (48)	0.22 (0.1)	0.13	[0.04, 0.23]	2	22	0.11	[0.02, 0.24]	3	25
19 (55)	0.19 (0.14)	0.13	[0.02, 0.24]	2	23	0.24	[0.05, 0.43]	1	4
33 (24)	0.14 (0.17)	0.12	[0.02, 0.24]	2	24	0.11	[0.01, 0.23]	3	23
20 (70)	0.19 (0.12)	0.12	[0.03, 0.22]	2	25	0.09	[0.01, 0.19]	3	39
21 (80)	0.19 (0.08)	0.12	[0.03, 0.2]	2	26	0.07	[0.01, 0.16]	3	57
23 (55)	0.18 (0.08)	0.11	[0.04, 0.2]	2	27	0.19	[0.07, 0.3]	2	7
24 (53)	0.17 (0.06)	0.11	[0.04, 0.19]	2	28	0.09	[0.03, 0.16]	3	34
93 (59)	-0.12 (0.24)	0.12	[0.01, 0.25]	2	29	0.13	[0.01, 0.28]	3	16
26 (54)	0.17 (0.08)	0.11	[0.03, 0.19]	2	30	0.08	[0.02, 0.16]	3	45
28 (49)	0.17 (0.08)	0.11	[0.03, 0.19]	2	31	0.09	[0.01, 0.17]	3	42
66 (55)	0.05 (0.17)	0.11	[0.01, 0.23]	2	32	0.21	[0.01, 0.42]	2	6
45 (48)	0.11 (0.14)	0.11	[0.02, 0.21]	2	33	0.11	[0.01, 0.24]	3	27
39 (57)	0.13 (0.12)	0.11	[0.02, 0.21]	2	34	0.11	[0.02, 0.22]	3	22
32 (51)	0.14 (0.11)	0.11	[0.02, 0.2]	2	35	0.08	[0.01, 0.16]	3	49
34 (57)	0.14 (0.11)	0.11	[0.02, 0.2]	2	36	0.10	[0.02, 0.21]	3	28
29 (59)	0.15 (0.06)	0.10	[0.04, 0.17]	2	37	0.09	[0.03, 0.17]	3	32
Continued on next page									

30 (55)	0.15 (0.06)	0.10	[0.04, 0.17]	2	38	0.16	[0.06, 0.26]	2	10
31 (58)	0.14 (0.06)	0.10	[0.04, 0.16]	2	39	0.08	[0.03, 0.15]	3	43
38 (75)	0.13 (0.09)	0.10	[0.02, 0.18]	2	40	0.07	[0.01, 0.15]	3	58
35 (53)	0.14 (0.05)	0.10	[0.04, 0.16]	2	41	0.08	[0.03, 0.14]	3	44
37 (57)	0.14 (0.06)	0.10	[0.03, 0.16]	2	42	0.09	[0.03, 0.16]	3	38
36 (59)	0.14 (0.06)	0.09	[0.03, 0.16]	2	43	0.09	[0.03, 0.16]	3	36
95 (60)	-0.19 (0.22)	0.10	[0.01, 0.22]	2	44	0.09	[0.01, 0.22]	3	37
40 (50)	0.12 (0.08)	0.09	[0.02, 0.17]	2	45	0.08	[0.01, 0.16]	3	51
47 (62)	0.1 (0.1)	0.09	[0.02, 0.18]	2	46	0.08	[0.01, 0.18]	3	47
53 (53)	0.09 (0.1)	0.09	[0.01, 0.18]	2	47	0.09	[0.02, 0.17]	3	33
44 (53)	0.11 (0.07)	0.09	[0.02, 0.16]	2	48	0.08	[0.02, 0.15]	3	40
42 (56)	0.12 (0.06)	0.09	[0.02, 0.15]	2	49	0.08	[0.02, 0.16]	3	41
43 (53)	0.12 (0.07)	0.09	[0.02, 0.15]	2	50	0.08	[0.02, 0.15]	3	46
41 (57)	0.12 (0.05)	0.09	[0.03, 0.14]	2	51	0.08	[0.03, 0.14]	3	52
49 (23)	0.09 (0.09)	0.09	[0.02, 0.16]	2	52	0.09	[0.02, 0.2]	3	35
69 (73)	0.03 (0.11)	0.08	[0.01, 0.17]	2	53	0.07	[0, 0.16]	3	64
50 (53)	0.09 (0.07)	0.08	[0.02, 0.15]	2	54	0.08	[0.02, 0.14]	3	50
46 (53)	0.1 (0.05)	0.08	[0.02, 0.14]	2	55	0.07	[0.02, 0.13]	3	54
51 (51)	0.09 (0.07)	0.08	[0.02, 0.14]	2	56	0.06	[0.01, 0.13]	3	67
48 (51)	0.1 (0.05)	0.08	[0.02, 0.13]	2	57	0.06	[0.01, 0.11]	3	72
96 (80)	-0.27 (0.2)	0.08	[0.01, 0.19]	2	58	0.08	[0, 0.19]	3	55
68 (56)	0.04 (0.09)	0.07	[0.01, 0.15]	2	59	0.08	[0.01, 0.17]	3	48
63 (63)	0.05 (0.08)	0.07	[0.01, 0.15]	2	60	0.06	[0.01, 0.14]	3	65
60 (50)	0.06 (0.08)	0.07	[0.01, 0.14]	2	61	0.07	[0.01, 0.14]	3	59
74 (54)	0 (0.11)	0.07	[0.01, 0.16]	2	62	0.07	[0.01, 0.15]	3	56
58 (24)	0.06 (0.07)	0.07	[0.01, 0.14]	2	63	0.07	[0.01, 0.14]	3	61
52 (58)	0.09 (0.04)	0.07	[0.02, 0.12]	2	64	0.06	[0.02, 0.11]	3	63
65 (60)	0.05 (0.08)	0.07	[0.01, 0.15]	2	65	0.06	[0.01, 0.14]	3	70
54 (54)	0.08 (0.05)	0.07	[0.02, 0.12]	2	66	0.06	[0.01, 0.11]	3	71
72 (42)	0.02 (0.09)	0.07	[0.01, 0.15]	2	67	0.05	[0, 0.12]	3	84
61 (51)	0.06 (0.07)	0.07	[0.01, 0.13]	2	68	0.06	[0.01, 0.12]	3	74
55 (52)	0.08 (0.03)	0.07	[0.02, 0.11]	2	69	0.05	[0.01, 0.1]	3	79
78 (52)	-0.01 (0.1)	0.07	[0.01, 0.15]	2	70	0.07	[0.01, 0.14]	3	62
71 (53)	0.02 (0.09)	0.07	[0.01, 0.14]	2	71	0.07	[0.01, 0.15]	3	53
56 (53)	0.07 (0.04)	0.06	[0.02, 0.11]	2	72	0.06	[0.02, 0.11]	3	66
64 (58)	0.05 (0.07)	0.06	[0.01, 0.13]	2	73	0.07	[0.01, 0.13]	3	60
59 (75)	0.06 (0.06)	0.06	[0.01, 0.12]	2	74	0.05	[0.01, 0.11]	3	80
57 (58)	0.06 (0.05)	0.06	[0.01, 0.11]	2	75	0.06	[0.01, 0.11]	3	69
81 (80)	-0.01 (0.09)	0.06	[0.01, 0.13]	2	76	0.06	[0, 0.13]	3	77
94 (80)	-0.18 (0.14)	0.06	[0, 0.15]	2	77	0.06	[0, 0.15]	3	73
79 (42)	-0.01 (0.08)	0.06	[0.01, 0.13]	2	78	0.04	[0, 0.11]	3	86
84 (63)	-0.02 (0.09)	0.06	[0.01, 0.13]	2	79	0.06	[0, 0.13]	3	75
85 (51)	-0.02 (0.09)	0.06	[0.01, 0.13]	2	80	0.06	[0.01, 0.12]	3	76

Continued on next page

62 (54)	0.06 (0.04)	0.05	[0.01, 0.1]	2	81	0.05	[0.01, 0.09]	3	81
73 (59)	0.01 (0.07)	0.05	[0.01, 0.11]	2	82	0.06	[0.01, 0.13]	3	68
67 (87)	0.04 (0.04)	0.05	[0.01, 0.1]	2	83	0.05	[0.01, 0.1]	3	82
83 (51)	-0.01 (0.06)	0.05	[0, 0.1]	3	84	0.05	[0, 0.1]	3	83
80 (53)	-0.01 (0.06)	0.05	[0, 0.1]	3	85	0.05	[0, 0.11]	3	78
88 (42)	-0.03 (0.07)	0.04	[0, 0.1]	3	86	0.03	[0, 0.09]	4	92
77 (52)	0 (0.05)	0.04	[0, 0.09]	3	87	0.04	[0, 0.09]	3	85
87 (42)	-0.03 (0.06)	0.04	[0, 0.1]	3	88	0.03	[0, 0.09]	4	93
70 (53)	0.02 (0.03)	0.03	[0, 0.07]	3	89	0.04	[0.01, 0.08]	4	87
82 (20)	-0.01 (0.04)	0.03	[0, 0.08]	3	90	0.03	[0, 0.08]	4	90
92 (70)	-0.11 (0.07)	0.03	[0, 0.09]	3	91	0.04	[0, 0.1]	4	88
75 (51)	0 (0.04)	0.03	[0, 0.07]	3	92	0.03	[0, 0.07]	4	89
86 (49)	-0.03 (0.04)	0.03	[0, 0.07]	3	93	0.03	[0, 0.08]	4	91
76 (54)	0 (0.03)	0.02	[0, 0.05]	3	94	0.03	[0, 0.06]	4	95
90 (75)	-0.05 (0.04)	0.02	[0, 0.06]	3	95	0.03	[0, 0.07]	4	94
91 (51)	-0.07 (0.04)	0.02	[0, 0.05]	3	96	0.02	[0, 0.06]	4	96
89 (73)	-0.03 (0.03)	0.02	[0, 0.04]	3	97	0.02	[0, 0.05]	4	97

Notes: This table reports estimated contact penalties and the results of Empirical Bayes and grading exercises for each firm in the sample. Firms are labeled by their raw contact penalty rank, with their industry (2-digit SIC code) shown in parentheses. The column θ_i reports contact penalties with a job-clustered standard error in parentheses. The remaining columns report posterior means (Post. mean), 95% credible intervals (Post. CI), assigned grades using $\lambda = .25$ (Grd), and Condorcet ranks (Cond. rank), which are grades under $\lambda = 1$, in the baseline model and the model with industry effects.

Figure E1: Contact rates, standard errors, and name grades

Notes: This figure plots the estimated contact rates for each name against its standard error. The shape and color of each point indicate the grade assigned to the name using the same specification as Figure 3.

Figure E2: Predictive power of grades name for race and sex labels
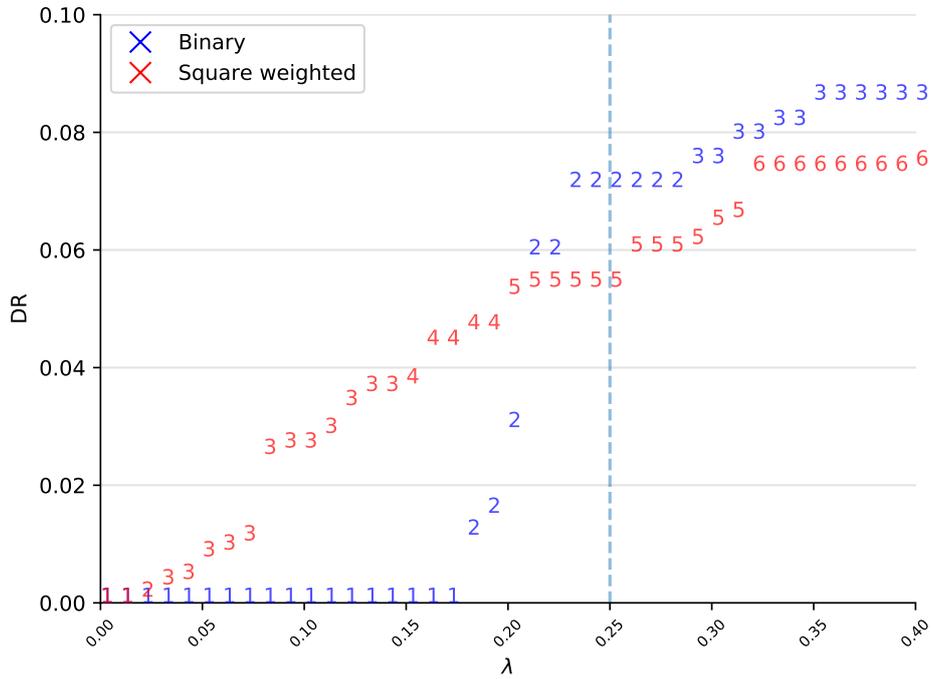
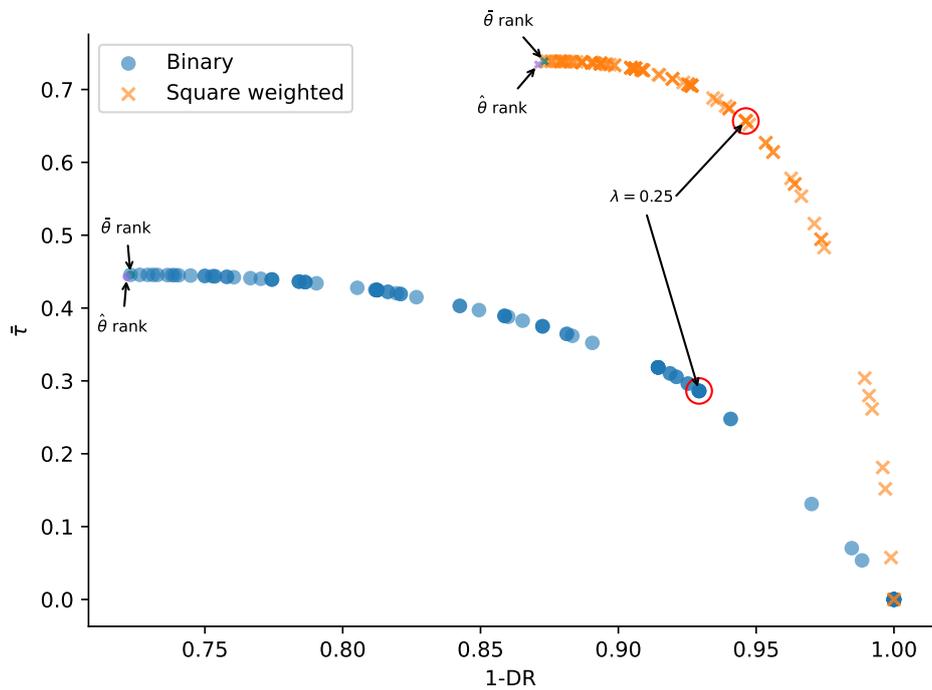a) Pseudo R^2
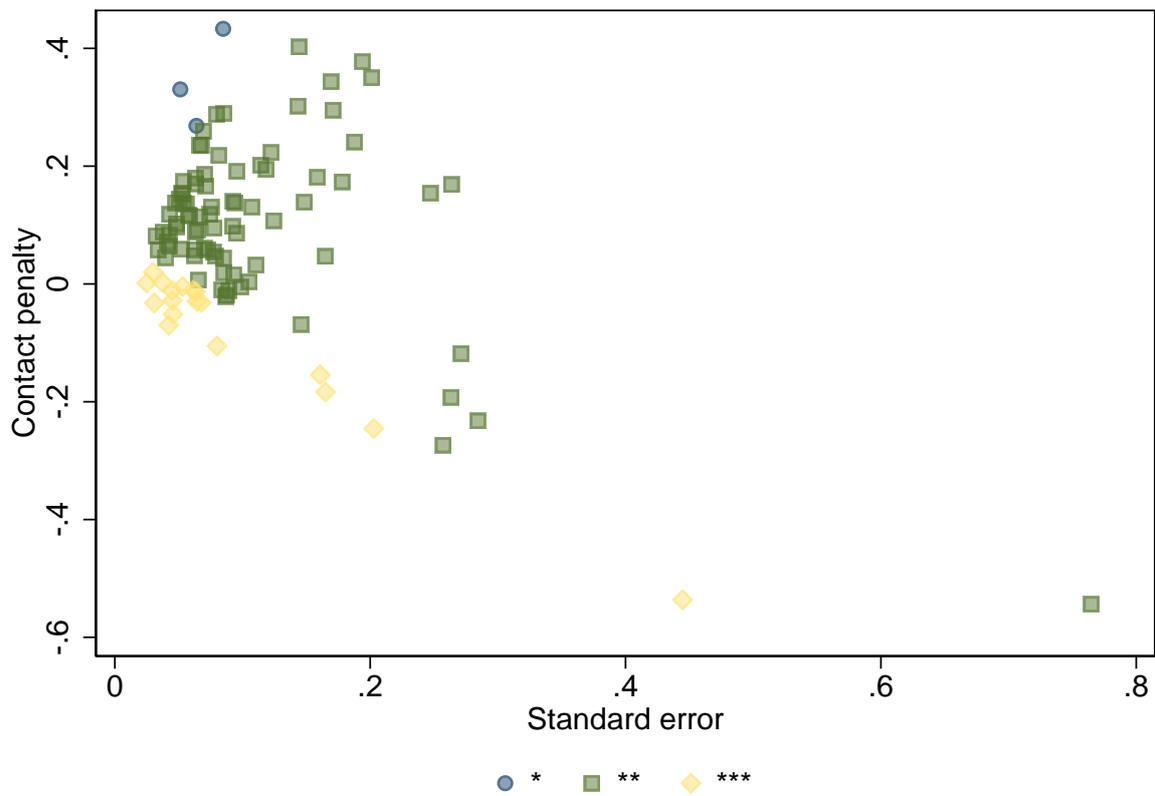
b) Area under the curve

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 λ. The number shown indicates the number of grades assigned.

Figure E3: Square weighted name ranking exercises

a) Grades and discordance

b) Reporting possibilities

Notes: This figure presents additional results from grading contact rates for names under binary and square-weighted loss. In both panels, posteriors are computed using the log-spline estimate plotted in Figure 1 as the prior. Panel (a) shows estimated Discordance Rates (DR) for an intermediate range of λ under binary and square-weighted loss. Panel (b) plots the expectation of Kendall's τ rank correlation between true contact rates and grades against Discordance Rates (DR) for a range of grades indexed by λ. Red circles highlight the DR and expected τ corresponding to $\lambda = 0.25$. "$\hat{\theta}$ rank" refers to ranks based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon Empirical Bayes posterior means.

Figure E4: Posterior means and grades for names under square-weighted loss

Notes: This figure shows posterior mean contact rates, 95% credible intervals, and assigned grades for names. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Names are ordered by their rank under $\lambda = 1$, when each name is assigned its own grade.

Figure E5: 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 λ. The number shown indicates the number of grades assigned.

Figure E6: Unadjusted and studentized contact gaps against standard errors

a) $\hat{\theta}_i$ vs. s_i

b) $\hat{T}_i$ vs. s_i

Notes: Panel (a) of this figure plots estimated contact penalties against their standard errors. The green line plots the conditional mean of θ_i given s_i implied by the GMM estimates. Panel (b) plots studentized contact gaps $\hat{T}_i$ against standard errors. The green line plots the relationship implied by the model.

Figure E7: Contact penalties, standard errors, and report card grades

Notes: This figure plots the estimated 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.

Figure E8: Grades and discordance as a function of λ

Notes: This figure shows estimated Discordance Rates (DR) as a function of λ for both binary and square-weighted loss. The number on each point indicates the number of unique grades in the underlying grading scheme. The vertical dashed line shows results for the benchmark case of $\lambda = 0.25$.

22

Figure E9: Reporting possibilities

Notes: This figure shows the expectation of Kendall's τ rank correlation between θ and assigned grades (labeled $\bar{\tau}$) against Discordance Rates (DR) for a range of grades indexed by λ and for both binary and square-weighted loss. Red circles highlight the DR and $\bar{\tau}$ corresponding to $\lambda = 0.25$. "$\hat{\theta}$ rank" refers to ranks based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon Empirical Bayes posterior means.
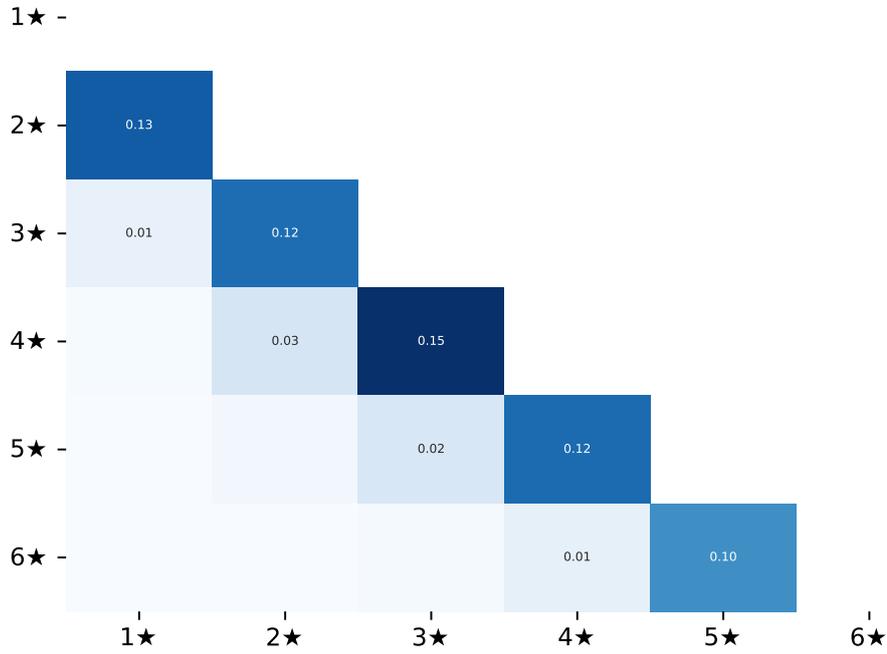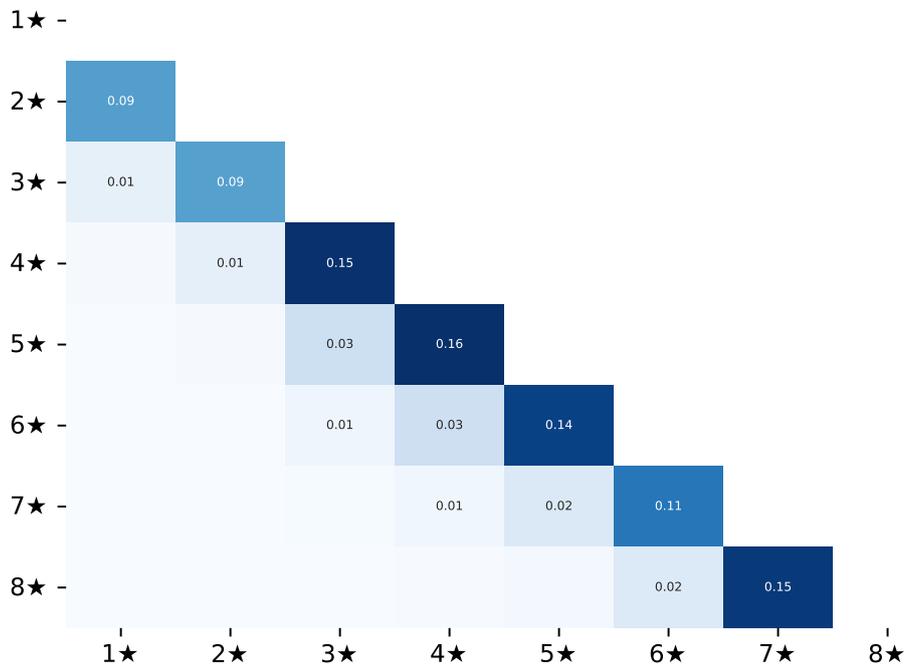
Figure E10: Square-weighted loss: Posterior means and grades

Notes: This figure shows posterior mean proportional contact penalties, 95% credible intervals, and assigned grades under square-weighted loss. Results are shown for $\lambda = 0.25$. 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. Firms labeled with black text are federal contractors, whereas firms in gray are not.

24

Figure E11: 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 model with industry effects under square-weighted loss. Results are shown for $\lambda = 0.25$. 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. Firms labeled with black text are federal contractors, whereas firms in gray are not.

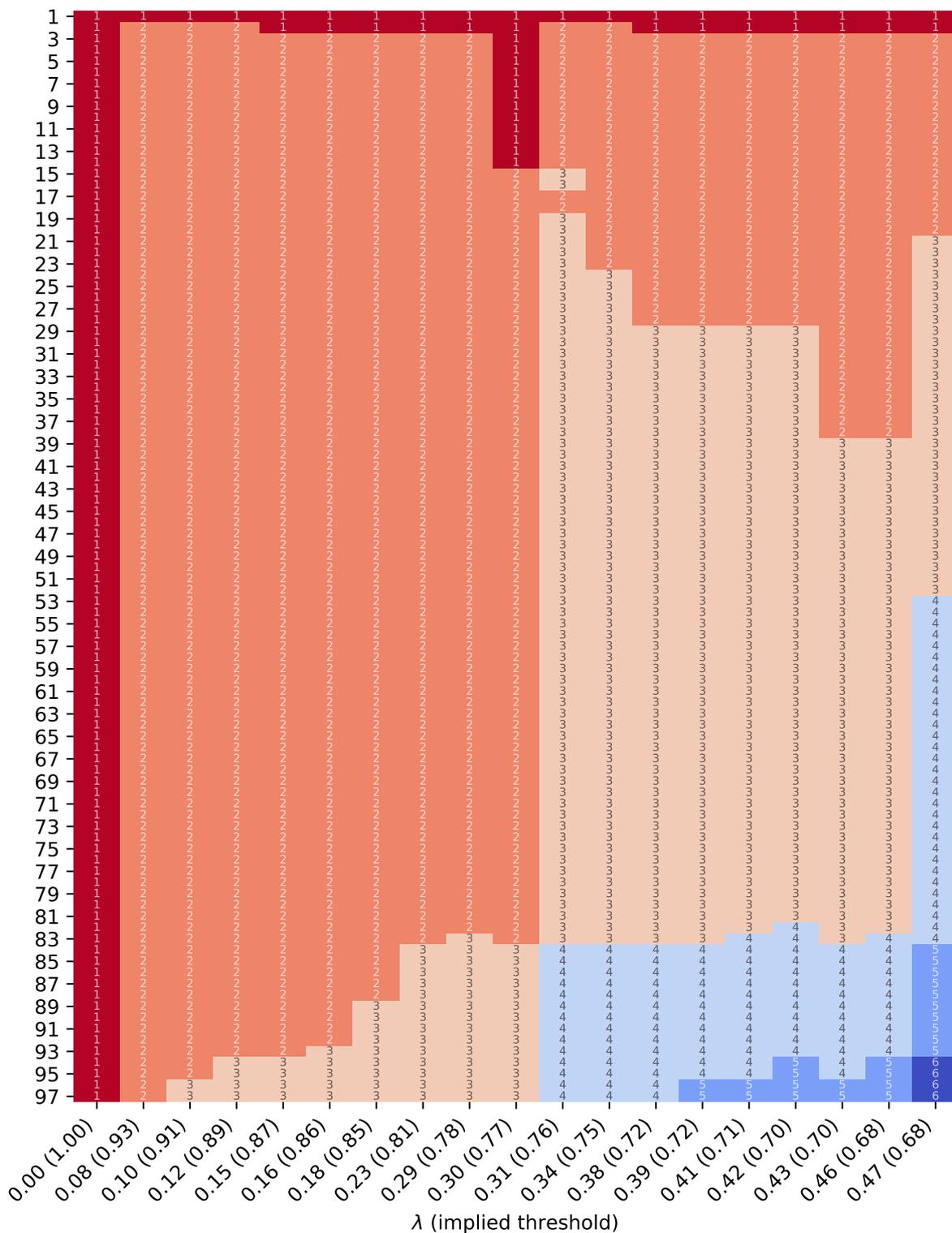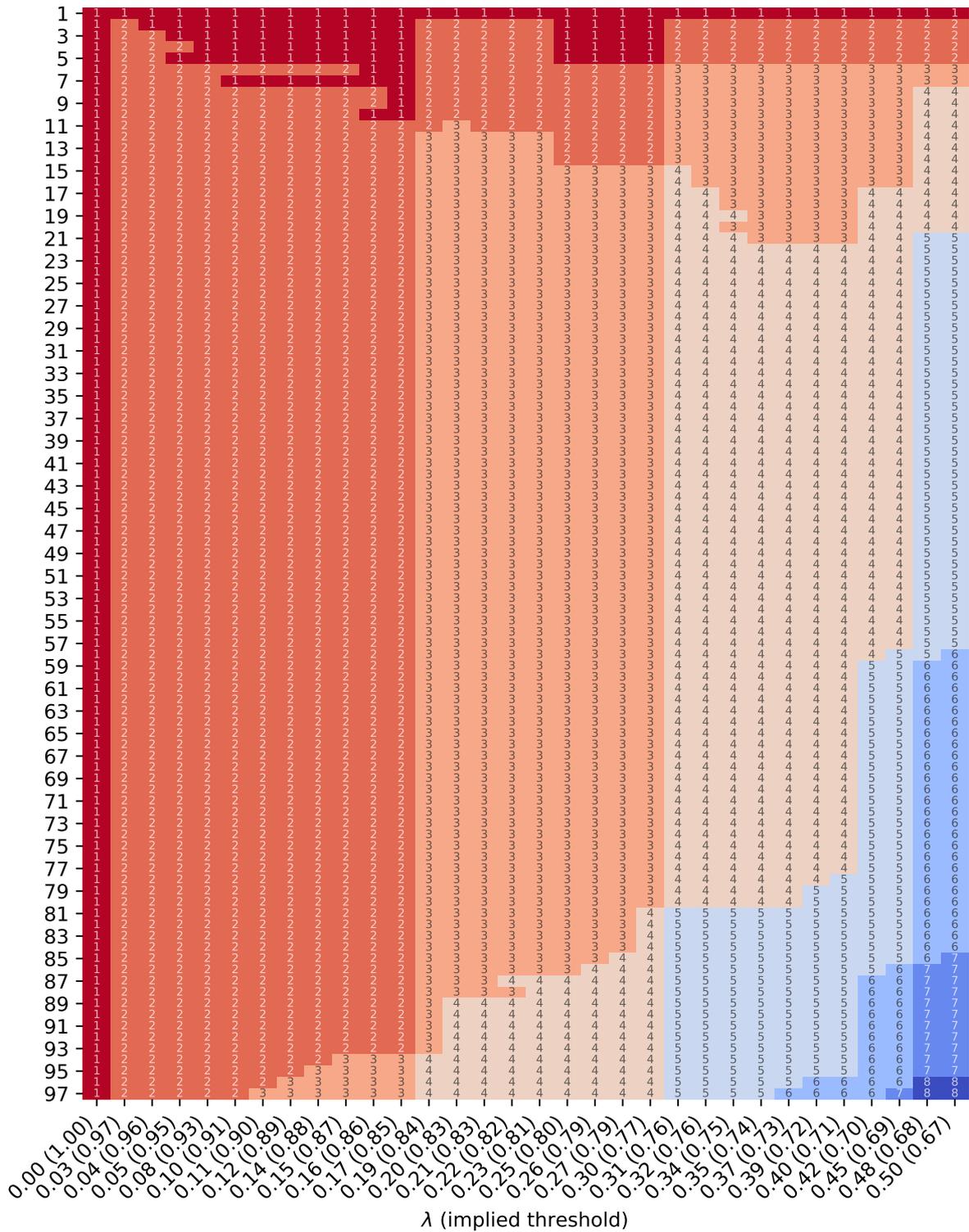Figure E12: DR for square-weighted loss

a) Binary

b) Industry effects

Notes: This figure shows mean Discordance Rates (DR) across grade pairs under square weighted loss. Panel a uses the baseline model, while panel b uses the model with industry effects. Only cells where DR is above 0.01 are annotated. In both panels, DR decays quickly when comparing non-adjacent grades.
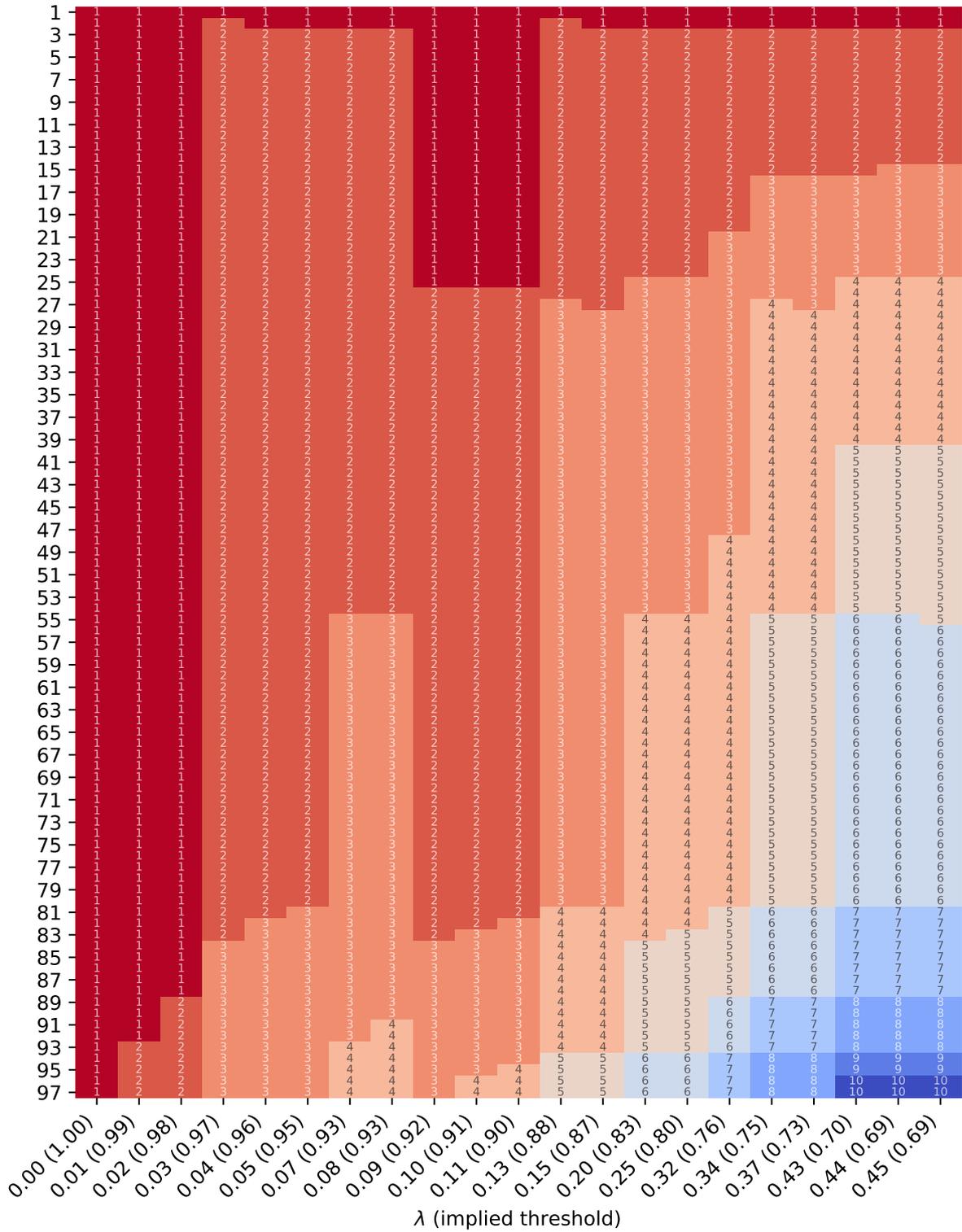
Figure E13: Binary loss: All grades

Notes: This figure shows grade assignments for each value of $\lambda <= 0.5$. To increase readability, only the smallest lambda that yields each unique set of grades is retained. The x-axis reports this λ and the value of $1/(1 + \lambda)$, which is the implied posterior threshold for pairwise ranking decisions. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade.

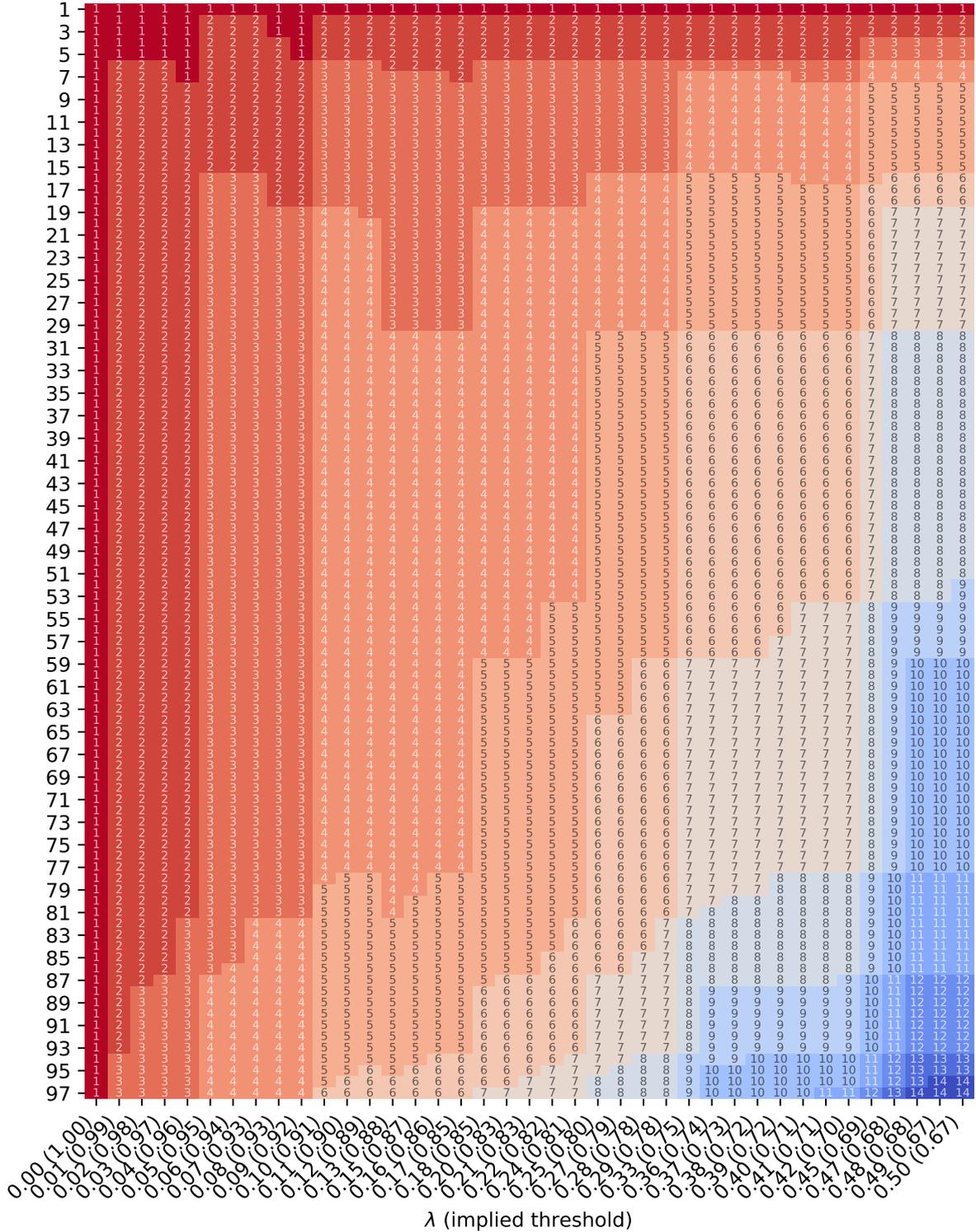Figure E14: Binary loss with industry effects: All grades

Notes: This figure shows grade assignments for each value of $\lambda \leq 0.5$. To increase readability, only the smallest lambda that yields each unique set of grades is retained. The x-axis reports this λ and the value of $1/(1+\lambda)$, which is the implied posterior threshold for pairwise ranking decisions. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade.

Figure E15: Square-weighted loss: All grades

Notes: This figure shows grade assignments for each value of $\lambda \leq 0.5$. To increase readability, only the smallest lambda that yields each unique set of grades is retained. The x-axis reports this λ and the value of $1/(1+\lambda)$, which is the implied posterior threshold for pairwise ranking decisions. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade.

Figure E16: Square-weighted loss with industry effects: All grades

Notes: This figure shows grade assignments for each value of $\lambda \leq 0.5$. To increase readability, only the smallest lambda that yields each unique set of grades is retained. The x-axis reports this λ and the value of $1/(1 + \lambda)$, which is the implied posterior threshold for pairwise ranking decisions. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade.