WORKING PAPER · NO. 2024-40

A Discrimination Report Card

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

APRIL 2024

THE UNIVERSITY OF CHICAGO
BECKER FRIEDMAN INSTITUTE
FOR ECONOMICS

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

A Discrimination Report Card

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

April 4, 2024

Abstract

We develop an empirical Bayes ranking procedure that assigns ordinal grades to noisy measurements, balancing the information content of the assigned grades against the expected frequency of ranking errors. Applying the method to a massive correspondence experiment, we grade the race and gender contact gaps of 97 U.S. employers, the identities of which we disclose for the first time. The grades are presented alongside measures of uncertainty about each firm's contact gap in an accessible report card that is easily adaptable to other settings where ranks and levels are of simultaneous interest.

Keywords: Discrimination, Empirical Bayes, Ranking, Correspondence Experiment

JEL Codes: J71, C11, C13

*Kline: UC Berkeley and NBER, pkline@econ.berkeley.edu; Rose: University of Chicago and NBER, ekrose@uchicago.edu; Walters: UC Berkeley and NBER, crwalters@econ.berkeley.edu. We thank Ben Scuderi for helpful feedback on an early draft of this paper and Hadar Avivi and Luca Adorni for outstanding research assistance. Seminar participants at Brown University, the 2022 California Econometrics Conference, Columbia University, CIREQ 2022 Montreal, Harvard University, Microsoft Research, Monash University, Peking University, Royal Holloway, UC Santa Barbara, UC Berkeley, The University of Virginia, the Cowles Econometrics Conference on Discrimination and Algorithmic Fairness, and The University of Chicago Interactions Conference provided useful comments. Routines for implementing the ranking procedures developed in this paper are available online at `https://github.com/ekrose/drrank`.

1 Introduction

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

Scholars, policymakers, and private businesses increasingly report simple "report cards" summarizing estimates of the quality or conduct of particular individuals, organizations, or places. Recent examples include assessments of the quality of 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, 2018; Chetty et al., 2018*a*). It is natural for readers to use such reports not only to assess the conduct of particular organizations but also to make comparisons between them. This "league table mentality," as Gu and Koenker (2020) have termed the phenomenon, forms a core element of the demand for report cards but is rarely incorporated directly into their construction.

This paper develops new empirical Bayes (EB) methods for grading units based upon noisy measures of conduct or performance while maintaining statistical guarantees on the reliability of the resulting grades. The information content of the grades is quantified by Kendall's τ measure of correlation (Kendall, 1938) between the implied (partial) ordering of units and the true ranking of latent conduct parameters. 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 gives the chances that the relative performance of a randomly selected pair of units is misordered.

We show that the tradeoff between these notions of information and reliability emerges naturally from a series of pairwise decisions in which an analyst guesses the ordering of parameters for each pair of units. When presented with 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 against the costs of misordering them. When $\lambda = 1$, it is optimal to assign each unit a unique grade to maximize the expected rank correlation with the true performance 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 unit outperforms another serving the role of a vote share.

When $\lambda < 1$ it is only optimal to strictly rank units 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. We show that setting $\lambda < 1$ can be motivated by a scientific reporting problem where a share of the audience is already informed about how the units should be ordered and an incorrect report will mislead them. Scientific communication is, of course, generally aided by transparency (Andrews and Shapiro, 2021) and we develop a reporting rubric that simultaneously communicates the "Condorcet ranks" that emerge when $\lambda = 1$ alongside the coarse grades associated with a chosen $\lambda < 1$. The proposed report card also summarizes information on performance levels, which can be especially important when assessing compliance with regulatory standards (Kline, 2023). Routines for implementing our EB grading procedure are available online at `https://github.com/ekrose/drrank`.

We use these methods to construct a *discrimination report card* that summarizes experimental evidence regarding the biases of a broad collection of Fortune 500 companies. Our analysis leverages a massive resume correspondence experiment, previously analyzed in Kline, Rose and Walters (2022), that sent up to 1,000 job applications to each of 108 firms, whose identities we disclose for the first time. These companies are familiar to most Americans and their conduct plausibly exerts a large influence on the U.S. labor market. The experiment conveyed race and gender to employers by randomly assigning distinctive names. Disparities in contact rates across race and gender categories provide noisy estimates of discriminatory conduct for each firm. To link our analysis to our earlier theory on ranking decisions, we use these estimates to construct empirically-grounded prior beliefs via empirical Bayes deconvolution methods and compute corresponding EB estimates of each firm's absolute and relative conduct.

As an introductory illustration of our method, we rank the contact rates of the first names used in the 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 strongly predict 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 for a sample of 97 firms 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 the parameter 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, which limits the expected share of firm pairs that are misranked to 3.9%. These grades capture roughly 25% 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 racial 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 affiliation 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 more than 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 70% of the variation in contact penalties across firms and yield a correlation with the latent ranks of 0.46, while limiting the expected share of firm pairs that are misranked to 5.6%.

Firms assigned the worst grade in this ranking contact white applicants 23% more often than Black applicants, similar to the lowest category in the ranking without industry effects. However, 9 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, the 11 firms receiving the best grade in the industry effects model 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 that scored more highly.

We also construct a report card scoring firm preferences for male versus female names. A four grade coding explains 44% of the variation in firms' proportional gender contact gaps. These grades exhibit a correlation of 0.12 with the latent ranks while limiting the expected share of firm pairs that are misranked to 1.8%. Four firms are assigned to two grades indicating a strong preference for male names and four are assigned a grade signaling a strong preference for female names. The magnitude of gender gaps in these three grades is large, with posterior mean estimates averaging more than 34 log points in absolute value. The remaining firms are assigned a grade with negligible average gender contact gaps.

Accounting for industry affiliation yields five gender report card grades. These grades explain 38% of the variation across firms in gender contact gaps, exhibit a correlation with the latent firm ranks of 0.16, and limit the expected share of misranked firm pairs to 1%. Incorporating industry affiliation nearly doubles the number of firms graded as

discriminating against men. However, the vast majority of firms continue to register negligible gender preferences, suggesting gender discrimination at the interview stage is rare and concentrated in particular industries. Grading the industry average gender contact gaps reveals that bias against male names is particularly concentrated in the apparel industry.

Our work extends a burgeoning literature on EB ranking methods. A large empirical literature ranks teachers, schools, hospitals, and neighborhoods using James-Stein style shrinkage rules (e.g., Chetty, Friedman and Rockoff, 2014; Chetty et al., 2018*b*). 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.

The estimates provided in our paper should not be construed as making a legal assessment that companies in our experiment violated anti-discrimination laws. However, 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 in our correspondence experiment receiving poor grades turn out to be federal contractors, suggesting this information may be of help in targeting future compliance efforts. While the legal ramifications of contact gaps in correspondence experiments remain unclear (U.S. EEOC, 1996; Onwuachi-Willig and Barnes, 2005; U.S. Equal Employment Opportunity Commission v. Target Corp, 460 F.3d 946 7th Cir. 2006), targeting investigations based on such experiments may yield additional actionable evidence.

Unfortunately, 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 empirical light on the problem of discrimination may have a more immediately salutary effect on corporate behavior 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. We then restricted attention to 108 firms with sufficient geographic variation in entry-level job postings and hiring platforms that were feasible to audit using our experimental methods. 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 managed by the same individual.

The experiment sampled job postings in a series of five waves, spanning the period from October 2019 to April 2021, 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 including 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. The final sample included roughly 84,000 applications 11,000 jobs at 108 firms.

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 came from the 2010 U.S. Census. We selected names with high race-specific shares among those that occur at least 10,000 times nationally. The full list of experimental names appears in Appendix Table F2.

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 within 30 days. 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. 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.

Beliefs regarding the likely values of the n discrimination levels $\theta_1, \ldots, \theta_n$ are represented by the distribution function $B : \mathbb{R}^n \to [0, 1]$, which is assumed to be continuously differentiable. In the empirical work to follow, B will take the form of a posterior distribution constructed using empirical Bayes methods, as detailed in the next section. For an analyst able to elicit B via introspection, what follows is a coherent account of how to translate these beliefs into an optimal ranking.

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. 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.

The posterior probability that θ_i is greater than θ_j can be written $\pi_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{x} dB_{ij}(t, x)$, where $B_{ij} : \mathbb{R}^2 \to [0, 1]$ denotes the bivariate distribution of beliefs over the pair (θ_i, θ_j). We assume beliefs are continuously distributed. Hence, ties are measure zero and $\pi_{ij} =$

$1 - \pi_{ji}$. This setup implies the expected utility of assigning grades $d = (d_1, d_2) \in \{1, 2\}^2$ to this pair of firms takes the form

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

The optimal grading 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.

A complementary interpretation of λ comes from viewing the grades as the solution to a scientific reporting problem. Suppose the grades are reported to an audience choosing between firms i and j. If they choose the firm with the lowest level of discrimination they receive payoff one. Otherwise, they obtain payoff zero. All members of the audience will choose whichever firm is recommended by the grades. However, a share $q \in (0, 1)$ of the audience is informed and will choose correctly between firms assigned the same grade, while the rest of the audience has chance $1/2$ of choosing correctly in the event of a tie.

This setup implies that deeming the pair a tie yields expected payoff $q + (1 - q)/2 = (1 + q)/2$, while properly ordering the firms generates payoff one and misordering them gives payoff zero. With this payoff structure, the expected utility of choosing grades d is now given by $\frac{1+q}{2} + \frac{1+q}{2} EU(\pi_{ij}, d; \frac{1-q}{1+q})$. Hence, the same λ-thresholding decision rule is optimal with $\lambda = \frac{1-q}{1+q} \in (0, 1)$ now a function of the audience's degree of sophistication. As the share q of the audience that is informed grows, λ falls, yielding greater 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 equals 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 $\theta = (\theta_1, \ldots, \theta_n)'$ denote the vector of latent biases and $d = (d_1, \ldots, d_n)'$ a vector of assigned grades, our loss function can be written:

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

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

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

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

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 strictly misranked according to their grades. Interpreting the decision problem as a series of tests of the null hypotheses that $\theta_i = \theta_j$ for each pair of firms, the Discordance Proportion may be seen as a directional (sometimes called type III) error rate – the share of null hypotheses that are rejected in favor of erroneous alternatives. This representation clarifies that the parameter λ trades off the desire to accurately classify firms by maximizing $\tau(d, \theta)$ against concerns about misclassifying them, as reflected by $DP(d, \theta)$.[1]

3.3 Risk function

While we 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. However, the expected values of both $\tau(d, \theta)$ and $DP(d, \theta)$ under beliefs B can be expressed in terms of the pairwise probabilities π_{ij}. The expected rank correlation

[1] Appendix A considers an extended family of loss functions that weight pairwise concordances and discordances by powers of the difference between the cardinal biases of the two firms, reflecting the notion that misranking firms with large differences in conduct is more costly than misordering firms with roughly equivalent conduct. This extension yields a tradeoff between weighted notions of rank correlation and the Discordance Proportion. An earlier version of this paper (Kline, Rose and Walters, 2023) reports the results of these rankings.

$\bar{\tau}(d) = \mathbb{E}_B[\tau(d,\theta)] = \int \tau(d,x)dB(x)$ is given by

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

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

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

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

$$\mathcal{R}(d;\lambda) = \mathbb{E}_B[L(d,\theta;\lambda)] = (1-\lambda)DR(d) - \lambda\bar{\tau}(d). \tag{3}$$

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:

$$d_{ij} + d_{jk} \leq 1 + d_{ik},\ d_{ik} + (1 - d_{jk}) \leq 1 + d_{ij},\ \text{and}\ e_{ij} + e_{jk} \leq 1 + e_{ik}. \tag{4}$$

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

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

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. Grades are then reconstructed from the solution $\{d_{ij}^*, e_{ij}^*\}_{(i,j)\in[n]^2}$ as $d_i^* = 1 + \sum_{j\in[n]} d_{ji}^*$.

3.4 Discordance rates

The reliability of the optimal grades is summarized by the Discordance Rate $DR(d^*)$, which gives the posterior expected frequency of discordances between all pairs of firms. 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 $\{\pi_{ij}\}_{i\neq j}$.

It is also useful to consider pairwise Discordance Rates between specific pairs of grades g and $g' < g$, defined as

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

The denominator of each pairwise rate is interpretable as 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. The pairwise DRs are symmetric ($DR_{g,g'} = DR_{g',g}$), making it convenient to report them as a lower triangular matrix. The overall DR is a weighted average of the pairwise rates with positive weight put on the on-diagonal terms $DR_{g,g}$, which are necessarily zero.

3.5 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 again useful to consider the task of ranking a single pair in the context of equation (5), ignoring cross-pair constraints. From section 3.1, when facing a single pair, the risk minimizing decision rule is

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

Hence, with $\lambda = 1$, it is optimal to choose $d_{ij} = 1\{\pi_{ij} > 1/2\}$, which can be seen as a MAP estimate of the pairwise rank. 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, independent normal beliefs)**.** Suppose $n = 3$ and we believe that $\theta_i \sim \mathcal{N}(\omega_i, 1)$ for $i \in \{1, 2, 3\}$. Moreover, our beliefs are independent across firms, implying $B_{ij} = \mathcal{N}(\omega_i, 1) \times \mathcal{N}(\omega_j, 1)$. It follows that

$$\pi_{ij} = \Phi\left(\frac{\omega_i - \omega_j}{\sqrt{2}}\right).$$

Let $\lambda = 1/4$, which implies $(1+\lambda)^{-1} = 0.8$. If $(\omega_1, \omega_3) = (2, 0)$, so that $\pi_{13} = \Phi(\sqrt{2}) = 0.92$ and $\pi_{31} = 1 - \pi_{13} = 0.08$, then it is optimal to choose $d_1 > d_3$. But if $\omega_2 \in (0.81, 1.19)$, it is optimal to set $d_1 = d_2$ and $d_2 = d_3$ because $\max\{\pi_{12}, \pi_{23}\} < 0.8$. By transitivity, this implies $d_1 = d_3$ which contradicts our earlier assertion that $d_1 > d_3$. □

Note that if we had set $\lambda = 1$ in the above example transitivity would have been satisfied because the beliefs 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 beliefs in this example, revealed by the fact that $\pi_{ij} > \pi_{ji} \iff \omega_i > \omega_j$. Sobel (1993) establishes the transitivity of beliefs in a broader exponential family subject to a corresponding index restriction. In general, however, such index representations are not guaranteed 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, it is also worth noting 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 set $d_1 > d_3$, $d_1 > d_2$, and $d_2 = d_3$. Thus pairwise thresholding yields two grades and no transitivity violations. Whether the transitivity constraints bind therefore depends on the structure of the pairwise beliefs.

3.6 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 can be immediately applied.[2] For example, Borda (1784)'s voting method simply ranks each firm i based on its number of pairwise election wins; i.e., based upon $\sum_{j \neq i} 1\{\pi_{ij} > 1/2\}$. If (as we have assumed) B 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

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

[2] 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.

Integrating out θ and noting that $\pi_{ij} = 1 - \pi_{ji}$ for all $i \neq j$ yields

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

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 and independence of remote alternatives.

The summand $\left(2\pi_{ij} - 1\right)\left(d_{ji} - d_{ij}\right)$ in (7) is minimized by the pairwise MAP thresholding rule $d_{ij} = 1\{\pi_{ij} > 1/2\}$.[3] When $\lambda = 1$, the 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. Accordingly, we will refer to the grades generated by our procedure with $\lambda = 1$ as *Condorcet ranks*. 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 $\lambda(1+\lambda)^{-1} < \pi_{ij} < (1+\lambda)^{-1}$.

Condorcet rankings satisfy 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 B. 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 $\lambda < 1$.

[3]Note that pairwise MAP thresholding need not yield the most likely global ordering. For example, with 3 firms, the modal ordering is $\arg\max_{(i,j,k):i\neq j\neq k} \pi_{ijk}$, where $\pi_{ijk} = \int_{-\infty}^{\infty}\int_{-\infty}^{x}\int_{-\infty}^{y} dB(x,y,z)$. Suppose that $\pi_{123} = \pi_{132} = \pi_{231} = 0.11$ and $(\pi_{312}, \pi_{213}) = (0.37, 0.30)$. Here, the modal ordering is $\{3,1,2\}$. By the law of total probability $\pi_{ij} = \pi_{ijk} + \pi_{ikj} + \pi_{kij}$, which implies $(\pi_{12}, \pi_{23}, \pi_{13}) = (0.59, 0.52, 0.52)$. Hence, pairwise MAP thresholding yields the ordering $\{1,2,3\}$.

Proposition 2 (λ-Smith criterion)**.** Let $\mathcal{S}$ denote a collection of firms with 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 B.

4 Empirical Bayes

The previous section described a Bayesian approach to ranking units given beliefs B. Suppose for each firm $i \in [n]$, we have a consistent estimate $\hat{\theta}_i$ of θ_i along with that estimate's asymptotic standard error s_i. We will use these measurements to provide an objective grounding to our beliefs B. To do so, we introduce assumptions about the data generating process giving rise to the measurements.

Assumption 1 (Normal Noise)**.** $\hat{\theta}_i | \theta_i, s_i \sim \mathcal{N}(\theta_i, s_i^2)$ for each $i \in [n]$.

Assumption 1 stipulates that the estimation error $\hat{\theta}_i - \theta_i$ is normally distributed with known variance equal to s_i^2. This assumption can be justified by conventional asymptotic approximations. In our main application, the $\hat{\theta}_i$ and s_i are computed from a large number of job applications to each firm, making such approximations likely to be accurate.

Assumption 2 (Independent Noise)**.** The $\{\hat{\theta}_i\}_{i\in[n]}$ are mutually independent conditional on $\{\theta_i, s_i\}_{i\in[n]}$.

Assumption 2 posits that the statistical noise in the estimates is independent across firms. This assumption is sensible in our main application, where the estimate for each firm comes from a separate experiment. It is straightforward to relax this assumption when the covariance structure of the noise has a known low-dimensional structure but we do not pursue such an extension here.

Assumption 3 (Random Effects)**.** $\theta_i | s_i \overset{iid}{\sim} G$.

Assumption 3 models the θ_i parameters as random draws from a larger population of firms. By treating the sampling process as *iid* we abstract from the fact that a finite number of Fortune 500 firms could have been sampled in our experiment. An alternate interpretation of this assumption is that the experiment could have sampled a different set of jobs from the same firms, resulting in a new collection of θ_i's.[4]

The *mixing distribution* $G : \mathbb{R} \to [0, 1]$ characterizes the distribution of discriminatory conduct in the population of firms, which allows us to make probability statements about the latent θ_i parameters. By treating the θ_i as identically distributed conditional on the standard errors s_i, Assumption 3 rules out the possibility of dependence between latent effect sizes and precision of the estimates. This independence restriction may require a transformation of the parameters to be plausible. The framework outlined here applies after implementing such transformations, as we discuss further in the empirical work to follow.

4.1 Identification and estimation of G

Assumptions 1-3 imply that each observed $\hat{\theta}_i$ is the sum of a draw from G and a normally distributed error with variance s_i^2. By the law of total probability, we can write the conditional distribution of $\hat{\theta}_i$ given s_i as

$$\Pr\left(\hat{\theta}_i < t | s_i = s\right) = \int \Phi\left(\frac{t - x}{s}\right) dG\left(x\right) \equiv F(t|s),$$

where Φ denotes the standard normal CDF. This equation links the distribution F of point estimates to the distribution G of latent parameters. Given a consistent estimate $\hat{F}$ of F, this integral equation can be solved to recover an estimate $\hat{G}$ of the mixing distribution G. There are many proposals for solving deconvolution problems of this nature. The recent literature (Efron, 2016; Gu and Koenker, 2020) focuses primarily on maximum likelihood estimators, an approach we follow here.

4.2 EB posteriors and grades

The empirical Bayes approach treats the estimate $\hat{G}$ as a prior in decisionmaking. We can use this prior to form posterior beliefs over the θ_i's given the available evidence $\{\hat{\theta}_i, s_i\}_{i \in [n]}$. When $\hat{G}$ is close to the true G, the empirical Bayes posterior and resulting decision rules will approximate the beliefs and decisions of an oracle that knows the population distribution of discrimination.

[4] Appendix D of Kline, Rose and Walters (2022) expands on this interpretation.

By Bayes' rule

$$\Pr\left(\theta_i < t \mid \hat{\theta}_i = \hat{t}, s_i = s\right) = \frac{\int_{-\infty}^{t} \frac{1}{s}\phi\left(\frac{\hat{t}-x}{s}\right) dG\left(x\right)}{\int_{-\infty}^{\infty} \frac{1}{s}\phi\left(\frac{\hat{t}-x}{s}\right) dG\left(x\right)} \equiv \mathcal{P}(t|\hat{t}, s; G),$$

where ϕ denotes the standard normal density. The empirical Bayes posterior distribution for firm i is $\mathcal{P}(t|\hat{\theta}_i, s_i; \hat{G})$. By plugging in $\hat{G}$ for G we "borrow strength" from other observations when interpreting the evidence for firm i (Efron and Morris, 1973; Morris, 1983). As detailed in Appendix C, we construct bivariate empirical Bayes posteriors over pairs (θ_i, θ_j) to form empirical pairwise contrast probabilities $\hat{\pi}_{ij}$. Grades are then generated by minimizing (5) subject to (4), substituting $\hat{\pi}_{ij}$ for each π_{ij}. Appendix E demonstrates via simulation that the expected loss generated by making decisions based upon the EB posteriors comes very close to the loss expected from an oracle that knows G (and hence the "true" π_{ij}'s).

To interpret the posterior contrast probabilities, it is helpful to consider the following hypothetical thought experiment. Imagine replicating our correspondence experiment an infinite number of times, in each instance drawing conduct parameters from the distribution G and noise according to Assumptions 1-2. Each π_{ij} gives the share of new firm pairs with realized evidence configuration $(\hat{\theta}_i, \hat{\theta}_j, s_i, s_j)$ among which the event $\theta_i > \theta_j$ occurs. In contrast, a frequentist test would consider the likelihood of the observed evidence in repeated draws of the noise conditional on the conduct parameters of the firms under study. While frequentist p-values allow retrospective assessments of null hypotheses, our EB estimates of the π_{ij}'s offer a best guess of what to expect when reporting any set of grades.

4.3 Reliability of grades

To summarize the reliability of the grades we report estimates of the Discordance Rate $DR(d^*)$, replacing the posterior contrast probabilities $\{\pi_{ij}\}_{i \neq j}$ in (2) with their EB analogues $\{\hat{\pi}_{ij}\}_{i \neq j}$. Likewise, $\bar{\tau}(d^*)$ is estimated by plugging in the relevant posterior contrast probabilities to arrive at a posterior mean estimate of the rank correlation between the assigned grades and the true ranks.

As noted earlier, the Discordance Rate gives the expected frequency of discordances between pairs of firms. Assumptions 1-3 clarify that in our EB framework, this expectation averages over both draws of firm-specific parameters from G and draws of the normal noise in each firm's estimate. Hence, the Discordance Rate answers the following question: if we were to rerun the entire experiment – sampling a new set of jobs from the same population G – and we happened to get the same collection of point estimates and standard errors, how many grading mistakes should we expect to make? The EB estimate

of $DR(d^*)$, which substitutes the estimated $\hat{G}$ for the unknown G, therefore provides an assessment of *average* grade reliability across experiments like ours.

The dependence of the optimized empirical grades on the $\{\hat{\pi}_{ij}\}_{i \neq j}$ generates a finite sample bias attributable to estimation error in $\hat{G}$. This bias will tend to yield overly optimistic assessments of both $DR(d^*)$ and $\bar{\tau}(d^*)$ when $\hat{G}$ is poorly estimated. We explore this issue further in Appendix E, finding in a Monte Carlo simulation calibrated to our leading application that these finite sample biases are small.

5 Ranking names

As an introductory illustration of the methods developed thus far, we now rank the employer contact rates 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 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. Column (4) displays test statistics and p-values from Wald tests of the hypothesis that contact rates are equal within each race and sex group. We cannot reject the null hypothesis that names with the same nominal race and sex are treated equally by employers ($p \geq 0.24$). Consistent with these test results, the bottom rows of Table 1 reveal that a bias-corrected estimate of the total variance in contact rates across names is approximately equal to the between-group variance explained by race and sex.[5] These findings indicate that employers treat names with the same nominal race and sex equally.

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 the socioeconomic status 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 maternal education of the first names

[5]The between-group variance is computed with 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. Applying this formula to the race and sex groups yields a variance of $(0.011)^2$, while applying it to the full set of name-specific contact rates produces a variance of $(0.010)^2$. It is, of course, logically impossible for the between-group variance to exceed the total variance across names, but this logical constraint is not imposed on the unbiased variance estimators used here.

considered in their experiment varied widely within race but was insignificantly related to contact rates.[6] 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 grading 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.

5.1 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 grades – that is, those with p_i closest to $1/2$ – will be estimated with the most noise, leading to a violation of Assumption 3.

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 $\theta_i = \sin^{-1}\sqrt{p_i}$ and the variance $(4N_i)^{-1}$ no longer depends on θ_i.

To estimate the distribution G of θ_i we first apply a non-parametric maximum likelihood (NPMLE) estimator (Koenker and Mizera, 2014; Koenker and Gu, 2017). The NPMLE estimates a discrete approximation to G 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 ($p = 0.17$).[7]

[6] Recent work by Crabtree et al. (2022) directly elicits perceptions of educational attainment and income by first name on a variety of online platforms. This study finds that the extent of variation in perceptions of social class across racially distinctive first names in the same race category is comparable to the variability between race categories (see their Figure 4).

[7] 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

A plot of the estimated marginal distribution $\hat{G}$ produced by the NPMLE appears in Figure 1. The bars correspond to histograms of $\hat{\theta}_i$ while the green spikes represent the estimated probability mass function $d\hat{G}$ of 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 the sample variance of $\hat{\theta}_i$ estimates minus the 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 the log density of the mixing distribution with a natural cubic spline with five knots. 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 to yield a $\hat{G}$ whose mean and variance comes as close as possible to the sample mean of the $\{\hat{\theta}_i\}_{i\in[n]}$ and their debiased variance estimate, as described further in Appendix D.

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. The two modes of the mixing distributions produced by both the NPMLE and log-spline approaches fall near the race-specific mean contact rates even though the race labels were not used in estimation.

The lower panel of Figure 1 converts these estimates back into probability points via the inverse transform $p_i = \sin(\theta_i)^2$. The NPMLE finds two large mass points at $p_i = 0.226$ and $p_i = 0.244$. 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 $p_i = 0.260$, which lies just above the estimated average contact rate for distinctively white female names.

The discrete $\hat{G}$ produced by the NPMLE is a data-dependent approximation to the mixing distribution. Even if differences in the treatment of names are driven primarily by employer perceptions of 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 what follows, we rely on the log-spline estimate of G which (as in the theoretical analysis of Section 3) implies that ties are measure zero.

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.

5.2 Reporting possibilities

The top left panel of Figure 2 depicts the EB posterior contrast probabilities $\hat{\pi}_{ij}$ (see the Appendix for computational details). Names are 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 tend to have $\hat{\pi}_{ij}$'s near 1/2, indicating little confidence in their relative order. We have accordingly defined each diagonal entry π_{ii} (quantities that are not used elsewhere) equal to 1/2 as a convention. Reassuringly, name pairs with distant ranks are associated with $\hat{\pi}_{ij}$'s near 0 or 1, implying that the experimental data are highly informative about the relative orderings 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 point representing each solution reports the number of distinct grades for that choice of λ. A sharp elbow emerges around $\lambda = 0.18$, above which the DR grows rapidly. The DR increases with λ even when the number of grades is constant because the set of firms assigned each grade has changed.

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 very 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 an estimated Discordance Rate of 0.28. For comparison, we also show the results of naively ranking based on $\hat{\theta}_i$ or the EB posterior mean $\bar{\theta}_i = \int_{-\infty}^{\infty} t d\mathcal{P}(t|\hat{\theta}_i, s_i; \hat{G})$. Remarkably, both naive approaches yield ranks with $\bar{\tau}$ and DR similar to those produced by our report card procedure when $\lambda = 1$. Essentially the same outcome results from a James-Stein type linear shrinkage estimator nominally predicated on normality of G.[8] Breaking the posterior mean $\bar{\theta}_i$ into quartiles or deciles yields results similar to setting $\lambda < 1$.

To improve on the reliability of the Condorcet grades, we set $\lambda = 0.25$, implying via equation (6) that, in the absence of transitivity considerations, we would abstain from strictly ranking pairs with posterior certainty less than 80%. This choice yields two grades that are both highly informative ($\bar{\tau} = 0.29$) and reliable ($DR = 0.07$). For comparison, lowering the implicit posterior threshold to 70% by setting $\lambda = 0.41$ would yield three grades and increase the estimated $\bar{\tau}$ by 11% (to $\bar{\tau} = 0.32$) at the expense of a 21% increase in the estimated DR. Conversely, requiring $\lambda < 0.18$ would generate only one grade, yielding both $\bar{\tau}$ and DR of zero by construction.

[8] The linear shrinkage estimator of θ_i can be written $\bar{\theta}_{i,lin} = \bar{\theta} + \frac{\hat{V}}{\hat{V}+s_i^2}\left(\hat{\theta}_i - \bar{\theta}\right)$, where $\bar{\theta} = n^{-1}\sum_{i\in[n]}\hat{\theta}_i$ and $\hat{V} = (n-1)^{-1}\sum_{i\in[n]}\left(\hat{\theta}_i - \bar{\theta}\right) - n^{-1}\sum_{i\in[n]} s_i^2$.

5.3 Grades and demographics

Figure 3 lists the first names according to their Condorcet ranks, 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. Approximately 72 (i.e., 95%) of these 76 intervals should be expected to contain their name's true latent contact rate.[9] 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.

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 ranks. In practice, however, names' Condorcet rankings are very nearly monotone in their posterior means. An exception is found in the name "Latoya" which exhibits a higher posterior mean, but a lower Condorcet rank, than the name "Maurice." This rank reversal reflects the greater posterior uncertainty associated with the name Latoya, which is evident in the name's wider credible interval. All else equal, a name whose posterior distribution is highly diffuse will tend to receive a middling rank.

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 maximize the expected rank correlation with the latent θ_i ranks. The coarse ranks that emerge when $\lambda < 1$ sacrifice rank correlation in exchange for fewer mistakes. Each name's color reflects its assigned grade. Appendix Figure F1 shows how these grades vary with name-specific contact rates and standard errors. As expected, names with higher sample contact rates tend to earn the top grade $\star\star$. However, heteroscedasticity in the estimates prevents the grades from being characterized by a single cutoff contact rate.

Though we estimated 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 C, 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

[9]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 $\theta \mapsto \sin(\theta)^2$.

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 F2 reports the pseudo-R^2 (McFadden, 1974) 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.

These findings demonstrate that our grades are strong predictors of a name's race but not its sex. Given that the overall gender gap in contact rates is statistically insignificant in our experiment, the failure to predict gender is not 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.

6 Ranking racial contact gaps

We turn now to ranking firms in their relative treatment of Black versus white names. We begin by defining a firm-specific bias measure θ_i that is scale invariant and then develop a statistical model of the dependence between θ_i and s_i that suggests a transformation of the data for which the precision independence requirement of Assumption 3 holds. Unlike in our study of names, this transformation takes the form of a residualization of $\hat{\theta}_i$ against s_i. We then deconvolve this estimated residual and study the reporting possibilities associated with grades based on the estimated distribution of contact gaps.

6.1 Defining θ_i

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.[10]

While our past work probed for discrimination by estimating the levels gap $p_{iw} - p_{ib}$, this measure is not ideal for ranking firm conduct as level gaps will mechanically be smaller for firms that contact fewer applications overall. To mitigate the influence of variation in overall contact rates on our measure of discrimination, we focus on the proportional bias against Black names at firm i:

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

which has the advantage of being scale invariant. We estimate θ_i with 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.[11]

In what follows, we exclude the eleven firms in the 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 favors white names by roughly 10%. Detailed point estimates and uncertainty measures for all 97 firms used in our analysis are provided in Appendix F5.

Twenty-one of the 97 estimated contact gaps are negative, indicating a preference for distinctively Black names. The firm-specific estimates are noisy, however, with 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 F3 plots $\hat{\theta}_i$

[10] To account for the fact that some job vacancies closed before we were able to send all four pairs of applications, we weight the sample contact rates inversely by the number of applications sent to each job. This weighting amounts to first computing the average contact rate at each job, then taking an unweighted average across jobs.

[11] 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 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 $\mathbb{V}\left[\hat{\theta}_i\right] \approx \frac{\mathbb{V}[\hat{p}_{iw}]}{p_{iw}^2} + \frac{\mathbb{V}[\hat{p}_{ib}]}{p_{ib}^2} - 2\frac{\mathbb{C}[\hat{p}_{iw}, \hat{p}_{ib}]}{p_{iw} p_{ib}}$.

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 ($p < 0.001$).

6.2 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:

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

The parameter β governs how the conditional distribution of bias varies with the standard error s_i. When β is positive, both the mean and variance of θ_i increase monotonically with s_i. The latent variable v_i captures heterogeneity in discrimination among firms with similar standard errors. We assume v_i is fully independent of s_i and follows a distribution $G_v : \mathbb{R}_+ \to [0, 1]$ with strictly positive support. In the framework of Section 4, this restriction replaces Assumption 3, or equivalently, suggests that it applies to the transformation θ_i / s_i^{β}.

To evaluate the plausibility of the model in equation (8), we scrutinize some of the moment conditions it implies. Letting $\mathbb{E}[v_i|s_i] = \mu > 0$ and $\mathbb{V}[v_i|s_i] = \sigma_v^2 > 0$, consider the following "studentized" version of $\hat{\theta}_i$:

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

Maintaining Assumptions 1 and 2, each estimate $\hat{\theta}_i$ is presumed to be centered at the true θ_i and normally distributed with variances given by s_i^2. Consequently, the model in (8) restricts T_i to have mean zero and variance one conditional on s_i. These restrictions, in turn, 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 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 T_i has 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 the conditional mean of θ_i is roughly proportional to $\sqrt{s_i}$. The large estimated value of σ_v reveals that discrimination varies substantially among firms with similar standard errors.

The top panel of Appendix Figure F3 superimposes the estimated conditional expectation function $\hat{\mathbb{E}}[\theta_i|s_i] = s_i^{\hat{\beta}}\hat{\mu}$ 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 F3 plots values of the estimated residual $\hat{T}_i = \dfrac{\hat{\theta}_i - s_i^{\hat{\beta}}\hat{\mu}}{\sqrt{s_i^{2\hat{\beta}}\hat{\sigma}_v^2 + s_i^2}}$ against s_i. In line with our model, $\hat{T}_i$ exhibits roughly constant variance and a mean near zero throughout the observed range of s_i.

6.3 Estimating G

To estimate the distribution function G_v, we deconvolve the residual $\hat{v}_i = \hat{\theta}_i / s_i^{\hat{\beta}}$. Assumption 1, in conjunction with Slutsky's Theorem, implies the following large-n approximation to the distribution of this residual:

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

Relying again on a variant of Efron (2016)'s log-spline estimator, we parametrize G_v as a natural cubic spline with five knots and 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 $(\hat{\mu}, \hat{\sigma}_v^2)$ 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 $\hat{G} : x \mapsto n^{-1} \sum_i \hat{G}_v(x/s_i^{\hat{\beta}})$ of the distribution of contact gaps.

The upper left panel of Figure 4 plots the log-spline estimate $\hat{G}_v$ overlaid against the histogram of $\hat{v}_i$. $\hat{G}_v$ is less dispersed than the histogram, reflecting the noise in the estimates. The upper right panel plots the corresponding estimate of $\hat{G}$ against the histogram of contact gap estimates $\{\hat{\theta}_i\}_{i=1}^n$. Unlike with our earlier analysis of names, the density $\hat{G}$ is unimodal but 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, peaking at values indicating modest bias against Black names. The NPMLE estimate of the variance of the θ_i's departs somewhat from both the log-spline estimate and the 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, with the NPMLE assigning little mass to negative values of θ_i. Since a discrete distribution with exact ties seems implausible, we rely again on the log-spline estimates in what follows. The EB posterior distribution 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.

6.4 Industry effects

In Kline, Rose and Walters (2022) we found large differences in the magnitude of contact gaps across 2-digit industries. Appendix Table F3 provides an updated list of 19 industry groupings designed to ensure that at least three of the 97 firms studied in this paper are present in each group.[12] Industries are assigned using the SIC codes of establishments reported in the 2019 InfoGroup Historical Datafiles (InfoGroup, 2019). In cases where firms operate in multiple industries, codes are assigned to best match the jobs sampled in the experiment.

Many of these industries have only 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:

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

$$\xi_i \mid s_i, \eta_{\mathrm{k}(i)} \overset{iid}{\sim} G_\xi, \quad i \in \{1, ..., n\}, \quad \eta_k \mid \mathbf{s}_k \overset{iid}{\sim} G_\eta, \quad k \in \{1, ..., K\},$$

where the function $\mathrm{k} : \{1, \ldots, n\} \to \{1, \ldots, K\}$ returns a firm's industry, $\mathbf{s}_k$ is the vector of standard errors for all firms with $\mathrm{k}(i) = k$, and the distribution functions $G_\eta : \mathbb{R}_+ \to [0, 1]$ and $G_\xi : \mathbb{R}_+ \to [0, 1]$ have strictly positive support. This hierarchical specification relaxes the *iid* restriction in Assumption 3: the industry effect $\eta_{\mathrm{k}(i)}$ captures correlation in discrimination among firms in the same industry, while the firm effect ξ_i captures departures from the industry average. These two effects are independent, both of precision levels and each other, implying the marginal distribution of v_i can be written $G_v : x \mapsto \int_0^\infty G_\xi\left(x/z\right) dG_\eta\left(z\right)$. We normalize $\mathbb{E}[\eta_k] = 1$, which implies $\mathbb{E}[\xi_i] = \mu$.

The marginal variance of v_i in this model is $\sigma_v^2 = \sigma_\eta^2\sigma_\xi^2 + \sigma_\eta^2\mu^2 + \sigma_\xi^2$, 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

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

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 $V_k \equiv \left(\sigma_\eta^2\sigma_\xi^2/n_k + \sigma_\eta^2\mu_v^2 + \sigma_\xi^2/n_k\right) + n_k^{-2}\sum_{i:\mathrm{k}(i)=k} s_i^{2(1-\beta)}$. Letting $\bar{s}_k = n_k^{-1}\sum_{i:\mathrm{k}(i)=k} s_i$ denote the average standard error in industry k, our two new moment

[12] The industry definitions from Kline, Rose and Walters (2022) yield 24 industry codes, three of which contain only one of our 97 firms. Report cards based on these legacy definitions are provided in the Appendix and an earlier version of this paper (Kline, Rose and Walters, 2023).

conditions can be written

$$\mathbb{E}\left[(\bar{v}_k - \mu_v)^2 - V_k\right] = 0, \quad \mathbb{E}\left[\left\{(\bar{v}_k - \mu_v)^2 - V_k\right\}\bar{s}_k\right] = 0. \tag{10}$$

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 more than 20 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 one third of the marginal variance in v_i stems from within industry variation.[13]

To identify the marginal distribution of θ_i, we assume that both G_η and G_ξ belong to the exponential family with log density parameterized by a five-knot natural cubic spline. Generalizing Efron (2016)'s log-spline estimator to the hierarchical case, these distributions are estimated by penalized maximum likelihood (see Appendix D 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 bottom left panel of Figure 4. Table F4 reports moments of the within- and between-industry distributions implied by the log-spline estimates as well as moments of the overall contact ratio $\theta_i = s_i^\beta \eta_{\mathrm{k}(i)} \xi_i$. The mean contact gap, between-industry standard deviation, and total standard deviation reported in Table F4 closely match the corresponding GMM estimates of these parameters in Table 3.

As can be seen in Figure 4, 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 right panel of Figure 4 shows that the implied distribution of θ_i is similar to the estimate from the model without industry effects in the top right panel, 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.

[13]The within industry variance is $\mathbb{E}[\mathbb{V}[v_i|\eta_{\mathrm{k}(i)}]] = \mathbb{E}[\sigma_\xi^2 \eta_{\mathrm{k}(i)}^2] = \sigma_\xi^2 \mathbb{E}[\eta_{\mathrm{k}(i)}^2] = \sigma_\xi^2 \left(\sigma_\eta^2 + 1\right)$. Hence, the within industry variance share evaluates to $(\sigma_\eta^2 + 1)\sigma_\xi^2/\sigma_v^2$.

6.5 Reporting possibilities

Figure 5 plots the pairwise posterior ranking probabilities $\hat{\pi}_{ij}$ 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_i. Firms of middling rank, on the other hand, are more difficult to distinguish from others. Including industry effects tightens the posteriors, which leads the $\hat{\pi}_{ij}$'s to become more dispersed around 1/2.

The pairwise probabilities that satisfy the naive thresholding rule $\hat{\pi}_{ij} > (1+\lambda)^{-1}$ when λ has been set to 0.25 have been bordered in red. 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.

Panel (a) of Figure 6 plots the number of distinct grades that result from minimizing our estimate of $\mathcal{R}(d;\lambda)$ 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 increases the DR to 5.6%.

Panel (b) of Figure 6 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 0.59 and lowers their DR to 0.20. 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.[14] Coarsening the posterior mean into deciles or quartiles lowers the DR somewhat, but at the cost of excessively large reductions in $\bar{\tau}$. We also report the results of ranking based upon linear shrinkage estimators in the

[14]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 $\hat{G}$.

James-Stein tradition. These ranks perform substantially worse than naively ranking the point estimates $\hat{\theta}_i$. This poor performance is an artifact of our earlier finding that more precise estimates tend to exhibit less bias, which suggests the noisiest estimates should be shrunk the least.

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. Resolving transitivity 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 5.6% in the hierarchical specification with industry effects. Fortunately, the resulting grades remain highly informative: $\bar{\tau}$ is 0.21 in our baseline specification and 0.46 when industry effects are included.

7 Racial discrimination report cards

Figure 7 provides a concise, low-dimensional summary of differences in racial discrimination across firms. This report card is based on the baseline specification without industry effects. The firms are ordered by their Condorcet ranks (i.e., their grades under $\lambda = 1$). 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.[15]

In addition to the report card grades, the Figure plots an empirical Bayes posterior mean estimate of each firm's bias θ_i. To arrive at these posterior means, the EB model effectively shrinks each point estimate towards the average $\hat{\theta}_i$ of firms with similar standard errors (see Appendix Figure F3). Bracketing the posterior mean estimates are EB 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.

Setting $\lambda = 0.25$ generates a report card with three grades, represented in Figure 7 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 the two firms with the worst Condorcet ranks, Genuine Parts (Napa Auto) and AutoNation, are assigned the grade of $\star$, suggesting they are the heaviest discriminators. Fourteen firms are assigned the score of $\star\star\star$, which indicates that this group is the least-biased against Black applicants. The firm receiving the best Condorcet rank is Charter/Spectrum.

While the Condorcet ranks, the ranks of the posterior means, and the ranks of the

[15]Contractor status as of September 2020 was obtained via a Freedom of Information Act request to OFCCP.

bias estimates are highly correlated, this correlation is not perfect. For example, Genuine Parts has the sixth largest proportional contact gap estimate (see Appendix Table F5 for the complete list) but is assigned a Condorcet rank of 1 and the largest posterior mean. In contrast, AutoNation has the largest proportional contact gap estimate but a Condorcet rank of 2 and the second largest posterior mean. This rank reversal reflects that AutoNation has a larger standard error than Genuine Parts, which leads to more shrinkage of its point estimate towards the center of the distribution.

Appendix Figure F5 depicts the relationship between report card grades and firm-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 F7 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 estimate 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.

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 suggest that two of the five 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.

7.1 Industry effects

Figure 8 displays a racial discrimination report card based on the model with industry effects. Each firm's industry code is listed in parentheses next to its name. 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 nine, while seventeen firms are now assigned the second-worst grade $\star\star$. Eleven firms are assigned the best grade of $\star\star\star\star$. Appendix Figure F8 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 0.23. 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. Appendix Figure F11 shows an alternate grading based upon the industry codes utilized in Kline, Rose and Walters (2022). Encouragingly, the results are broadly similar, though fewer firms are assigned the worst grade because that grouping of industries is a less powerful predictor of firm conduct.

The small number of grades generated by our grading procedure explain a substantial portion of the total variance 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 7, implying an R^2 of roughly 25%. Adding industry boosts the R^2 to 70%. In other words, the four categories displayed in Figure 8 explain more than two thirds 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 9 plots posterior estimates of industry mean contact penalties

$\eta_k n_k^{-1} \sum_{i:\mathrm{k}(i)=k} s_i^{\hat{\beta}}$. The industry with the greatest estimated bias against Black names is SIC 55, "Auto dealers / services / parts," with a posterior mean contact penalty of 22%, while the industry with the smallest estimated bias is SIC 54, "Food Stores," which has a posterior mean of roughly 5%. In an industry grading scheme with $\lambda = 0.25$ (which yields four total grades), SIC 55 and SIC 59 ("Other retail") are assigned the worst grade of $\star$. SIC 56 ("Apparel Stores") receives the unique grade of $\star\star$. A group of eleven industries receives the best grade of $\star\star\star\star$ and exhibits an average posterior mean contact gap of roughly 6%. 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.

These substantial industry differences explain the more informative firm-level posteriors generated by the report card incorporating industry effects. For example, Disney has a negative point estimate (-0.12) but a large standard error (0.24), leading to an intermediate classification of $\star\star$ in the baseline report card in Figure 7. Disney's industry classification is SIC 59 because the Disney jobs in our sample are primarily at retail stores. Due to the substantial discrimination in this industry depicted in Figure 9, the report card with industry effects places Disney in the most discriminatory category of $\star$ (see Figure 8). This change reflects the strong within-industry correlation in conduct present in our data, which leads to substantial weight on the industry average for firms with noisy contact gap estimates. While such industry-based shrinkage will tend to increase the accuracy of grades and posterior mean predictions on average, it may worsen predictions for firms that are atypical of the industries in which they operate.

7.2 Misclassification

Figure 10 assesses the reliability of report card grades by reporting the lower-triangular matrix of estimated between-grade Discordance Rates in our baseline model that omits industry effects. Panel (a) reveals that 11% of the firm comparisons across grades $\star$ and $\star\star$ are expected to be misordered. The DR 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 estimated DR's between 11 and 14%, while the discordance rate for non-adjacent grades is estimated to be only 0.8%.

Panel (b) of Figure 10 summarizes the reliability of the grades obtained when conditioning on industry effects. Discordance Rates between adjacent grades are estimated to range from 11% to 17%. DR's for grades separated by two categories are estimated to fall below 3%, and the estimated DR between the worst grade ($\star$) and the best grade ($\star\star\star\star$) is 0.4%. These findings suggest that a comparison of the best- and worst-performers in Figure 8 isolates firms with large differences in discriminatory conduct while yielding few misclassifications.

8 Ranking gender contact gaps

We turn now to studying firms' gender preferences. Though gender does not seem to be an important aspect of the average treatment of names, the firms in our experiment vary enormously in their propensity to contact names of different genders, with some firms preferring women and others preferring men. In what follows, we build a statistical model of this "bidirectional" discrimination and study the reporting possibilities offered by our ranking procedure.

8.1 Defining θ_i

Paralleling our analysis of race, gender contact gaps are defined proportionally as $\theta_i = \ln p_{im} - \ln p_{if}$, where p_{im} and p_{if} refer to average contact rates for male and female names at firm i. These gender gaps are estimated by plugging in sample contact rates $\hat{p}_{im}$ and $\hat{p}_{if}$ to form the estimator $\hat{\theta}_i = \ln \hat{p}_{im} - \ln \hat{p}_{if}$.

As Table 2 reveals, the mean value of $\hat{\theta}_i$ is nearly zero. However, the bias-corrected standard deviation of gender contact gaps is 0.194, nearly three times the corresponding estimate for race (0.069). A zero average gender gap coupled with substantial dispersion across firms implies that some firms favor male applications, while others favor female applications. This finding is consistent both with our past analysis of levels gaps in this experiment (Kline, Rose and Walters, 2022) and analysis of other correspondence experiments (Kline and Walters, 2021; Schaerer et al., 2023).

8.2 A model of precision dependence

Inspection of the relationship between $\hat{\theta}_i$ and s_i (depicted in Appendix Figure F4) suggests the variance, but not the level, of θ_i depends on s_i. Accordingly, we work with a linear model taking the form:

$$\theta_i = \mu + s_i^{\beta} v_i, \quad v_i \mid s_i \sim G_v,$$

where the distribution function $G_v : \mathbb{R} \to [0, 1]$ has unrestricted support, the constant μ measures average gender bias in the population, $\mathbb{E}[v_i|s_i] = 0$, and $\mathbb{V}[v_i|s_i] = \sigma_v^2 > 0$. Note that this specification is essentially a recentered version of (8) that allows θ_i to take on negative values. Defining the relevant studentized contact gap measure as $T_i = \dfrac{\hat{\theta}_i - \mu}{\sqrt{s_i^{2\beta}\sigma_v^2 + s_i^2}}$, the parameters (β, μ, σ_v) are estimated by GMM using the moment conditions in (9).

The GMM estimates are reported in the third column of Table 3. The parameter μ is statistically indistinguishable from zero, suggesting that the average firm treats male and female names equally. The parameter β is estimated to exceed one and is easily

distinguishable from zero, indicating that more precise estimates are associated with gender bias of smaller absolute magnitude. The estimated value of σ_v implies a standard deviation of θ_i of 1.83 percentage points, which is very close to the aforementioned estimate of 1.94 percentage points obtained by debiasing the sample variance. Our model provides an excellent fit to the data: the GMM J statistic is below its expected value and the scatterplot of $\hat{T}_i$ against s_i (depicted in the bottom panel of Appendix Figure F4) is homoscedastic and centered around zero.

8.3 Estimating G

We use the GMM parameter estimates $(\hat{\mu}, \hat{\beta})$ to form the residual $\hat{v}_i = (\hat{\theta}_i - \hat{\mu})/s_i^{\hat{\beta}}$. Appealing again to Slutsky's Theorem, we assume $\hat{v}_i \mid v_i, s_i \sim \mathcal{N}(v_i, s_i^{2(1-\beta)})$. As in our analysis of racial gaps, we estimate the distribution G of proportional contact gaps by first deconvolving $\hat{v}_i$ using the log-spline estimator to obtain the estimated distribution $\hat{G}_v$. We then estimate G with $\hat{G} : x \mapsto n^{-1} \sum_i \hat{G}_v((x - \hat{\mu})/s_i^{\hat{\beta}})$.

The results are shown in the upper panel of Figure 11. The estimated density of θ_i is peaked near zero indicating most firms have very weak gender preferences. However, the heavy tails suggest a small minority of firms have strong gender preferences. For comparison we show NPMLE estimates derived from the GLVmix procedure of Koenker and Gu (2017), which assumes θ_i is independent of sample size N_i. Reassuringly, the NPMLE estimates of G align closely with the log-spline estimates.

8.4 Industry effects

To allow for industry effects, we decompose v_i into additively separable industry and firm components:

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

$$\xi_i \mid s_i, \eta_{\mathrm{k}(i)} \overset{iid}{\sim} G_\xi, \quad i \in \{1, ..., n\}, \quad \eta_k \mid \mathbf{s}_k \overset{iid}{\sim} G_\eta, \quad k \in \{1, ..., K\},$$

where $\eta_{\mathrm{k}(i)}$ is a mean zero industry effect with variance σ_η^2, ξ_i is a mean zero firm effect with variance σ_ξ^2, and the distribution functions $G_\xi : \mathbb{R} \to [0, 1]$ and $G_\eta : \mathbb{R} \to [0, 1]$ have unrestricted support. We assume these components, and therefore v_i itself, are fully independent of s_i. Letting $\bar{v}_k = n_k^{-1} \sum_{i:\mathrm{k}(i)=k} \hat{v}_i$ the model implies $\mathbb{V}[\hat{v}_i] = \sigma_\eta^2 + n_k^{-1}\sigma_\xi^2 + n_k^{-2} \sum_{i:\mathrm{k}(i)=k} s_i^{2(1-\beta)} \equiv V_k$.

Using these definitions, we add the moment conditions in (10) to our GMM system, which yields estimates of the variance components $(\sigma_\eta, \sigma_\xi)$. The fourth column of Table 3 reveals that this hierarchical model fits well, again yielding a J-statistic below its expected value. The estimated marginal distribution of θ_i suggests the average firm in our experiment has a gender bias of exactly zero. The standard deviation of θ_i is roughly

15 percentage points. Between-industry variation is estimated to account for nearly half of the variation in proportional gender contact gaps.

As in our analysis of race gaps, we estimate the distributions G_ξ and G_η with a hierarchical generalization of Efron (2016)'s log-spline procedure. The resulting densities are shown in the lower panel of Figure 11. Both the within and between industry components exhibit substantial variability but are not especially peaked near zero. However, the implied marginal distribution of gender bias closely matches that produced by the baseline model that ignores industry affiliation. The mean and variance implied by this density are close to simple unbiased estimators of these moments.

8.5 Reporting possibilities

Figure 12 plots the pairwise posterior ranking probabilities $\hat{\pi}_{ij}$ for gender, with firms ordered by their rank under $\lambda = 1$. While substantial information is available regarding relative ranks, pairwise thresholding with $\lambda = 0.25$ would again yield numerous transitivity violations. Imposing transitivity yields four grades, with a large middle category of $\star\star\star$. The hierarchical model with industry effects yields starker posterior contrasts. Yet pairwise thresholding continues to yield rampant transitivity violations. Imposing transitivity yields five grades.

Figure 13 shows the tradeoffs between DR and $\bar{\tau}$ estimated to arise when ranking firms' gender preferences. Setting $\lambda = 0.25$ yields an estimated Discordance rate of roughly 2% in our baseline specification and roughly 1% when including industry effects. The bottom panel of the figure reveals that the Condorcet grades obtained by setting $\lambda = 1$ would be very informative about relative gender discrimination, yielding a rank correlation with the underlying firm discrimination parameters in excess of 0.4 regardless of whether industry affiliation is taken into account. The Condorcet grades are not particularly reliable, however, yielding Discordance rates approaching 30%.

As was the case with race, ranking gender bias based upon posterior means results in grades with informativeness and reliability similar to the Condorcet ranks. Unlike with race, we also obtain similar gender results when naively ranking firms based upon their unadjusted point estimates $\hat{\theta}_i$. This finding is a reflection of the kurtosis in the distribution of gender contact gaps, which suggests that when firms have gender preferences, those preferences manifest in large point estimates, making it easy to distinguish such firms from their gender-neutral counterparts. Ranking on linear shrinkage estimates performs more poorly than ranking on point estimates, which owes again to the fat tails of G and the fact that standard James-Stein type estimators ignore dependence of the mixing distribution on precision. As with race, ad hoc coarsenings of point estimates into deciles or quartiles lie well within the reporting possibilities frontier, indicating they are dominated by our grading procedure.

The reporting frontier for the model with industry effects lies only slightly above that of the baseline model. However, the grades produced when setting $\lambda = 0.25$ are substantially more informative with industry effects ($\bar{\tau} = 0.16$ vs 0.12) while being slightly more reliable ($DR = 0.01$ vs $DR = 0.018$).

9 Gender discrimination report cards

Figure 14 provides a report card for gender discrimination using the same rubric as was used for race: firms are sorted by their Condorcet ranks and posterior means $\bar{\theta}_i$ are listed along with credible intervals. Here, the posterior means shrink point estimates towards zero, with substantially greater shrinkage factors for less precise observations (see Appendix Figure F4). Consistent with the estimated distribution of θ_i reported in Figure 11, the posterior means suggest most firms have negligible gender preferences. However, firms with the highest Condorcet ranks (e.g., Builders FirstSource and LKQ Auto) are estimated to strongly prefer male applicants, while firms with the lowest Condorcet ranks (e.g., Ascena and Nationwide) are estimated to strongly prefer female applicants.

Unlike in our previous examples, the grades that emerge when setting $\lambda = 0.25$ are not a strict coarsening of the Condorcet ranks. Two firms—State Farm and Aramark—whose Condorcet ranks suggest bias against male applicants, receive a middling grade of $\star\star\star$ as a result of the relative imprecision of their estimates. Appendix Figure F6 depicts the relationship between the gender report card grades and firm-specific contact gap estimates and standard errors. As was the case with race, classification boundaries are mildly nonlinear in $(\hat{\theta}_i, s_i)$ space, with large standard errors tending to yield mediocre grades. Firms assigned the grade $\star\star\star$ are estimated to exhibit negligible gender preferences, with an average posterior mean of -0.01. The four depicted grades are estimated to explain 44% of the variation in proportional contact gaps.

Six firms (Builders Firstsource, LKQ Auto, State Farm, CBRE, Nationwide, and Ascena) have absolute gender bias estimates well above the 4/5th's rule standard. The firm VFC, while receiving the grade $\star\star\star$, exhibits a posterior mean just below this threshold. Three of the four firms that received grades of $\star$ or $\star\star$, indicating a preference for male names, are federal contractors. Two of the four firms graded as $\star\star\star\star$, indicating a preference for female names, are federal contractors.

9.1 Industry effects

Figure 15 updates the gender report card to account for industry affiliation. The Condorcet ranks that result from the model with industry effects are very similar to those produced by the baseline model reported in Figure 14. Seven firms (Builders Firstsource, LKQ Auto, Victoria's Secret, Gap, Foot Locker, VFC, and Ascena) with extremal Con-

dorcet ranks have posterior mean biases exceeding the 4/5th's rule standard.

Setting $\lambda = 0.25$ yields five grades that are a strict coarsening of the Condorcet ranks. Appendix Figure F10 lists the grades that result from all possible choices of λ. The depicted grades with $\lambda = 0.25$ are estimated to explain 38% of the variation in proportional gender contact gaps. Builders FirstSource is the only firm to receive a grade of $\star$: its posterior mean suggests a bias against distinctively female names of 67%. The grade $\star\star\star$ is comprised of firms with roughly gender neutral conduct, with an average posterior mean θ_i of 0.005. In contrast, the average posterior bias against male names among firms assigned the grade $\star\star\star\star\star$ is -40%. Appendix Figure F12 reports an alternate grading based upon the industry codes used in Kline, Rose and Walters (2022). Those codes, which are less informative, yield only three grades but lead similar firms to be assigned to categories indicating strong gender preferences.

Figure 16 displays grades of industry average conduct. Only two grades emerge when setting $\lambda = 0.25$. Apparel stores (SIC 56) is the sole industry to receive a grade of $\star\star$, reflecting what appears to be a strong preference for female names. The magnitude of the posterior mean estimate is substantial, suggesting a roughly 33 log point advantage for female names in this sector. In contrast, auto dealers / services / parts (SIC 55), which registered large biases against Black names in Figure 9, is estimated to exhibit a negligible bias against female names.

9.2 Misclassification

Figure 17 summarizes the reliability of the gender report card in terms of Discordance Rates between grades. In our baseline model, the estimated share of firm pairs expected to be misclassified between adjacent grades ranges from 8% to 13%. Between non-adjacent grades the expected misclassification probability is estimated to be small: on the order of 1-3%.

When accounting for industry, five grades are present, with $\star\star\star$ indicating gender neutral conduct. The expected share of firms misclassified between $\star$, which suggests discrimination against female names, and $\star\star\star$ is estimated to be 5.5%. However, the expected share of firms graded as $\star\star\star\star\star$ that are less biased against men than a firm receiving a grade of $\star\star\star$ is estimated to be 0.8%. Hence, the chances of erroneously being classified as discriminating against women are higher than the chances of erroneously being classified as discriminating against men.

10 Conclusion

We have proposed a new empirical Bayes method for ranking noisy measurements and used it to grade the discriminatory conduct of firms in a large-scale correspondence ex-

periment. The experiment is shown to contain a wealth of information about the relative conduct of firms: our most granular (Condorcet) grades of discrimination against Black names that take into account industry affiliation yield an expected correlation with the true firm ranks of 0.59. 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 report cards with three or four grades, depending on whether the model conditions on industry. These coarse grades turn out to be substantially more reliable than the Condorcet ranks, lowering the estimated share of firm pairs that are misordered to less than 6%. These grades are also highly informative, offering an estimated correlation with the true firm ranks of 0.2 or greater. In addition to conveying information about the ranking of firm conduct, the grades capture important differences in conduct levels. Firms assigned the worst grade are estimated to favor white applicants over Black applicants by more than 20%, while racial gaps in callbacks among firms assigned the best grade are negligible. Similarly stark differences emerged in a ranking of firms' gender preferences: firms assigned extremal grades exhibit gender contact gaps on the order of 40%, while the vast majority of firms received a middle grade signaling minimal gender differences.

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 (Kline, Rose and Walters, 2022; Berson, Laouenan and Valat, 2020; Challe et al., 2022; Mocanu, 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. 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 mitigating hiring discrimination.

References

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, 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 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)*.

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*.

Chetty, Raj, and Nathaniel Hendren. 2018. "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.

Efron, Bradley, and Carl Morris. 1973. "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach." *Journal of the American Statistical Association*, 68(341): 117–130.

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.

InfoGroup. 2019. "Historical Datafiles."

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.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2023. "A Discrimination Report Card." *arXiv preprint arXiv:2306.13005*.

Kline, Patrick M, and Christopher R Walters. 2021. "Reasonable doubt: Experimental detection of job-level employment discrimination." *Econometrica*, 89(2): 765–792.

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.

McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." *Frontiers in Econometrics*.

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

Mocanu, Tatiana. 2022. "Designing gender equity: Evidence from hiring practices and committees." Working paper.

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).

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

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

Onwuachi-Willig, Angela, and Mario L. Barnes. 2005. "By any other name: on being regarded as Black, and why Title VII should apply even if Lakisha and Jamal are white." *Wisconsin Law Review*, 1283.

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.

Schaerer, Michael, Christilene du Plessis, My Hoang Bao Nguyen, Robbie CM van Aert, Leo Tiokhin, Daniël Lakens, Elena Giulia Clemente, Thomas Pfeiffer, Anna Dreber, Magnus Johannesson, et al. 2023. "On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions." *Organizational Behavior and Human Decision Processes*, 179: 104280.

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.

U.S. EEOC. 1996. "Enforcement Guidance: Whether "testers" can file charges and litigate claims of employment discrimination." EEOC Notice No. N-915.002. `https://www.eeoc.gov/laws/guidance/enforcement-guidance-whether-testers-can-file-charges-and-litigate-claims-employment`.

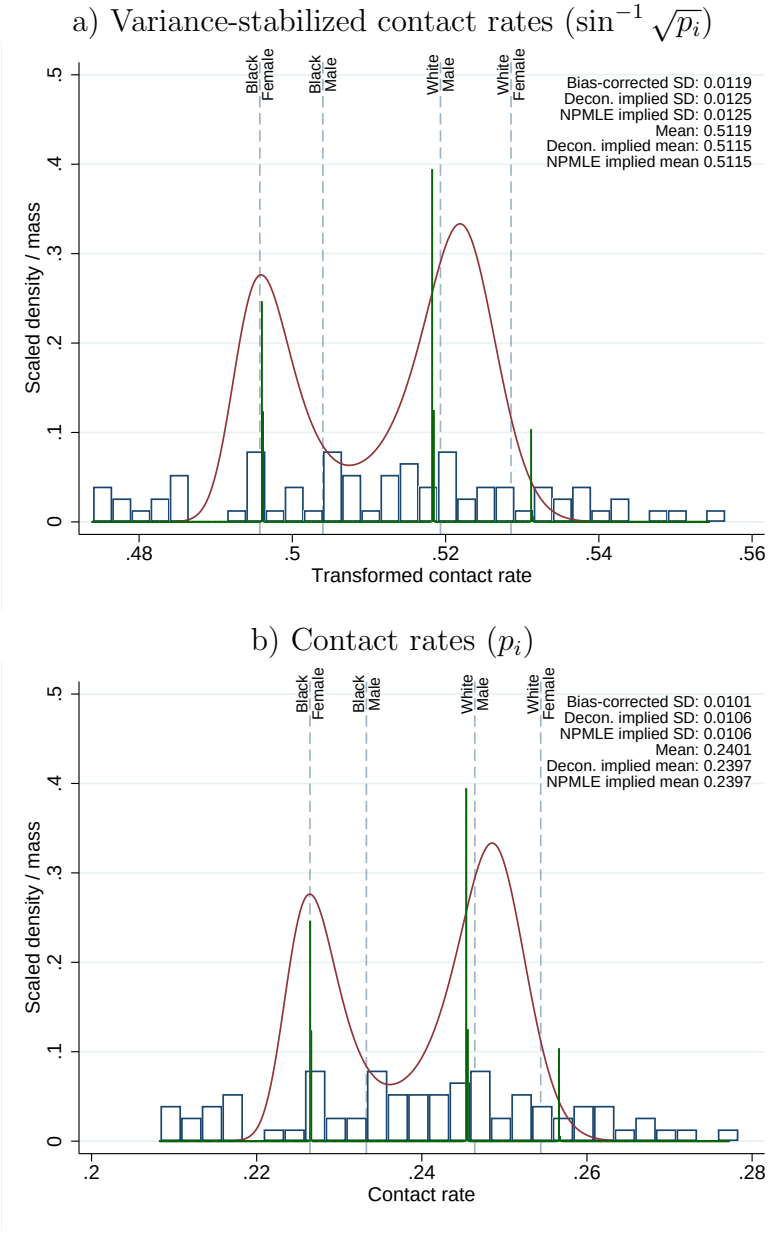
U.S. Equal Employment Opportunity Commission v. Target Corp. 460 F.3d 946 7th Cir. 2006. `https://casetext.com/case/us-eeoc-v-target-corp`.

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

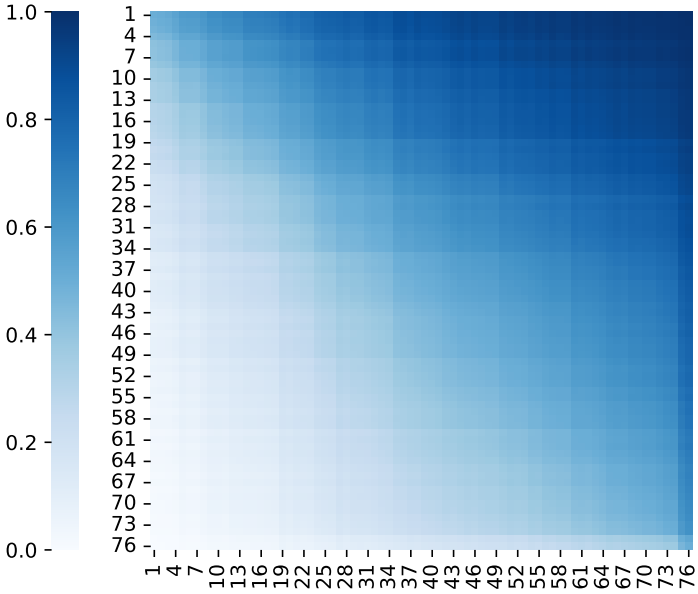
Figure 1: Deconvolution estimates of name-specific contact rate distributions



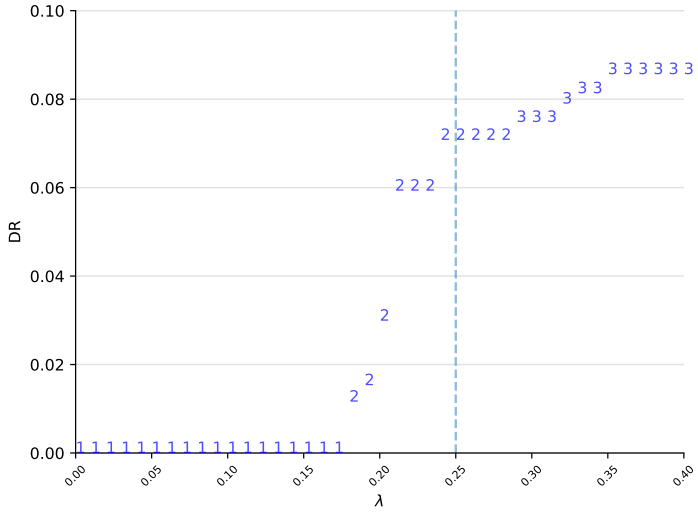
Notes: This figure presents non-parametric estimates of the distribution of name-specific contact rates. Panel (a) deconvolves transformed contact rates $\hat{\theta}_i = \sin^{-1}\left(\sqrt{\hat{p}_i}\right)$, where $\hat{p}_i$ is the contact rate for applications sent with first name i. The hollow 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 as a cubic spline with five knots. 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

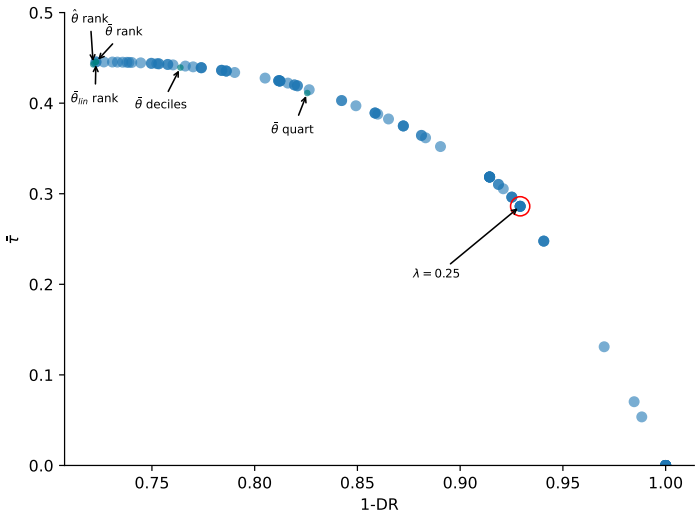
a) Pairwise posterior contrasts



b) Grades and discordance

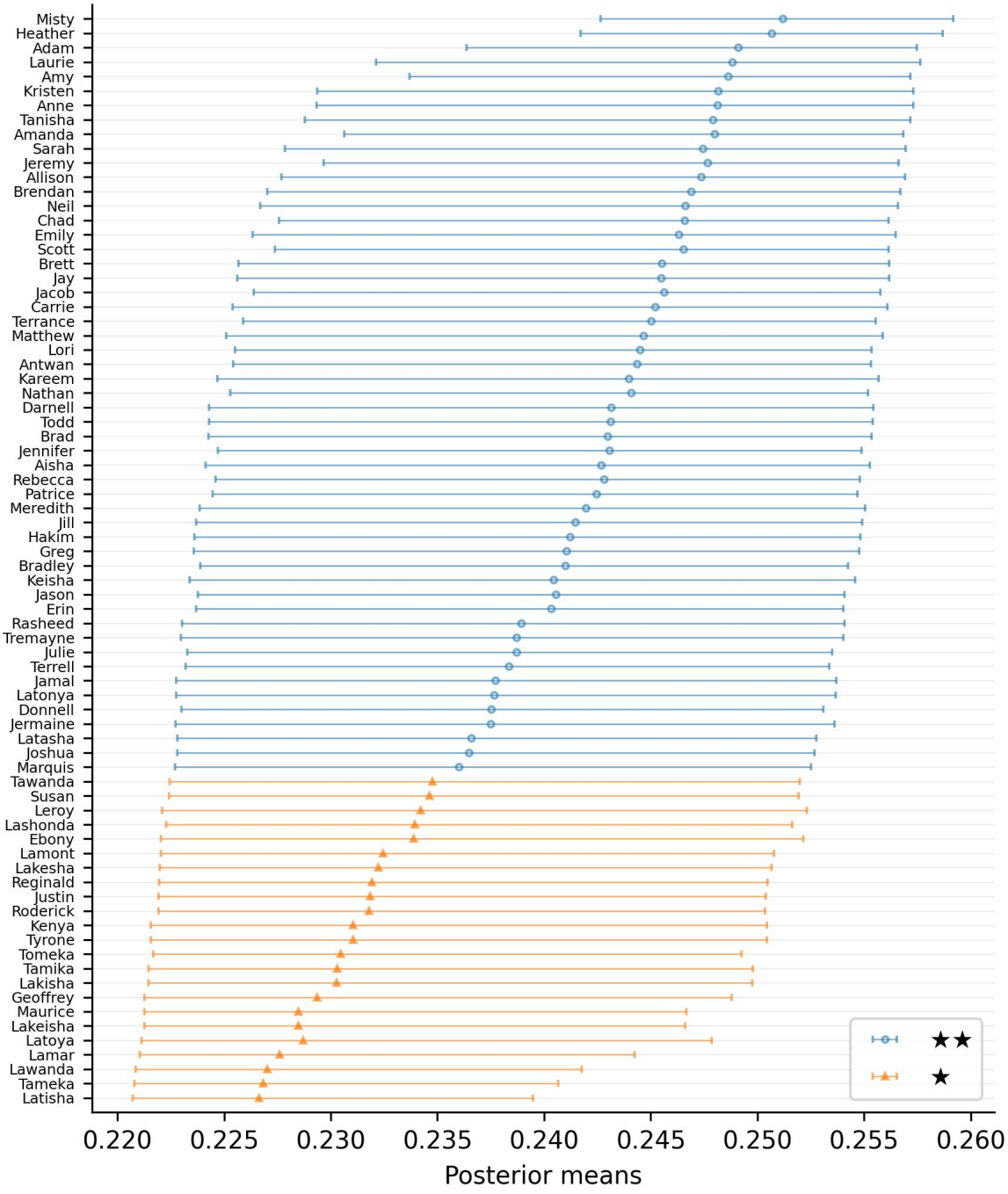


c) Reporting possibilities



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. "$\bar{\theta}_{dec}$" and "$\bar{\theta}_{quart}$" refer to grades corresponding to deciles and quartiles of these empirical Bayes posterior means. "$\bar{\theta}_{lin}$ rank" refers to ranks based on linear shrinkage estimates.

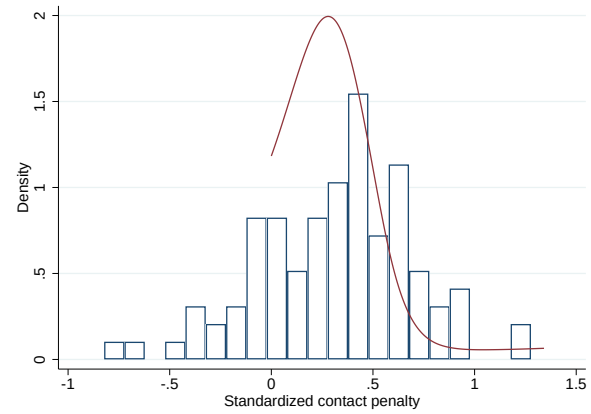
Figure 3: Posterior means and grades of first names



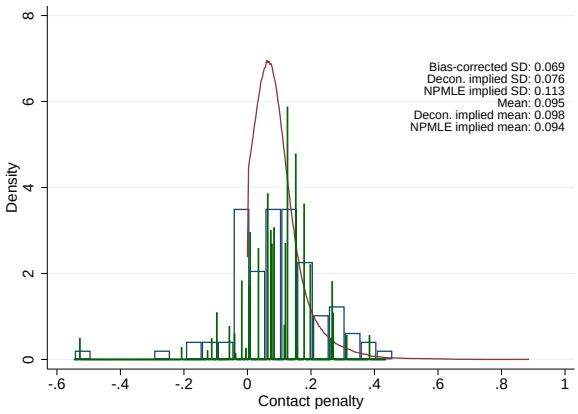
Notes: This figure shows posterior mean contact rates, 95% credible intervals, and assigned grades for names. Results are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Names are ordered by their rank under $\lambda = 1$, when each name is assigned its own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

Figure 4: Deconvolution estimates of race contact penalty distributions

a) No industry effects

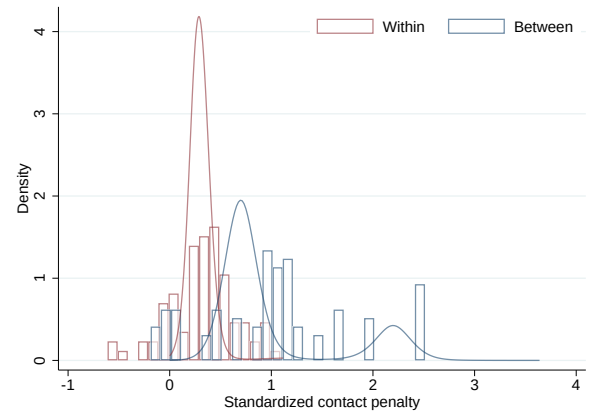


i) Distribution of v_i

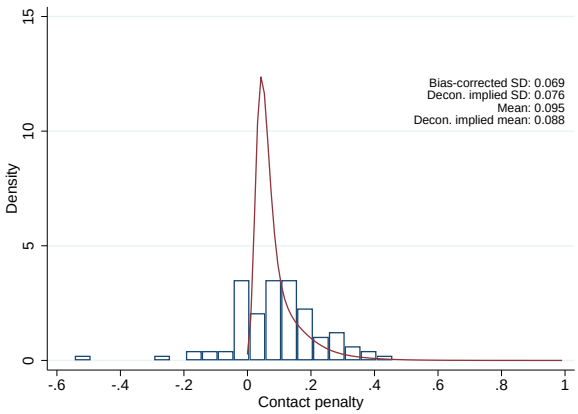


ii) Distribution of θ_i

b) With industry effects



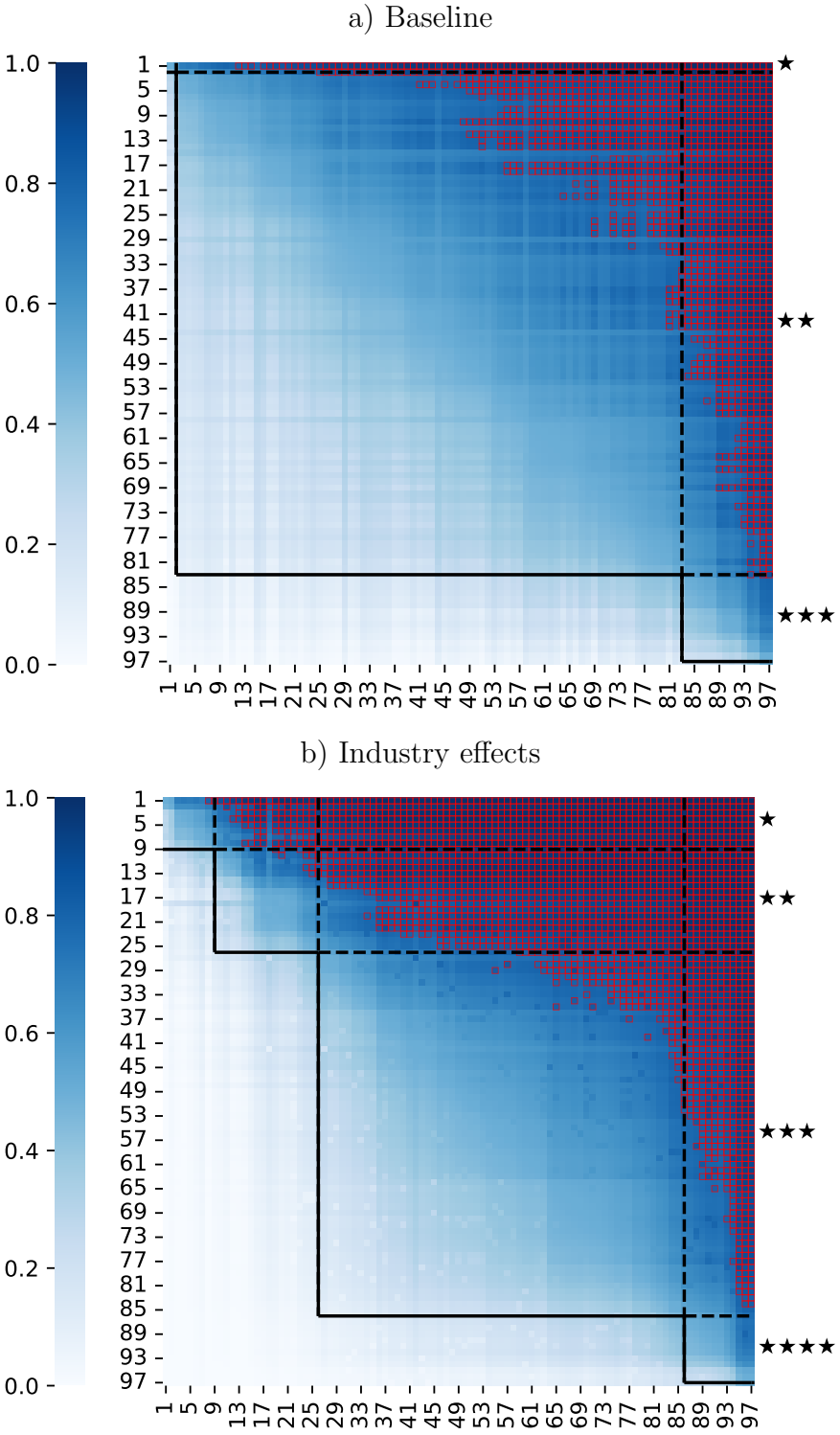
i) Distributions of η_k and ξ_i



ii) Distribution of θ_i

Notes: This figure presents non-parametric deconvolution estimates of the distribution of firm-specific race contact penalties along with corresponding histograms of firm-specific estimates. Estimates are based on the model $\theta_i = s_i^{\beta} v_i$, where θ_i is the proportional contact gap in favor of distinctively white names and s_i is the standard error of the estimate $\hat{\theta}_i$. Blues bars in part (i) of Panel (a) show a histogram of estimates $\hat{v}_i = \hat{\theta}_i / s_i^{\hat{\beta}}$, where $\hat{\beta}$ is the GMM estimate of β. The histogram is overlaid with the estimated distribution of v_i computed with the log-spline deconvolution procedure described in the Appendix. Part (ii) of Panel (a) plots a histogram of $\hat{\theta}_i$ along with the corresponding log-spline and non-parametric maximum likelihood (NPMLE) estimates of the distribution of θ_i. Panel (b) decomposes the standardized contact gap into within- and between-industry components, so that $v_i = \eta_{\mathrm{k}(i)} \xi_i$, where $\mathrm{k}(i)$ is the industry of firm i and the mean of the between-industry component η_k is normalized to 1. Blue bars in part (ii) of Panel (b) show a histogram of estimates $\bar{v}_k$, computed as the industry mean of $\hat{v}_i$. Red bars show a histogram of within-industry estimates $\hat{\xi}_i = \hat{v}_i / \bar{v}_{\mathrm{k}(i)}$. Blue and red curves display hierarchical log-spline estimates of the distributions of η_k and ξ_i. Part (ii) of Panel (b) overlays the histogram of $\hat{\theta}_i$ with the marginal distribution of θ_i implied by the hierarchical log-spline estimates. Bias-corrected standard deviation estimates are computed by subtracting the average squared standard error from the sample variance of estimated contact penalties, then taking the square root.

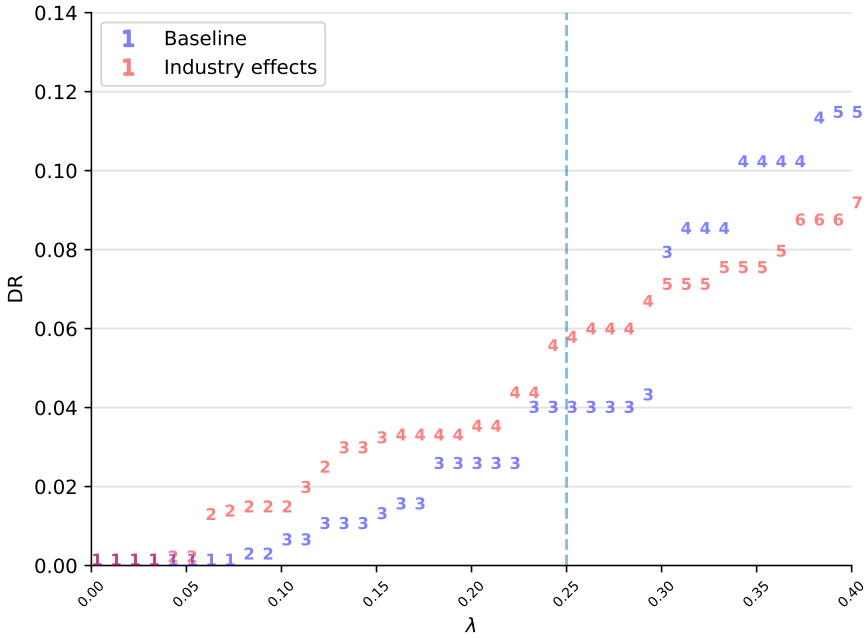
Figure 5: Posterior contrasts for race



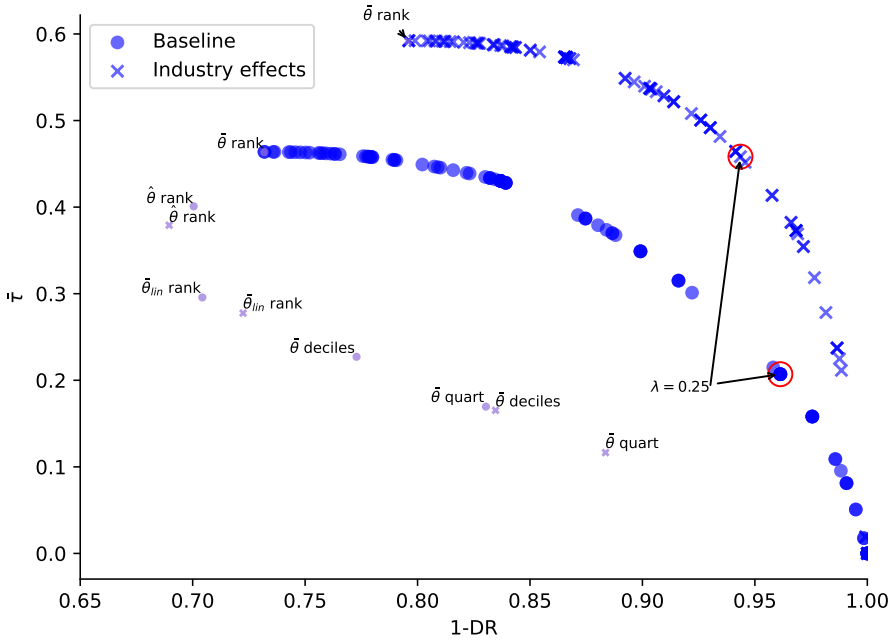
Notes: This figure plots pairwise posterior contrast probabilities for firm-specific contact penalties. Firms are ordered by their ranks under $\lambda = 1$, with the rank implying the largest θ_i is denoted by 1. Shading indicates the posterior probability that the contact penalty for the firm on the vertical axis exceeds the contact penalty for the firm on the horizontal axis. Firm pairs where $\hat{\pi}_{ij} > 1/(1 + 0.25)$ are bordered in red, indicating that pairwise optimal decision would rank the firm on the horizontal axis below the firm on the vertical axis when $\lambda = 0.25$. The black lines define optimal grades for this λ for the firms in the rows. Panel (a) shows results for a baseline model without industry effects, while Panel (b) reports results from a model with industry effects.

Figure 6: Grades, discordance, and reporting possibilities for race

a) Grades and discordance vs. λ

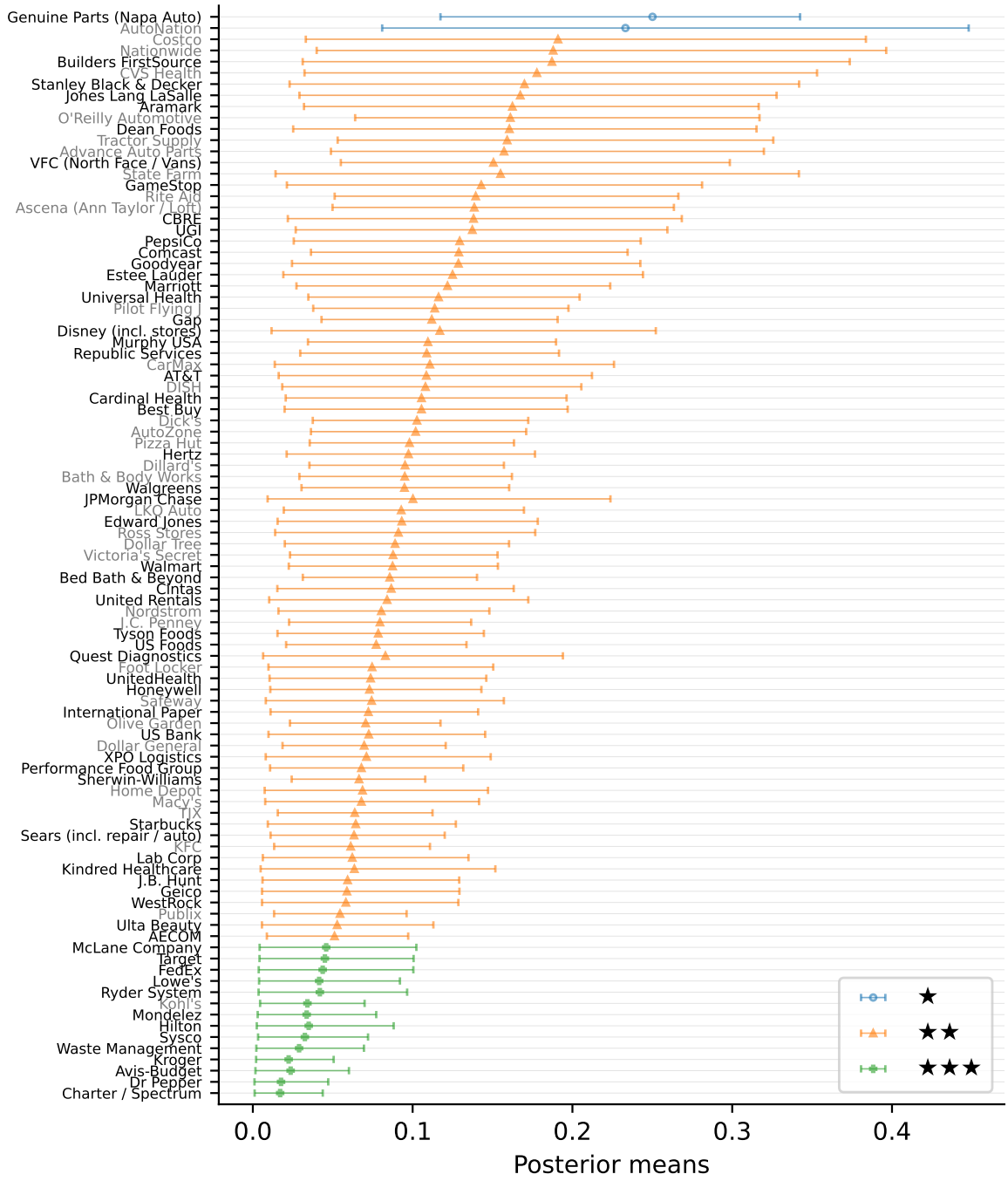


b) Reporting possibilities



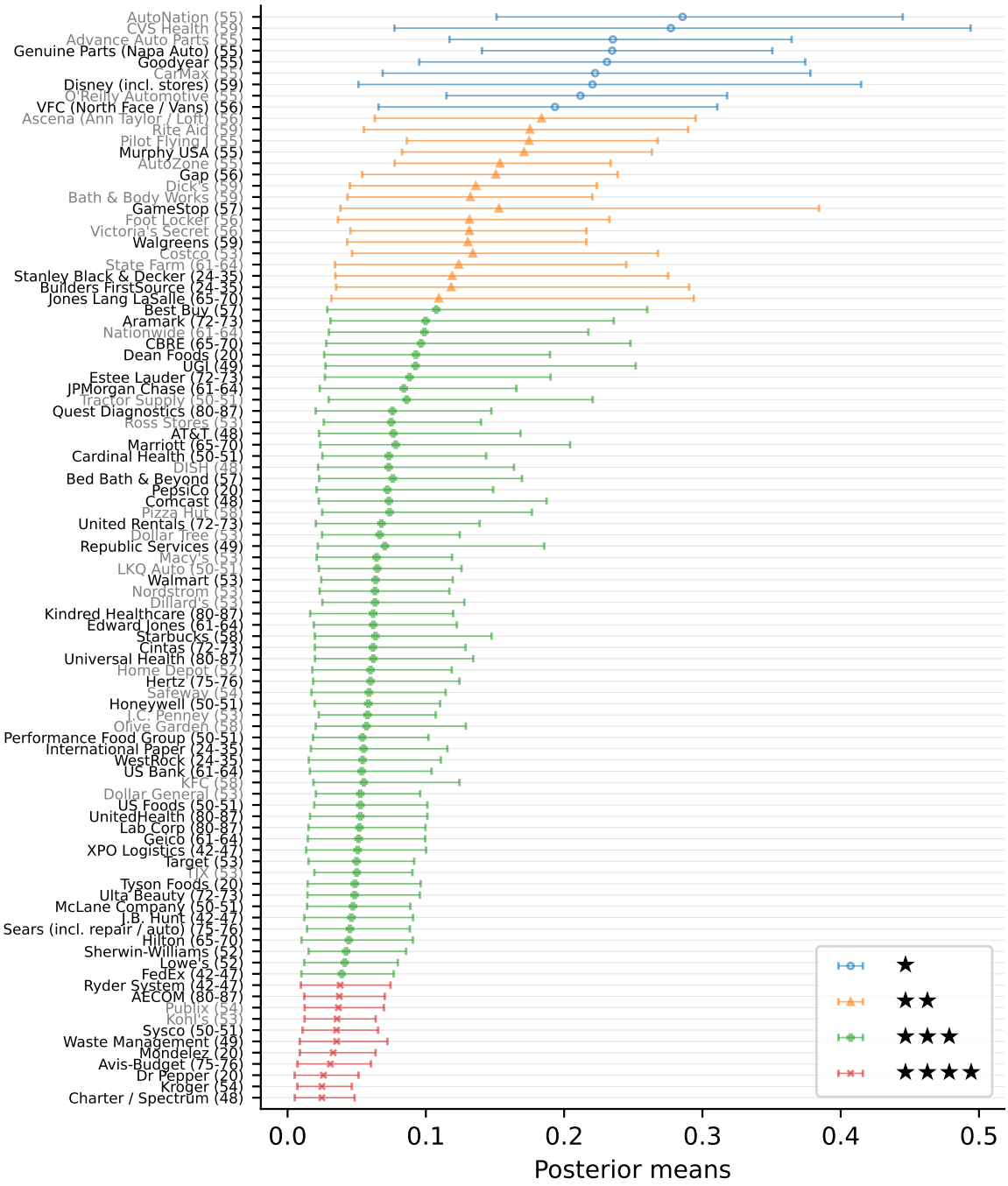
Notes: This figure summarizes informativeness and reliability of report card grades for race. Panel (a) shows estimated Discordance Rates (DR) as a function of λ. The number on each point indicates the number of unique grades in the underlying grading scheme. The vertical dashed line shows results for the benchmark case of $\lambda = 0.25$. Panel (b) shows the expectation of Kendall's τ rank correlation between θ and assigned grades against the estimated 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" plots the $\bar{\tau}$ and DR associated with ranking firms based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon empirical Bayes posterior means. "$\bar{\theta}_{dec}$" and "$\bar{\theta}_{quart}$" refer to grades corresponding to deciles and quartiles of these empirical Bayes posterior means. "$\bar{\theta}_{lin}$ rank" refers to ranks based on linear shrinkage estimates.

Figure 7: Race report card: posterior means and grades of firms (baseline)



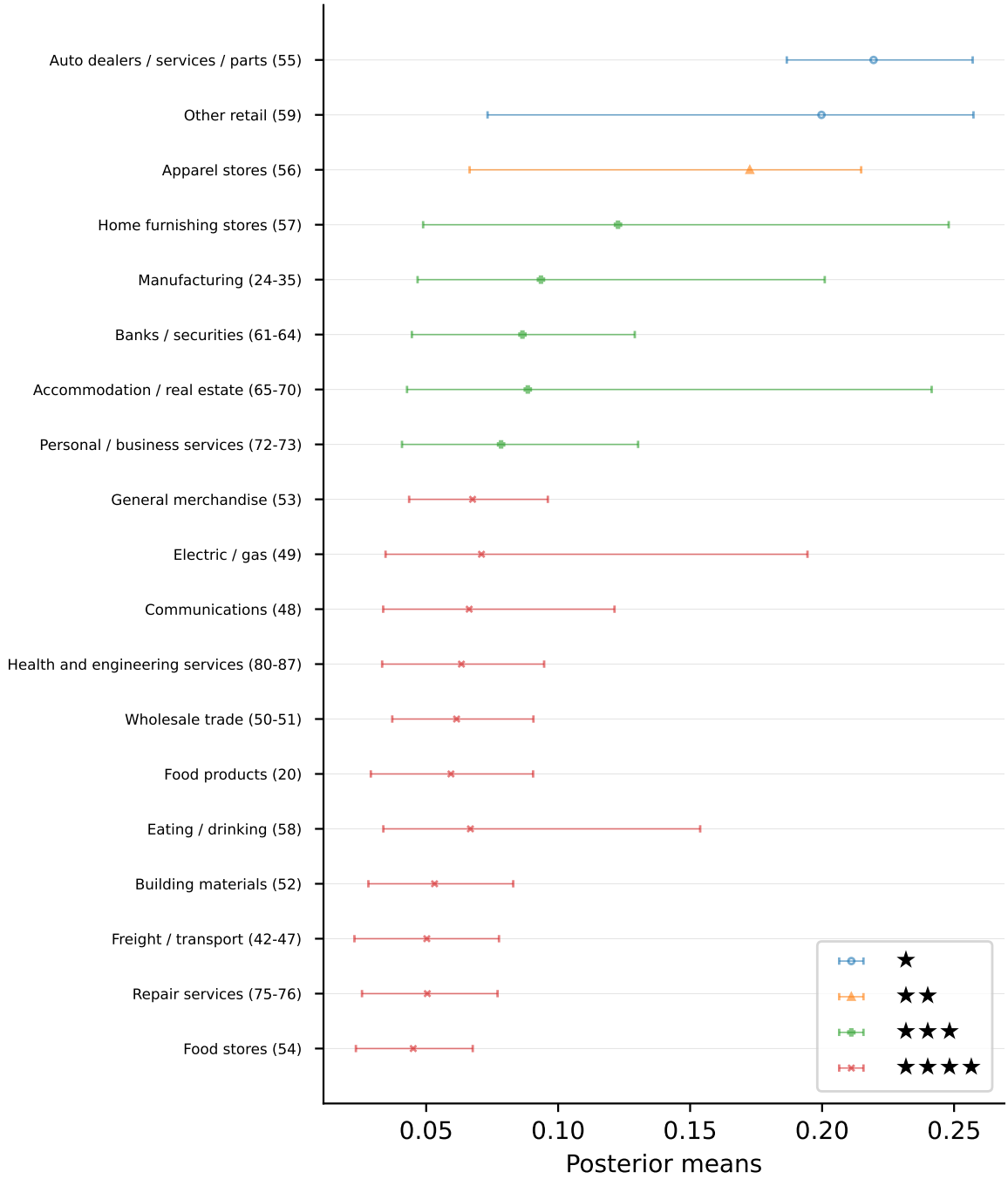
Notes: This figure shows posterior mean proportional contact penalties for distinctively Black names, 95% credible intervals, and assigned grades. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Posterior estimates come from a baseline model without industry effects. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade. Firms labeled with black text are federal contractors, whereas firms in gray are not.

Figure 8: Race report card: posterior means and grades of firms (industry effects)



Notes: This figure shows posterior mean proportional contact penalties for distinctively Black names, 95% credible intervals, and assigned grades from the industry random effect model. Grades 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. Industry codes listed in parentheses next to firm names. Firms labeled with black text are federal contractors, whereas firms in gray are not.

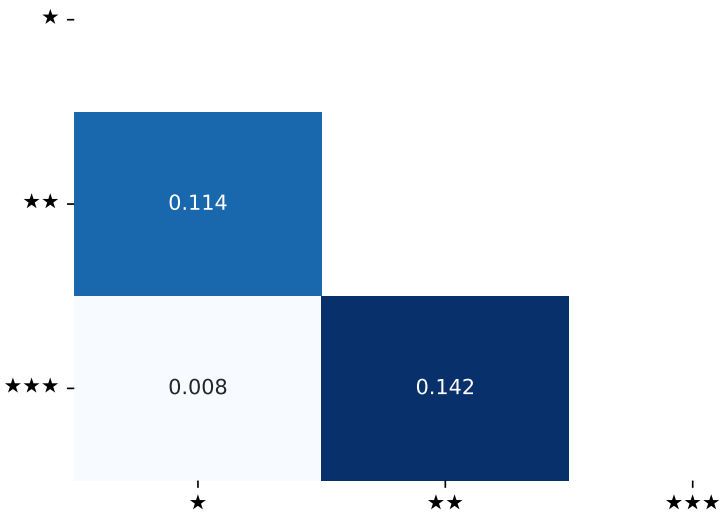
Figure 9: Race report card: posterior means and grades of industries



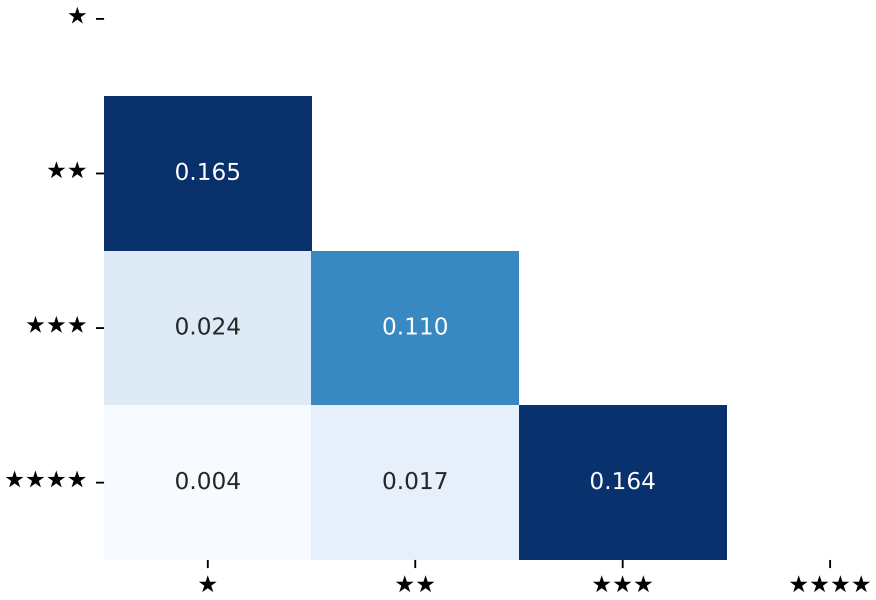
Notes: This figure shows posterior means, 95% credible intervals, and assigned grades for industry mean proportional contact penalties for distinctively Black names. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Each industry is labeled by its name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

Figure 10: Race report card: DR in baseline and industry effects model

a) Baseline



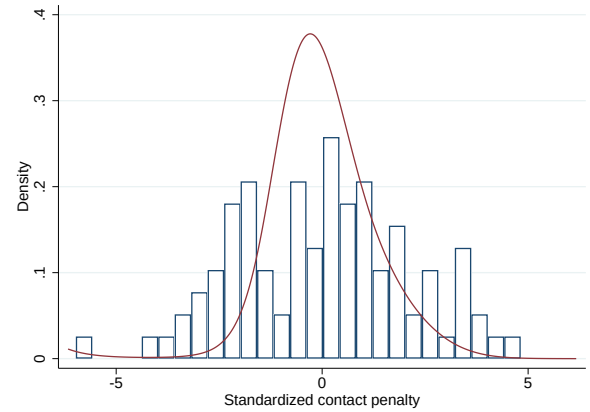
b) Industry effects



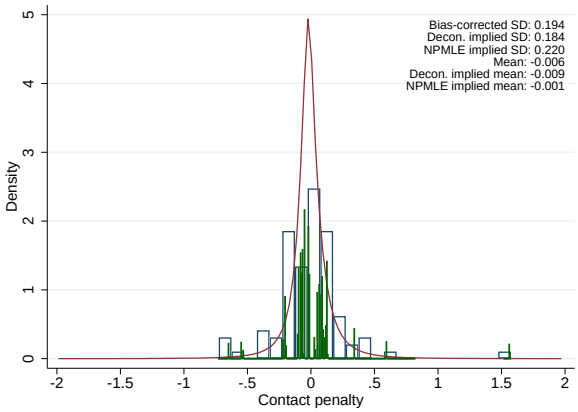
Notes: This figure shows mean Discordance Rates (DR) across grade pairs for the baseline model and the model with industry effects for race. In both panels, $DR_{g,g'}$ is the expected share of pairwise comparisons between firms in grades g and g' where the ordering implied by the grades differs from the true ordering of conduct.

Figure 11: Deconvolution estimates of gender contact penalty distributions

a) No industry effects

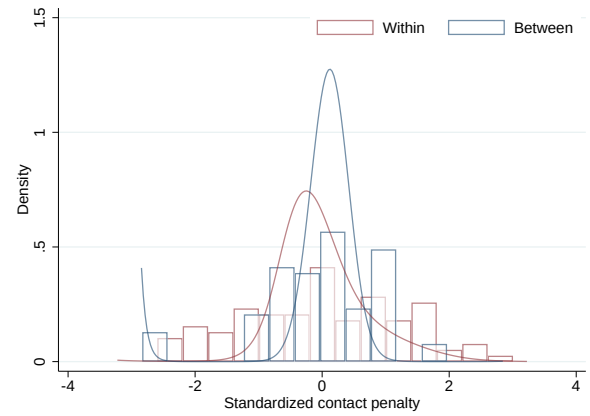


i) Distribution of v_i

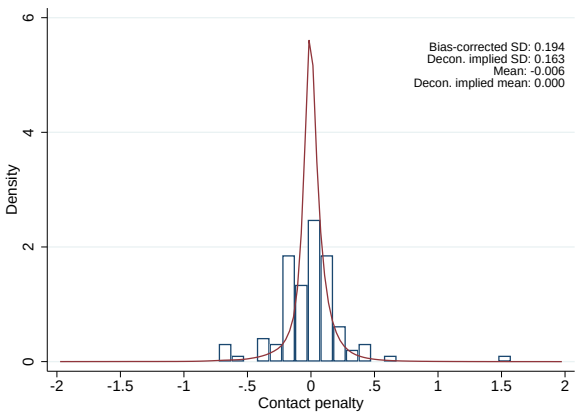


ii) Distribution of θ_i

b) With industry effects



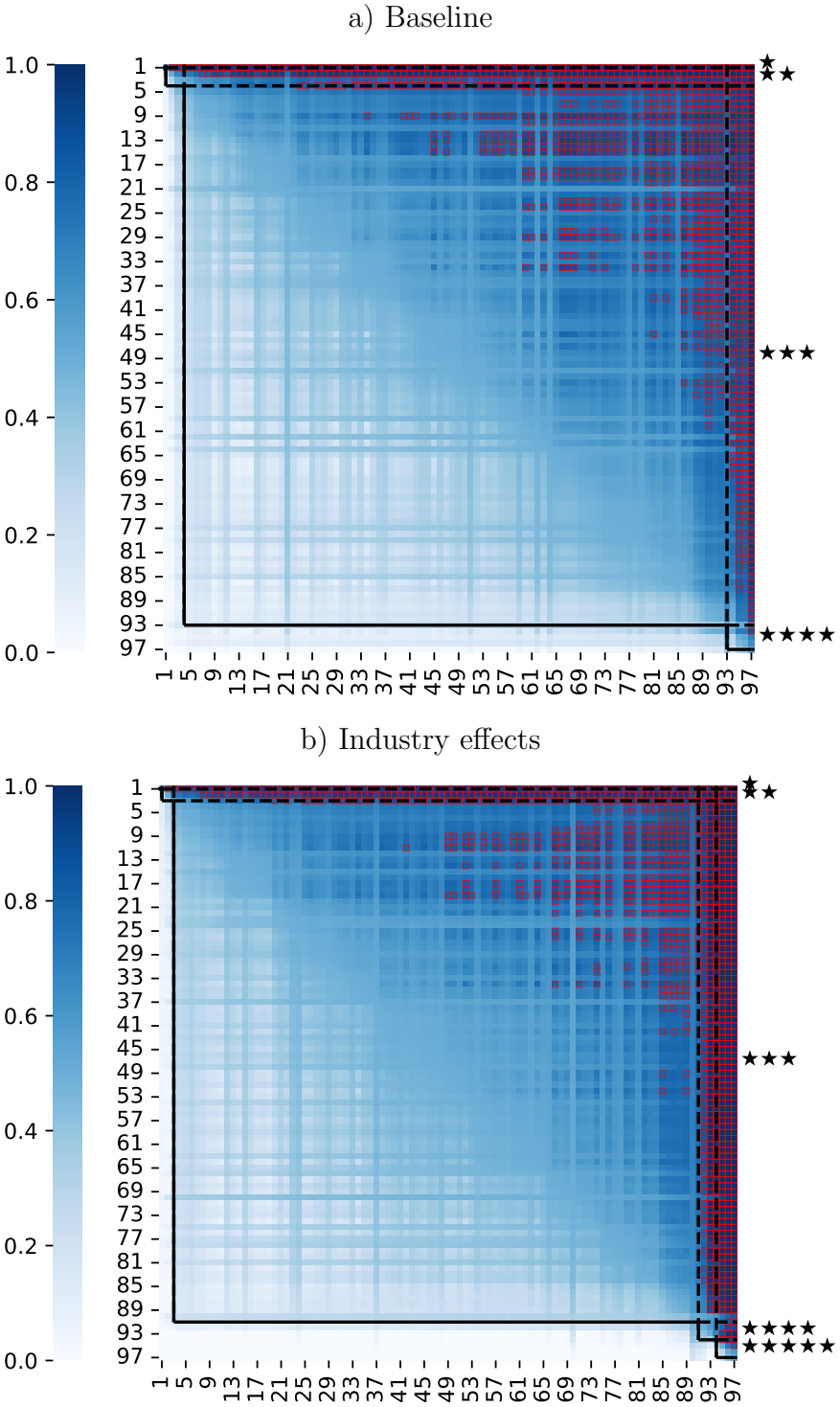
i) Distributions of η_k and ξ_i



ii) Distribution of θ_i

Notes: This figure presents non-parametric deconvolution estimates of the distribution of firm-specific gender contact penalties along with corresponding histograms of firm-specific estimates. Estimates are based on the model $\theta_i = \mu + s_i^{\beta} v_i$, where θ_i is the proportional contact gap in favor of distinctively male names, s_i is the standard error of the estimate $\hat{\theta}_i$, and $E[v_i] = 0$. Blue bars in part (i) of Panel (a) show a histogram of estimates $\hat{v}_i = (\hat{\theta}_i - \hat{\mu})/s_i^{\hat{\beta}}$, where $\hat{\mu}$ and $\hat{\beta}$ are the GMM estimates of μ and β. The histogram is overlaid with the estimated distribution of v_i computed with the log-spline deconvolution procedure described in the Appendix. Part (ii) of Panel (a) plots a histogram of $\hat{\theta}_i$ along with the corresponding log-spline and non-parametric maximum likelihood (NPMLE) estimates of the distribution of θ_i. Panel (b) decomposes the standardized contact gap into within- and between-industry components, so that $v_i = \eta_{\mathrm{k}(i)} + \xi_i$, where $\mathrm{k}(i)$ is the industry of firm i and the means of both components are normalized to zero. Blue bars in part (ii) of Panel (b) show a histogram of estimates $\bar{v}_k$, computed as the industry mean of $\hat{v}_i$. Red bars show a histogram of within-industry estimates $\hat{\xi}_i = \hat{v}_i - \bar{v}_{\mathrm{k}(i)}$. Blue and red curves display hierarchical log-spline estimates of the distributions of η_k and ξ_i. Part (ii) of Panel (b) overlays the histogram of $\hat{\theta}_i$ with the marginal distribution of θ_i implied by the hierarchical log-spline estimates.. Bias-corrected standard deviation estimates are computed by subtracting the average squared standard error from the sample variance of estimated contact penalties, then taking the square root.

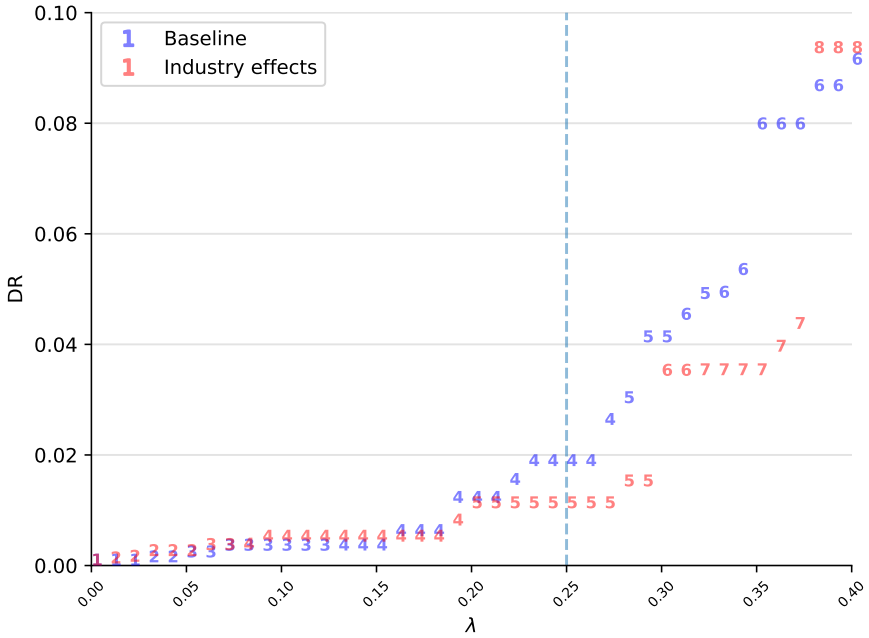
Figure 12: Posterior contrasts for gender



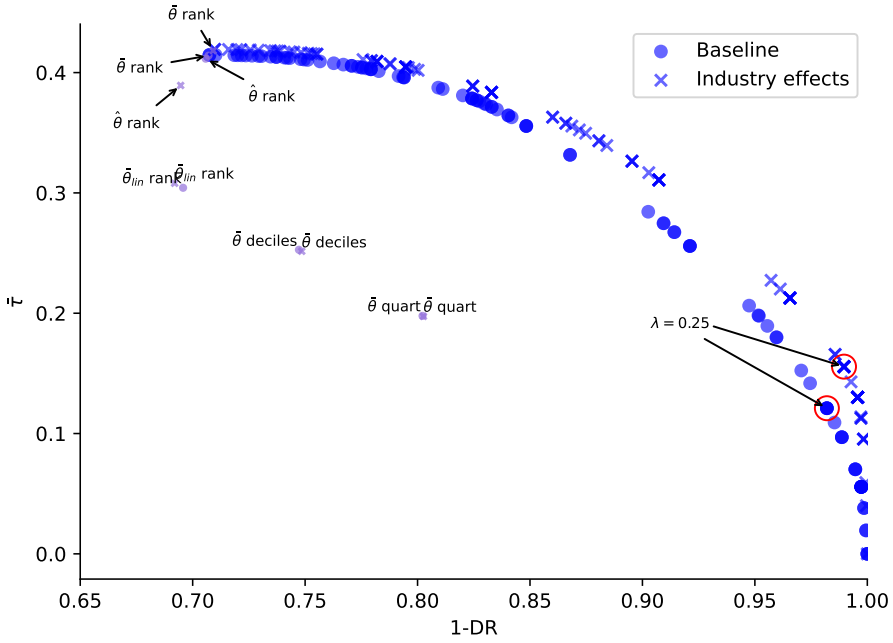
Notes: This figure plots pairwise posterior contrast probabilities for firm-specific gender contact differences. Firms are ordered by their ranks under $\lambda = 1$ within ranks for $\lambda = 0.25$, with the rank implying the largest θ_i is denoted by 1. Shading indicates the posterior probability that the contact penalty for the firm on the vertical axis exceeds the contact difference for the firm on the horizontal axis. Firm pairs where $\hat{\pi}_{ij} > 1/(1 + 0.25)$ are bordered in red, indicating that pairwise optimal decision would rank the firm on the horizontal axis below the firm on the vertical axis when $\lambda = 0.25$. The black lines define optimal grades for this λ for the firms in the rows. Panel (a) shows results for a baseline model without industry effects, while Panel (b) reports results from a model with industry effects.

Figure 13: Grades, discordance, and reporting possibilities for gender

a) Grades and discordance vs. λ

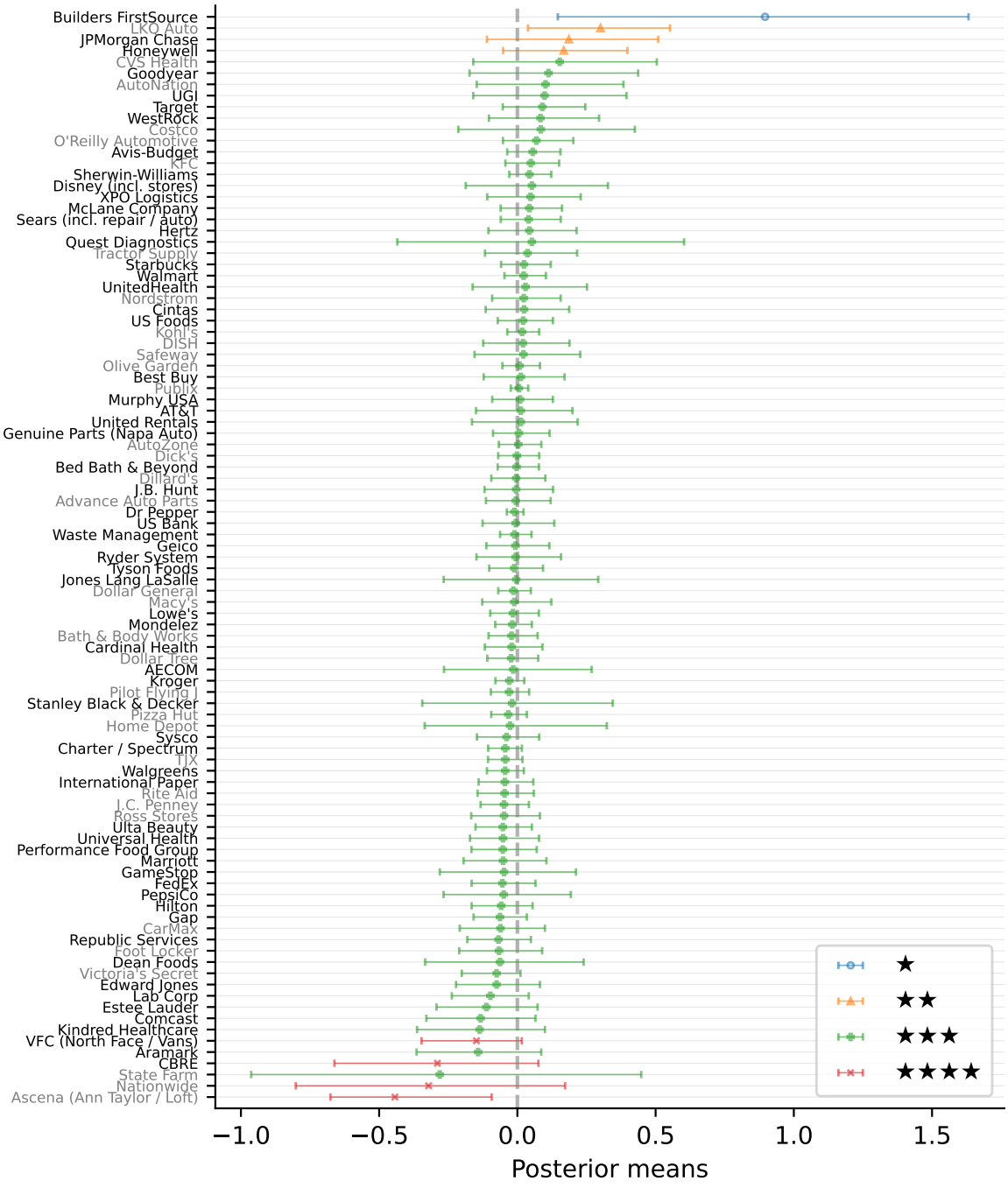


b) Reporting possibilities



Notes: This figure summarizes informativeness and reliability of report card grades for gender. Panel (a) 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 $\lambda = 0.25$. Panel (b) shows the expectation of Kendall's τ rank correlation between θ and assigned grades against the estimated 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" plots the $\bar{\tau}$ and DR associated with ranking firms based upon point estimates. "$\bar{\theta}$ rank" refers to ranks based upon empirical Bayes posterior means. "$\bar{\theta}_{dec}$" and "$\bar{\theta}_{quart}$" refer to grades corresponding to deciles and quartiles of these empirical Bayes posterior means. "$\bar{\theta}_{lin}$ rank" refers to ranks based on linear shrinkage estimates.

Figure 14: Gender report card: posterior means and grades of firms (baseline)

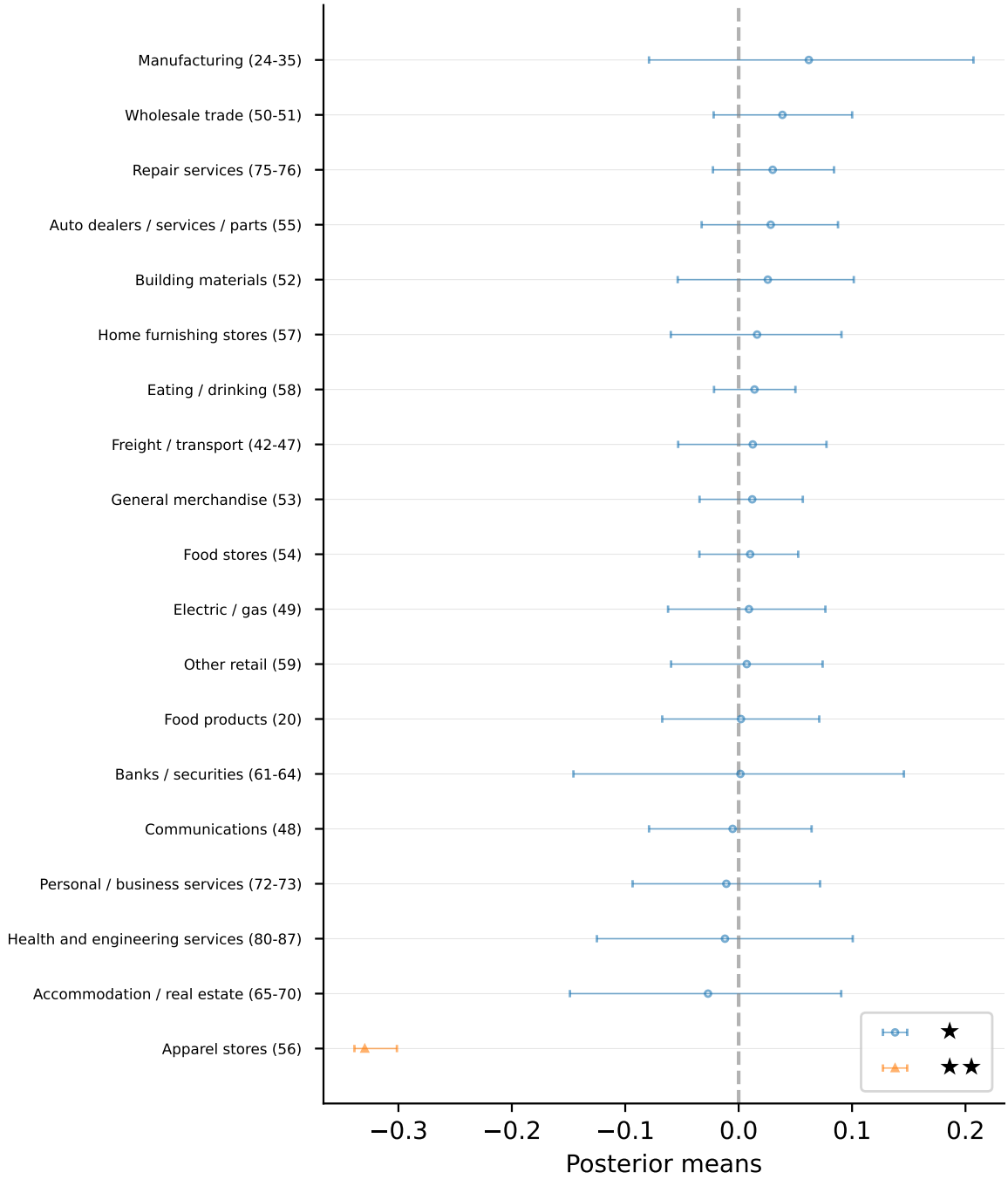


Notes: This figure shows posterior mean proportional gender contact differences between distinctively male and female names, 95% credible intervals, and assigned grades. Negative differences imply favoring female applications on average, while positive differences imply favoring men. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Posterior estimates come from a baseline model without industry effects. Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade. Firms labeled with black text are federal contractors, whereas firms in gray are not.

Figure 15: Gender report card: posterior means and grades of firms (industry effects)

Notes: This figure shows posterior mean proportional gender contact differences between distinctively male and female names, 95% credible intervals, and assigned grades from the industry random effect model. Negative differences imply favoring female applications on average, while positive differences imply favoring men. Grades 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. Industry codes listed in parentheses next to firm names. Firms labeled with black text are federal contractors, whereas firms in gray are not.

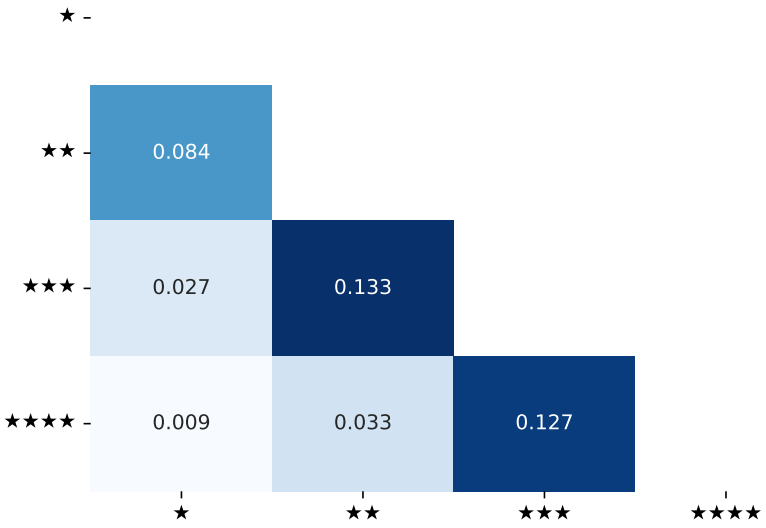
Figure 16: Gender report card: Posterior means and grades of industries



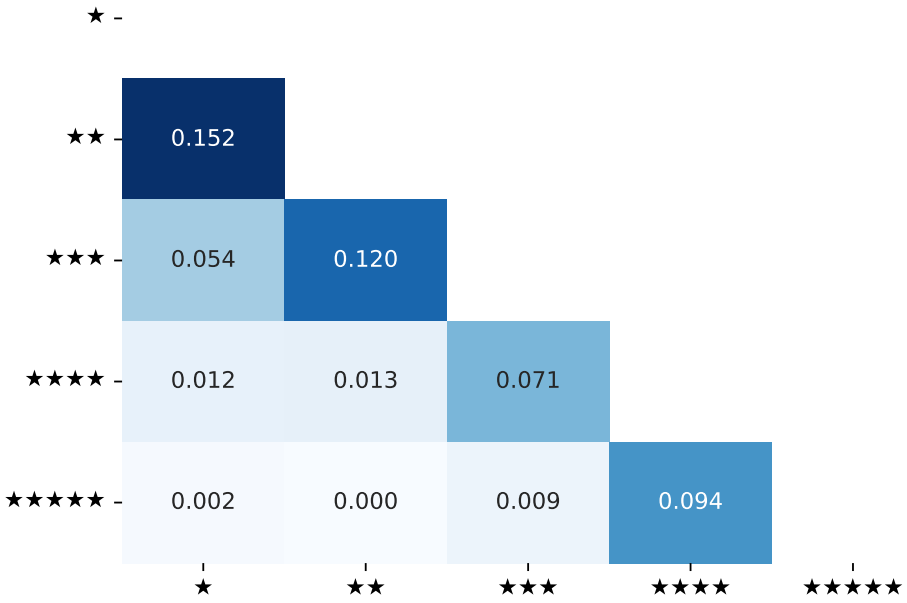
Notes: This figure shows posterior means, 95% credible intervals, and assigned grades for industry mean proportional gender contact differences between distinctively male and female names. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Each industry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

Figure 17: Gender report card: 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 for gender. In both panels, $DR_{g,g'}$ is the expected share of pairwise comparisons between firms in grades g and g' where the ordering implied by the grades differs from the true ordering of conduct.

Tables

Table 1: Summary statistics for first names sample

	Contact rate (1)	# apps (2)	# first names (3)	Wald test of heterogeneity (4)
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 reweighted 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 F2 for a list of first names used in the analysis.

Table 2: Summary statistics for firm sample

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

Notes: This table presents summary statistics for firm contact penalties. "White" and "Black" refer to average firm-level contact rates for white and Black applications. "Male" and "Female" refer to averages for male and female applications. Difference is the average contact rate difference (White minus Black, and Male minus Female). Log difference is the average of the primary contact penalty measure $\hat{\theta}_i$ used in the analysis. Standard errors in parentheses.

Table 3: GMM estimates of contact penalty parameters

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

Notes: This table reports generalized method of moments (GMM) estimates of the parameters of race and gender contact penalty distributions. Panel (a) shows GMM estimates of parameters from models for the race or gender contact penalty θ_i, while Panel (b) reports moments of the distribution of θ_i implied by the model estimates, with standard errors computed by the Delta method. Estimates for race in column (1) are based on the model $\theta_i = s_i^\beta v_i$, where θ_i is the proportional contact gap in favor of distinctively-white names, $\mathbb{E}[v_i|s_i] = \mu$, $\mathbb{V}[v_i|s_i] = \sigma_v^2$, and s_i is the standard error of $\hat{\theta}_i$. Column (2) allows an industry component of the form $v_i = \eta_{\mathrm{k}(i)}\xi_i$, where $\mathrm{k}(i)$ is the industry of firm i and $E[\eta_k] = 1$. Estimates for gender in column (3) are based on the model $\theta_i = \mu + s_i^\beta v_i$, where θ_i is the proportional contact gap in favor of distinctively-male names, $\mathbb{E}[v_i|s_i] = 0$, and $\mathbb{V}[v_i|s_i] = \sigma_v^2$. Column (4) allows an industry component of the form $v_i = \eta_{\mathrm{k}(i)} + \xi_i$, where $\mathbb{E}[\eta_k] = \mathbb{E}[\xi_i] = 0$. Estimates come from two-step optimally-weighted GMM with an identity weighting matrix in the first step. Variance matrices in column (2) and (4) are clustered by industry. The within share is $\frac{\mathbb{E}[\mathbb{V}[v_i|\eta_{\mathrm{k}(i)}]]}{\mathbb{V}[v_i]}$, which equals $\frac{(\sigma_\eta^2+1)\sigma_\xi^2}{\sigma_\eta^2\sigma_\xi^2+\sigma_\eta^2\mu^2+\sigma_\xi^2}$ in column (2) and $\frac{\sigma_\xi^2}{\sigma_\eta^2+\sigma_\xi^2}$ in column (4).

Online Appendix

A Discrimination Report Card

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

Appendix A Extension to weighted loss

Ranking mistakes may be more costly when the magnitude of the mistake is larger. To capture such concerns, we consider a family of loss functions that weight pairwise concordances and discordances by the p'th power of the difference between the cardinal biases of the two firms.

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

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

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

where $\mu_{ij}^p=\mathbb{E}_B\left[\max\{(\theta_i-\theta_j),0\}^p\right]$. 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.

An earlier version of this paper (Kline, Rose and Walters, 2023) reports rankings of both first names and firms for the case where $p=2$ ("square-weighted loss"). These rankings tended to yield more grades at each value of λ than arise under binary $(p=0)$ loss. This phenomenon arises because under square weighting finer classifications yield only small mistakes on average, which give rise to correspondingly small expected losses.

When working with p-weighted loss a corresponding weighted version of the conditional discordance rate can be employed:

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

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

Appendix B Proofs of propositions

This section provides proofs of the propositions discussed in Section 3.6, which are restated 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. □

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 $\mathcal{S}$. □

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. □

Appendix C Computing posteriors

This appendix details computation of posterior distributions for the firm contact gap analysis of Section 7. Computation of posteriors for the name contact rate analysis of Section 5 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 $v_i = \theta_i / s_i^\beta$ given $Y_i = (\hat{\theta}_i, s_i)$ can be written

$$\begin{aligned} f_v(x|Y_i; G_v, \beta) =& \frac{\mathcal{L}(\hat{\theta}_i|v_i = x, s_i; \beta) dG_v(x)}{\int \mathcal{L}(\hat{\theta}_i|v_i = u, s_i; \beta) dG_v(u)}, \\ \mathcal{L}(\hat{\theta}_i|v_i = x, s_i; \beta) =& \frac{1}{s_i^{1-\beta}} \phi\left(\frac{(\hat{\theta}_i / s_i^\beta) - x}{s_i^{1-\beta}} \right). \end{aligned}$$

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 $\mathcal{P}(t|\hat{\theta}_i, s_i; \hat{G}_v) = \int_{-\infty}^{t/s_i^{\hat{\beta}}} f_v(x|Y_i; \hat{G}_v, \hat{\beta}) dx$.

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

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

We plug $\hat{G}_v$ and $\hat{\beta}$ into these formulas to construct empirical Bayes posterior contrast probabilities $\hat{\pi}_{ij}$ by numerical integration. Substituting with $\hat{\pi}_{ij}$ for π_{ij} in (5), the grades are computed by minimizing the posterior expected loss subject to the constraints in (4) using Gurobi.

C.1 Industry effects

Posteriors for the hierarchical industry effects model of Section 6.4 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 $\boldsymbol{\xi}_k = \mathbf{z} = (z_1, ..., z_{n_k})'$ is given by:

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

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.

C.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 $\mathbb{E}\left[\theta_i|Y_i\right]$ within grade. The posterior mean estimate of each $\bar{\theta}_g^2$ is slightly harder to compute because

$$\begin{aligned}\bar{\theta}_g^2 &= \frac{\sum_{i=1}^n D_{ig}\theta_i^2}{\left(\sum_{i=1}^n D_{ig}\right)^2} + \frac{\sum_{i=1}^n \sum_{i' \neq i} D_{ig} D_{i'g} \theta_i \theta_{i'}}{\left(\sum_{i=1}^n D_{ig}\right)^2} \\ &= (nw_g)^{-2} \left\{\sum_{i=1}^n D_{ig}\theta_i^2 + \sum_{i=1}^n \sum_{i' \neq i} D_{ig} D_{i'g} \theta_i \theta_{i'}\right\}.\end{aligned}$$

Our posterior mean estimate of this quantity is computed analogously as

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

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

Appendix D Hierarchical log-spline estimator

We extend the empirical Bayes log-spline deconvolution approach from Efron (2016), and corresponding penalized maximum likelihood estimator, to separately estimate within- and between-industry distributions of race and gender contact gaps. 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 5×1 vector of values of natural cubic spline basis functions for point m (as detailed in Efron 2016) and α_η is a 5×1 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 5×1 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 under our model for race contact gaps in Section 6.4 is given by:

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

where $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{s}_k$ are vectors collecting the $\hat{\theta}_i$ and s_i for firms with $\mathrm{k}(i) = k$. The likelihood function for gender gaps adapts this expression to the alternative model outlined in Section 8.

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:

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

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 $\bar{\eta}_1 = 1$ for race and $\bar{\eta}_1 = 0$ for gender, so that η_k has a degenerate distribution at unity (or zero for gender).

The upper limit of the support for each component is set equal to the maximum of the empirical distribution of corresponding estimates or five GMM-estimated standard deviations above the GMM-estimated mean, whichever is larger. For gender, the lower limit of the support is similarly set equal to the minimum of the empirical distribution and five standard deviations below the mean; for race we set the lower support limits equal to zero. To limit the influence of outliers, we truncate the support of each component in each model to not exceed seven GMM-estimated standard deviations from the GMM-estimated mean. Since the scales of the two mixing distributions are not separately identified we impose the constraint $\sum_m g_{\eta,m}(\alpha_m)\bar{\eta}_m = 1$ for race. For gender we impose corresponding constraints normalizing the means of both ξ_i and η_k to zero.

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. Specifically, we compute the log-spline estimator for a grid of values of the penalty parameters and compute model-implied moments of the resulting distribution, then compute the quadratic distance between log-spline and GMM moment estimates (scaled by the inverse variance matrix of the GMM estimates). We then choose the value of the penalty parameter that minimizes this distance. 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 (or just the variance for gender, since the mean of the standardized gender gap is normalized to zero). In practice all parameters match well, as can be seen by comparing Tables 3 and F4.

Appendix E Monte Carlo evaluation of grades

To evaluate the composite performance of the grading procedure, a Monte Carlo exercise was conducted that conditions on the standard errors and industries of the 97 firms used in the racial discrimination report card. Each simulation draws a new θ_i and $\hat{\theta}_i$ for each firm from the models described in Section 6. Optimal grades are then computed under $\lambda = 0.25$, both when treating the distribution G_v as known (to evaluate oracle risk) and when re-estimating G_v in each simulation (to evaluate empirical risk).

Table E1 reports the results of 250 simulations. Column (1) assumes G_v obeys the baseline log-spline form described in Section 6.3 and reported in the upper panel of Figure 4. The oracle grading rule d^* computes the optimal grades given knowledge of the true G_v underlying the simulation. The empirical rule $\hat{d}^*$ computes the optimal grades using an estimated G_v obtained by applying GMM followed by the log-spline procedure in each simulation draw.[16] Column (2) fits a log-normal G_v via the method of moments (i.e., matching the mean and bias corrected variance) and simulates from this distribution. The empirical rule relies on a GMM step in each simulation draw followed by a corresponding method of moments step that recovers the log-normal distribution parameters from the mean and bias corrected variance of estimated residuals. Column (3) uses estimates from the model with industry effects in Section 6.4 as the prior and the empirical rule re-estimates the model in each simulation via GMM followed by the hierarchical log-spline method.

For both the oracle and empirical procedures, we report the expected rank correlation, discordance proportion, and loss evaluated under the true mixing distribution G_v. Regret is expressed as the average difference between the losses produced by the empirical and oracle rules. In all cases the regret is small (ranging from 0.012 to 0.017) indicating that the EB grades are nearly Bayes optimal. The slightly larger regret generated by the industry effects specification reflects that it is more difficult to adapt to the richer parameter space entertained by the hierarchical model.

The final panel of Table E1 reports the average expected rank correlation and discordance proportion of the empirical and oracle procedures when evaluated under the estimated $\hat{G}_v$. The posterior mean estimates of the τ and DP of the oracle rule produced under $\hat{G}_v$ are roughly unbiased, differing only slightly from their estimates under G_v.[17] This finding suggests the $\hat{\pi}_{ij}$ provide accurate estimates of the π_{ij}. In contrast, the posterior expected τ and DP of the empirical rule, when computed under $\hat{G}_v$, tend to be

[16] GMM estimation was initialized at the true parameters and optimized using a trust region reflective search procedure. In both the initial (unweighted) step and the second (optimally weighted) step we capped the procedure at 10 iterations.

[17] In the industry effects specification the estimates under $\hat{G}_v$ are slightly pessimistic, suggesting a lower τ and higher DP for the oracle rule than actually prevails under G_v. This phenomenon emerges because in roughly 12% of the simulations GMM finds no within industry component. We throw such simulations out, leading to a mild selection bias.

overly optimistic. This optimism bias results from the fact that the empirical grades are highly nonlinear functions of the $\hat{\pi}_{ij}$. For reference, we also report the standard deviation of these biases across Monte Carlo simulations. Dividing these standard deviations by $\sqrt{250} \approx 15.8$ yields a pair of standard errors on the expected bias that can be used to assess whether the average biases are distinguishable from simulation error.

In the baseline specification, the $\bar{\tau}$ of the empirical grades is biased up by about 0.036, while the DR is biased down by roughly 0.018. While the standard deviation of each bias is large, the average biases are both statistically distinguishable from simulation error at the 1% level. These biases do not seem to be driven by over-parameterization of the log-spline: the more parsimonious log-normal model yields almost identical mean biases and greater variability of bias. The model with industry effects exhibits a similar degree of over-optimism but is more precise than the simpler one level models, yielding an average rank correlation estimate that is biased up by 0.023 and an estimated discordance rate that is biased down by 0.011. The standard deviation of both biases is roughly 1/3 smaller than found with the baseline procedure, suggesting that the industry effects model yields error rate estimates that are both more accurate and precise.

Notably, the optimism bias in the estimated DR is driven in part by its lower bound of zero, which generates a skewed distribution of estimation errors $\mathbb{E}_{\hat{G}}[\tau(\hat{d}^*, \theta)] - \mathbb{E}_G[\tau(\hat{d}^*, \theta)]$. In the industry model the 25th, 50th, and 75th percentiles of these errors are -0.017, -0.006, and -0.001 respectively. Hence, the median bias of the DR estimate is 0.006, which is roughly half its mean bias of 0.011.

In sum, the grades produced by EB procedure appear to be nearly optimal in the sense of producing risk close to that of a Bayesian oracle. The EB estimates of reliability and informativeness are somewhat over-optimistic. However, this optimism bias tends to be small, particularly for the industry effects model. If the bias in our data were the same as the average bias in our Monte Carlo DGP, then the estimated Discordance Rate of 5.6% in our industry effects model of race gaps would need to be adjusted up to 6.7%.

Table E1: Monte Carlo simulations

	Baseline	Log-normal	Industry effects
Oracle risk			
$\mathbb{E}[\mathbb{E}_G[\tau(d^*,\theta)]]$	0.191	0.209	0.385
$\mathbb{E}[\mathbb{E}_G[DP(d^*,\theta)]]$	0.042	0.047	0.053
$\mathbb{E}[\mathbb{E}_G[\mathcal{R}(d^*,\theta;\lambda=0.25)]]$	-0.016	-0.017	-0.056
Empirical risk			
$\mathbb{E}[\mathbb{E}_G[\tau(\hat{d}^*,\theta)]]$	0.212	0.203	0.316
$\mathbb{E}[\mathbb{E}_G[DP(\hat{d}^*,\theta)]]$	0.063	0.060	0.058
$\mathbb{E}[\mathbb{E}_G[\mathcal{R}(\hat{d}^*,\theta;\lambda=0.25)]]$	-0.006	-0.006	-0.035
Regret			
$\mathbb{E}[\mathbb{E}_G[\mathcal{R}(\hat{d}^*,\theta;\lambda=0.25)-\mathcal{R}(d^*,\theta;\lambda=0.25)]]$	0.010	0.011	0.021
Estimated risk components			
$\mathbb{E}[\mathbb{E}_{\hat{G}}[\tau(d^*,\theta)]]$	0.192	0.203	0.333
$\mathbb{E}[\mathbb{E}_{\hat{G}}[\tau(\hat{d}^*,\theta)]]$	0.248	0.238	0.339
$\mathbb{V}(\mathbb{E}_{\hat{G}}[\tau(\hat{d}^*,\theta)]-\mathbb{E}_G[\tau(\hat{d}^*,\theta)])^{0.5}$	0.059	0.064	0.054
$\mathbb{E}[\mathbb{E}_{\hat{G}}[DP(d^*,\theta)]]$	0.042	0.050	0.080
$\mathbb{E}[\mathbb{E}_{\hat{G}}[DP(\hat{d}^*,\theta)]]$	0.045	0.042	0.047
$\mathbb{V}(\mathbb{E}_{\hat{G}}[DP(\hat{d}^*,\theta)]-\mathbb{E}_G[DP(\hat{d}^*,\theta)])^{0.5}$	0.029	0.032	0.027

Notes: This table reports the results of 250 Monte Carlo evaluations of the performance of the grading procedure. Here $\mathbb{E}_G$ denotes integration against the posterior distribution of θ given the oracle prior G and $\hat{\theta}$ while $\mathbb{E}$ denotes integration against simulated draws of $\hat{\theta}$. $\mathbb{E}_{\hat{G}}$ denotes integration against the posterior distribution of θ given the estimated prior $\hat{G}$ and $\hat{\theta}$. $\mathbb{V}$ denotes the variance across simulation draws. The first panel reports the expected rank correlation, discordance proportion, and risk of an oracle rule that forms grades using $\lambda = 0.25$. The second panel reports the same statistics for a rule that relies on an estimated prior in each simulation. Regret is the expected difference in risk between the empirical rule and the oracle rule. The final panel reports average expected rank correlations and discordance proportions of the empirical rule evaluated under the estimated $\hat{G}$ instead of the true G, as well as the standard deviation of the difference between the two evaluations for the empirical rule. Column (1) simulates data from the prior estimated in Section 6.3. Column (2) is the same but assumes that G_v is log-normal, both for simulating data and estimating G in each simulation. Column (3) simulates data from the model including industry effects described in Section 6.4.

Appendix F Additional Figures and Tables

Table F2: First names assigned by race and gender

	Black male		White male		Black female		White female	
	Name	Source	Name	Source	Name	Source	Name	Source
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
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 F3: Industries represented in firm sample

	# Firms (1)	# Jobs (2)	# Apps (3)
2-digit SIC industry (code)			
Food products (20)	5	470	3,333
Manufacturing (24-35)	4	382	2,931
Freight / transport (42-47)	4	458	3,300
Communications (48)	4	407	2,855
Electric / gas (49)	3	320	2,419
Wholesale trade (50-51)	8	817	6,186
Building materials (52)	3	377	2,755
General merchandise (53)	12	1,355	10,231
Food stores (54)	3	305	2,316
Auto dealers / services / parts (55)	9	1,016	7,857
Apparel stores (56)	5	550	4,303
Home furnishing stores (57)	3	351	2,708
Eating / drinking (58)	4	500	4,000
Other retail (59)	6	715	5,482
Banks / securities (61-64)	6	575	4,280
Accommodation / real estate (65-70)	4	397	3,024
Personal / business services (72-73)	5	550	4,177
Repair services (75-76)	3	340	2,551
Health and engineering services (80-87)	6	568	4,202

Notes: This table describes the number of firms in each two-digit SIC industry in the firm sample, along with the total number of jobs sampled and applicants sent. Industries were assigned using the most commonly reported SIC code of establishments listed in the InfoGroup Historical Datafiles database for 2019. In cases where InfoGroup reports a large share of establishments in multiple industries, we use the code that best reflects the jobs sampled in the experiment and ensures peer firms are grouped together. The resulting codes differ in 19 cases from those used in Kline, Rose and Walters (2022), which used SIC codes assigned before the experiment was conducted. Some industry codes are grouped to ensure that each category includes at least three firms. Labels for grouped industries were chosen to reflect the 2-digit codes of the firms actually included, rather than all potential industries in the grouping.

Table F4: Deconvolution estimates of random effect distributions

	No industry effects	With industry effects		
	Contact penalty (θ_i) (1)	Industry effect (η_k) (2)	Firm effect (ξ_i) (3)	Contact penalty (θ_i) (4)
		a) Race estimates		
Mean	0.098	1.000	0.300	0.088
	(0.010)	-	(0.053)	(0.016)
Std. Dev.	0.076	0.619	0.115	0.076
	(0.012)	(0.186)	(0.038)	(0.019)
Skewness	2.027	1.365	1.611	2.885
	(0.457)	(0.800)	(0.865)	(0.918)
Excess kurtosis	7.610	0.384	8.320	15.369
	(3.799)	(2.406)	(6.045)	(12.445)
		b) Gender estimates		
Mean	-0.009	0.000	0.000	0.000
	(0.000)	-	-	(0.008)
Std. Dev.	0.184	0.644	0.686	0.163
	(0.037)	(0.257)	(0.191)	(0.035)
Skewness	-0.469	-3.094	0.702	-1.296
	(2.378)	(1.597)	(1.116)	(1.657)
Excess kurtosis	29.680	11.316	1.654	22.735
	(24.073)	(5.802)	(3.137)	(7.822)

Notes: This table reports estimated moments of the distributions of industry and firm effects for race and gender contact gaps. Results are derived from hierarchical log-spline deconvolution estimates, with spline parameters estimated by penalized maximum likelihood. Panel (a) displays results for race, while Panel (b) shows results for gender. Standard errors come from 1,000 iterations of a parametric bootstrap procedure that resamples from the estimated mixing distribution. Each bootstrap trial takes a draw of the latent parameters from the full-sample mixing distribution estimate and draws normally-distributed estimation error using firm-specific standard errors. We then re-estimate the mixing distribution in each trial and compute moments of the resulting estimate. Standard errors are standard deviations of these moment estimates across bootstrap trials.

Table F5: Race discrimination: Detailed results by firm

					Baseline model				Industry effect model			
Firm (SIC group)	# apps	$\hat{p}_w$	$\hat{p}_b$	$\hat{\theta}_i$	Post. Mean	Post. CI	Grd	Cond. rank	Post. Mean	Post. CI	Grd	Cond. rank
Genuine Parts (Napa Auto) (55)	966	0.33 (0.03)	0.24 (0.03)	0.33 (0.07)	0.25	[0.12, 0.34]	1	1	0.23	[0.14, 0.35]	1	4
AutoNation (55)	869	0.14 (0.03)	0.09 (0.02)	0.43 (0.13)	0.23	[0.08, 0.45]	1	2	0.29	[0.15, 0.44]	1	1
Costco (53)	1000	0.07 (0.02)	0.05 (0.01)	0.38 (0.28)	0.19	[0.03, 0.38]	2	3	0.13	[0.05, 0.27]	2	22
Nationwide (61-64)	455	0.09 (0.03)	0.06 (0.02)	0.4 (0.22)	0.19	[0.04, 0.4]	2	4	0.10	[0.03, 0.22]	3	29
Builders FirstSource (24-35)	581	0.07 (0.02)	0.05 (0.02)	0.35 (0.29)	0.19	[0.03, 0.37]	2	5	0.12	[0.04, 0.29]	2	25
CVS Health (59)	787	0.05 (0.02)	0.04 (0.01)	0.34 (0.24)	0.18	[0.03, 0.35]	2	6	0.28	[0.08, 0.49]	1	2
Stanley Black & Decker (24-35)	790	0.05 (0.02)	0.04 (0.02)	0.17 (0.31)	0.17	[0.02, 0.34]	2	7	0.12	[0.03, 0.28]	2	24
Jones Lang LaSalle (65-70)	577	0.07 (0.02)	0.05 (0.02)	0.3 (0.23)	0.17	[0.03, 0.33]	2	8	0.11	[0.03, 0.29]	2	26
Aramark (72-73)	935	0.07 (0.02)	0.05 (0.02)	0.3 (0.19)	0.16	[0.03, 0.32]	2	9	0.10	[0.03, 0.24]	3	28
O'Reilly Automotive (55)	973	0.34 (0.03)	0.26 (0.03)	0.27 (0.08)	0.16	[0.06, 0.32]	2	10	0.21	[0.11, 0.32]	1	8
Dean Foods (20)	295	0.14 (0.05)	0.11 (0.04)	0.24 (0.24)	0.16	[0.03, 0.32]	2	11	0.09	[0.03, 0.19]	3	31
Tractor Supply (50-51)	943	0.2 (0.03)	0.15 (0.03)	0.29 (0.11)	0.16	[0.05, 0.33]	2	12	0.09	[0.03, 0.22]	3	35
Advance Auto Parts (55)	967	0.28 (0.03)	0.21 (0.03)	0.29 (0.11)	0.16	[0.05, 0.32]	2	13	0.24	[0.12, 0.36]	1	3
VFC (North Face / Vans) (56)	791	0.18 (0.04)	0.14 (0.03)	0.26 (0.09)	0.15	[0.06, 0.3]	2	14	0.19	[0.07, 0.31]	1	9
State Farm (61-64)	481	0.05 (0.02)	0.08 (0.04)	-0.54 (0.44)	0.16	[0.01, 0.34]	2	15	0.12	[0.03, 0.24]	2	23
GameStop (57)	790	0.06 (0.02)	0.05 (0.02)	0.17 (0.21)	0.14	[0.02, 0.28]	2	16	0.15	[0.04, 0.38]	2	18
Rite Aid (59)	962	0.22 (0.03)	0.17 (0.03)	0.24 (0.08)	0.14	[0.05, 0.27]	2	17	0.18	[0.06, 0.29]	2	11
Ascena (Ann Taylor / Loft) (56)	590	0.35 (0.04)	0.28 (0.04)	0.24 (0.09)	0.14	[0.05, 0.26]	2	18	0.18	[0.06, 0.3]	2	10
CBRE (65-70)	597	0.04 (0.02)	0.03 (0.02)	0.18 (0.19)	0.14	[0.02, 0.27]	2	19	0.10	[0.03, 0.25]	3	30
UGI (49)	546	0.11 (0.03)	0.09 (0.03)	0.22 (0.15)	0.14	[0.03, 0.26]	2	20	0.09	[0.03, 0.25]	3	32
PepsiCo (20)	916	0.05 (0.02)	0.04 (0.02)	0.2 (0.14)	0.13	[0.03, 0.24]	2	21	0.07	[0.02, 0.15]	3	43
Comcast (48)	231	0.42 (0.07)	0.34 (0.06)	0.22 (0.1)	0.13	[0.04, 0.23]	2	22	0.07	[0.02, 0.19]	3	44
Goodyear (55)	387	0.08 (0.04)	0.07 (0.03)	0.19 (0.14)	0.13	[0.02, 0.24]	2	23	0.23	[0.1, 0.37]	1	5
Estee Lauder (72-73)	579	0.14 (0.03)	0.12 (0.03)	0.14 (0.17)	0.13	[0.02, 0.24]	2	24	0.09	[0.03, 0.19]	3	33
Marriott (65-70)	964	0.16 (0.03)	0.13 (0.03)	0.19 (0.12)	0.12	[0.03, 0.22]	2	25	0.08	[0.02, 0.2]	3	39
Universal Health (80-87)	586	0.32 (0.04)	0.27 (0.04)	0.19 (0.08)	0.12	[0.03, 0.2]	2	26	0.06	[0.02, 0.13]	3	58
Pilot Flying J (55)	993	0.36 (0.03)	0.3 (0.03)	0.18 (0.08)	0.11	[0.04, 0.2]	2	27	0.17	[0.09, 0.27]	2	12
Gap (56)	996	0.33 (0.04)	0.27 (0.04)	0.17 (0.06)	0.11	[0.04, 0.19]	2	28	0.15	[0.05, 0.24]	2	15

Continued on next page

Disney (incl. stores)	858	0.09	0.1	-0.12	0.12	[0.01,	2	29	0.22	[0.05,	1	7
(59)		(0.02)	(0.03)	(0.24)		0.25]				0.41]		
Murphy USA (55)	927	0.3	0.25	0.17	0.11	[0.03,	2	30	0.17	[0.08,	2	13
		(0.03)	(0.03)	(0.08)		0.19]				0.26]		
Republic Services (49)	943	0.22	0.19	0.17	0.11	[0.03,	2	31	0.07	[0.02,	3	48
		(0.03)	(0.03)	(0.08)		0.19]				0.19]		
CarMax (55)	775	0.14	0.14	0.05	0.11	[0.01,	2	32	0.22	[0.07,	1	6
		(0.03)	(0.03)	(0.17)		0.23]				0.38]		
AT&T (48)	893	0.13	0.11	0.11	0.11	[0.02,	2	33	0.08	[0.02,	3	38
		(0.02)	(0.02)	(0.14)		0.21]				0.17]		
DISH (48)	771	0.28	0.25	0.13	0.11	[0.02,	2	34	0.07	[0.02,	3	41
		(0.04)	(0.04)	(0.12)		0.21]				0.16]		
Cardinal Health	974	0.23	0.2	0.14	0.11	[0.02,	2	35	0.07	[0.03,	3	40
(50-51)		(0.03)	(0.03)	(0.11)		0.2]				0.14]		
Best Buy (57)	920	0.18	0.16	0.14	0.11	[0.02,	2	36	0.11	[0.03,	3	27
		(0.03)	(0.03)	(0.11)		0.2]				0.26]		
Dick's (59)	975	0.38	0.32	0.15	0.10	[0.04,	2	37	0.14	[0.05,	2	16
		(0.04)	(0.03)	(0.06)		0.17]				0.22]		
AutoZone (55)	1000	0.38	0.33	0.15	0.10	[0.04,	2	38	0.15	[0.08,	2	14
		(0.04)	(0.03)	(0.06)		0.17]				0.23]		
Pizza Hut (58)	1000	0.42	0.36	0.14	0.10	[0.04,	2	39	0.07	[0.03,	3	45
		(0.04)	(0.04)	(0.06)		0.16]				0.18]		
Hertz (75-76)	786	0.24	0.21	0.13	0.10	[0.02,	2	40	0.06	[0.02,	3	60
		(0.04)	(0.03)	(0.09)		0.18]				0.12]		
Dillard's (53)	925	0.34	0.3	0.14	0.10	[0.04,	2	41	0.06	[0.03,	3	53
		(0.03)	(0.03)	(0.05)		0.16]				0.13]		
Bath & Body Works	990	0.31	0.27	0.14	0.10	[0.03,	2	42	0.13	[0.04,	2	17
(59)		(0.03)	(0.03)	(0.06)		0.16]				0.22]		
Walgreens (59)	910	0.41	0.35	0.14	0.09	[0.03,	2	43	0.13	[0.04,	2	21
		(0.04)	(0.04)	(0.06)		0.16]				0.22]		
JPMorgan Chase	981	0.06	0.07	-0.19	0.10	[0.01,	2	44	0.08	[0.02,	3	34
(61-64)		(0.02)	(0.02)	(0.22)		0.22]				0.17]		
LKQ Auto (50-51)	587	0.23	0.2	0.12	0.09	[0.02,	2	45	0.06	[0.02,	3	50
		(0.04)	(0.04)	(0.08)		0.17]				0.13]		
Edward Jones (61-64)	965	0.12	0.11	0.1	0.09	[0.02,	2	46	0.06	[0.02,	3	55
		(0.02)	(0.02)	(0.1)		0.18]				0.12]		
Ross Stores (53)	650	0.22	0.2	0.09	0.09	[0.01,	2	47	0.07	[0.03,	3	37
		(0.03)	(0.03)	(0.1)		0.18]				0.14]		
Dollar Tree (53)	998	0.28	0.25	0.11	0.09	[0.02,	2	48	0.07	[0.02,	3	47
		(0.03)	(0.03)	(0.07)		0.16]				0.12]		
Victoria's Secret (56)	931	0.38	0.34	0.12	0.09	[0.02,	2	49	0.13	[0.05,	2	20
		(0.04)	(0.04)	(0.06)		0.15]				0.22]		
Walmart (53)	400	0.64	0.57	0.12	0.09	[0.02,	2	50	0.06	[0.02,	3	51
		(0.05)	(0.05)	(0.07)		0.15]				0.12]		
Bed Bath & Beyond	998	0.36	0.32	0.12	0.09	[0.03,	2	51	0.08	[0.02,	3	42
(57)		(0.04)	(0.04)	(0.05)		0.14]				0.17]		
Cintas (72-73)	747	0.23	0.21	0.09	0.09	[0.02,	2	52	0.06	[0.02,	3	57
		(0.04)	(0.03)	(0.09)		0.16]				0.13]		
United Rentals (72-73)	917	0.12	0.12	0.03	0.08	[0.01,	2	53	0.07	[0.02,	3	46
		(0.02)	(0.02)	(0.11)		0.17]				0.14]		
Nordstrom (53)	941	0.21	0.19	0.09	0.08	[0.02,	2	54	0.06	[0.02,	3	52
		(0.03)	(0.03)	(0.07)		0.15]				0.12]		
J.C. Penney (53)	994	0.31	0.28	0.1	0.08	[0.02,	2	55	0.06	[0.02,	3	63
		(0.04)	(0.04)	(0.05)		0.14]				0.11]		
Tyson Foods (20)	797	0.36	0.33	0.09	0.08	[0.02,	2	56	0.05	[0.01,	3	78
		(0.04)	(0.04)	(0.07)		0.14]				0.1]		
US Foods (50-51)	961	0.29	0.27	0.1	0.08	[0.02,	2	57	0.05	[0.02,	3	71
		(0.03)	(0.03)	(0.05)		0.13]				0.1]		
Quest Diagnostics	907	0.02	0.03	-0.27	0.08	[0.01,	2	58	0.08	[0.02,	3	36
(80-87)		(0.01)	(0.01)	(0.2)		0.19]				0.15]		
Foot Locker (56)	995	0.15	0.14	0.04	0.07	[0.01,	2	59	0.13	[0.04,	2	19
		(0.03)	(0.03)	(0.09)		0.15]				0.23]		

Continued on next page

UnitedHealth (80-87)	942	0.1	0.1	0.05	0.07	[0.01,	2	60	0.05	[0.02,	3	72
		(0.03)	(0.03)	(0.08)		0.15]				0.1]		
Honeywell (50-51)	556	0.16	0.15	0.06	0.07	[0.01,	2	61	0.06	[0.02,	3	62
		(0.04)	(0.04)	(0.08)		0.14]				0.11]		
Safeway (54)	429	0.25	0.25	0	0.07	[0.01,	2	62	0.06	[0.02,	3	61
		(0.05)	(0.05)	(0.11)		0.16]				0.11]		
International Paper	954	0.22	0.2	0.06	0.07	[0.01,	2	63	0.06	[0.02,	3	66
(24-35)		(0.03)	(0.03)	(0.07)		0.14]				0.12]		
Olive Garden (58)	1000	0.4	0.37	0.09	0.07	[0.02,	2	64	0.06	[0.02,	3	64
		(0.04)	(0.04)	(0.04)		0.12]				0.13]		
US Bank (61-64)	966	0.18	0.17	0.05	0.07	[0.01,	2	65	0.05	[0.02,	3	68
		(0.03)	(0.03)	(0.08)		0.15]				0.1]		
Dollar General (53)	787	0.48	0.45	0.08	0.07	[0.02,	2	66	0.05	[0.02,	3	70
		(0.04)	(0.04)	(0.05)		0.12]				0.1]		
XPO Logistics (42-47)	861	0.16	0.16	0.02	0.07	[0.01,	2	67	0.05	[0.01,	3	75
		(0.03)	(0.03)	(0.09)		0.15]				0.1]		
Performance Food	520	0.35	0.33	0.06	0.07	[0.01,	2	68	0.05	[0.02,	3	65
Group (50-51)		(0.05)	(0.05)	(0.07)		0.13]				0.1]		
Sherwin-Williams (52)	980	0.48	0.44	0.08	0.07	[0.02,	2	69	0.04	[0.02,	3	84
		(0.04)	(0.04)	(0.03)		0.11]				0.09]		
Home Depot (52)	987	0.06	0.06	-0.01	0.07	[0.01,	2	70	0.06	[0.02,	3	59
		(0.02)	(0.02)	(0.1)		0.15]				0.12]		
Macy's (53)	851	0.19	0.19	0.02	0.07	[0.01,	2	71	0.06	[0.02,	3	49
		(0.03)	(0.03)	(0.09)		0.14]				0.12]		
TJX (53)	767	0.53	0.49	0.07	0.06	[0.02,	2	72	0.05	[0.02,	3	77
		(0.04)	(0.04)	(0.04)		0.11]				0.09]		
Starbucks (58)	1000	0.3	0.28	0.05	0.06	[0.01,	2	73	0.06	[0.02,	3	56
		(0.03)	(0.03)	(0.07)		0.13]				0.15]		
Sears (incl. repair /	968	0.3	0.29	0.06	0.06	[0.01,	2	74	0.05	[0.01,	3	82
auto) (75-76)		(0.04)	(0.04)	(0.06)		0.12]				0.09]		
KFC (58)	1000	0.35	0.33	0.06	0.06	[0.01,	2	75	0.06	[0.02,	3	69
		(0.04)	(0.04)	(0.05)		0.11]				0.12]		
Lab Corp (80-87)	826	0.14	0.14	-0.01	0.06	[0.01,	2	76	0.05	[0.02,	3	73
		(0.02)	(0.03)	(0.09)		0.13]				0.1]		
Kindred Healthcare	567	0.11	0.14	-0.18	0.06	[0, 0.15]	2	77	0.06	[0.02,	3	54
(80-87)		(0.03)	(0.03)	(0.14)						0.12]		
J.B. Hunt (42-47)	877	0.25	0.25	-0.01	0.06	[0.01,	2	78	0.05	[0.01,	3	81
		(0.04)	(0.04)	(0.08)		0.13]				0.09]		
Geico (61-64)	432	0.43	0.44	-0.02	0.06	[0.01,	2	79	0.05	[0.01,	3	74
		(0.06)	(0.06)	(0.09)		0.13]				0.1]		
WestRock (24-35)	606	0.21	0.21	-0.02	0.06	[0.01,	2	80	0.05	[0.02,	3	67
		(0.04)	(0.04)	(0.09)		0.13]				0.11]		
Publix (54)	947	0.76	0.72	0.06	0.05	[0.01,	2	81	0.04	[0.01,	4	89
		(0.03)	(0.03)	(0.04)		0.1]				0.07]		
Ulta Beauty (72-73)	999	0.24	0.23	0.01	0.05	[0.01,	2	82	0.05	[0.01,	3	79
		(0.03)	(0.03)	(0.07)		0.11]				0.1]		
AECOM (80-87)	374	0.12	0.12	0.04	0.05	[0.01,	2	83	0.04	[0.01,	4	88
		(0.05)	(0.05)	(0.04)		0.1]				0.07]		
McLane Company	704	0.4	0.4	-0.01	0.05	[0, 0.1]	3	84	0.05	[0.01,	3	80
(50-51)		(0.04)	(0.04)	(0.06)						0.09]		
Target (53)	974	0.2	0.2	-0.01	0.05	[0, 0.1]	3	85	0.05	[0.02,	3	76
		(0.03)	(0.03)	(0.06)						0.09]		
FedEx (42-47)	648	0.19	0.2	-0.03	0.04	[0, 0.1]	3	86	0.04	[0.01,	3	86
		(0.04)	(0.04)	(0.07)						0.08]		
Lowe's (52)	788	0.36	0.36	0	0.04	[0, 0.09]	3	87	0.04	[0.01,	3	85
		(0.04)	(0.04)	(0.05)						0.08]		
Ryder System (42-47)	914	0.18	0.19	-0.03	0.04	[0, 0.1]	3	88	0.04	[0.01,	4	87
		(0.03)	(0.03)	(0.06)						0.07]		
Kohl's (53)	944	0.53	0.52	0.02	0.03	[0, 0.07]	3	89	0.04	[0.01,	4	90
		(0.04)	(0.04)	(0.03)						0.06]		
Mondelez (20)	788	0.44	0.44	-0.01	0.03	[0, 0.08]	3	90	0.03	[0.01,	4	93
		(0.04)	(0.04)	(0.04)						0.06]		

Continued on next page

Hilton (65-70)	886	0.24	0.26	-0.11	0.03	[0, 0.09]	3	91	0.04	[0.01,	3	83
		(0.04)	(0.04)	(0.07)						0.09]		
Sysco (50-51)	941	0.18	0.18	0	0.03	[0, 0.07]	3	92	0.04	[0.01,	4	91
		(0.03)	(0.03)	(0.04)						0.07]		
Waste Management	930	0.45	0.46	-0.03	0.03	[0, 0.07]	3	93	0.04	[0.01,	4	92
(49)		(0.04)	(0.04)	(0.04)						0.07]		
Kroger (54)	940	0.46	0.46	0	0.02	[0, 0.05]	3	94	0.02	[0.01,	4	96
		(0.04)	(0.04)	(0.03)						0.05]		
Avis-Budget (75-76)	797	0.31	0.32	-0.05	0.02	[0, 0.06]	3	95	0.03	[0.01,	4	94
		(0.04)	(0.04)	(0.04)						0.06]		
Dr Pepper (20)	537	0.88	0.94	-0.07	0.02	[0, 0.05]	3	96	0.03	[0.01,	4	95
		(0.04)	(0.02)	(0.04)						0.05]		
Charter / Spectrum	960	0.45	0.46	-0.03	0.02	[0, 0.04]	3	97	0.02	[0.01,	4	97
(48)		(0.04)	(0.04)	(0.03)						0.05]		

Notes: This table reports estimated contact penalties and the results of empirical Bayes and grading exercises for race. Each firm's industry (2-digit SIC code group) is shown in parentheses. The next column reports the total number of applications sent to this firm. The columns $\hat{p}_w$ and $\hat{p}_b$ give estimates of the probability that a white and Black application (respectively) is contacted at the average job sampled from the firm in question. The column $\hat{\theta}_i$ reports contact penalties (with positive values indicating discrimination against Black applicants). Job-clustered standard errors are reported in parentheses. The remaining columns report posterior means (Post. mean), 95% credible intervals (Post. CI), assigned grades using $\lambda = 0.25$ (Grd), and Condorcet ranks (Cond. rank), which are grades under $\lambda = 1$, in the baseline model and the model with industry effects.

Table F6: Gender discrimination: Detailed results by firm

					Baseline model				Industry effect model			
Firm (SIC group)	# apps	$\hat{p}_m$	$\hat{p}_f$	$\hat{\theta}_i$	Post. Mean	Post. CI	Grd	Cond. rank	Post. Mean	Post. CI	Grd	Cond. rank
Builders FirstSource	581	0.11	0.02	1.57	0.90	[0.15,	1	1	0.67	[-0.09,	1	1
(24-35)		(0.03)	(0.01)	(0.55)		1.63]				1.44]		
LKQ Auto (50-51)	587	0.29	0.15	0.66	0.30	[0.04,	2	2	0.26	[-0.01,	2	2
		(0.05)	(0.04)	(0.21)		0.55]				0.54]		
JPMorgan Chase	981	0.08	0.05	0.45	0.19	[-0.11,	2	3	0.13	[-0.16,	3	4
(61-64)		(0.02)	(0.02)	(0.26)		0.51]				0.48]		
Honeywell (50-51)	556	0.19	0.12	0.42	0.17	[-0.05,	2	4	0.15	[-0.06,	2	3
		(0.05)	(0.03)	(0.19)		0.4]				0.41]		
CVS Health (59)	787	0.05	0.04	0.38	0.15	[-0.16,	3	5	0.11	[-0.19,	3	6
		(0.02)	(0.01)	(0.28)		0.5]				0.49]		
Goodyear (55)	387	0.08	0.06	0.3	0.11	[-0.17,	3	6	0.11	[-0.15,	3	5
		(0.04)	(0.03)	(0.27)		0.44]				0.46]		
AutoNation (55)	869	0.13	0.1	0.28	0.10	[-0.15,	3	7	0.10	[-0.13,	3	7
		(0.03)	(0.02)	(0.24)		0.38]				0.41]		
UGI (49)	546	0.11	0.08	0.27	0.10	[-0.16,	3	8	0.07	[-0.17,	3	12
		(0.03)	(0.03)	(0.25)		0.39]				0.4]		
Target (53)	974	0.23	0.18	0.26	0.09	[-0.05,	3	9	0.08	[-0.06,	3	9
		(0.04)	(0.03)	(0.14)		0.25]				0.26]		
WestRock (24-35)	606	0.24	0.19	0.24	0.08	[-0.1,	3	10	0.08	[-0.1,	3	8
		(0.05)	(0.04)	(0.18)		0.3]				0.32]		
Costco (53)	1000	0.07	0.05	0.24	0.08	[-0.21,	3	11	0.07	[-0.2,	3	15
		(0.02)	(0.01)	(0.29)		0.43]				0.44]		
O'Reilly Automotive	973	0.33	0.26	0.22	0.07	[-0.05,	3	12	0.07	[-0.05,	3	10
(55)		(0.03)	(0.03)	(0.12)		0.2]				0.22]		
Avis-Budget (75-76)	797	0.34	0.29	0.19	0.06	[-0.04,	3	13	0.06	[-0.03,	3	11
		(0.04)	(0.04)	(0.09)		0.16]				0.18]		
KFC (58)	1000	0.37	0.31	0.17	0.05	[-0.04,	3	14	0.05	[-0.04,	3	17
		(0.04)	(0.04)	(0.1)		0.15]				0.17]		
Sherwin-Williams	980	0.5	0.42	0.16	0.04	[-0.03,	3	15	0.05	[-0.03,	3	19
(52)		(0.04)	(0.04)	(0.08)		0.12]				0.14]		
Disney (incl. stores)	858	0.1	0.08	0.17	0.05	[-0.19,	3	16	0.04	[-0.18,	3	23
(59)		(0.03)	(0.02)	(0.24)		0.33]				0.34]		
XPO Logistics	861	0.17	0.14	0.16	0.05	[-0.11,	3	17	0.04	[-0.11,	3	21
(42-47)		(0.03)	(0.03)	(0.16)		0.23]				0.25]		
McLane Company	704	0.43	0.37	0.15	0.04	[-0.06,	3	18	0.05	[-0.05,	3	14
(50-51)		(0.05)	(0.05)	(0.11)		0.16]				0.19]		
Sears (incl. repair /	968	0.32	0.27	0.15	0.04	[-0.06,	3	19	0.05	[-0.05,	3	18
auto) (75-76)		(0.04)	(0.04)	(0.11)		0.16]				0.19]		
Hertz (75-76)	786	0.25	0.21	0.15	0.04	[-0.1,	3	20	0.06	[-0.08,	3	13
		(0.04)	(0.04)	(0.16)		0.21]				0.25]		
Quest Diagnostics	907	0.03	0.03	0.17	0.05	[-0.43,	3	21	0.00	[-0.43,	3	69
(80-87)		(0.02)	(0.01)	(0.46)		0.6]				0.59]		
Tractor Supply	943	0.19	0.17	0.13	0.04	[-0.12,	3	22	0.06	[-0.09,	3	16
(50-51)		(0.03)	(0.03)	(0.16)		0.22]				0.25]		
Starbucks (58)	1000	0.31	0.28	0.1	0.02	[-0.06,	3	23	0.03	[-0.05,	3	22
		(0.04)	(0.03)	(0.09)		0.12]				0.14]		
Walmart (53)	400	0.63	0.57	0.1	0.02	[-0.05,	3	24	0.03	[-0.04,	3	25
		(0.05)	(0.05)	(0.08)		0.1]				0.12]		
UnitedHealth (80-87)	942	0.11	0.09	0.11	0.03	[-0.16,	3	25	0.01	[-0.17,	3	48
		(0.03)	(0.03)	(0.2)		0.25]				0.26]		
Nordstrom (53)	941	0.21	0.19	0.1	0.02	[-0.09,	3	26	0.03	[-0.08,	3	28
		(0.04)	(0.03)	(0.13)		0.16]				0.18]		
Cintas (72-73)	747	0.23	0.21	0.1	0.02	[-0.11,	3	27	0.01	[-0.13,	3	41
		(0.04)	(0.04)	(0.15)		0.19]				0.2]		
US Foods (50-51)	961	0.29	0.27	0.09	0.02	[-0.07,	3	28	0.04	[-0.05,	3	20
		(0.04)	(0.03)	(0.1)		0.13]				0.16]		

Continued on next page

Kohl's (53)	944	0.54	0.5	0.09	0.02	[-0.04,	3	29	0.02	[-0.03,	3	26
		(0.04)	(0.04)	(0.06)		0.08]				0.1]		
DISH (48)	771	0.28	0.26	0.09	0.02	[-0.12,	3	30	0.01	[-0.13,	3	43
		(0.04)	(0.04)	(0.16)		0.19]				0.2]		
Safeway (54)	429	0.26	0.24	0.09	0.02	[-0.15,	3	31	0.03	[-0.14,	3	29
		(0.06)	(0.05)	(0.19)		0.23]				0.25]		
Olive Garden (58)	1000	0.4	0.38	0.06	0.01	[-0.05,	3	32	0.02	[-0.05,	3	31
		(0.04)	(0.04)	(0.08)		0.08]				0.11]		
Best Buy (57)	920	0.18	0.17	0.07	0.01	[-0.12,	3	33	0.02	[-0.11,	3	33
		(0.03)	(0.03)	(0.15)		0.17]				0.2]		
Publix (54)	947	0.76	0.72	0.05	0.01	[-0.02,	3	34	0.01	[-0.02,	3	34
		(0.03)	(0.03)	(0.04)		0.04]				0.06]		
Murphy USA (55)	927	0.28	0.27	0.06	0.01	[-0.09,	3	35	0.03	[-0.07,	3	27
		(0.04)	(0.03)	(0.12)		0.13]				0.16]		
AT&T (48)	893	0.12	0.12	0.07	0.01	[-0.15,	3	36	0.00	[-0.15,	3	54
		(0.02)	(0.02)	(0.18)		0.2]				0.21]		
United Rentals	917	0.13	0.12	0.07	0.01	[-0.16,	3	37	0.00	[-0.17,	3	62
(72-73)		(0.02)	(0.03)	(0.19)		0.22]				0.23]		
Genuine Parts (Napa	966	0.29	0.28	0.05	0.01	[-0.09,	3	38	0.02	[-0.07,	3	30
Auto) (55)		(0.04)	(0.03)	(0.11)		0.12]				0.15]		
AutoZone (55)	1000	0.36	0.35	0.04	0.00	[-0.07,	3	39	0.02	[-0.05,	3	32
		(0.04)	(0.04)	(0.09)		0.09]				0.11]		
Dick's (59)	975	0.36	0.35	0.02	0.00	[-0.07,	3	40	0.00	[-0.06,	3	42
		(0.04)	(0.04)	(0.09)		0.08]				0.1]		
Bed Bath & Beyond	998	0.34	0.34	0.02	0.00	[-0.07,	3	41	0.01	[-0.06,	3	38
(57)		(0.04)	(0.04)	(0.09)		0.08]				0.1]		
Dillard's (53)	925	0.32	0.32	0.01	0.00	[-0.09,	3	42	0.01	[-0.08,	3	39
		(0.04)	(0.04)	(0.11)		0.1]				0.13]		
J.B. Hunt (42-47)	877	0.25	0.25	0.01	0.00	[-0.12,	3	43	0.01	[-0.11,	3	44
		(0.04)	(0.04)	(0.13)		0.13]				0.15]		
Advance Auto Parts	967	0.25	0.25	0.01	-0.01	[-0.11,	3	44	0.02	[-0.09,	3	35
(55)		(0.03)	(0.03)	(0.13)		0.12]				0.15]		
Dr Pepper (20)	537	0.9	0.92	-0.02	-0.01	[-0.04,	3	45	0.00	[-0.03,	3	52
		(0.03)	(0.03)	(0.04)		0.02]				0.04]		
US Bank (61-64)	966	0.17	0.17	0.01	-0.01	[-0.13,	3	46	0.00	[-0.12,	3	59
		(0.03)	(0.03)	(0.14)		0.13]				0.15]		
Waste Management	930	0.45	0.46	-0.01	-0.01	[-0.06,	3	47	0.00	[-0.06,	3	49
(49)		(0.04)	(0.04)	(0.07)		0.05]				0.07]		
Geico (61-64)	432	0.44	0.44	0.01	-0.01	[-0.11,	3	48	0.00	[-0.11,	3	56
		(0.06)	(0.06)	(0.13)		0.12]				0.13]		
Ryder System	914	0.18	0.18	0.01	-0.01	[-0.15,	3	49	0.01	[-0.13,	3	46
(42-47)		(0.03)	(0.03)	(0.16)		0.16]				0.19]		
Tyson Foods (20)	797	0.34	0.34	-0.01	-0.01	[-0.1,	3	50	0.00	[-0.09,	3	58
		(0.04)	(0.04)	(0.11)		0.09]				0.11]		
Jones Lang LaSalle	577	0.06	0.06	0.03	0.00	[-0.27,	3	51	-0.03	[-0.29,	3	77
(65-70)		(0.02)	(0.03)	(0.27)		0.29]				0.3]		
Dollar General (53)	787	0.46	0.47	-0.03	-0.01	[-0.07,	3	52	0.00	[-0.06,	3	50
		(0.05)	(0.05)	(0.07)		0.05]				0.07]		
Macy's (53)	851	0.19	0.19	-0.01	-0.01	[-0.13,	3	52	0.00	[-0.11,	3	47
		(0.03)	(0.03)	(0.14)		0.12]				0.15]		
Lowe's (52)	788	0.35	0.36	-0.03	-0.02	[-0.1,	3	53	0.00	[-0.08,	3	45
		(0.04)	(0.04)	(0.1)		0.08]				0.11]		
Mondelez (20)	788	0.43	0.45	-0.04	-0.02	[-0.08,	3	54	-0.01	[-0.07,	3	61
		(0.04)	(0.05)	(0.08)		0.05]				0.07]		
Bath & Body Works	990	0.29	0.3	-0.05	-0.02	[-0.1,	3	55	-0.01	[-0.09,	3	62
(59)		(0.04)	(0.03)	(0.11)		0.07]				0.1]		
Cardinal Health	974	0.21	0.22	-0.04	-0.02	[-0.12,	3	56	0.01	[-0.08,	3	36
(50-51)		(0.03)	(0.03)	(0.12)		0.09]				0.13]		
Dollar Tree (53)	998	0.26	0.27	-0.05	-0.02	[-0.11,	3	57	0.00	[-0.09,	3	55
		(0.03)	(0.04)	(0.11)		0.08]				0.1]		
AECOM (80-87)	374	0.12	0.12	0	-0.02	[-0.27,	3	58	-0.02	[-0.25,	3	74
		(0.05)	(0.05)	(0.26)		0.27]				0.28]		

Continued on next page

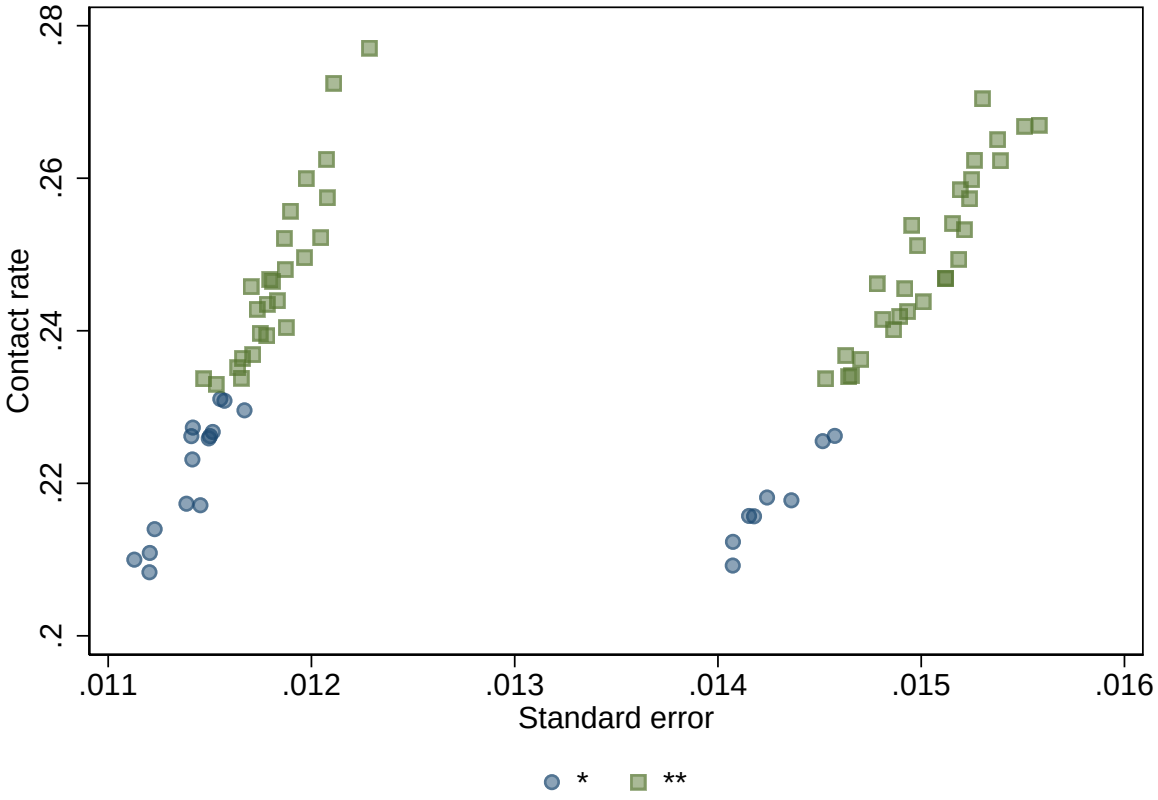
Kroger (54)	940	0.44	0.48	-0.09	-0.03	[-0.08,	3	59	-0.01	[-0.07,	3	63
		(0.04)	(0.04)	(0.07)		0.03]				0.05]		
Pilot Flying J (55)	993	0.31	0.34	-0.09	-0.03	[-0.1,	3	60	0.00	[-0.07,	3	53
		(0.03)	(0.03)	(0.09)		0.04]				0.07]		
Stanley Black &	790	0.04	0.04	0	-0.02	[-0.34,	3	61	0.04	[-0.26,	3	24
Decker (24-35)		(0.02)	(0.02)	(0.33)		0.35]				0.41]		
Pizza Hut (58)	1000	0.37	0.41	-0.1	-0.03	[-0.09,	3	62	-0.01	[-0.07,	3	57
		(0.04)	(0.04)	(0.08)		0.03]				0.07]		
Home Depot (52)	987	0.06	0.06	-0.02	-0.03	[-0.34,	3	63	0.02	[-0.26,	3	37
		(0.02)	(0.02)	(0.32)		0.32]				0.38]		
Sysco (50-51)	941	0.17	0.19	-0.11	-0.04	[-0.15,	3	64	0.01	[-0.1,	3	40
		(0.03)	(0.04)	(0.13)		0.08]				0.13]		
Charter / Spectrum	960	0.42	0.49	-0.15	-0.04	[-0.11,	3	65	-0.03	[-0.09,	3	73
(48)		(0.04)	(0.05)	(0.08)		0.02]				0.04]		
TJX (53)	767	0.48	0.55	-0.15	-0.04	[-0.11,	3	66	-0.02	[-0.09,	3	66
		(0.04)	(0.04)	(0.08)		0.02]				0.05]		
Walgreens (59)	910	0.35	0.41	-0.15	-0.04	[-0.11,	3	67	-0.02	[-0.09,	3	70
		(0.04)	(0.04)	(0.09)		0.02]				0.05]		
International Paper	954	0.2	0.22	-0.13	-0.05	[-0.14,	3	68	-0.01	[-0.11,	3	60
(24-35)		(0.03)	(0.04)	(0.12)		0.06]				0.1]		
Rite Aid (59)	962	0.18	0.21	-0.14	-0.05	[-0.14,	3	69	-0.02	[-0.12,	3	71
		(0.03)	(0.03)	(0.12)		0.06]				0.09]		
J.C. Penney (53)	994	0.27	0.32	-0.15	-0.05	[-0.13,	3	70	-0.02	[-0.11,	3	68
		(0.04)	(0.04)	(0.11)		0.04]				0.07]		
Ross Stores (53)	650	0.2	0.22	-0.13	-0.05	[-0.17,	3	71	-0.02	[-0.13,	3	67
		(0.03)	(0.03)	(0.15)		0.08]				0.12]		
Ulta Beauty (72-73)	999	0.22	0.25	-0.16	-0.05	[-0.15,	3	72	-0.04	[-0.14,	3	78
		(0.03)	(0.04)	(0.12)		0.05]				0.07]		
Universal Health	586	0.27	0.32	-0.15	-0.05	[-0.17,	3	73	-0.04	[-0.15,	3	79
(80-87)		(0.05)	(0.05)	(0.15)		0.08]				0.1]		
Performance Food	520	0.32	0.37	-0.15	-0.05	[-0.17,	3	74	0.00	[-0.11,	3	51
Group (50-51)		(0.05)	(0.05)	(0.14)		0.07]				0.12]		
Marriott (65-70)	964	0.14	0.16	-0.13	-0.05	[-0.19,	3	75	-0.05	[-0.2,	3	83
		(0.03)	(0.03)	(0.17)		0.11]				0.12]		
GameStop (57)	790	0.05	0.06	-0.09	-0.05	[-0.28,	3	76	-0.01	[-0.23,	3	65
		(0.02)	(0.02)	(0.25)		0.21]				0.26]		
FedEx (42-47)	648	0.18	0.21	-0.16	-0.05	[-0.17,	3	77	-0.02	[-0.13,	3	72
		(0.04)	(0.04)	(0.14)		0.07]				0.1]		
PepsiCo (20)	916	0.05	0.05	-0.1	-0.05	[-0.27,	3	78	-0.03	[-0.23,	3	76
		(0.02)	(0.02)	(0.24)		0.19]				0.23]		
Hilton (65-70)	886	0.23	0.27	-0.18	-0.06	[-0.17,	3	79	-0.05	[-0.17,	3	81
		(0.04)	(0.04)	(0.13)		0.06]				0.08]		
Gap (56)	996	0.27	0.33	-0.2	-0.06	[-0.16,	3	80	-0.25	[-0.34,	4	93
		(0.04)	(0.04)	(0.12)		0.03]				-0.11]		
CarMax (55)	775	0.13	0.15	-0.16	-0.06	[-0.21,	3	81	-0.01	[-0.15,	3	64
		(0.02)	(0.03)	(0.17)		0.1]				0.15]		
Republic Services	943	0.19	0.23	-0.21	-0.07	[-0.18,	3	82	-0.03	[-0.15,	3	75
(49)		(0.03)	(0.04)	(0.14)		0.05]				0.09]		
Foot Locker (56)	995	0.13	0.16	-0.18	-0.07	[-0.21,	3	83	-0.34	[-0.5,	5	94
		(0.03)	(0.03)	(0.17)		0.09]				-0.12]		
Dean Foods (20)	295	0.12	0.13	-0.12	-0.06	[-0.33,	3	84	-0.04	[-0.29,	3	80
		(0.05)	(0.05)	(0.29)		0.24]				0.27]		
Victoria's Secret (56)	931	0.32	0.4	-0.23	-0.07	[-0.2,	3	85	-0.22	[-0.29,	4	92
		(0.04)	(0.04)	(0.1)		0.01]				-0.12]		
Edward Jones	965	0.1	0.13	-0.21	-0.08	[-0.22,	3	86	-0.04	[-0.19,	3	82
(61-64)		(0.02)	(0.02)	(0.17)		0.08]				0.12]		
Lab Corp (80-87)	826	0.12	0.16	-0.29	-0.10	[-0.24,	3	87	-0.06	[-0.2,	3	84
		(0.02)	(0.03)	(0.16)		0.04]				0.08]		
Estee Lauder (72-73)	579	0.11	0.15	-0.31	-0.11	[-0.29,	3	88	-0.08	[-0.25,	3	85
		(0.03)	(0.03)	(0.21)		0.07]				0.11]		
Comcast (48)	231	0.31	0.45	-0.37	-0.13	[-0.33,	3	89	-0.08	[-0.28,	3	86
		(0.07)	(0.08)	(0.22)		0.07]				0.11]		

Continued on next page

Kindred Healthcare (80-87)	567	0.1 (0.03)	0.15 (0.04)	-0.36 (0.25)	-0.14	[-0.36, 0.1]	3	90	-0.09	[-0.3, 0.14]	3	87
VFC (North Face / Vans) (56)	791	0.12 (0.03)	0.19 (0.04)	-0.42 (0.18)	-0.15	[-0.35, 0.02]	4	91	-0.42	[-0.55, -0.22]	5	95
Aramark (72-73)	935	0.05 (0.01)	0.07 (0.02)	-0.38 (0.25)	-0.14	[-0.36, 0.09]	3	92	-0.09	[-0.31, 0.13]	3	88
CBRE (65-70)	597	0.02 (0.01)	0.05 (0.02)	-0.72 (0.38)	-0.29	[-0.66, 0.08]	4	93	-0.19	[-0.61, 0.16]	4	91
State Farm (61-64)	481	0.05 (0.03)	0.08 (0.04)	-0.53 (0.66)	-0.28	[-0.96, 0.45]	3	94	-0.15	[-0.75, 0.54]	3	89
Nationwide (61-64)	455	0.05 (0.02)	0.1 (0.04)	-0.73 (0.48)	-0.32	[-0.8, 0.17]	4	95	-0.17	[-0.61, 0.29]	3	90
Ascena (Ann Taylor / Loft) (56)	590	0.21 (0.04)	0.42 (0.05)	-0.66 (0.17)	-0.44	[-0.68, -0.09]	4	96	-0.44	[-0.57, -0.3]	5	96

Notes: This table reports estimated contact differences and the results of empirical Bayes and grading exercises for gender. Each firm's industry (2-digit SIC code group) is shown in parentheses. The next column reports the total number of applications sent to this firm. The columns $\hat{p}_w$ and $\hat{p}_b$ give estimates of the probability that a male and female application (respectively) is contacted at the average job sampled from the firm in question. The column $\hat{\theta}_i$ reports contact differences (with positive values indicating favoring male applicants). Job-clustered standard errors are reported in parentheses. The remaining columns report posterior means (Post. mean), 95% credible intervals (Post. CI), assigned grades using $\lambda = 0.25$ (Grd), and Condorcet ranks (Cond. rank), which are grades under $\lambda = 1$, in the baseline model and the model with industry effects.

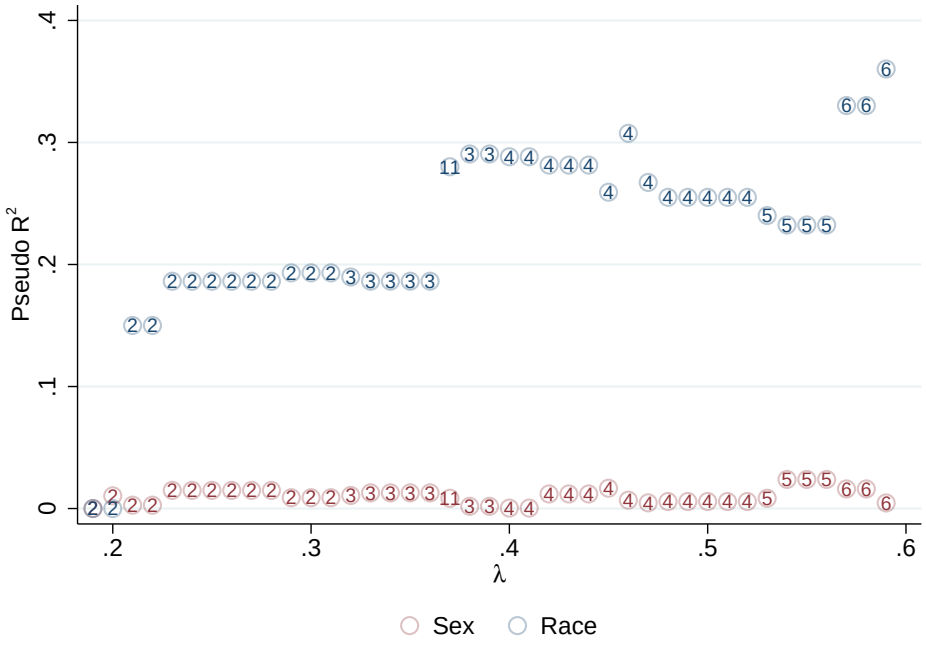
Figure F1: Contact rates, standard errors, and name grades



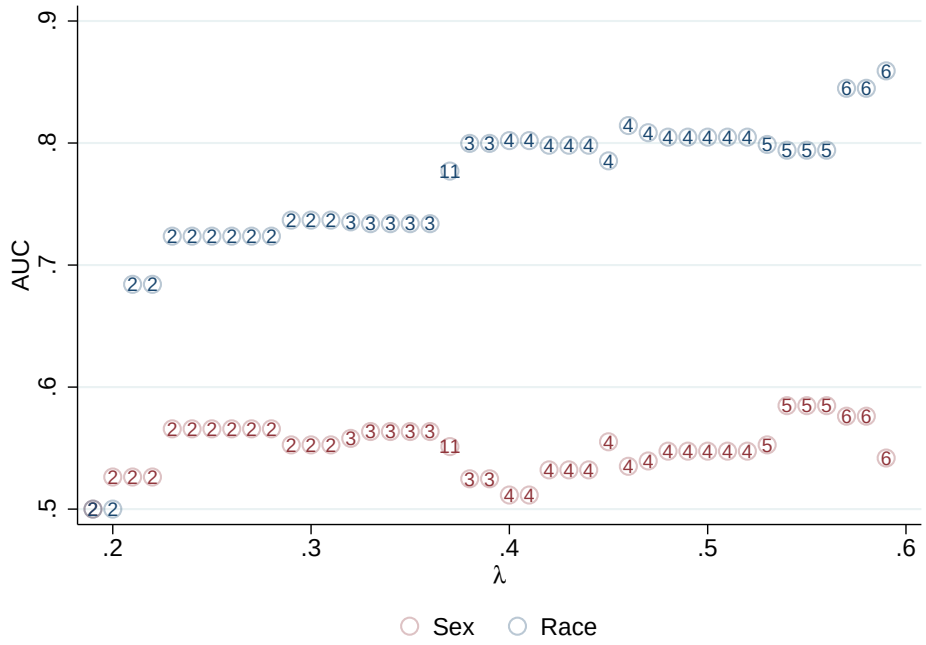
Notes: This figure plots the estimated contact rates for each name against its standard error. The shape and color of each point indicate the grade assigned to the name using the same specification as Figure 3.

Figure F2: Predictive power of grades name for race and sex labels

a) Pseudo R^2



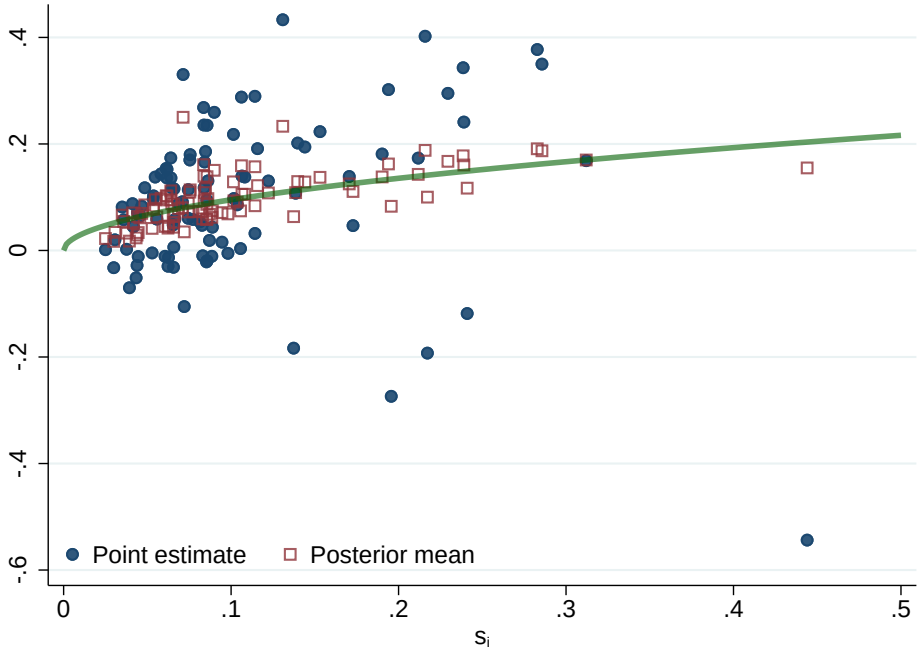
b) Area under the curve



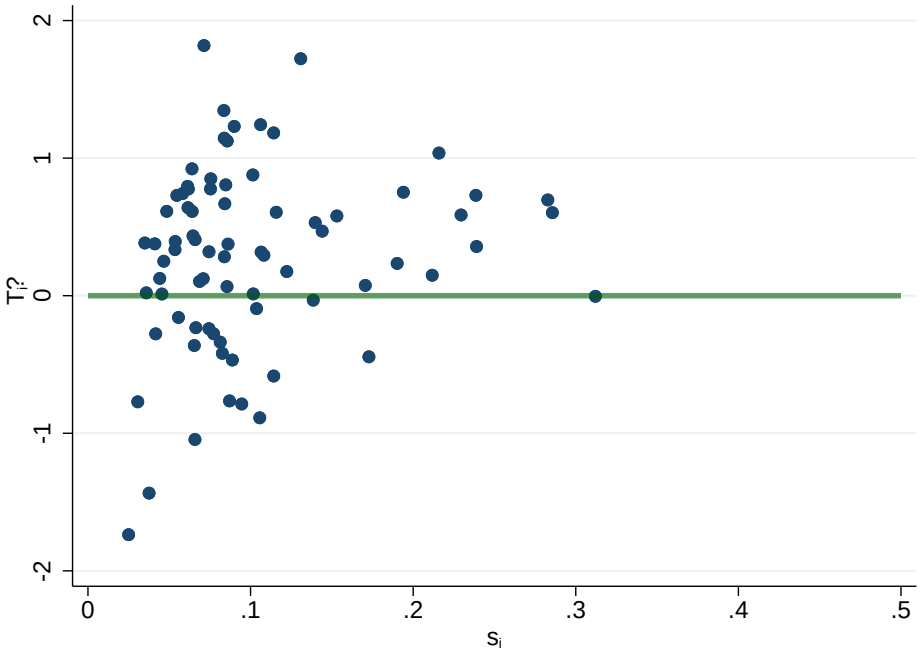
Notes: This figure plots the psuedo-R^2 (Panel (a)) and AUC (Panel (b)) for a series of logistic regressions using an indicator for the race or sex of the name as the outcome and dummies for assigned grades as the explanatory variables for an intermediate range of λ. The number shown indicates the number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

Figure F3: Unadjusted and studentized racial contact gaps against standard errors

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



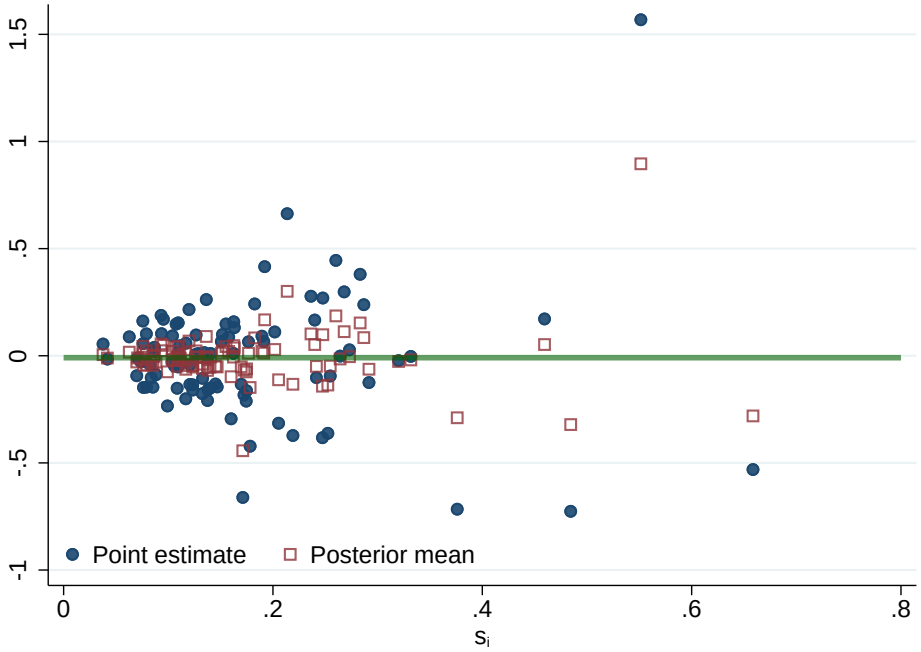
b) $\hat{T}_i$ vs. s_i



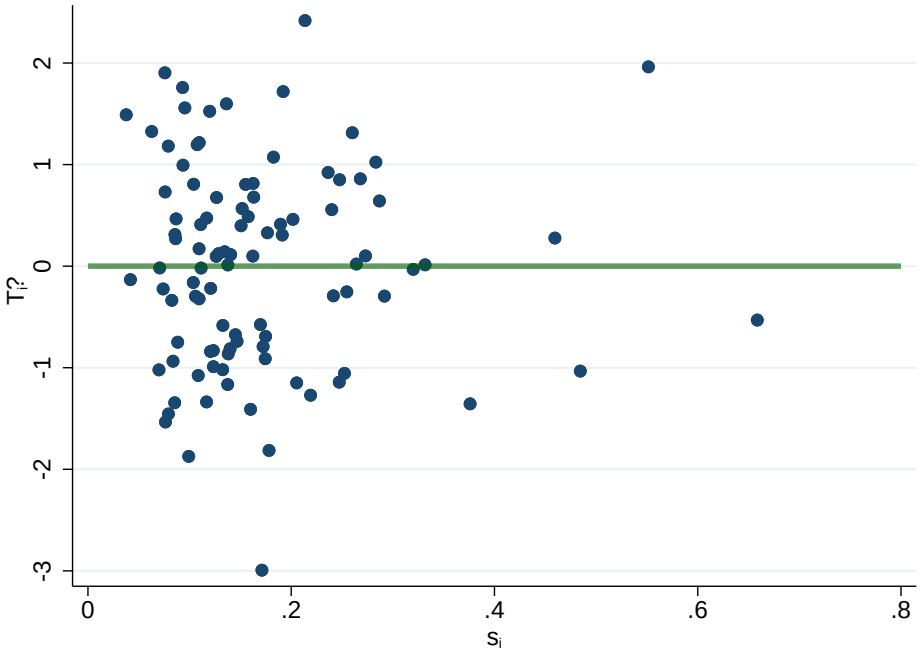
Notes: Panel (a) of this figure plots estimated race contact gaps against their standard errors. The green line plots the conditional mean of θ_i given s_i implied by the GMM estimates. Posterior mean estimates $\bar{\theta}_i$ from the baseline model are superimposed on this panel to illustrate EB shrinkage of contact gaps towards the conditional mean. Panel (b) plots studentized contact gaps $\hat{T}_i$ against standard errors. The green line plots the relationship implied by the model.

Figure F4: Unadjusted and studentized gender contact gaps against standard errors

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

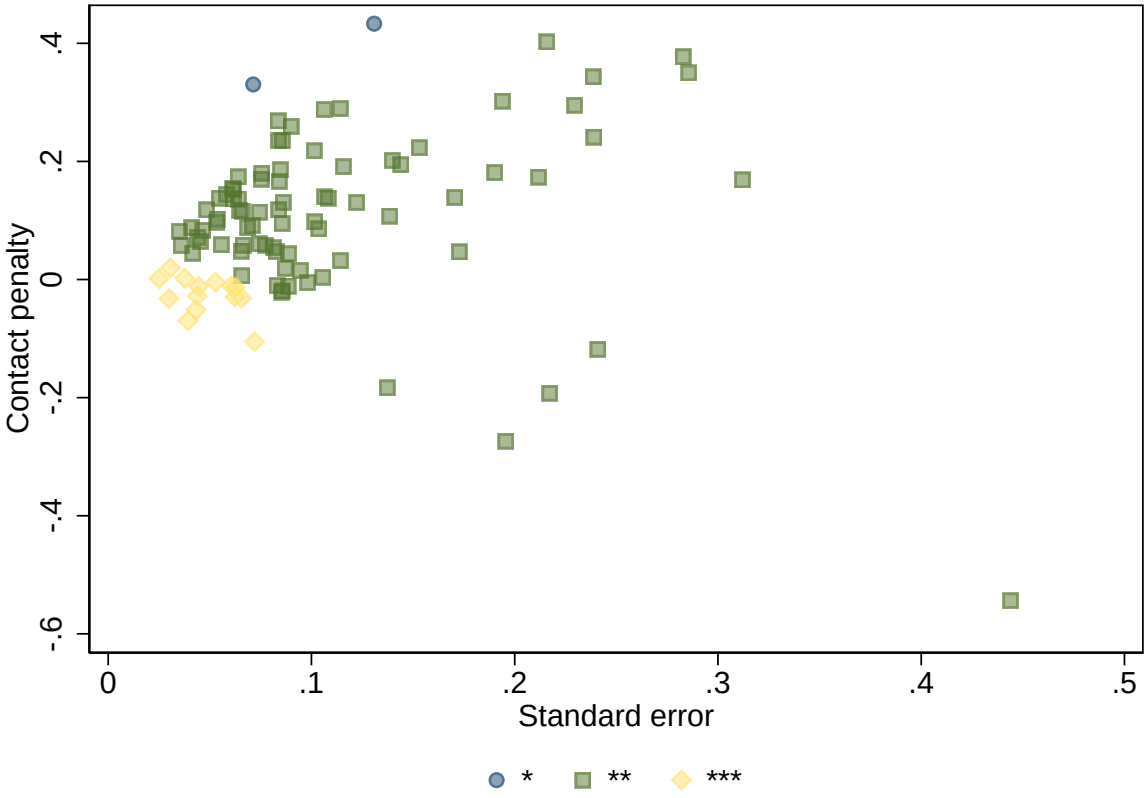


b) $\hat{T}_i$ vs. s_i



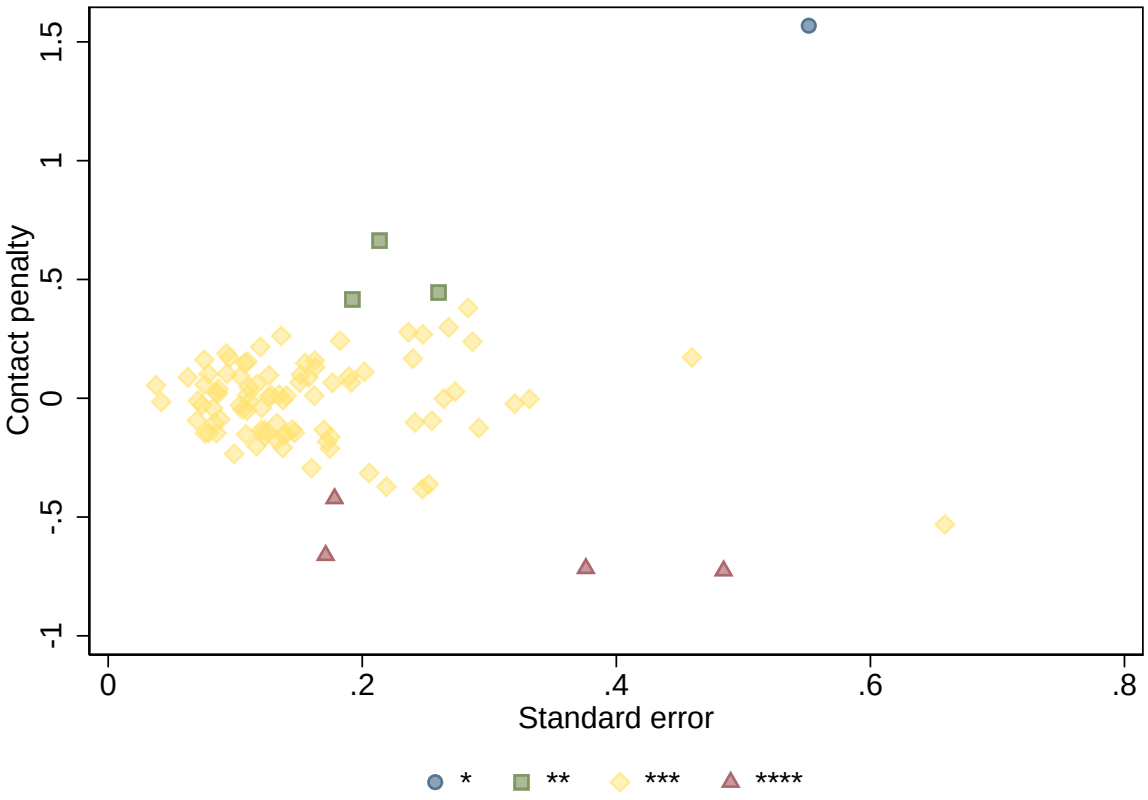
Notes: Panel (a) of this figure plots estimated gender contact gaps against their standard errors. The green line plots the conditional mean of θ_i given s_i implied by the GMM estimates. Posterior mean estimates $\bar{\theta}_i$ from the baseline model are superimposed on this panel to illustrate EB shrinkage of contact gaps towards the conditional mean. Panel (b) plots studentized contact gaps $\hat{T}_i$ against standard errors. The green line plots the relationship implied by the model.

Figure F5: Race: Contact penalties, standard errors, and report card grades



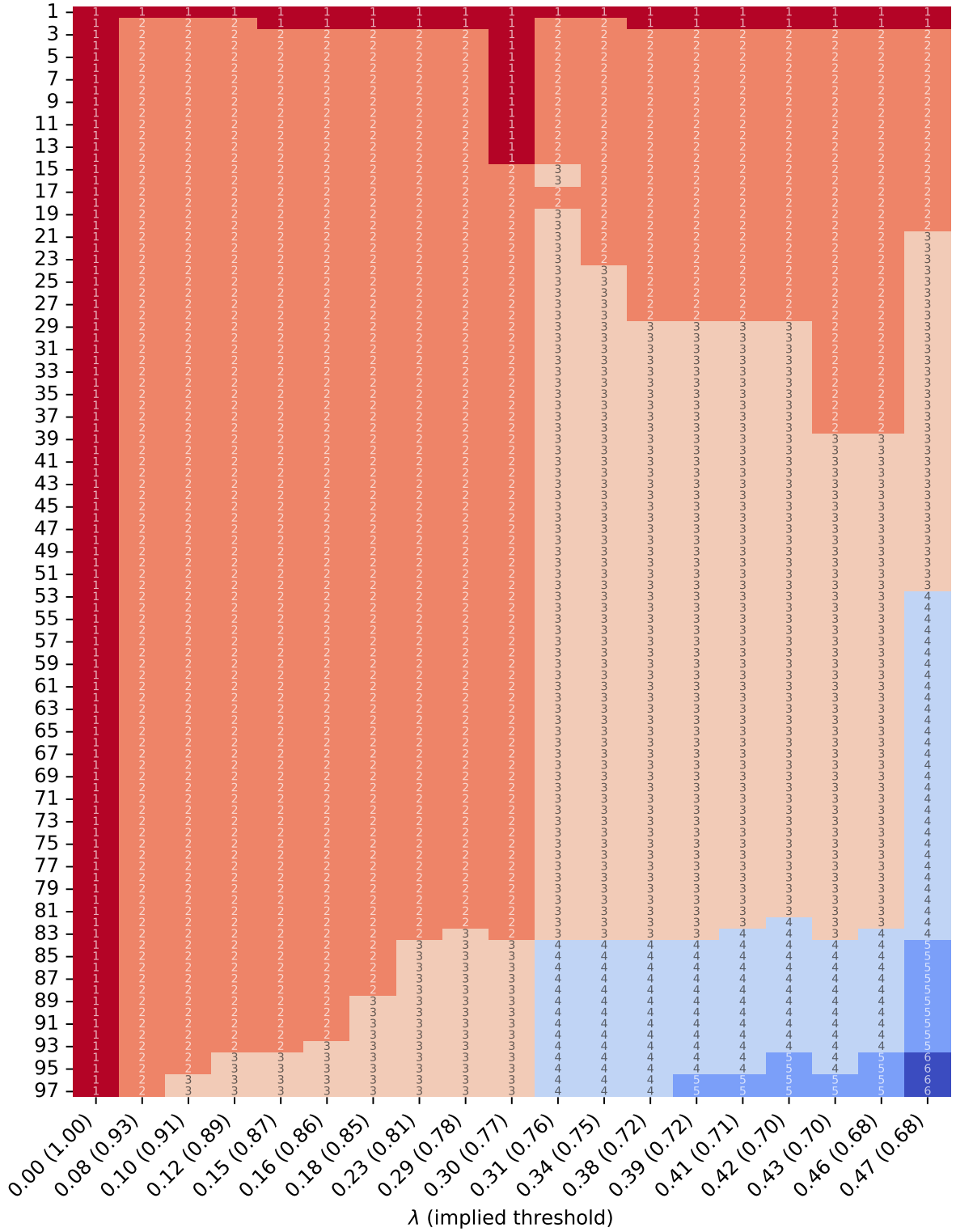
Notes: This figure plots the estimated contact penalty for a Black name at each firm against the standard error of the contact penalty estimate. The shape and color of each point indicate the grade assigned to the firm using the same specification as Figure 7.

Figure F6: Gender: Contact penalties, standard errors, and report card grades



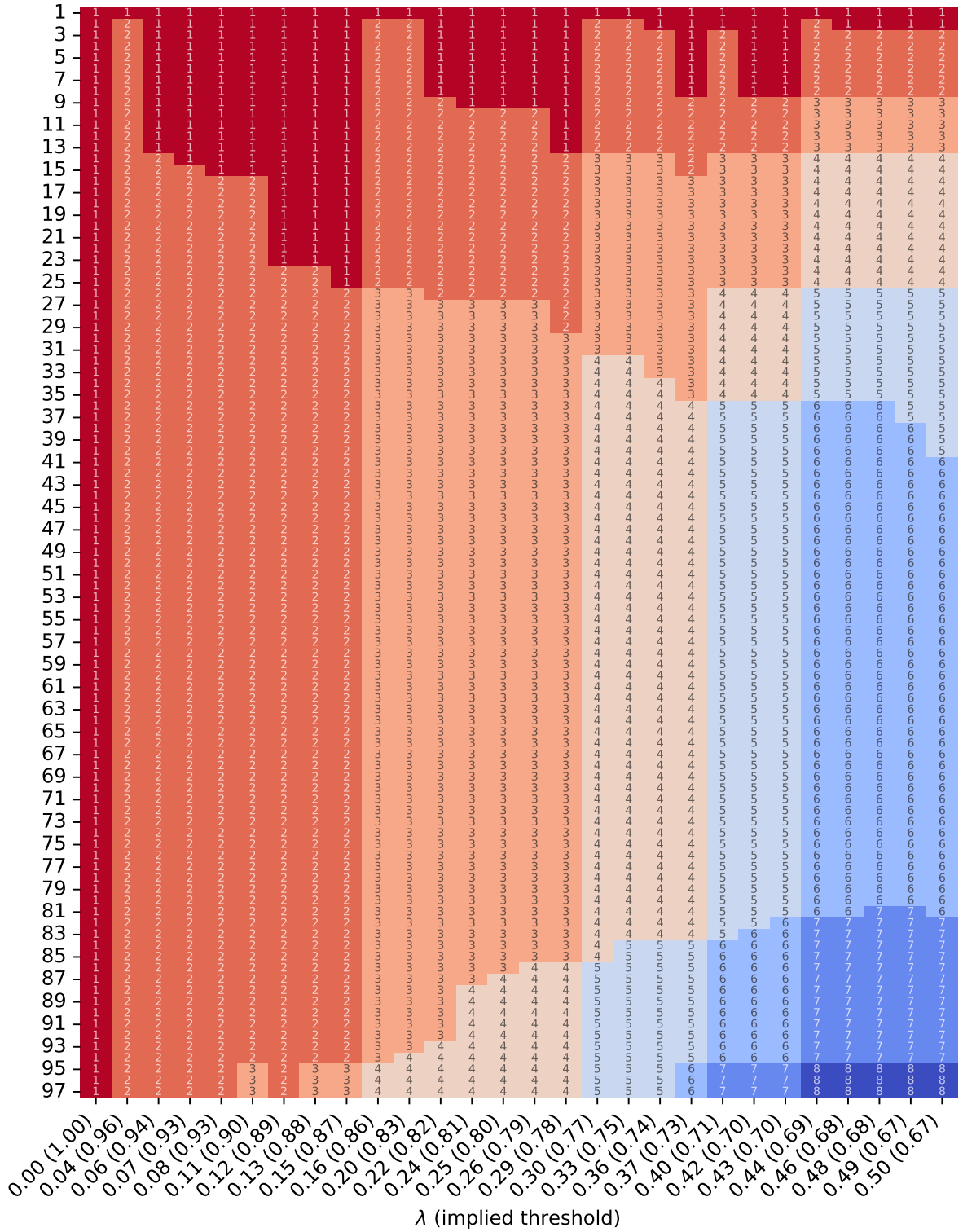
Notes: This figure plots the estimated gender contact difference for each firm against the standard error of the contact difference estimate. The shape and color of each point indicate the grade assigned to the firm using the same specification as Figure 14.

Figure F7: Race: All firm grades (baseline)



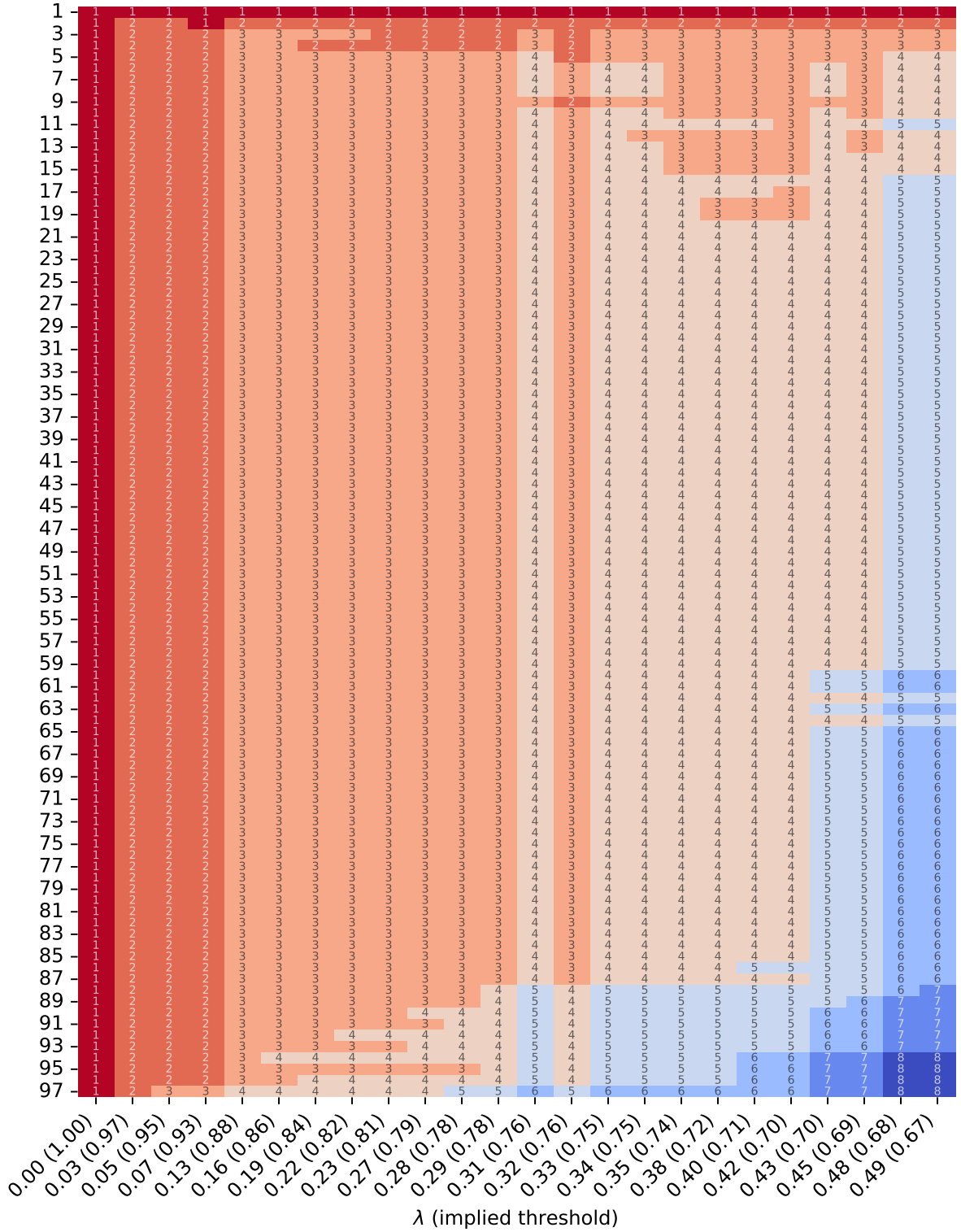
Notes: This figure shows race grade assignments for each value of $\lambda <= 0.5$ from a baseline model without industry effects. To increase readability, only the smallest λ that yields each unique set of grades is retained. The horizontal axis reports this λ and the corresponding value of $1/(1+\lambda)$, 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 F8: Race: All firm grades (industry effects)



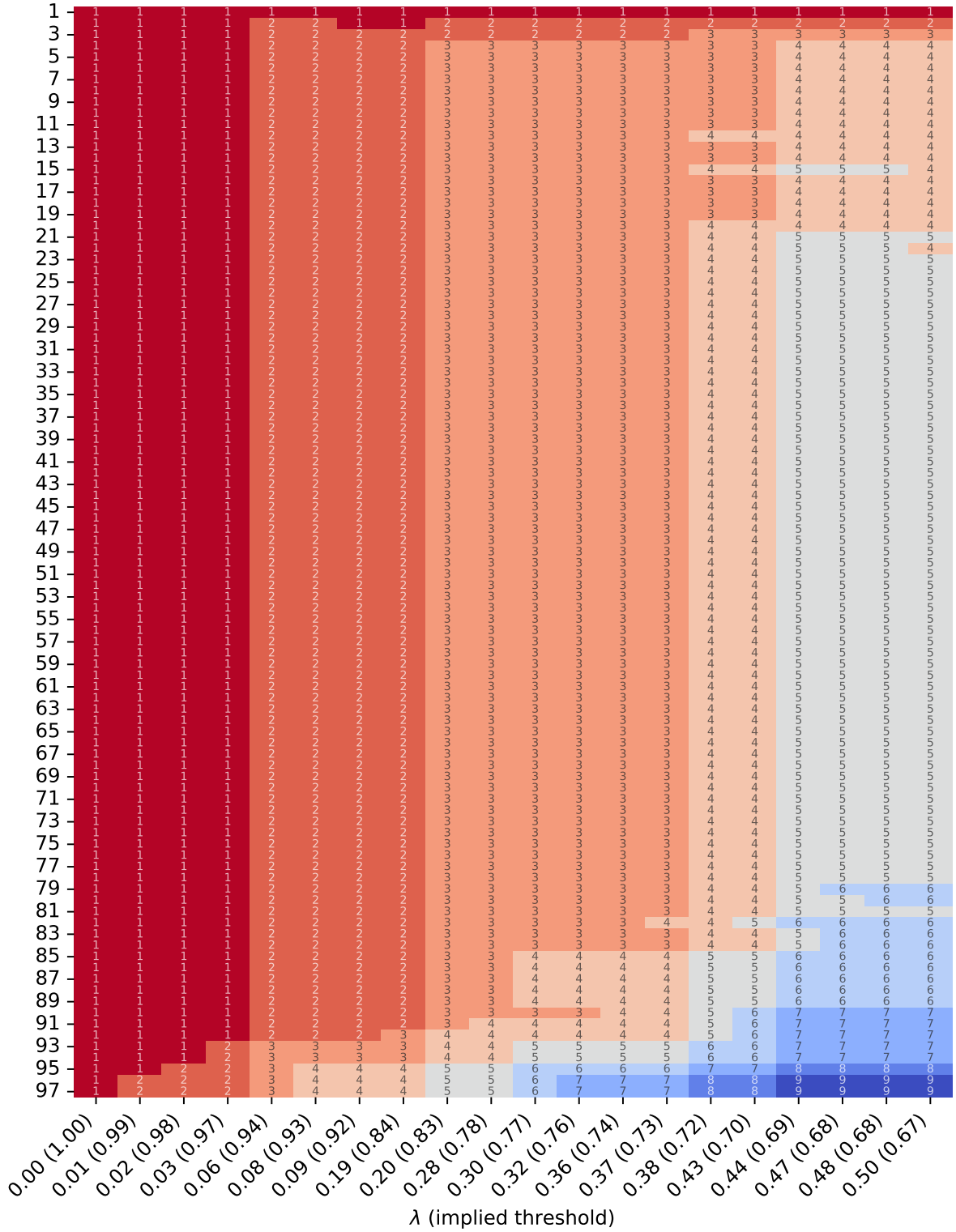
Notes: This figure shows race grade assignments for each value of $\lambda \leq 0.5$ from a model with industry effects. To increase readability, only the smallest λ that yields each unique set of grades is retained. The horizontal axis reports this λ and the corresponding value of $1/(1+\lambda)$, which 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 F9: Gender: All firm grades (baseline)



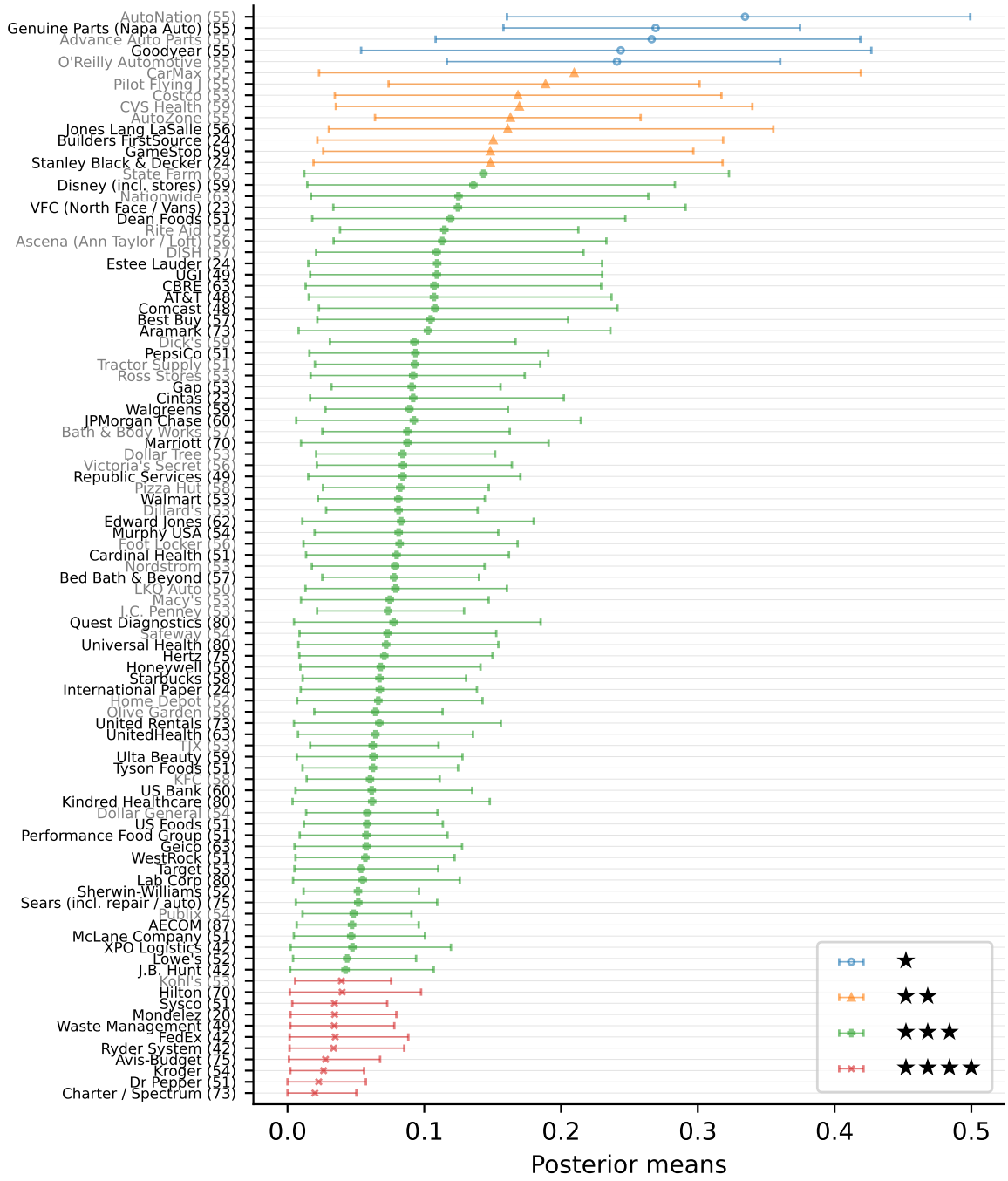
Notes: This figure shows gender grade assignments for each value of $\lambda <= 0.5$ from a baseline model without industry effects. To increase readability, only the smallest λ that yields each unique set of grades is retained. The horizontal axis reports this λ and the corresponding value of $1/(1+\lambda)$, 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 F10: Gender: All firm grades (industry effects)



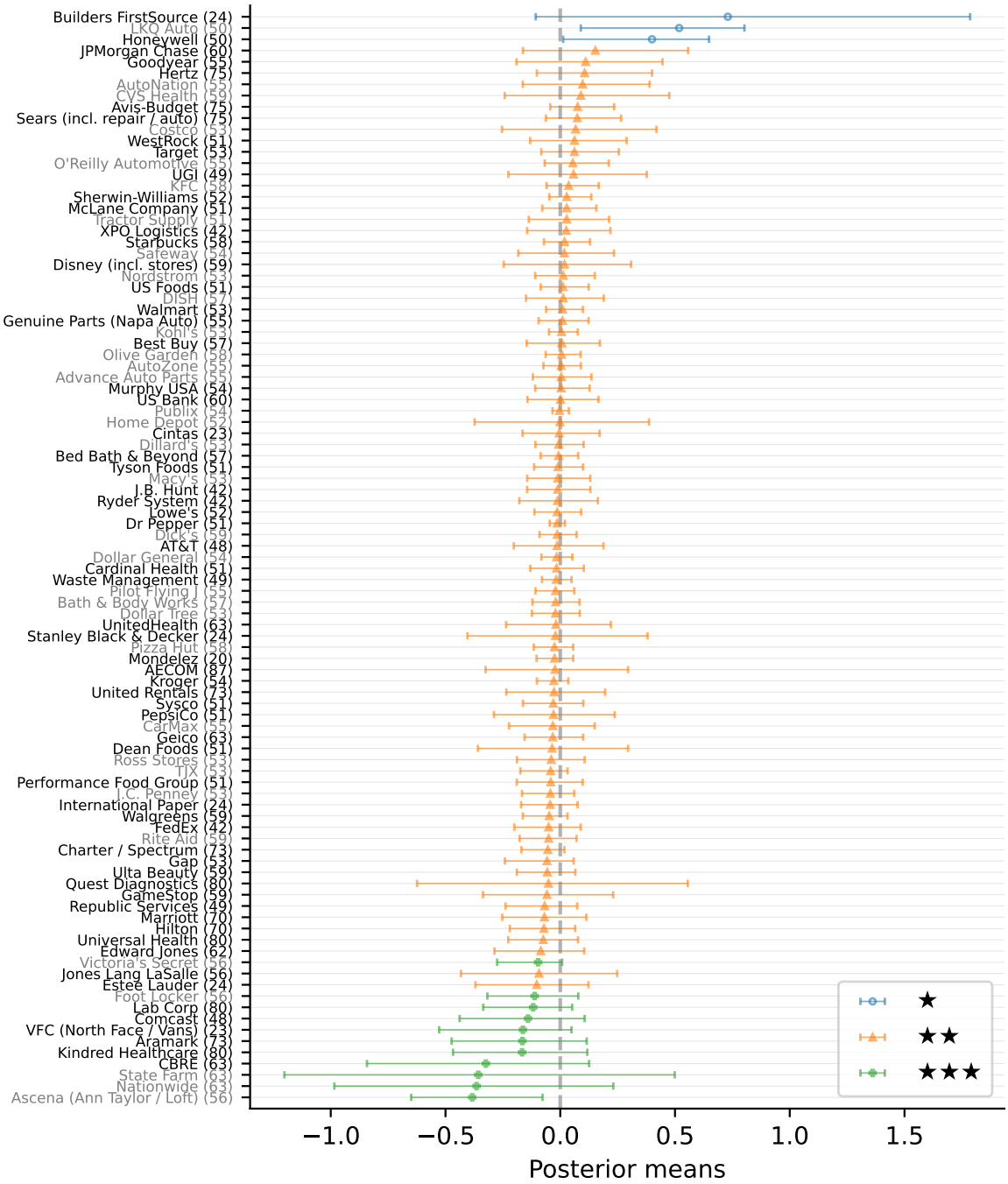
Notes: This figure shows gender grade assignments for each value of $\lambda \leq 0.5$ from a model with industry effects. To increase readability, only the smallest λ that yields each unique set of grades is retained. The horizontal axis reports this λ and the corresponding value of $1/(1+\lambda)$, 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 F11: Race report card using alternative industry codings



Notes: This figure shows posterior mean proportional contact penalties for distinctively Black names, 95% credible intervals, and assigned grades from the industry random effect model. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Posterior estimates come from a model with industry effects using the same industry assignments and groupings as in Kline, Rose and Walters (2022). Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade. Firms labeled with black text are federal contractors, whereas firms in gray are not.

Figure F12: Gender report card using alternative industry codings



Notes: This figure shows posterior mean proportional gender contact differences between distinctively male and female names, 95% credible intervals, and assigned grades from the industry random effect model. Negative differences imply favoring female applications on average, while positive differences imply favoring men. Grades are shown for $\lambda = 0.25$, implying an 80% threshold for posterior ranking probabilities. Posterior estimates come from a model with industry effects using the same industry assignments and groupings as in Kline, Rose and Walters (2022). Firms are ordered by their rank under $\lambda = 1$, when each firm is assigned its own grade. Firms labeled with black text are federal contractors, whereas firms in gray are not.