American Economic Review 2024, 114(8): 2472–2525
https://doi.org/10.1257/aer.20230700

A Discrimination Report Card[†]

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

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 US 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. (*JEL* C11, D22, J15, J16, J71)

> *Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.*[1]
>
> —Louis Brandeis

Scholars, policymakers, and private businesses increasingly produce 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). 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

*Kline: UC Berkeley and NBER (email: pkline@econ.berkeley.edu); Rose: University of Chicago and NBER (email: ekrose@uchicago.edu); Walters: UC Berkeley and NBER (email: crwalters@econ.berkeley.edu). Isaiah Andrews was the coeditor for this article. 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. The experiment used in our analysis was funded by a grant from J-PAL North America's Social Policy Research Initiative. Routines for implementing the ranking procedures developed in this paper are available online at https://github.com/ekrose/drrank.

[†]Go to https://doi.org/10.1257/aer.20230700 to visit the article page for additional materials and author disclosure statements.

[1] Brandeis, Louis, D. 1913. "What Publicity Can Do." *Harper's Weekly*, December 20.

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 in our data deposit (Kline, Rose, and Walters 2024).

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 US 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 nonparametric 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 the task 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 percent 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 percent. These grades capture roughly 25 percent 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 percent, while the gap among firms awarded the best grade is only 3 percent.

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's (2016) 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 percent 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 percent.

Firms assigned the worst grade in this ranking contact White applicants 23 percent more often than Black applicants, similar to the lowest category in the ranking without industry effects. However, nine firms receive this label in the model with industry effects compared to only two in the baseline model, an indication of the extra information conveyed by industry. Similar to the specification without industry, the eleven 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

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 percent 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 percent. 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 percent 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 percent. 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). 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 nonparametric 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's (2022) 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 (US EEOC 1996; Onwuachi-Willig and Barnes 2005; and *US EEOC v. Target* 2006[2]), 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.

I. 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 US county. This restriction was intended to ensure nationwide 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 to 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 nine unique names for each race and gender group. This list was supplemented with ten additional names per group from a

[2] *US Equal Employment Opportunity Commission v. Target Corp*, 460 F.3d 946 7th Cir. 2006.

database of speeding tickets issued in North Carolina between 2006 and 2018. We classified a name as racially distinctive if more than 90 percent 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 US 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 online 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 email 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 email address or phone number. Further details on the experimental design are available in Kline, Rose, and Walters (2022).

II. 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.

A. *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

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

The optimal grading policy is a simple posterior threshold rule:

- Set $(d_i = 2, d_j = 1)$ if and only if $\pi_{ij} > 1/(1+\lambda)$.
- Set $(d_i = 1, d_j = 2)$ if and only if $\pi_{ji} > 1/(1+\lambda)$.
- 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 $(1+q)/2 + [(1+q)/2]EU(\pi_{ij},d;(1-q)/(1+q))$. Hence, the same λ-thresholding decision rule is optimal with $\lambda = (1-q)/(1+q) \in (0,1)$ 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.

B. *Compound Loss*

Now consider the case where we can gamble on the relative rank of all $\binom{n}{2}$ pairs of firms. Kendall's (1938) 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ denote the vector of latent biases and $\mathbf{d} = (d_1, \ldots, d_n)'$ a vector of assigned grades, our loss function can be written,

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

While every discordant pair yields a loss of one, 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 $\mathbf{d}$ and $\boldsymbol{\theta}$, which we denote $\tau(\mathbf{d}, \boldsymbol{\theta})$. When $\lambda < 1$, ranking mistakes are more costly than forgone concordances, which creates an incentive to declare ties.

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

$$L(\mathbf{d}, \boldsymbol{\theta}; \lambda) = (1 - \lambda) DP(\mathbf{d}, \boldsymbol{\theta}) - \lambda \tau(\mathbf{d}, \boldsymbol{\theta}),$$

where the quantity $DP(\mathbf{d}, \boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \big(\mathbf{1}\{\theta_i > \theta_j, d_i < d_j\} + \mathbf{1}\{\theta_i < \theta_j, d_i > d_j\} \big)$ 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 $\tau(\mathbf{d}, \boldsymbol{\theta})$ against concerns about misclassifying them, as reflected by $DP(\mathbf{d}, \boldsymbol{\theta})$.[3]

C. *Risk Function*

While we would ideally like to choose grades $\mathbf{d}$ that balance the rank correlation $\tau(\mathbf{d}, \boldsymbol{\theta})$ against the discordance proportion $DP(\mathbf{d}, \boldsymbol{\theta})$, these quantities are not directly observed. However, the expected values of both $\tau(\mathbf{d}, \boldsymbol{\theta})$ and $DP(\mathbf{d}, \boldsymbol{\theta})$ under beliefs B can be expressed in terms of the pairwise probabilities π_{ij}. The expected rank correlation $\bar{\tau}(\mathbf{d}) = E_B[\tau(\mathbf{d}, \boldsymbol{\theta})] = \int \tau(\mathbf{d}, x) dB(x)$ is given by

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

[3] Online 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.

Likewise, the expected value of $DP(\mathbf{d},\boldsymbol{\theta})$, a quantity we term the *discordance rate* (DR), is

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

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

$$\mathcal{R}(\mathbf{d};\lambda) \;=\; E_B\left[L(\mathbf{d},\boldsymbol{\theta};\lambda)\right] \;=\; (1-\lambda)DR(\mathbf{d}) - \lambda\,\bar{\tau}(\mathbf{d}). \tag{3}$$

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

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

Hence, we can rewrite the problem of choosing $\mathbf{d} \in [n]^n$ 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 $d_i^* = 1 + \sum_{j\in[n]} d_{ji}^*$.

D. *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

$$DR_{g,g'} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{1}\{d_i^* = g\}\mathbf{1}\{d_j^* = g'\}E_B\left[\mathbf{1}\{\theta_i < \theta_j\}\right]}{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{1}\{d_i^* = g\}\mathbf{1}\{d_j^* = g'\}}$$

$$= \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{1}\{d_i^* = g\}\mathbf{1}\{d_j^* = g'\}\pi_{ji}}{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{1}\{d_i^* = g\}\mathbf{1}\{d_j^* = g'\}}.$$

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.

E. *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 IIA, when facing a single pair, the risk minimizing decision rule is

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

Hence, with $\lambda = 1$, it is optimal to choose $d_{ij} = \mathbf{1}\{\pi_{ij} > 1/2\}$, 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 $\pi_{ij} > \pi_{ji} \Leftrightarrow \omega_i > \omega_j$. 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.

F. *Connections to Social Choice*

The literature on ranking methods bears a close connection to problems of social choice. If we reinterpret π_{ij} as the share of votes for firm i over firm j in a pairwise election then a number of standard preference aggregation schemes can be applied.[4] For example, Borda's (1784) voting method simply ranks each firm i based on its number of pairwise election wins: i.e., based upon $\sum_{j \neq i} \mathbf{1}\{\pi_{ij} > 1/2\}$. 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 IIIC turns out to be closely tied to Condorcet's (1785) voting scheme. To develop this connection, it is useful to define the Kemeny (1959) distance between the vectors $\boldsymbol{\theta}$ and $\mathbf{d}$, which can be written

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

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

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

Young and Levenglick (1978) show that Condorcet's (1785) voting scheme is equivalent to choosing a ranking $\mathbf{d}$ 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 $(2\pi_{ij} - 1)(d_{ji} - d_{ij})$ in (7) is minimized by the pairwise MAP thresholding rule $d_{ij} = \mathbf{1}\{\pi_{ij} > 1/2\}$.[5] 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.

[4] 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.

[5] Note that pairwise MAP thresholding need not yield the most likely global ordering. For example, with three 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 $\{1, 2, 3\}$.

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 online 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 nonempty subset of candidates such that every candidate in the subset is majority-preferred over every candidate not in the subset. The following proposition establishes a corresponding property of our grades in the case where $\lambda < 1$.

PROPOSITION 2 (λ-Smith Criterion): *Let $\mathcal{S}$ denote a collection of firms 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 online 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 nonpreferred 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 online Appendix B.

III. 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): $\hat{\theta}_i \,|\, \theta_i, s_i \sim \mathcal{N}\left(\theta_i, s_i^2\right)$ *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* $\left\{\hat{\theta}_i\right\}_{i\in[n]}$ *are mutually independent conditional on* $\left\{\theta_i, s_i\right\}_{i\in[n]}$.

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.[6]

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.

A. *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

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

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

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.

B. *EB Posteriors and Grades*

The empirical Bayes approach treats the estimate $\hat{G}$ as a prior in decision-making. 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.

By Bayes' rule,

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

where ϕ denotes the standard normal density. The empirical Bayes posterior distribution for firm i is $\mathcal{P}\left(t \,|\, \hat{\theta}_i, s_i; \hat{G}\right)$. 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 online 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}. Online 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 and 2. Each π_{ij} gives the share of new firm pairs with realized evidence configuration $\left(\hat{\theta}_i, \hat{\theta}_j, s_i, s_j\right)$ 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.

C. *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

Table 1—Summary Statistics for First Names Sample

	Contact rate (1)	Number of apps (2)	Number of 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 online Appendix Table F2 for a list of first names used in the analysis.

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 online Appendix E, finding in a Monte Carlo simulation calibrated to our leading application that these finite sample biases are small.

IV. 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. The COVID-19 epidemic and other disruptions led to minor imbalances in the sample counts across name categories. 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.[7] These findings suggest 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 considered in their experiment varied widely within race but was insignificantly related to contact rates.[8] 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.

[7] The between-group variance is computed with the formula $\left[(G-1)/G\right]\left(S^2-\bar{s}^2\right)$, 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 $\bar{s}^2$ 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.

[8] 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).

A. *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 of 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 $\big(d/dx\big)\sin^{-1}\sqrt{x} = \Big[2\sqrt{x(1-x)}\Big]^{-1}$. Consequently, the delta method implies $\hat{\theta}_i$ has asymptotic distribution $\mathcal{N}\big(\theta_i,(4N_i)^{-1}\big)$, where $\theta_i \;=\; \sin^{-1}\sqrt{p_i}$ and the variance $(4N_i)^{-1}$ no longer depends on θ_i.[9]

To estimate the distribution G of θ_i we first apply a nonparametric maximum likelihood (NPMLE) estimator (Koenker and Mizera 2014; Koenker and Gu 2017). The NPMLE estimates a discrete approximation to G assuming that $\hat{\theta}_i \,|\, \theta_i, N_i \;\sim\; \mathcal{N}(\theta_i,(4N_i)^{-1})$. 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 ($p \;=\; 0.17$).[10]

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's (2016) 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 $\hat{G}$ with mean and variance as close as possible to the sample mean of the $\big\{\hat{\theta}_i\big\}_{i\in[n]}$ and their debiased variance estimate, as described further in online Appendix D.

[9] To account for imbalance across waves, we replace the ratio C_i/N_i with an average contact probability $\hat{p}_i$ that weights the data in inverse proportion to the number of applications sent in that wave with that first name. Standard errors are derived by applying the delta method to $\hat{\theta}_i \;=\; \sin^{-1}(\sqrt{\hat{p}_i})$ and using a heteroscedasticity robust estimate of $\mathrm{var}\,(\hat{p}_i)$. This turns out to yield a standard error nearly indistinguishable from $(4N_i)^{-1/2}$ (correlation = .98).

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

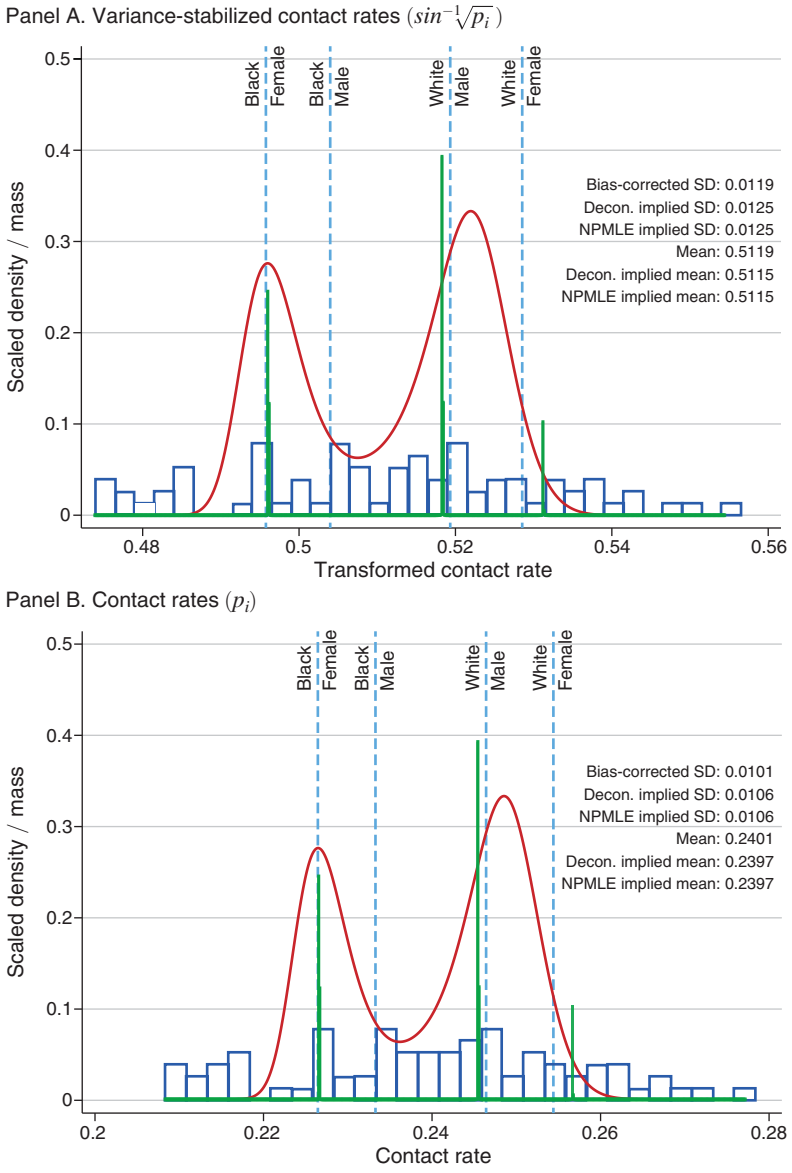

FIGURE 1. DECONVOLUTION ESTIMATES OF NAME-SPECIFIC CONTACT RATE DISTRIBUTIONS

Notes: This figure presents nonparametric 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 nonparametric 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.

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 in 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.

Panel B 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 II) implies that ties are measure zero.

B. *Reporting Possibilities*

Panel A of Figure 2 depicts the EB posterior contrast probabilities $\hat{\pi}_{ij}$ (see the online 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.

Panel B of Figure 2 depicts the discordance rate that arises from minimizing $\mathcal{R}(\mathbf{d};\lambda)$—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.

Panel C 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 $\bar{\theta}_i = \int_{-\infty}^{\infty} t d\mathcal{P}\left(t \mid \hat{\theta}_i, s_i; \hat{G}\right)$. 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

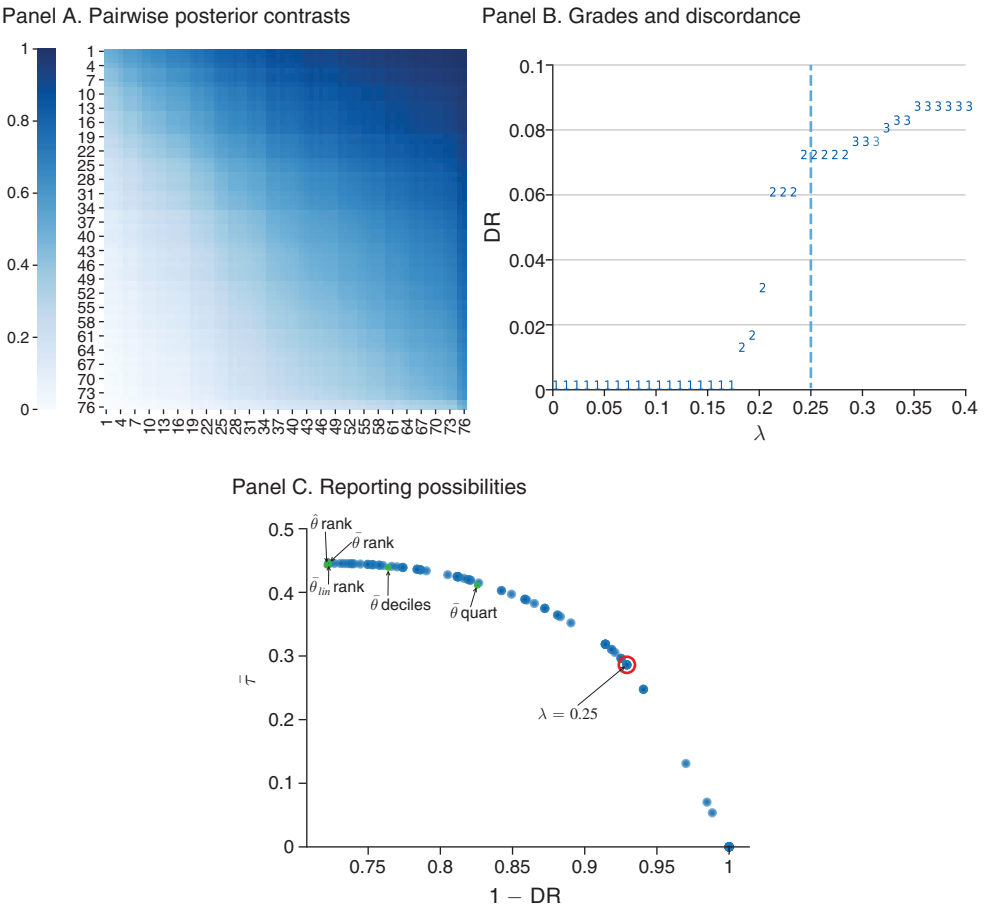

FIGURE 2. NAME RANKING EXERCISES

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. "$\bar{\theta}_{deciles}$" 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.

on normality of G.[11] 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 percent.

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

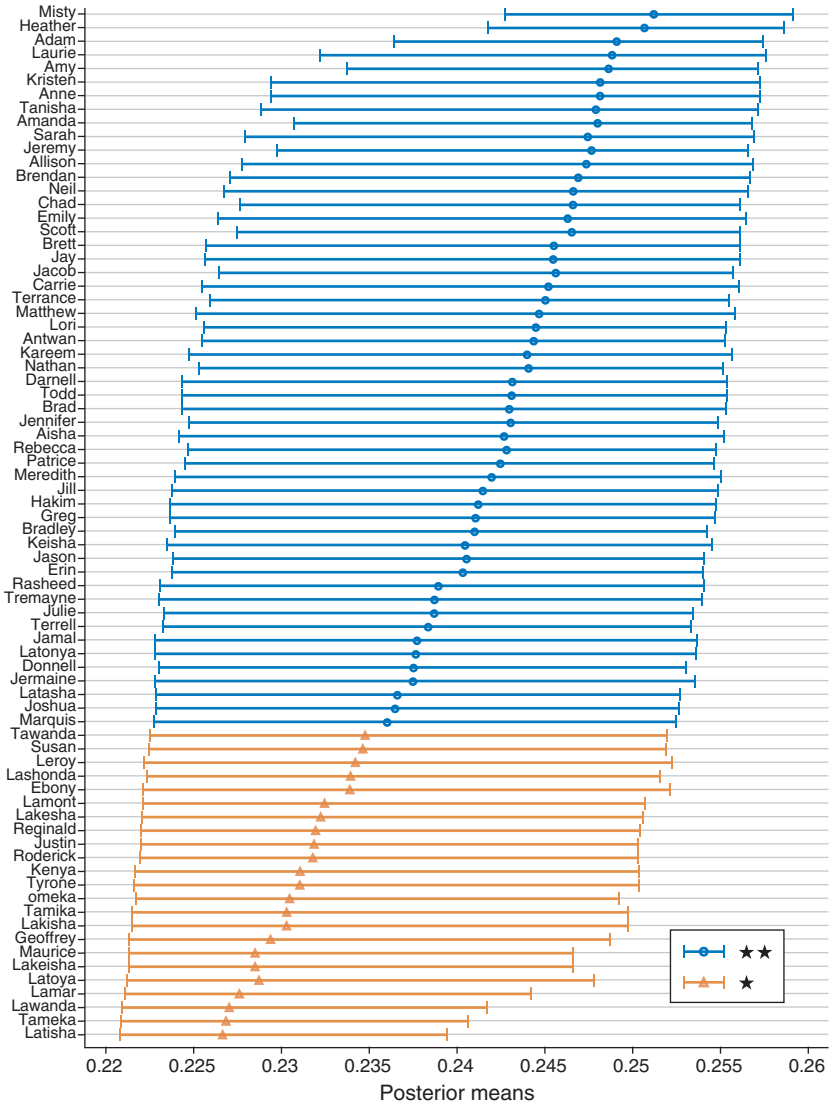

FIGURE 3. POSTERIOR MEANS AND GRADES OF FIRST NAMES

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

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 percent by setting $\lambda = 0.41$ would yield three grades and increase the estimated $\bar{\tau}$ by 11 percent (to $\bar{\tau} = 0.32$) at the expense of a 21 percent 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.

C. *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 percent) of these 76 intervals should be expected to contain their name's true latent contact rate.[12] 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 nonmonotonicity 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 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. Online 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 online 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 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 percent.

[12] 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 $\boldsymbol{\theta} \mapsto \sin(\boldsymbol{\theta})^2$.

The coarse grades that emerge from our procedure continue to align closely with our race labels: 35 of the 53 names (66 percent) in the top grade are distinctively White, while just 3 of the 23 names (13 percent) in the second grade are distinctively 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. Online 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.

V. 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.

A. *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.[13]

[13] 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.

TABLE 2—SUMMARY STATISTICS FOR FIRM SAMPLE

	Race		Gender	
	White (1)	Black (2)	Male (3)	Female (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)	
Number of firms	97			
Number of jobs	10,453			
Number of 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.

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.[14]

In what follows, we exclude the eleven firms in the experiment with callback rates below 3 percent 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 percent. Detailed point estimates and uncertainty measures for all 97 firms used in our analysis are provided in online 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

[14] Specifically, a second-order Taylor expansion of $\hat{p}_{iw}/\hat{p}_{ib} = \exp(\hat{\theta}_i)$ around the point (p_{iw}, p_{ib}) yields the approximation $\text{var}(\hat{p}_{iw}/\hat{p}_{ib}) \approx \theta_i^2\{\text{var}(\hat{p}_{iw})/p_{iw}^2 + \text{var}(\hat{p}_{ib})/p_{ib}^2 - 2[\text{cov}(\hat{p}_{iw}, \hat{p}_{ib})/(p_{iw}p_{ib})]\}$. Consequently, the delta method implies that $\text{var}[\hat{\theta}_i] \approx \text{var}(\hat{p}_{iw})/p_{iw}^2 + \text{var}(\hat{p}_{ib})/p_{ib}^2 - 2[\text{cov}(\hat{p}_{iw}, \hat{p}_{ib})/(p_{iw}p_{ib})]$.

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 depend mechanically on θ_i, it is possible for θ_i and s_i to be correlated. The top panel of online Appendix Figure F3 plots $\hat{\theta}_i$ against s_i, revealing that firms with more precise estimates tend to show less bias against Black names. The Spearman correlation between $\hat{\theta}_i$ and s_i is 0.36 ($p < 0.001$).

B. *A Model of Precision Dependence*

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

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

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 III, 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 $E[v_i | s_i] = \mu > 0$ and $\text{var}(v_i | s_i) = \sigma_v^2 > 0$, 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:

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

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

TABLE 3—GMM ESTIMATES OF CONTACT PENALTY PARAMETERS

	Race		Gender	
	No industry effects (1)	With industry effects (2)	No industry effects (3)	With industry effects (4)
Panel 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)
Panel B. Contact penalty distributions				
Mean of θ_i	0.092	0.093	−0.009	0.000
	(0.011)	(0.013)	(0.015)	(0.017)
SD 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, $E[v_i|s_i] = \mu$, $\text{var}(v_i|s_i) = \sigma_v^2$, and s_i is the standard error of $\hat{\theta}_i$. Column 2 allows an industry component of the form $v_i = \eta_{k(i)} \xi_i$, where $k(i)$ 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, $E[v_i|s_i] = 0$, and $\text{var}(v_i|s_i) = \sigma_v^2$. Column 4 allows an industry component of the form $v_i = \eta_{k(i)} + \xi_i$, where $E[\eta_k] = E[\xi_i] = 0$. 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 $E\left[\text{var}\left(v_i|\eta_{k(i)}\right)\right]/\text{var}(v_i)$, which equals $\left[\left(\sigma_\eta^2+1\right)\sigma_\xi^2\right]/\left(\sigma_\eta^2\sigma_\xi^2+\sigma_\eta^2\mu^2+\sigma_\xi^2\right)$ in column 2 and $\sigma_\xi^2/\left(\sigma_\eta^2+\sigma_\xi^2\right)$ in column 4.

($p = 0.97$). 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 online Appendix Figure F3 superimposes the estimated conditional expectation function $\hat{E}[\theta_i|s_i] = s_i^{\hat{\beta}}\hat{\mu}$ 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 online Appendix Figure F3 plots values of the estimated residual $\hat{T}_i = \left(\hat{\theta}_i - s_i^{\hat{\beta}}\hat{\mu}\right)/\sqrt{s_i^{2\hat{\beta}}\hat{\sigma}_v^2 + s_i^2}$ 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.

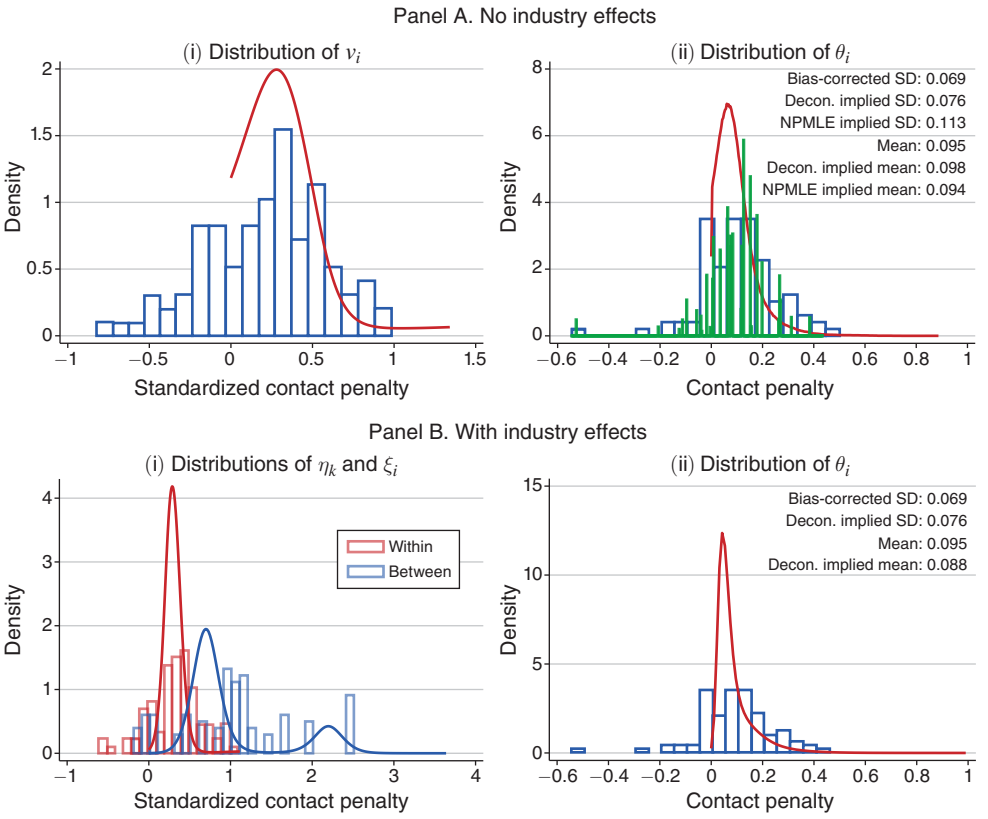

FIGURE 4. DECONVOLUTION ESTIMATES OF RACE CONTACT PENALTY DISTRIBUTIONS

Notes: This figure presents nonparametric 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 online Appendix. Part (ii) of panel A plots a histogram of $\hat{\theta}_i$ along with the corresponding log-spline and nonparametric 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 $v_i = \eta_{k(i)} \xi_i$, where $k(i)$ 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 $\hat{\xi}_i = \hat{v}_i / \bar{v}_{k(i)}$. 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.

C. *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:

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

Relying again on a variant of Efron's (2016) 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 $\left(\hat{\mu}, \hat{\sigma}_v^2\right)$ 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 $\hat{G} : x \mapsto n^{-1}\sum_i \hat{G}_v(x/s_i^{\beta})$ of the distribution of contact gaps.

Panel A(i) 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. Panel A(ii) 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 percent. 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 $\left(\theta_i, N_i s_i^2\right)$ 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 $n^{-1}\sum_i \left[\left(\hat{\theta}_i - \bar{\theta}\right)^2 - s_i^2\right]$. 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.

D. *Industry Effects*

In Kline, Rose, and Walters (2022) we found large differences in the magnitude of contact gaps across two-digit industries. Online 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.[15] 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 three firms, precluding a fixed effects approach to incorporating industry affiliation into the model. We therefore employ a hierarchical random effects specification of v_i taking the form

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

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

[15] 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 online Appendix and an earlier version of this paper (Kline, Rose, and Walters 2023).

where the function $k:\{1,\dots,n\}\to\{1,\dots,K\}$ returns a firm's industry, $\mathbf{s}_k$ is the vector of standard errors for all firms with $k(i)=k$, 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 $\eta_{k(i)}$ 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 $G_v:x\mapsto\int_0^\infty G_\xi(x/z)dG_\eta(z)$. We normalize $E[\eta_k]=1$, which implies $E[\xi_i]=\mu$.

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

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

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

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

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.[16]

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's (2016) log-spline estimator to the hierarchical case, these distributions are estimated by penalized maximum likelihood (see online 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 panel B(i) of Figure 4. Online Appendix Table F4 reports moments of the within- and between-industry distributions

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

implied by the log-spline estimates as well as moments of the overall contact ratio $\theta_i = s_i^{\beta}\eta_{k(i)}\xi_i$. The mean contact gap, between-industry standard deviation, and total standard deviation reported in online Appendix 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 $E[\eta_k] = 1$. Panel B(ii) of Figure 4 shows that the implied distribution of θ_i is similar to the estimate from the model without industry effects in panel A(ii), 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.

E. *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 no.9 cannot be distinguished from firm no.4 or firm no.49, suggesting each of these pairs in isolation would be labeled a tie. However, firm no.49 is clearly distinguishable from firm no.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 $\mathcal{R}(\mathbf{d};\lambda)$ 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 percent. Introducing industry effects yields four groups and increases the DR to 5.6 percent.

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

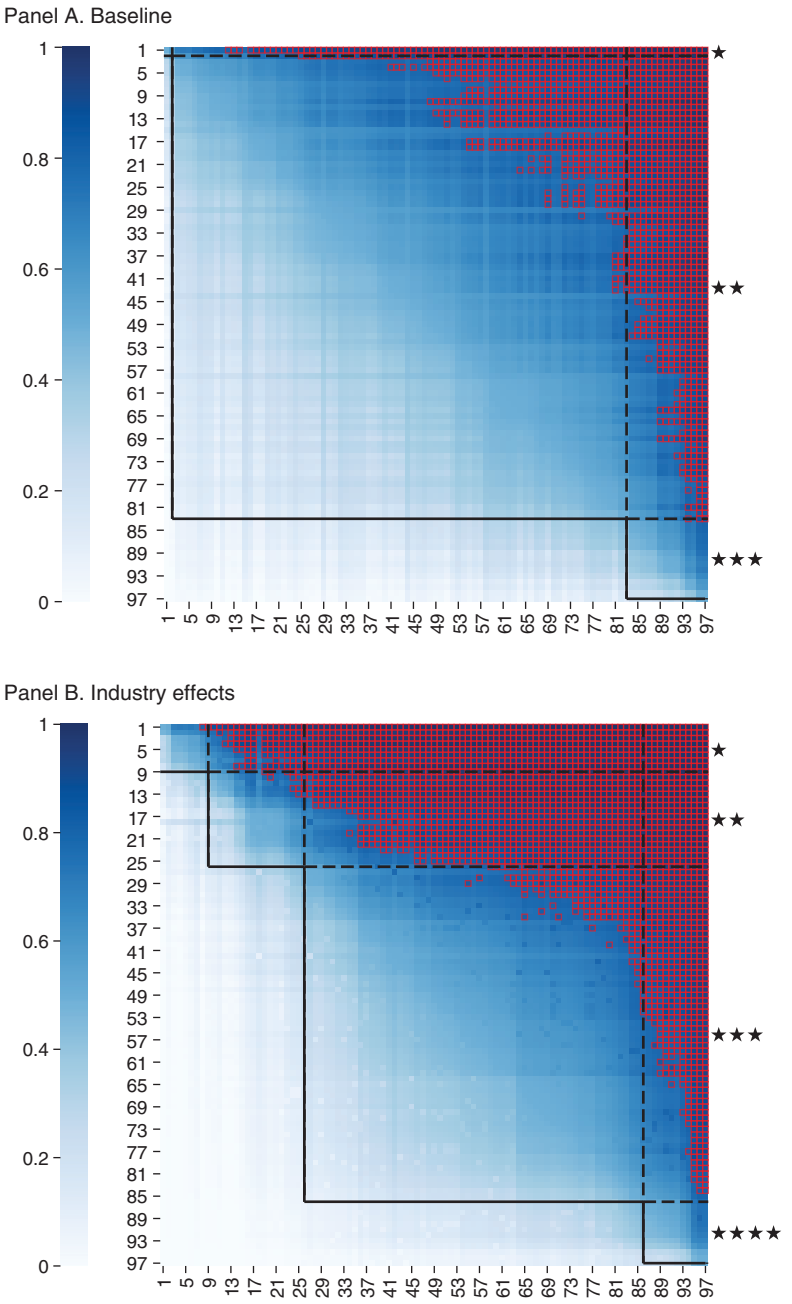

FIGURE 5. POSTERIOR CONTRASTS FOR RACE

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.

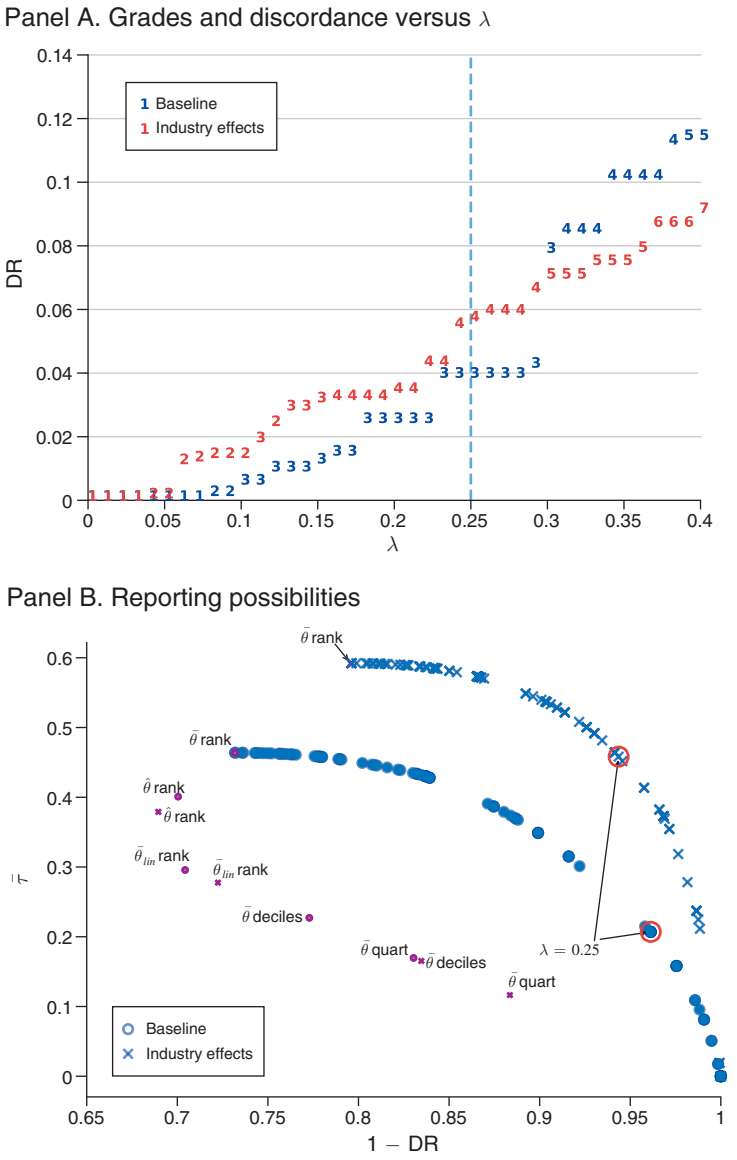

FIGURE 6. GRADES, DISCORDANCE, AND REPORTING POSSIBILITIES FOR RACE

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 $\boldsymbol{\theta}$ 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. "$\bar{\theta}_{deciles}$" 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.

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.[17] 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 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 percent to make pairwise ranking decisions. Resolving transitivity violations raises the required posterior certainty above 80 percent in most instances, yielding a discordance rate of only 3.9 percent in the baseline specification without industry effects and 5.6 percent 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.

VI. 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.[18]

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 online Appendix Figure F3). Bracketing the posterior mean estimates are EB 95 percent 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.

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

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

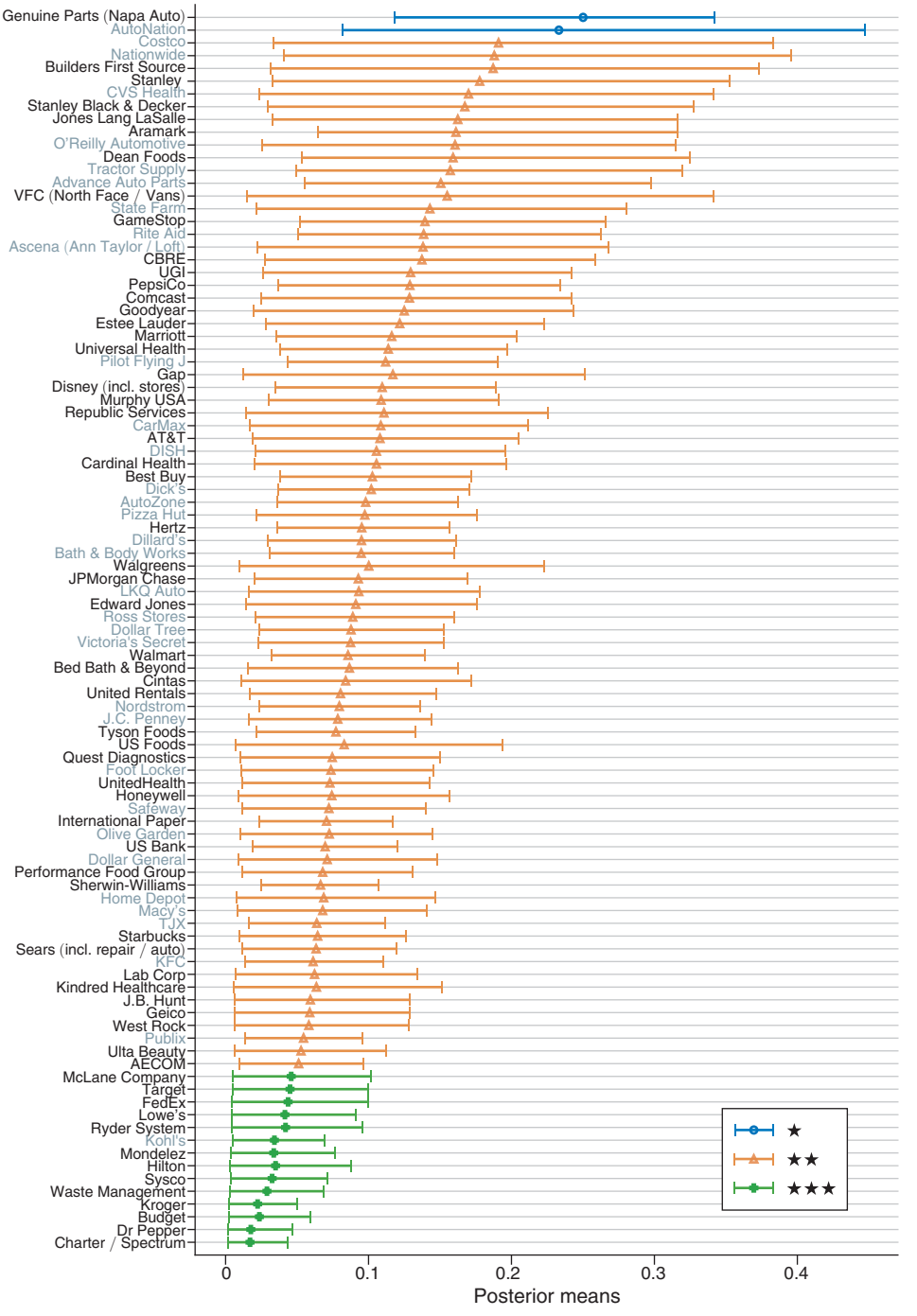

FIGURE 7. RACE REPORT CARD: POSTERIOR MEANS AND GRADES OF FIRMS (BASELINE)

Notes: This figure shows posterior mean proportional contact penalties for distinctively Black names, 95 percent credible intervals, and assigned grades. Grades are shown for $\lambda = 0.25$, implying an 80 percent 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.

While the Condorcet ranks, the ranks of the posterior means, and the ranks of the bias estimates are highly correlated, this correlation is not perfect. For example, Genuine Parts has the sixth largest proportional contact gap estimate (see online 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.

Online 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. Online 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. 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 percent 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 80 percent) of the rate for the group with the highest rate will generally be

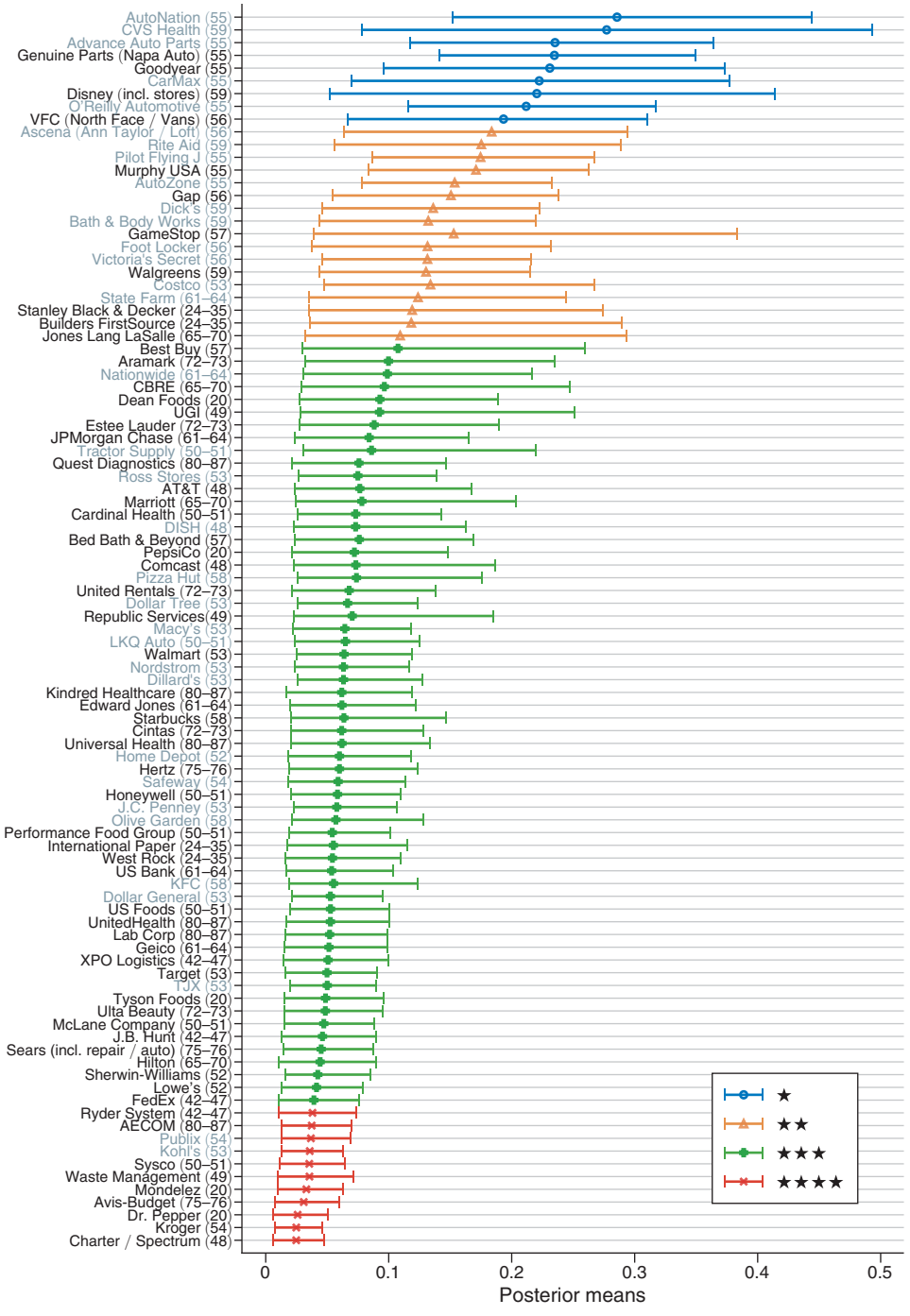

FIGURE 8. RACE REPORT CARD: POSTERIOR MEANS AND GRADES OF FIRMS (INDUSTRY EFFECTS)

Notes: This figure shows posterior mean proportional contact penalties for distinctively Black names, 95 percent credible intervals, and assigned grades from the industry random effect model. Grades are shown for $\lambda = 0.25$, implying an 80 percent 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.

regarded by the federal enforcement agencies as evidence of adverse impact" (p. 213). Our estimates suggest the contact rates for fictitious applicants in our experiment may have violated this standard.

A. *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 $\star\star\star\star$. Online 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. Online 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 online 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 percent. Adding industry boosts the R^2 to 70 percent. 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 $\eta_k n_k^{-1} \sum_{i:k(i)=k} \hat{s_i^{\beta}}$. 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 percent, while the industry with the smallest estimated bias is SIC 54, "food stores," which has a posterior mean of roughly 5 percent. 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 percent. The role of common industry-level practices in generating the

FIGURE 9. RACE REPORT CARD: POSTERIOR MEANS AND GRADES OF INDUSTRIES

Notes: This figure shows posterior means, 95 percent 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 percent threshold for posterior ranking probabilities. Each industry is labeled by its name and two-digit SIC code.

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

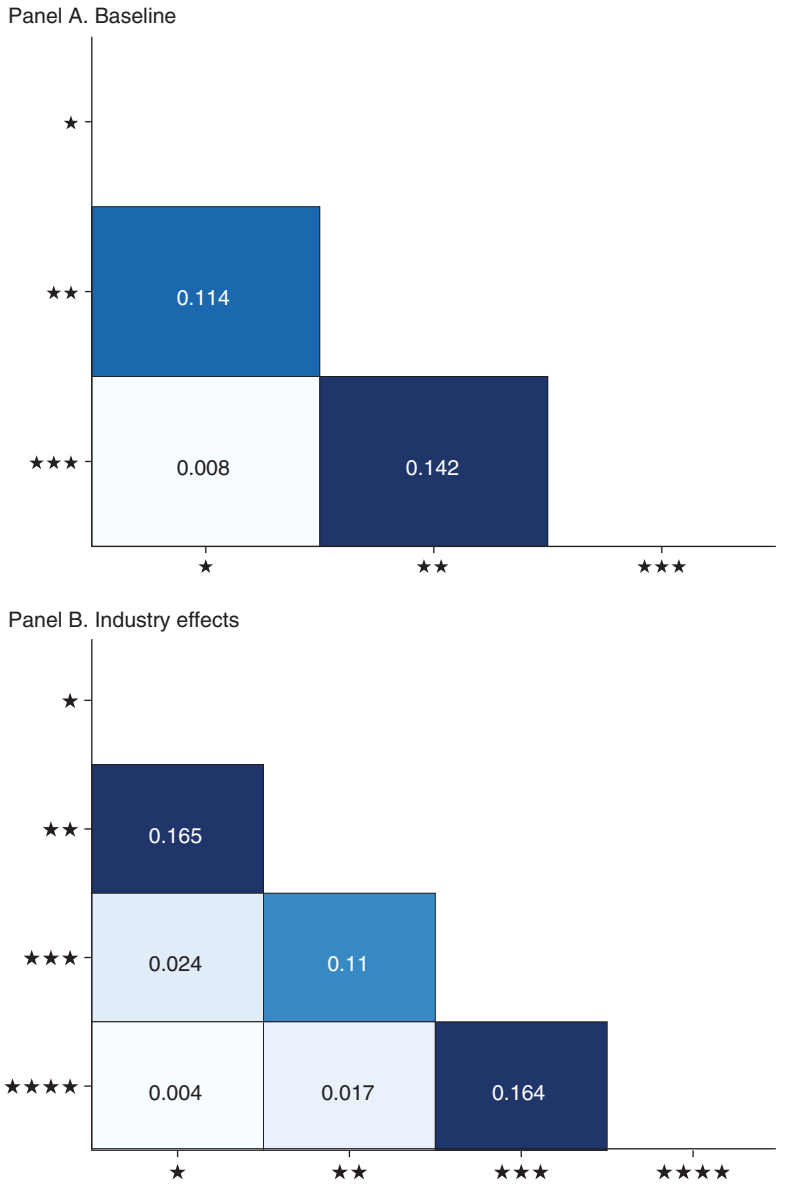

Figure 10. Race Report Card: DR in Baseline and Industry Effects Model

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.

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.

B. *Misclassification*

Figure 10 assesses the reliability of our 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 percent of the firm comparisons across grades $\star$ and $\star\star$ are expected to be misordered. The DR naturally declines when comparing nonadjacent grades. The expected share of misordered comparisons across grades $\star$ and $\star\star\star$ is below 1 percent. Adjacent grades have estimated DRs between 11 and 14 percent, while the discordance rate for nonadjacent grades is estimated to be only 0.8 percent.

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 percent to 17 percent. DRs for grades separated by two categories are estimated to fall below 3 percent, and the estimated DR between the worst grade $(\star)$ and the best grade $(\star\star\star\star)$ is 0.4 percent. 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.

VII. 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.

A. *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).

B. *A Model of Precision Dependence*

Inspection of the relationship between $\hat{\theta}_i$ and s_i (depicted in online 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:

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

where the distribution function $G_v : \mathbb{R} \to [0,1]$ has unrestricted support, the constant μ measures average gender bias in the population, $E[v_i \,|\, s_i] = 0$, and $\text{var}(v_i \,|\, s_i) = \sigma_v^2 > 0$. 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 $T_i = \left(\hat{\theta}_i - \mu\right)/\sqrt{s_i^{2\beta}\sigma_v^2 + s_i^2}$, 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 online Appendix Figure F4) is homoskedastic and centered around zero.

C. *Estimating G*

We use the GMM parameter estimates $\left(\hat{\mu}, \hat{\beta}\right)$ to form the residual $\hat{v}_i = \left(\hat{\theta}_i - \hat{\mu}\right)/s_i^{\hat{\beta}}$. Appealing again to Slutsky's Theorem, we assume $\hat{v}_i \,|\, v_i, s_i \sim \mathcal{N}\left(v_i, s_i^{2(1-\beta)}\right)$. 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 $\hat{G} : x \;\mapsto\; n^{-1}\sum_i \hat{G}_v\left((x - \hat{\mu})/s_i^{\hat{\beta}}\right)$.

The results are shown in panel A 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.

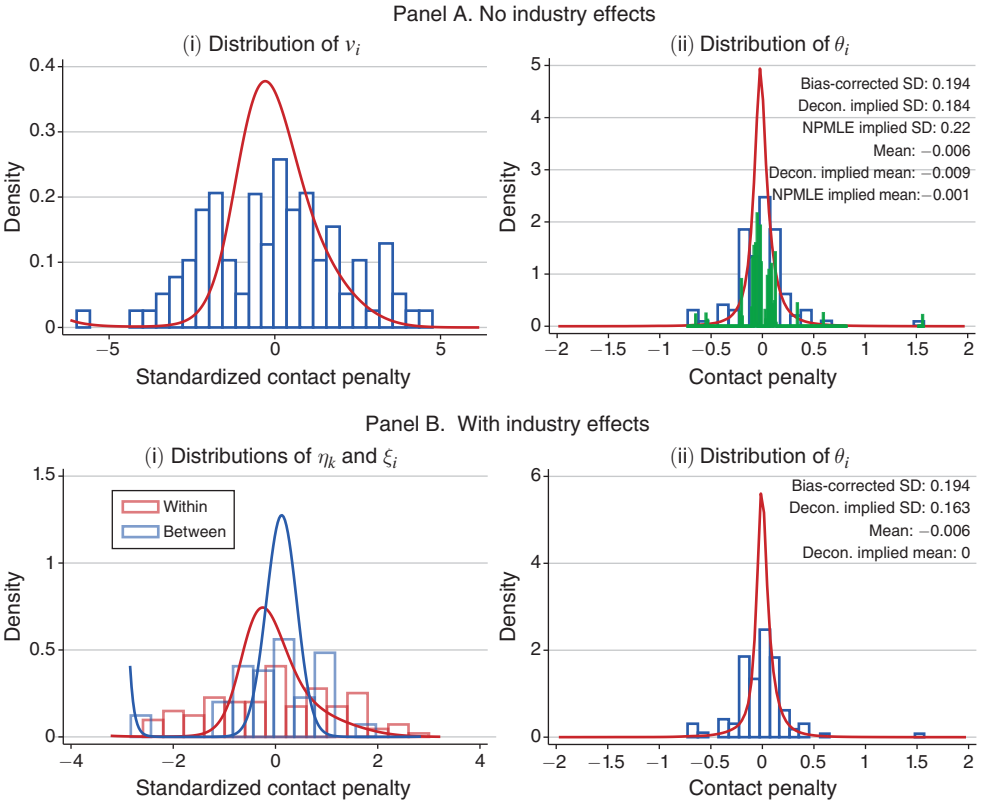

FIGURE 11. DECONVOLUTION ESTIMATES OF GENDER CONTACT PENALTY DISTRIBUTIONS

Notes: This figure presents nonparametric 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 online Appendix. Part (ii) of panel A plots a histogram of $\hat{\theta}_i$ along with the corresponding log-spline and nonparametric 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 $v_i = \eta_{k(i)} + \xi_i$, where $k(i)$ 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 $\hat{\xi}_i = \hat{v}_i - \bar{v}_{k(i)}$. 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.

D. *Industry Effects*

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

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

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

where $\eta_{k(i)}$ is a mean zero industry effect with variance σ^2_η, ξ_i is a mean zero firm effect with variance σ^2_ξ, 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 $\bar{v}_k = n_k^{-1} \sum_{i:k(i)=k} \hat{v}_i$, the model implies $\text{var}(\hat{v}_i) = \sigma^2_\eta + n_k^{-1}\sigma^2_\xi + n_k^{-2} \sum_{i:k(i)=k} s_i^{2(1-\beta)} \equiv V_k$.

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's (2016) log-spline procedure. The resulting densities are shown in the panel B 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.

E. *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 ⋆⋆⋆. 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 percent in our baseline specification and roughly 1 percent when including industry effects. Panel B 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 percent.

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

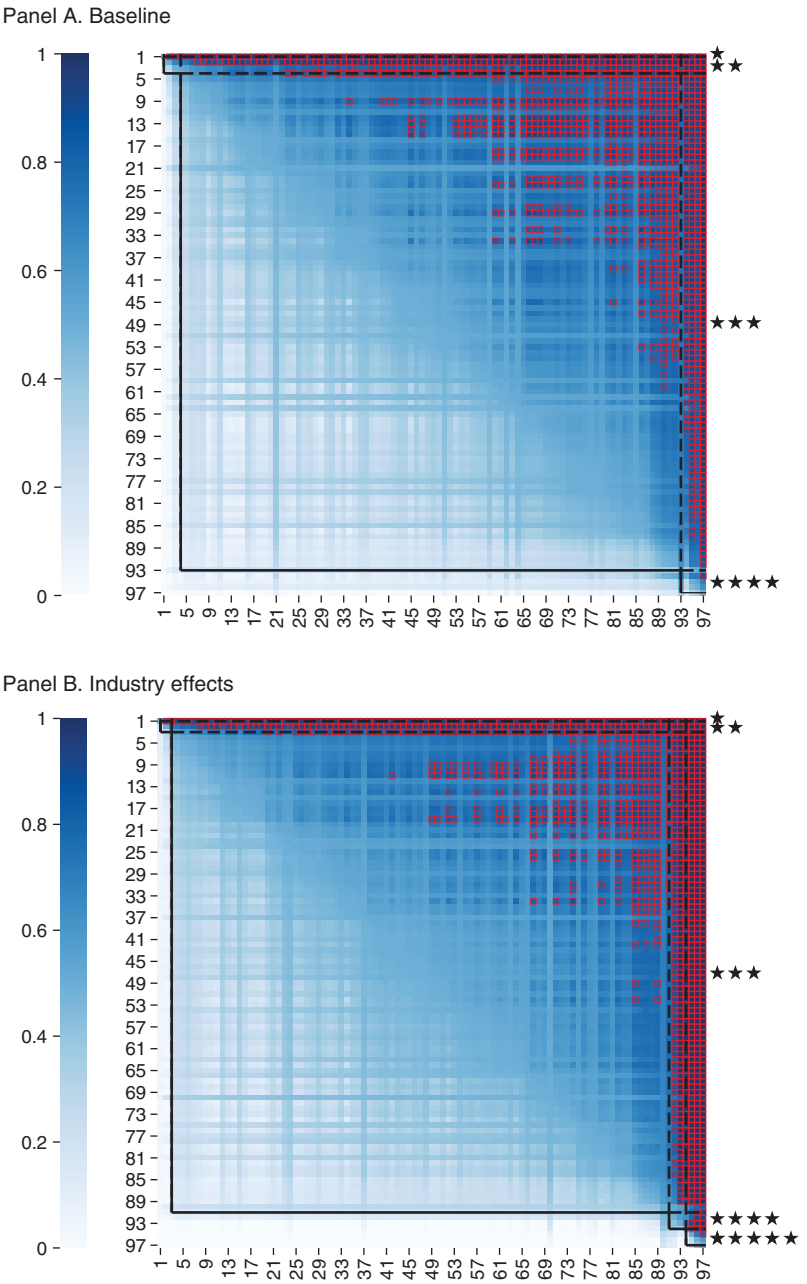

FIGURE 12. POSTERIOR CONTRASTS FOR GENDER

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.

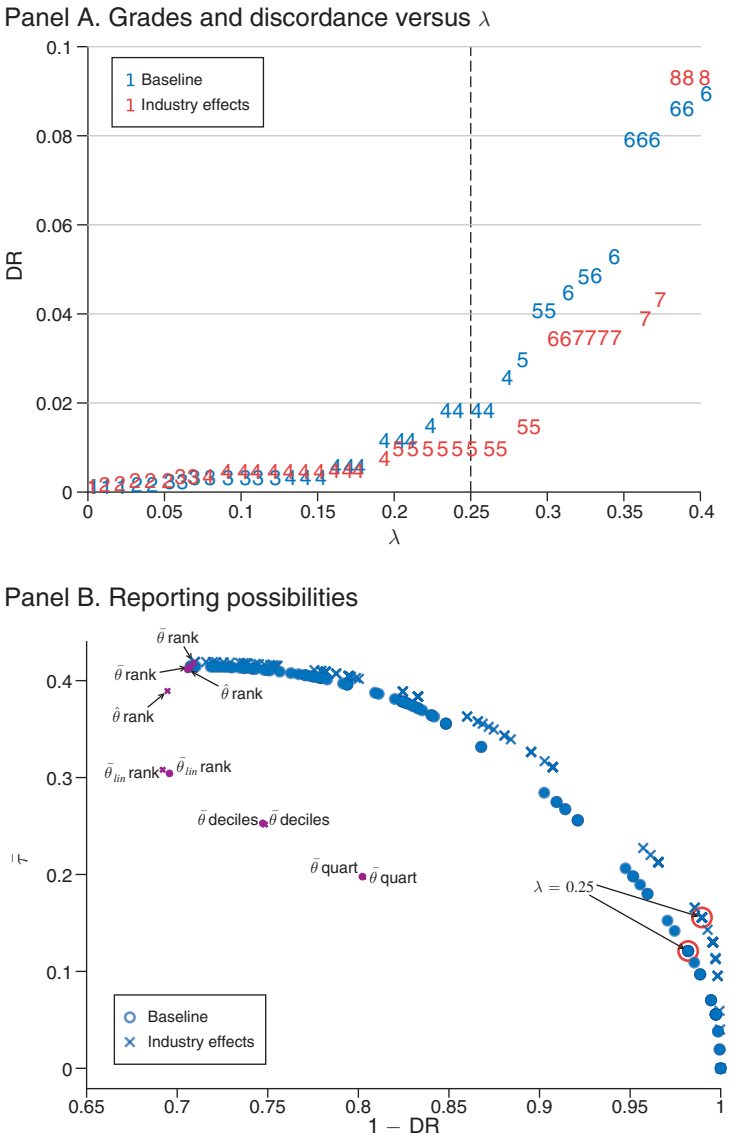

FIGURE 13. GRADES, DISCORDANCE, AND REPORTING POSSIBILITIES FOR GENDER

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 $\boldsymbol{\theta}$ 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.

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

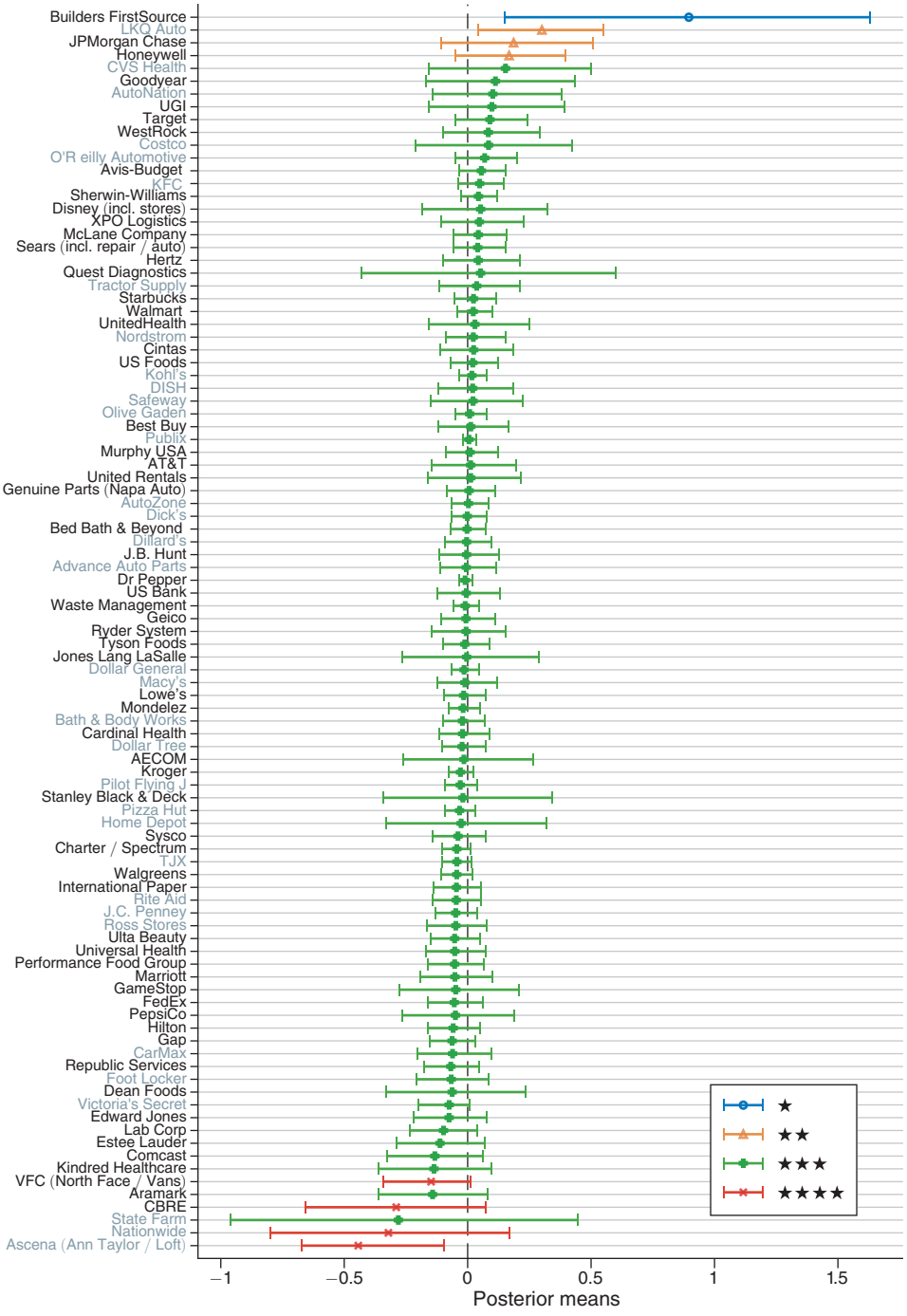

FIGURE 14. GENDER REPORT CARD: POSTERIOR MEANS AND GRADES OF FIRMS (BASELINE)

Notes: This figure shows posterior mean proportional gender contact differences between distinctively male and female names, 95 percent 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 percent 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.

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 ($\bar{\tau} = 0.16$ versus 0.12) while being slightly more reliable ($DR = 0.01$ versus $DR = 0.018$).

VIII. 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 online 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. Online 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 $\left(\hat{\theta}_i, s_i\right)$ 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 percent 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 four-fifths' rule standard. The firm VFC, while receiving the grade $\star\star\star\star$, 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.

A. *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 Condorcet ranks have posterior mean biases exceeding the four-fifths' 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

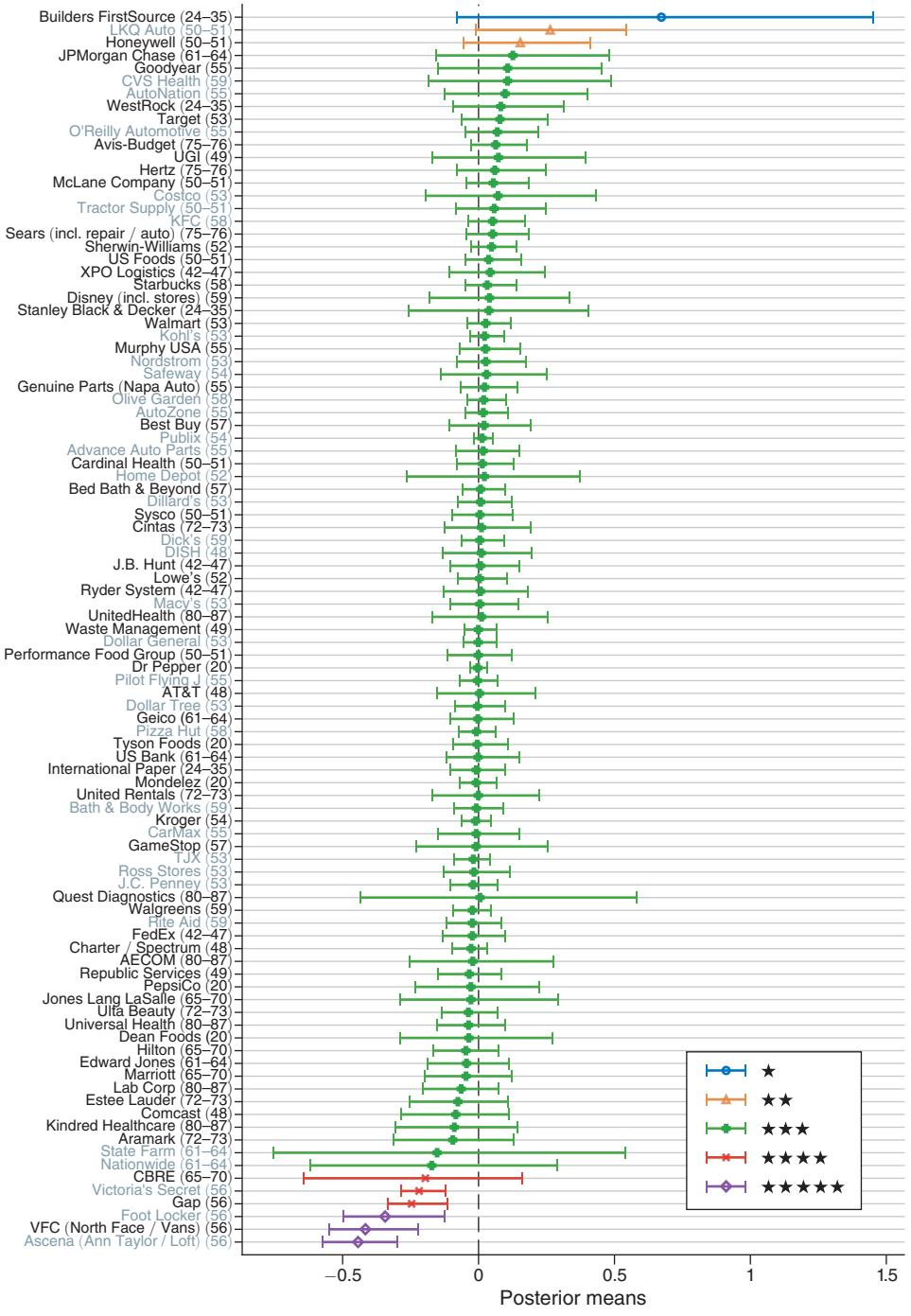

FIGURE 15. GENDER REPORT CARD: POSTERIOR MEANS AND GRADES OF FIRMS (INDUSTRY EFFECTS)

Notes: This figure shows posterior mean proportional gender contact differences between distinctively male and female names, 95 percent 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 percent 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.

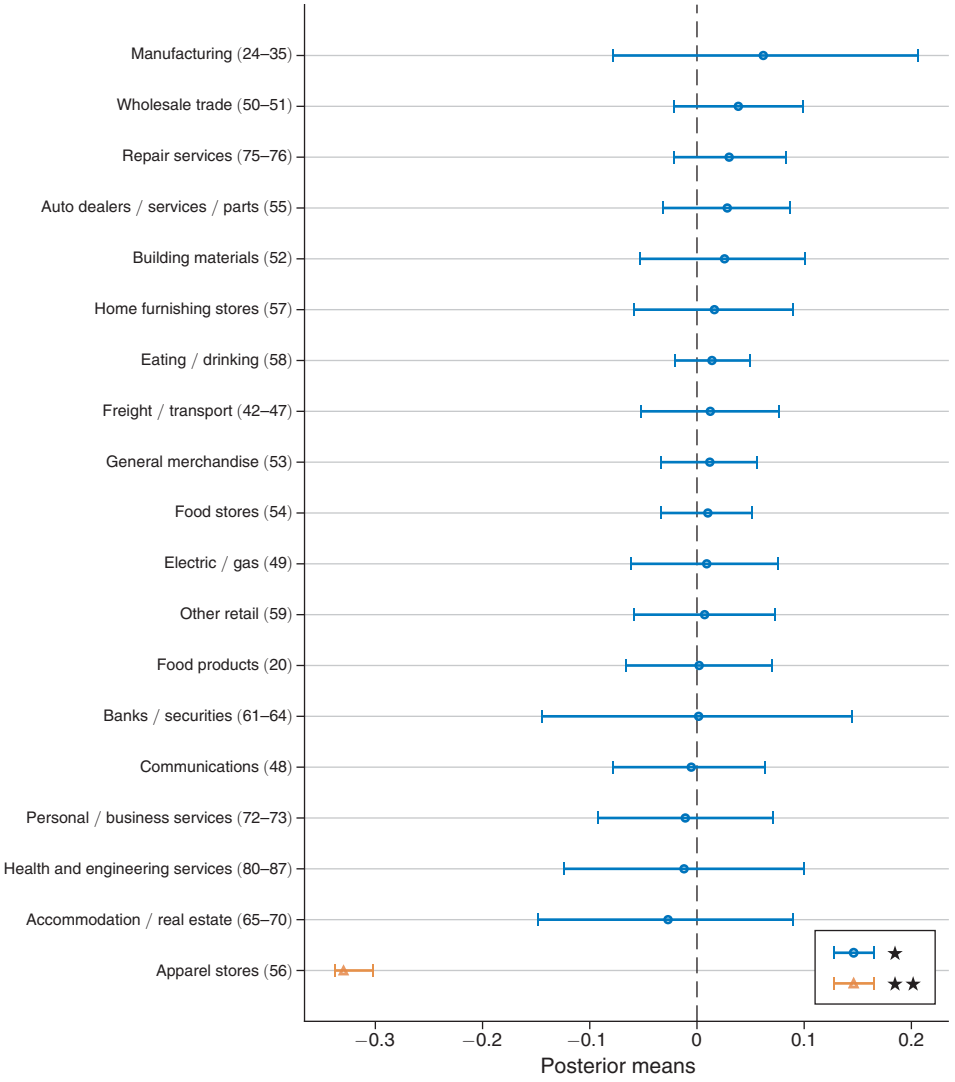

FIGURE 16. GENDER REPORT CARD: POSTERIOR MEANS AND GRADES OF INDUSTRIES

Notes: This figure shows posterior means, 95 percent 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 percent threshold for posterior ranking probabilities. Each industry is labeled by its name and two-digit SIC code.

of λ. The depicted grades with $\lambda = 0.25$ are estimated to explain 38 percent 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 percent. 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 percent. Online Appendix Figure F12 reports an alternate grading based upon the industry codes used in Kline, Rose, and Walters (2022). Those codes, which

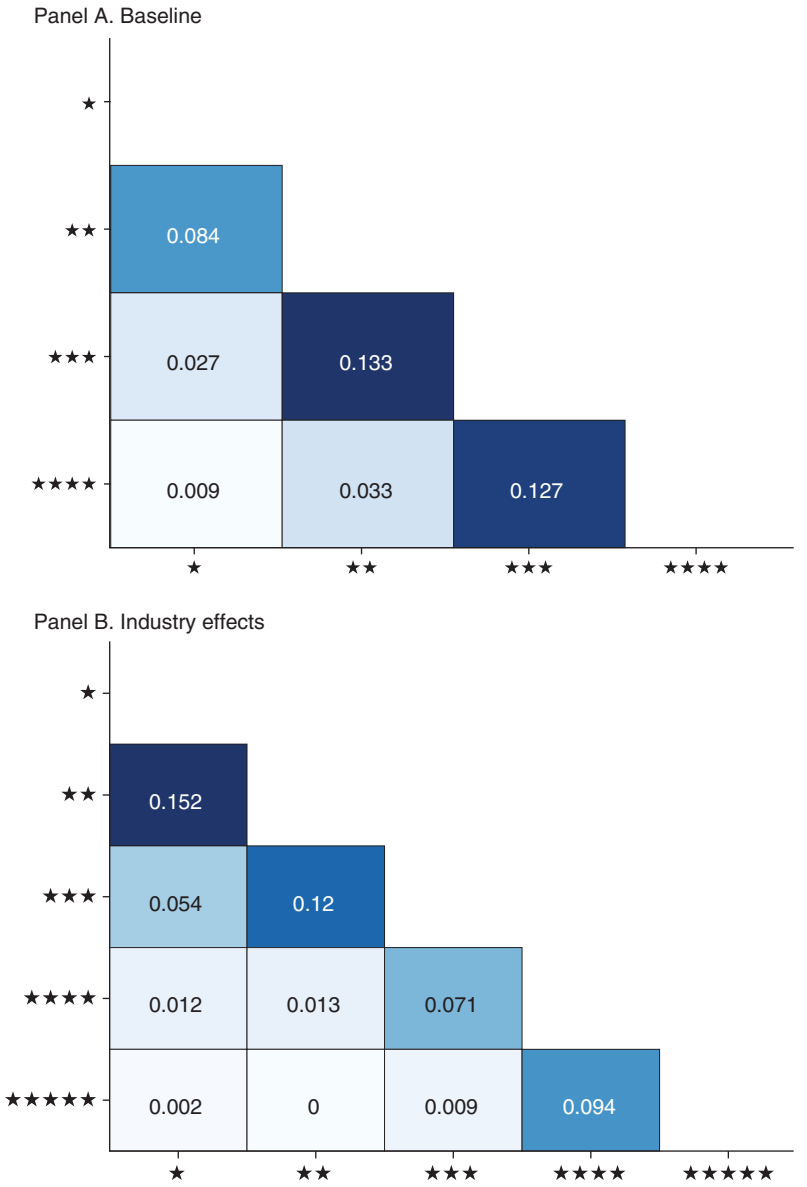

FIGURE 17. GENDER REPORT CARD: DR IN BASELINE AND INDUSTRY EFFECTS MODEL

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.

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.

B. *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 percent to 13 percent. Between nonadjacent grades the expected misclassification probability is estimated to be small: on the order of 1–3 percent.

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 percent. 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 percent. Hence, the chances of erroneously being classified as discriminating against women are higher than the chances of erroneously being classified as discriminating against men.

IX. 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 experiment. 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 percent 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 percent. 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 percent, 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 percent, 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's (2017) 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." NBER Working Paper 30981.

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey. 2019. "Inference on Winners." NBER Working Paper 25456.

Andrews, Isaiah, and Jesse M. Shapiro. 2021. "A Model of Scientific Communication." *Econometrica* 89 (5): 2117–42.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2021. "Race and the Mismeasure of School Quality." NBER Working Paper 29608.

Bai, Yuehao, Andres Santos, and Azeem M. Shaikh. 2021. "A Two-Step Method for Testing Many Moment Inequalities." *Journal of Business and Economic Statistics* 40 (3): 1070–80.

Bartlett, M.S. 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 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.

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.

Bergman, Peter, Eric W. Chan, and Adam Kapor. 2020. "Housing Search Frictions: Evidence from Detailed Search Data and a Field Experiment." NBER Working Paper 27209.

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: 104–13.

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, J.C. 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 (¾): 324–45.

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." Unpublished.

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie Jones, and Sonya Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper 25147.

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 23618.

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

Commission, Equal Employment Opportunity. 1978. "Uniform Guidelines on Employee Selection Procedures." *Federal Register* 43 (166): 38290–315.

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: 454–72.

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): 117–30.

Fryer Jr, Roland G., and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *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): 469–89.

Gu, Jiaying, and Roger Koenker. 2020. "Invidious Comparisons: Ranking and Selection as Compound Decisions." Unpublished.

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

InfoGroup. 2019. "Historical Datafiles." via UC Berkeley Libraries. https://search.library.berkeley.edu/permalink/01UCS_BER/1thfj9n/alma991049603029706532

Kemeny, John G. 1959. "Mathematics without Numbers." *Daedalus* 88 (4): 577–91.

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 U.S. Employers." *Quarterly Journal of Economics* 137 (4): 1963–2036.

Kline, Patrick, Evan K. Rose, and Christopher R. Walters. 2023. "A Discrimination Report Card." Unpublished.

Kline, Patrick, Evan K. Rose, Christopher R. Walters. 2024. "Replication Data for: A Discrimination Report Card." American Economic Association [Publisher], Inter-university Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/E198284V1.

Kline, Patrick M., and Christopher R. Walters. 2021. "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination." *Econometrica* 89 (2): 765–92.

Koenker, Roger, and Jiaying Gu. 2017. "REBayes: An R package for Empirical Bayes Mixture Methods." *Journal of Statistical Software* 82 (8): 1–26.

Koenker, Roger, and Ivan Mizera. 2014. "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules." *Journal of the American Statistical Association* 109 (506): 674–85.

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, and Dylan Ellis. 2013. *Using Administrative Data to Address Federal Contractor Violations of Equal Employment Opportunity Laws*. Oakland, CA: 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): 318–53.

McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics,* edited by Paul Zarembka.

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." Unpublished.

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 26883.

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): 1154–65.

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): 256–69.

Schaerer, Michael, Christilene du Plessis, My Hoang Bao Nguyen, Robbie C.M. van Aert, Leo Tiokhin, Daniël Lakens, Elena Giulia Clemente, 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* 41 (6): 1027–41.

Sobel, Marc J. 1990. "Complete Ranking Procedures with Appropriate Loss Functions." *Communications in Statistics-Theory and Methods* 19 (12): 4525–44.

Sobel, Marc J. 1993. "Bayes and Empirical Bayes Procedures for Comparing Parameters." *Journal of the American Statistical Association* 88 (422): 687–93.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 64 (3): 479–98.

US EEOC. 1996. *Enforcement Guidance: Whether "testers" can file charges and litigate claims of employment discrimination*. Washington, DC: U.S. Equal Employment Opportunity Commission.

Young, H. Peyton. 1986. "Optimal Ranking and Choice from Pairwise Comparisons." In *Information Pooling and Group Decision Making,* edited by Bernard Grofman and Guillermo Owen, 113–22. Bingley, UK: Emerald.

Young, H. Peyton, and Arthur Levenglick. 1978. "A Consistent Extension of Condorcet's Election Principle." *SIAM Journal on Applied Mathematics* 35 (2): 285–300.

This article has been cited by:

1. Christopher Walters. Empirical Bayes methods in labor economics 183-260. [Crossref]