SUPPLEMENTARY MATERIAL

A Proofs of Propositions

Corollary 3.2.1 (Underestimation of racial disparity)**.** *Let $y \in \mathcal{Y}$. If race is binary ($\mathcal{R} = \{0, 1\}$), and $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$, then*

$$\hat{\mu}^{(wtd)}_{Y|R}(y \mid 1) - \hat{\mu}^{(wtd)}_{Y|R}(y \mid 0) < \mathbb{P}(Y = y \mid R = 1) - \mathbb{P}(Y = y \mid R = 0).$$

Proof. For notational simplicity, let $\mu_r = \mathbb{P}(Y = y \mid R = r)$ and $\hat{\mu}_r = \hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r)$ for $r \in \{0, 1\}$. Since $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, necessarily $\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = 1\} \mid G, X, S)] > 0$ and $\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = 0\} \mid G, X, S)] < 0$. We note that the corollary could be stated under this more general condition, but was not for expositional clarity. Thus by Theorem 3.2, $\hat{\mu}_1 - \mu_r < 0$ and $\hat{\mu}_1 - \mu_r > 0$. Then

$$\begin{aligned}
\hat{\mu}_1 - \hat{\mu}_0 &= \hat{\mu}_1 - \mu_1 + \mu_1 - \mu_0 + \mu_0 - \hat{\mu}_0 \\
&= (\hat{\mu}_1 - \mu_1) - (\hat{\mu}_0 - \mu_0) + (\mu_1 - \mu_0) \\
&< \mu_1 - \mu_0,
\end{aligned}$$

as claimed. □

Corollary 3.2.2 (Unbiasedness of weighting under Assumption CI-YR)**.** *Let $y \in \mathcal{Y}$. If race is binary (so $\mathcal{R} = \{0, 1\}$), and Assumption CI-YR holds, then as $N \to \infty$, $\hat{\mu}^{(wtd)}_{Y|R}(y \mid r) - \mathbb{P}(Y = y \mid R = r) \xrightarrow{a.s.} 0$.*

Proof. If Assumption CI-YR holds, then $\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = 1\} \mid G = g, X = x, S = s) = 0$ for all g, x, and s. Consequently the right-hand side of the result of Theorem 3.2 is 0. □

Theorem 4.1 (Identification)**.** *For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then, under Assumption CI-YS, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, G = g, X = x)$ are identified if and only if both $\mathbf{P}$ and the augmented matrix $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$ have rank $|\mathcal{R}|$.*

Proof. Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y \mid R, G, X$, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned}
&\mathbb{P}(Y = y \mid G = g, X = x, S = s) \\
&\quad = \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x, S = s)\, \mathbb{P}(R = r \mid G = g, X = x, S = s) \\
&\quad = \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x)\, \mathbb{P}(R = r \mid G = g, X = x, S = s).
\end{aligned}$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, G = g, X = x)$. These parameters are identified if and only if this system has a unique solution, i.e. if the coefficient matrix $\mathbf{P}$ has rank $|\mathcal{R}|$ and so does the augmented matrix $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$. □

Theorem 4.2 (Unbiasedness of OLS Estimator)**.** *If Assumptions CI-SG, ACC, and CI-YS hold, and the identification conditions in Theorem 4.1 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,*

$$\mathbb{E}[\hat{\mu}^{(p\text{-}ols)}_{Y|R}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

Proof. Fix $y \in \mathcal{Y}$ and define $m_{gxr} = \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x)]$. Then under Assumptions CI-SG, ACC, and CI-YS,

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, S = s] \\
&\quad = \sum_{r \in \mathcal{R}} \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x)]\, \mathbb{P}(R = r \mid G = g, X = x, S = s) \\
&\quad = \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r,
\end{aligned}$$

where as in the main text $\hat{\mathbf{p}}$ is the (random) vector of BISG probabilities. In fact, since the right-hand side depends on S only through $\hat{\mathbf{p}}$, we have

$$\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, \hat{\mathbf{p}}] = \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r.$$

So the conditional expectation of $\mathbf{1}\{Y = y\}$ given X, G, and the BISG probabilities $\hat{\mathbf{p}}$ is linear in those probabilities, with coefficients m_{gxr}. Consequently, the OLS estimate $\hat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ will be unbiased for m_{gxr}, by the standard results.

Now, we can expand $\mathbb{P}(Y = y \mid R = r)$ as

$$\begin{aligned}\mathbb{P}(Y = y \mid R = r) &= \sum_{x \in \mathcal{X}, g \in \mathcal{G}} \mathbb{P}(Y = y \mid R = r, G = g, X = x)\, \mathbb{P}(G = g, X = x \mid R = r) \\ &= \sum_{x \in \mathcal{X}, g \in \mathcal{G}} m_{gxr} q_{gx|r}.\end{aligned}$$

Since $\hat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ is unbiased for m_{gxr}, by the linearity of expectation the poststratified estimator $\hat{\mu}^{(\text{p-ols})}_{Y|R}(y \mid r)$ is unbiased for $\mathbb{P}(Y = y \mid R = r)$. □

Theorem 4.3 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators)**.** *For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\hat{\mu}^{(wtd)}_{Y|R}(y \mid \cdot) = \hat{\mu}^{(ols)}_{Y|R}(y \mid \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j \mid G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k \mid G_i, X_i, S_i) = 0$ and vice versa) or $\hat{\mu}^{(wtd)}_{Y|R}(y \mid j) = \hat{\mu}^{(wtd)}_{Y|R}(y \mid k)$.*

Proof. Fix a $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$. The weighting estimator of $\mathbb{P}(Y = y \mid R = r)$ within the set of individuals with $G_i = g$ and $X_i = x$ may be written

$$\hat{\mu}^{(\text{wtd})}_{Y|RGX}(y \mid r, g, x) = \frac{\hat{\mathbf{P}}^{\top}_{\mathcal{J}(xg)r} \mathbf{1}\{\mathbf{Y}_{\mathcal{J}(xg)} = y\}}{\hat{\mathbf{P}}^{\top}_{\mathcal{J}(xg)r} \mathbf{1}} = \frac{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{J}(xg)} = y\}) \right\|}{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}}(\mathbf{1}) \right\|},$$

the ratio of the projected length of the outcome vector $\mathbf{1}\{\mathbf{Y} = y\}$ and the constant vector $\mathbf{1}$ onto $\hat{\mathbf{P}}_{\cdot r}$. We can write the OLS estimator as

$$\hat{\mu}^{(\text{ols})}_{Y|R} = (\hat{\mathbf{P}}^{\top}_{\mathcal{J}(xg)} \hat{\mathbf{P}}_{\mathcal{J}(xg)})^{-1} \hat{\mathbf{P}}^{\top}_{\mathcal{J}(xg)} \mathbf{1}\{\mathbf{Y}_{\mathcal{J}(xg)} = y\} = \text{coord}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)}}(\text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{J}(xg)} = y\}),$$

where $\text{coord}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)}}$ is the function that returns the coordinates of its input vector in the $\hat{\mathbf{P}}_{\mathcal{J}(xg)}$ basis (by assumption $\hat{\mathbf{P}}_{\mathcal{J}(xg)}$ has rank $|\mathcal{R}|$ and so its columns are linearly independent). To make the comparison even easier, notice that we can break the projection $\text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}}$ into two steps, writing it instead as $\text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}} = \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}} \circ \text{proj}_{\hat{\mathbf{P}}}$. Letting $\mathbf{Y}_{\text{proj}} = \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{J}(xg)} = y\})$, then, we can rewrite our estimators as

$$\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) = \frac{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}}(\mathbf{Y}_{\text{proj}}) \right\|}{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)r}}(\mathbf{1}) \right\|} \quad \text{and} \quad \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid r) = \text{coord}_{\hat{\mathbf{P}}_{\mathcal{J}(xg)}}(\mathbf{Y}_{\text{proj}})_r.$$

Now, since the individual BISG probabilities are nonnegative and sum to 1, a pair $j, k \in \mathcal{R}$ of races has perfectly discriminating BISG probabilities if and only if the corresponding columns of $\hat{\mathbf{P}}$ are orthogonal, i.e., $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = I$. Begin by writing $\mathbf{Y}_{\text{proj}}$ in terms of the $\hat{\mathbf{P}}_{\mathcal{J}(xg)}$ basis, so

$$\mathbf{Y}_{\text{proj}} = \sum_{j \in \mathcal{R}} c_j \hat{\mathbf{P}}_{\mathcal{J}(xg)j},$$

and thus $\hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j) = c_j$. Without loss of generality, suppose the c_j are numbered as $c_1 \geq c_2 \geq \cdots \geq c_{|\mathcal{R}|}$. We can also expand $\mathbf{1}$ in the same basis. Since the individual probabilities must sum to one, in fact we have $\mathbf{1} = \sum_{j \in \mathcal{R}} \hat{\mathbf{P}}_{\mathcal{J}(xg)j}$.

For the forward direction, we assume $\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid j) = \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j) = c_j$; multiplying out the denominator of the weighting estimator, we have $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \mathbf{Y} = c_j \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \mathbf{1}$ for all j; substituting the basis expansions of $\mathbf{Y}_{\text{proj}}$ and $\mathbf{1}$, this yields

$$\sum_{k \in \mathcal{R}} c_k \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = \sum_{k \in \mathcal{R}} c_j \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k}, \quad \text{so} \quad \sum_{k \in \mathcal{R}} (c_j - c_k) \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = 0.$$

Now fix $j \in J_1 = \arg\max_j c_j$; this relation still holds, but now every term in the sum is nonnegative and in particular $c_j > c_k$ for all $k \notin J_1$. Therefore we must have $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = 0$ for all $k \notin J_1$. Then fix $j \in J_2 = \arg\max_{j \notin J_1} c_j$; since $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)l} = 0$ for all $l \in J_1$, every term in the sum is still nonnegative and in particular $c_j > c_k$ for all $k \notin J_1 \cup J_2$. Therefore we must have $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = 0$ for all $k \notin J_1 \cup J_2$. Proceeding this way through all sets of common values in the c_j we find that for all $j, k \in \mathcal{R}$, either $c_j = c_k$ or $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = 0$.

For the reverse direction, fix $j \in \mathcal{R}$ and let $J = \{k \in \mathcal{R} : c_k = c_j\}$, so that by assumption $\hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k} = 0$ for all $k \notin J$. Then by the above basis expansion, $\hat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid j) = c_j$, and

$$\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid j) = \frac{\sum_{k \in \mathcal{R}} c_k \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k}}{\sum_{k \in \mathcal{R}} \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k}} = \frac{c_j \sum_{k \in J} \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k}}{\sum_{k \in J} \hat{\mathbf{P}}_{\mathcal{J}(xg)j}^\top \hat{\mathbf{P}}_{\mathcal{J}(xg)k}} = c_j = \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j). \quad \square$$

Theorem 4.4 (Nonparametric Identification Under Assumption CI-YSF)**.** *Let $f : \mathcal{S} \to \mathbb{R}^d$, $d < |\mathcal{S}|$, with range $f(\mathcal{S})$. For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then under Assumption CI-YSF,*

and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$*, the conditional probabilities* $\mathbb{P}(Y = y \mid R, f(S) = z, G = g, X = x)$ *are identified if and only if both* $\mathbf{P}$ *and the augmented matrix* $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$ *have rank* $|\mathcal{R}|$.

Proof. The argument is identical to the proof of Theorem 4.1.

Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y \mid f(S), R, G, X$, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned}\mathbb{P}(Y = y \mid G = g, X = x, S = s) &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, f(S) = f(s), G = g, X = x, S = s) \\ &\quad \times \mathbb{P}(R = r \mid G = g, X = x, S = s) \\ &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, f(S) = f(s), G = g, X = x) \\ &\quad \times \mathbb{P}(R = r \mid G = g, X = x, S = s).\end{aligned}$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $z \in f(\mathcal{S})$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, f(S) = s, G = g, X = x)$. These parameters are identified if and only if this system has a unique solution, i.e. if the coefficient matrix $\mathbf{P}$ has rank $|\mathcal{R}|$ and so does the augmented matrix $\left(\mathbf{P} \quad \mathbf{b}\right)$. □

B Estimation

This appendix provides more discussion of estimation under the identifying assumptions.

B.1 Possible BIRDiE Models

Complete-pooling model. The simplest possible model is one in which the relationship between Y and R does not vary with G or X. This model is parametrized by $\Theta = \{\theta_r\}_{r \in \mathcal{R}}$, which describe the distribution of Y within every level of R:

$$\begin{aligned} Y_i \mid R_i, G_i, X_i, \Theta &\sim \text{Categorical}_y(\theta_{R_i}) \\ \theta_r &\overset{iid}{\sim} \text{Dirichlet}(\alpha), \end{aligned}$$

where Categorical$_{\mathcal{Y}}$ denotes a discrete (categorical) distribution on the set $\mathcal{Y}$. With known $\mathbf{R}$, the posterior of θ_r is conjugate, a fact which will make computation under the EM scheme described in Section B.2 extremely efficient. Of course, this efficiency comes at the cost of a restrictive model that allows for no role of G and X. If in reality $\mathbb{P}(Y \mid R)$ does vary along these dimensions, it is possible that the posterior of θ_r will not accurately estimate $\mathbb{P}(Y \mid R)$. In any case, if the analyst is interested in subgroup or small-area estimates of $\mathbb{P}(Y \mid R)$, the complete-pooling model will be of little use.

Saturated (no-pooling) model. At the other end of the spectrum from the complete-pooling model is a *saturated* or no-pooling model, which estimates a different distribution of $Y \mid R$ within every level of G and X:

$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Categorical}_y(\theta_{R_i G_i X_i})$$
$$\theta_{rgx} \overset{iid}{\sim} \text{Dirichlet}(\alpha).$$

This model is closest to the OLS estimator, though it ensures that all probability estimates lie in $[0, 1]$. As with the complete-pooling model, the posterior of θ_{rgx} is conjugate to its prior, and so computation can be made efficient. Additionally, this model allows for any arbitrary relationship between Y, R, G, and X. Since it is fully nonparametric, the posterior will converge to the true $\mathbb{P}(Y \mid R, G, X)$ with enough data in each (G, X) cell. However, in practice, the model can suffer from the curse of dimensionality: the number of (G, X) cells may be relatively large compared to the amount of available data, or even exceed it, especially since G can be quite large, covering many blocks or ZIP codes. In these cases, the prior will dominate the data in each cell, which could have a large biasing effect even on overall inferences about $\mathbb{P}(Y \mid R)$.

General mixed-effects model. As a compromise between the complete-pooling and no-pooling model, a partial pooling approach based on a multinomial mixed-effects model can be used. Properly specified, the mixed-effects model maintains the flexibility of the saturated model

while avoiding its high bias and variance in finite samples.

$$
\begin{aligned}
Y_i \mid R_i, G_i, X_i, \Theta &\sim \text{Categorical}_y(g^{-1}(\mu_{rgx})) \\
\mu_{rgxy} &= \mathbf{W}\beta_{ry} + \mathbf{Z}\mathbf{u}_{ry} \\
\mathbf{u}_{ry} \mid \phi_{ry} &\sim \mathcal{N}\Big(0, \Sigma(\phi_{ry})\Big) \\
\beta_{ry} \overset{iid}{\sim} f_r^{(\beta)}, &\quad \phi_{ry} \overset{iid}{\sim} f_r^{(\phi)},
\end{aligned}
$$

where g^{-1} is a softmax or other link function, $\mathbf{W}$ and $\mathbf{Z}$ are matrices that encode the fixed and random effects, respectively,[3] ϕ is a vector of random-effect parameters, and $f^{(\beta)}$ and $f^{(\phi)}$ are some priors for the super-scripted parameters. Some fixed or random effects could be shared across combinations of R and Y, though this could complicate computation. In practice, we recommend including X and especially G in the model as random effects, with hierarchical structure as appropriate. Such a structure partially pools estimates of $\mathbb{P}(Y \mid R, X, G)$ towards an overall estimate of $\mathbb{P}(Y \mid R)$, allowing the model to share information between geographic areas. This should prove especially useful in cases where some areas have few or no observations for certain racial groups.

We also recommend including group-level covariates as fixed effects, which will help share information across random effects and significantly improve generalization performance to unseen random effect levels (Buttice & Highton 2013). For example, if G records counties, analysts could include racial and socioeconomic variables measured at the county level as predictors. This would help produce more accurate estimates of $\mathbb{P}(Y \mid R, G)$ to the extent that variation in these probabilities is associated with these racial and socioeconomic variables. Ultimately the structure of this general model will have to be chosen based on the data and the relevant research question.

[3]The matrix $\mathbf{W}$ is the standard covariate matrix for a linear regression. The matrix $\mathbf{Z}$ maps the random effects to observations. In the case of a random intercept model, for instance, $\mathbf{Z}$ would have a column for every group and a row for every observation, and Z_{ij} would be an indicator for row i belonging to group j, and thus receiving random effect u_j.

B.2 Computation

The posterior in Equation (3) contains the high-dimensional discrete nuisance parameter $\mathbf{R}$, which poses a challenge for computation. We suggest two approaches for handling $\mathbf{R}$, one suited to small sample sizes, and one suited to large sample sizes. We also discuss uncertainty quantification for conjugate complete-data models.

Small samples: Inference directly on the marginal likelihood. Since $\mathbf{R}$ is discrete, we can marginalize it out as follows:

$$\pi(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) = \sum_{\mathbf{r} \in \mathcal{R}^N} \pi(\Theta, \mathbf{r} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \propto \pi(\Theta) \prod_{i=1}^{N} \sum_{r \in \mathcal{R}} \pi(Y_i \mid r, G_i, X_i, \Theta) \hat{P}_{ir}. \quad (4)$$

This decouples the total number of parameters from the sample size. Equation (4) has only continuous parameters, and so can be used with any general Bayesian inference procedure such as Markov chain Monte Carlo (MCMC). In practice, however, the moderate-to-high dimensionality (even after integrating out $\mathbf{R}$) and the complex likelihood due to the sum nested within the outer product, make MCMC algorithms computationally too expensive beyond small datasets.

Large samples: Expectation-Maximization. When the number of individuals exceeds a thousand or so, we use an Expectation-Maximization (EM) algorithm (Dempster et al. 1977) to compute the maximum *a posteriori* (MAP) estimate of Θ. The EM algorithm alternates between an E-step which calculates the expected log posterior density Q, averaging over the missing $\mathbf{R}$, and an M-step which maximizes Q over values of Θ. Specifically, given a current parameter estimate $\Theta^{(t)}$, the expected log posterior density can be written as

$$\begin{aligned} Q(\Theta^{(t+1)} \mid \Theta^{(t)}) &= \mathbb{E}_t[\log \pi(\Theta^{(t+1)}, \mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})] \\ &= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \Big\{ \Big(\log \pi(Y_i \mid r, G_i, , X_i, \Theta^{(t+1)}) + \log \hat{P}_{ir} \Big) \\ &\qquad \times \pi(R_i = r \mid \Theta^{(t)}, Y_i, G_i, X_i, S_i) \Big\} \\ &= C + \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \tilde{P}^{(t)}_{ir|Y} \log \pi(Y_i \mid r, G_i, X_i, \Theta^{(t+1)}), \end{aligned}$$

where C is a constant that does not depend on $\Theta^{(t+1)}$, and $\tilde{\mathbf{P}}^{(t)}_{|Y} = \pi(\mathbf{R} \mid \Theta^{(t)}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ are the BISG probabilities updated with Bayes' rule using the outcome $\mathbf{Y}$:

$$\tilde{P}^{(t)}_{ir|Y} = \frac{\pi(Y_i \mid r, G_i, , X_i, \Theta^{(t)})\hat{P}_{ir}}{\sum_{r' \in \mathcal{R}} \pi(Y_i \mid r', G_i, , X_i, \Theta^{(t)})\hat{P}_{ir'}}. \tag{5}$$

At the M-step, $Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is straightforward to maximize, since it is just the log complete-data posterior, with likelihood weights given by the $\tilde{P}^{(t)}_{ir|Y}$. Additionally, if Θ can be partitioned into parameters which only affect individuals in each racial group (as is the case with all the models in Section 4.2), the maximization can be performed separately on each group of individuals.

A critical advantage of this EM scheme over working directly with the marginal likelihood is that the maximization in the M-step can be performed using sufficient statistics calculated as part of the E-step, rather than on all of the individual entries in the data. Since the M-step is usually the bottleneck in the computation, this is enormously helpful—the problem size scales with $|\mathcal{Y}| \times |\mathcal{X}| \times |\mathcal{G}|$ rather than with N. Specifically, notice that we can rewrite $Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ (dropping the unnecessary constant) as

$$\begin{aligned} Q(\Theta^{(t+1)} \mid \Theta^{(t)}) &= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \tilde{P}^{(t)}_{ir|Y} \log \pi(Y_i \mid r, G_i, X_i, \Theta^{(t+1)}) \\ &= \log \pi(\Theta^{(t+1)}) + \sum_{r \in \mathcal{R}} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \log \pi(y \mid r, g, x, \Theta^{(t+1)}) \left(\sum_{i \in \mathcal{I}(yxg)} \tilde{P}^{(t)}_{ir|Y} \right), \end{aligned}$$

where $\mathcal{I}(yxg)$ is the set of individuals with $Y_i = y$, $X_i = x$, and $G_i = g$.

For BIRDiE models where the complete-data likelihood is conjugate to the prior, such as the complete- and no-pooling models, these sufficient statistics are used in the M-step anyway, and can be efficiently calculated during the E-step. In combination with the acceleration scheme described next, this allows the entire EM algorithm to be run to convergence on data with hundreds of thousands or millions of individuals in a matter of seconds.

While EM algorithms are stable, due to their monotonic increasing of the marginal likelihood, they are also often slow to converge (Laird 1993). To address this, we propose, and include in our open-source software implementation, the use of fixed-point iteration accelerators such as Ander-

son acceleration or SQUAREM (Varadhan & Roland 2008). These techniques can substantially reduce the overall computational time without meaningfully affecting the stability of inference.

Uncertainty quantification via blocked Gibbs sampling. While computationally efficient, the EM algorithm does not provide any uncertainty quantification. However, in large samples, sampling and model-based uncertainty are dominated by biases caused by even small violations of the underlying assumptions, a problem we discuss in Section 4.5 and the accompanying appendix. Even so, it is often useful to have some measure of sampling uncertainty. This is possible using blocked Gibbs sampling for some BIRDiE models with conjugate complete-data posteriors, such as the complete- and no-pooling models described above.

Our Gibbs sampling strategy is to alternate sampling from the model-updated BISG probabilities $\pi(\mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}, \Theta)$ and sampling from the complete-data posterior $\pi(\Theta \mid \mathbf{Y}, \mathbf{R}, \mathbf{G}, \mathbf{X}, \mathbf{S})$. The first step involves the same calculations as the E-step in Equation (5). The second step is computationally tractable in medium-to-large samples when the complete-data likelihood is conjugate to the prior, as is the case for the Categorical-Dirichlet pooling models proposed above.

Categorical-Dirichlet models with latent discrete variables, like the pooling BIRDiE models, are often tackled using a collapsed Gibbs sampler where Θ has been marginalized out. We found that approach to be unsuccessful here, however, since the one-by-one updating of each individual's R_i meant that the sampler was unable to traverse to the correct region of parameter space and got stuck near the BISG initialization. In contrast, the blocked Gibbs sampler is vectorized over individuals and rapidly moves to the mode identified by the EM algorithm.

For non-conjugate BIRDiE models with few parameters, bootstrapping the EM procedure is computationally feasible and can approximate the covariance matrix of the MAP estimate.

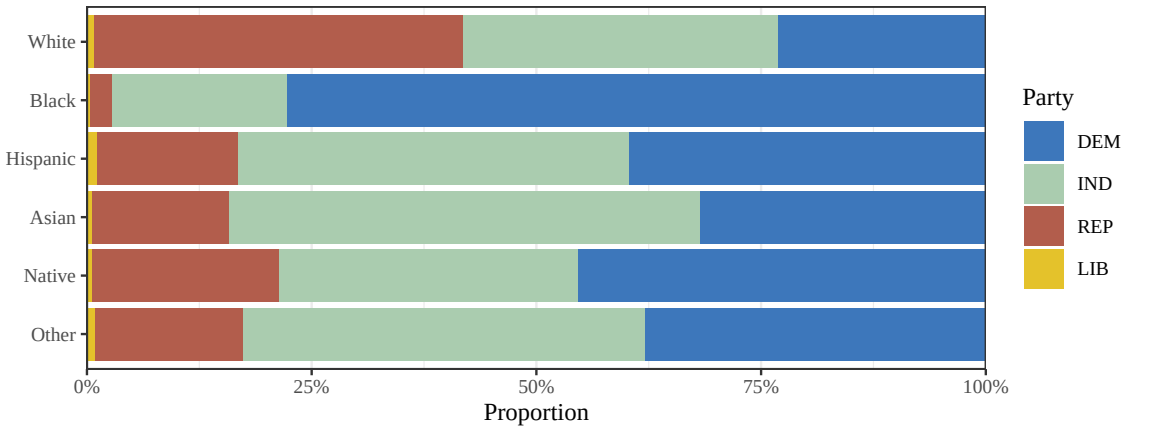

Figure 5: Distribution of party registration by race for a sample of 1,000,000 North Carolina voters. Parties are Libertarian (LIB), Republican (REP), Independent (IND), and Democratic (DEM).

C North Carolina Validation

C.1 Data details

The overall merged voter file contains 5,754,912 voters, 71.1% of which are White, 21.1% of which are Black, and 7.8% of which belong to another race. To reduce computational burden, we further subsampled this file by selecting 1,000,000 records at random without replacement. This sample size is large enough to ensure that sampling error in the estimates is negligibly small.

Figure 5 shows the distribution of party registration by self-reported race in this subsample. White voters disproportionately register as Republicans, while Black voters disproportionately register as Democrats. This serves as the ground-truth in our validation analysis.

C.2 Modeling details

For the BIRDiE models and OLS estimator, we use geographic effects matching the geographic level used in the BISG probabilities (e.g., county effects for the county-level BISG probabilities), except for the block-level probabilities. Due to the large number of individual census blocks, we use tract-level effects instead for this particular model. For the mixed-effects BIRDiE model, in addition to these geographic effects we add two geography-level covariates: the White and Black

fraction of the population in each individual's geography of residence. These covariates should help further regularize and share information among the individual geographic effects.

We use noninformative or weakly informative priors for both BIRDiE models. For the saturated model, we set all the Dirichlet hyperparameters to $\frac{1}{2}$, which is the Jeffreys prior. For the mixed model, the prior on the fixed effects β_r for each racial group is a weakly informative Normal with standard deviation $2p_r$, where p_r is the share of the racial group in the sample; this encourages more shrinkage for groups where less data is available. The overall global intercept received a $\mathcal{N}(0, 5^2)$ prior. The prior on the random intercept scale is Inv-Gamma$(4, 1.5)$, designed to support a range of possible heterogeneity of the outcome-race relationship across geographic levels while discouraging the M step from finding a mode at zero; it places 95% of its mass between 0.17 and 1.376. The random intercept correlation matrix (across levels of Y) received an LKJ prior with shape parameter 2. Variations of these prior choices did not noticeably affect the top-line estimates, however, due to the large quantity of data overall. To give an idea of the computational efficiency of the proposed method, the maximum runtime of the saturated of the BIRDiE models fit using the EM algorithm was 6.9 seconds (on a modern laptop with 8GB RAM), and the maximum runtime for the mixed model estimates was 216 seconds. Gibbs sampling for the saturated model took a maximum of 88 seconds and did not vary much with the number of parameters.

C.3 Total Variation distance results

For a more comprehensive look at the error in the estimated partisanship-by-race distributions, we turn to the total variation (TV) distance, which is calculated as

$$d_{\text{TV}}(\hat{\mu}_{Y|R}, \mu_{Y|R}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \sum_{r \in \mathcal{R}} |\hat{\mu}_{Y,R}(y, r) - \mu_{Y,R}(y, r)|,$$

where $\mu_{Y,R}$ is the joint distribution of Y and R. The TV distance is an upper bound on the error in *any* probability calculated from the estimated joint distribution, and as such is a useful general-purpose measure of estimation error. The left plot of Figure 6 shows the TV distance for each

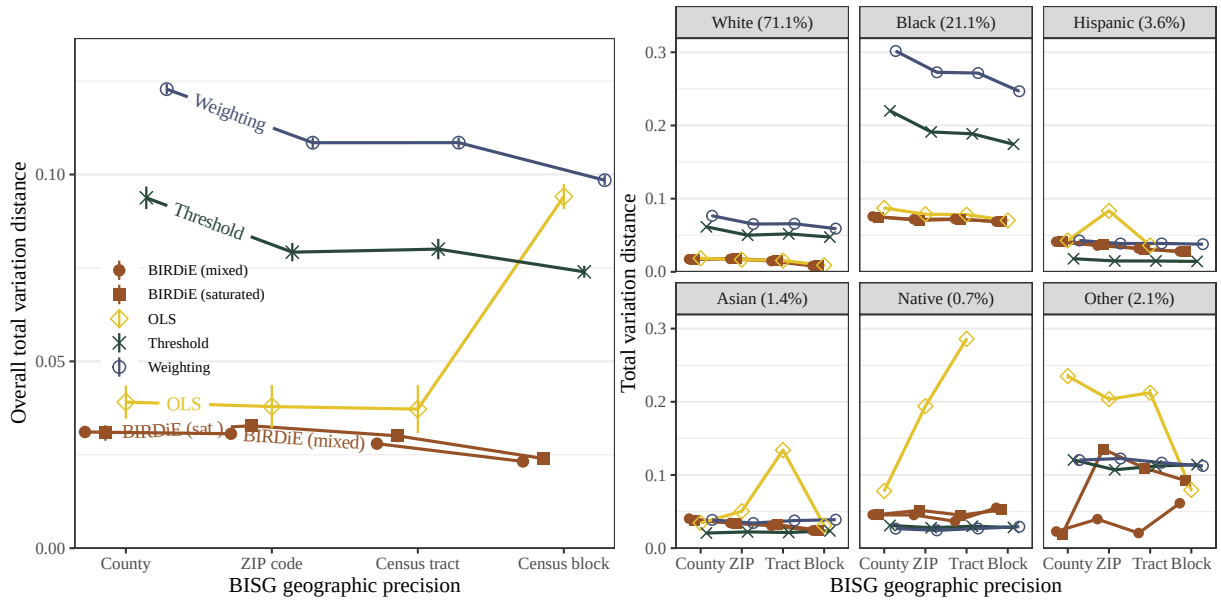

Figure 6: Total variation distance between the estimated and actual distribution of party registration, by estimation method and level of geographic detail used in the BISG predictions. The left plot shows the overall total variation distance, while the right plot decomposes it by racial group. 95% confidence intervals, where available, are indicated by the vertical lines.

estimator, not just for the county-level BISG estimates used in Figure 2 but also across the range of geographic levels used in the BISG calculation (x-axis). We also measure the TV distance for each conditional distribution by race:

$$d_{\text{TV}}^{(r)}(\hat{\mu}_{Y|r}, \mu_{Y|r}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} |\hat{\mu}_{Y,r}(y, r) - \mu_{Y,r}(y, r)|.$$

The right plot of Figure 6 shows these within-race TV distances, to illuminate how the estimators perform on each subgroup.

We find that both BIRDiE models substantially outperform every alternative method overall at every geographic level. In general, the estimates based on the BIRDiE models exhibit an overall total variation distance whose magnitude is about one third and one fourth of that for the thresholding and weighting estimators, respectively, and slightly but consistently lower than the OLS estimator. The weighting and thresholding estimators perform particularly poorly for Black voters. The error of OLS estimator spikes dramatically for smaller geographies and smaller racial groups, leading it to underperform thresholding in aggregate at the block level. This is due to

the small sample size per poststratification cell and the inability of the OLS estimator to enforce known bounds on the parameters. All of the methods except the OLS estimator perform more similarly for Hispanic, Asian, and Native registration estimates. As before, the saturated and mixed-effects BIRDiE models perform similarly across the board, though the saturated model lags the mixed-effects model for estimates of Other-race registration for smaller geographies. When fitting the saturated model with the EM algorithm, this difference was not present, so we suspect it is due to differences between the posterior mode and mean for the posterior distribution within each geography-race-party cell.

According to the TV distance measure, we find that finer geographic data provide only minor improvements in accuracy for the BIRDiE or conventional estimates. While possibly counter-intuitive, this finding underscores the fact that calibrated BISG probabilities, rather than highly precise probabilities, are all that is needed for accurate disparity estimation.

Of course, both calibrated and precise probabilities are to be preferred to imprecise but calibrated probabilities. In practice, however, there may be a tradeoff between the two. For example, including first and/or middle names in the BISG predictions may increase their precision. But, first and middle names names may lead to worse calibration, since BISG methods which use these names make a somewhat unrealistic conditional independence assumption, and data on first and middle names by race come from non-census sources (Tzioumis 2018, Rosenman et al. 2023). Additionally, unlike surnames, first and middle names (which are usually chosen by parents) can be more correlated with socioeconomic variables, leading to violations of Assumptions CI-YS.

C.4 OLS Behavior and Variance Inflation Metric

C.4.1 Effect of sample size on estimates

To demonstrate the role that the sample size plays on the reliability of the OLS estimator, we applied the estimator for a logarithmically-spaced range of sample sizes from 1,000 to 1,000,000 (the full sample). Since the underlying sample was randomly selected from the voter file, for

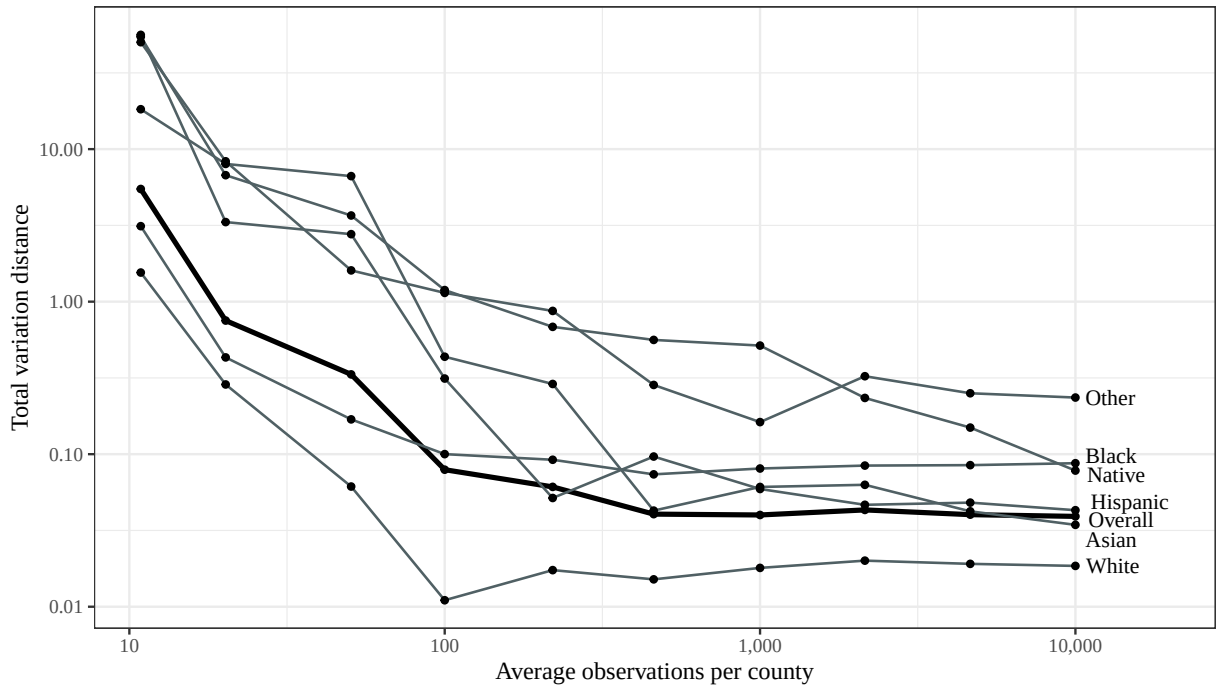

Figure 7: Accuracy of the OLS estimates by race and overall, as measured by the total variation distance, as a function of the average sample size in each county.

each sample size n we simply took the first n records from the sample. For each sample size we then calculated the average number of records in each county. Figure 7 shows the error in the county-level OLS estimates, as measured by the total variation distance, for each racial group and overall, plotted against the average number of records in each county. As seen in the main text, the OLS estimator performs relatively well for all racial groups and overall when the full 1,000,000 records are used. However, at smaller sample sizes, when the number of records per county is less than 100—meaning the actual number of individuals in smaller racial groups may be in the single digits or fewer—the OLS estimator performs much worse.

C.4.2 Variance inflation

Practitioners might be concerned about the possibility of a "weak surname instrument," where $\mathbb{P}(R \mid S, G, X) \approx \mathbb{P}(R \mid G, X)$. Of course, if $S \perp\!\!\!\perp R \mid G, X$, then Assumption CI-YS does not hold and the BIRDiE model is not identified. But if surnames only weakly predict race for a certain (G, X) or overall, then estimates of $\mathbb{P}(Y \mid R, G, X)$ may become highly variable.

This is the same problem that users of instrumental variables (IV) estimators face when the instruments are weak predictors of the endogenous variables. For IV estimators, due to sampling error in the first stage, weak instruments not only increase variance but can lead to heavy-tailed sampling distributions that invalidate standard confidence intervals. Here, the "first-stage" estimates are the BISG probabilities, which are fixed and have no sampling error. Thus, the only consequence of a weak instrument is high variability.

To help practitioners detect cases of weak instruments, we propose a variance inflation metric. Specifically, we suggest comparing the variance of the estimate of $\mathbb{P}(Y \mid R, G, X)$ for each (R, G, X) against the variance that would be obtained if the BISG estimates were perfect—i.e., if individual race were known exactly. While individual race is of course not observed, it is possible to consistently estimate the increase in the *variance* compared to the known-race estimate based only on the BISG probabilities.

Let $\hat{P}$ be the matrix of BISG probabilities for observations taking particular values (G, X), and P_{01} be the corresponding binary matrix encoding the true values of individual race. Then, the known-race estimate can be obtained by least-squares regression of Y on P_{01} and has variance $\sigma^2(P_{01}^\top P_{01})^{-1}$, where σ^2 is the residual variance. Since P_{01} is binary and the racial categories are mutually exclusive, $P_{01}^\top P_{01}$ is a diagonal matrix with entries counting the number of observations in each racial group. Because the BISG probabilities are accurate by Assumption ACC, these entries can be consistently estimated as the column sums of $\hat{P}$, i.e., $\mathbf{1}^\top \hat{P}$. The variance of the OLS estimate based on the BISG probabilities is $\sigma^2(\hat{P}^\top \hat{P})^{-1}$. Thus, the variance inflation estimator for racial group j is given by

$$\hat{\text{VI}}_j = \frac{(\hat{P}^\top \hat{P})^{-1}_{jj}}{(\mathbf{1}^\top \hat{P})_j},$$

which consistently estimates the true relative efficiency of the two estimates.

Figure 8 shows the distribution of the variance inflation metric for each racial group across county submodels (i.e., the OLS estimator applied to each county cell) in the North Carolina data. The top set of histograms is weighted by the population of each racial group, so that, for example,

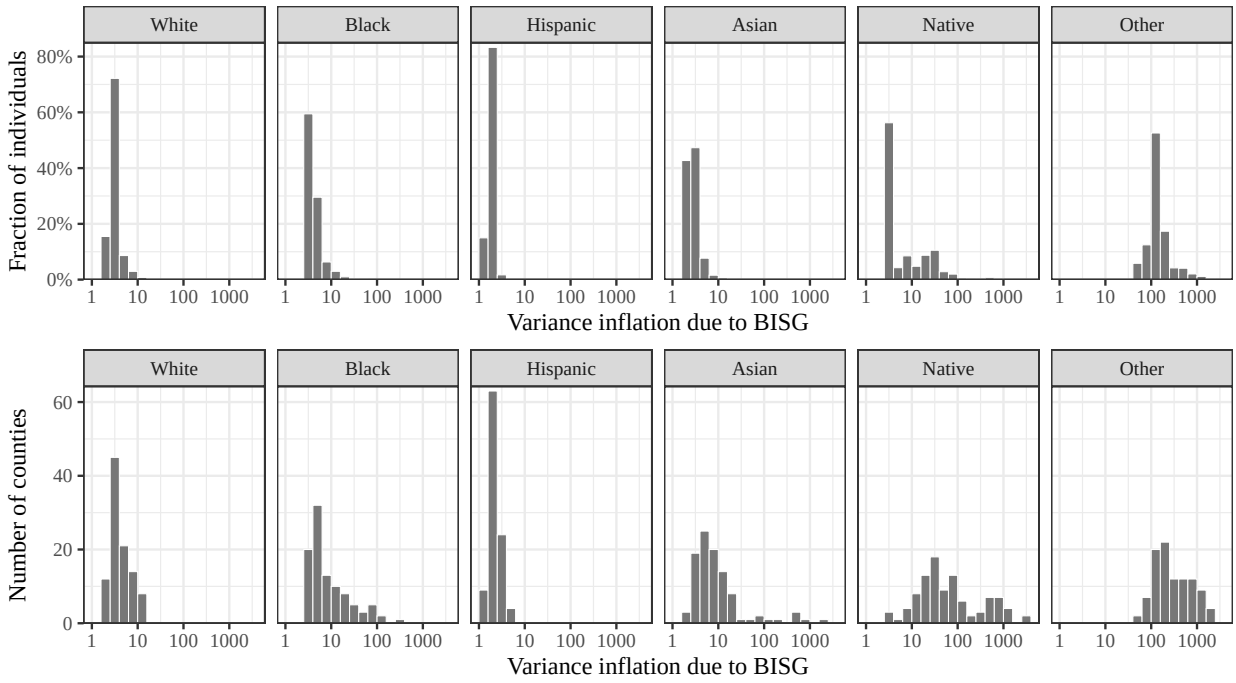

Figure 8: Variance inflation above a model with known individual race, by county submodel. The top set of histograms is weighted according to the population of each racial group; the bottom set is unweighted.

a county with a high variance inflation metric for the Native group but few Native voters but will not be weighted as much as a county with many Native voters. The bottom set of histograms is unweighted.

Figure 8 shows that surnames are a strong instrument for White, Black, Asian, and especially Hispanic voters, but a weaker instrument for Native voters and especially Other voters. This tracks with the measured precision of the BISG probabilities (see Appendix C.6).

C.5 Small-area Estimation

An advantage of the BIRDiE methodology is its explicit modeling of $\mathbb{P}(Y \mid R, G, X)$, which produces not only estimates of the marginal $\mathbb{P}(Y \mid R)$ but also subgroup estimates of how these conditional distributions vary across covariates and geographic areas. This section examines the accuracy of the saturated and mixed-effects BIRDiE models in recovering small-area relationships between party registration and race, compared to standard methodology that simply applies the weighting and thresholding estimators within each geographic area, as well as to the OLS es-

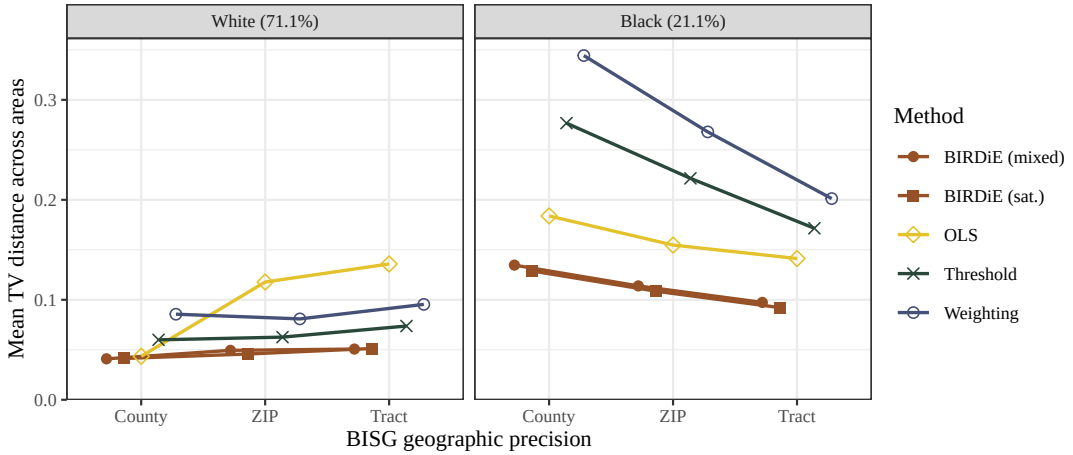

Figure 9: Accuracy of small-area estimates by race, as measured by the average total variation distance.

timator. We study accuracy at the county, ZIP code, and tract level, using the BISG probabilities and BIRDiE models that were applied to each. Since in fitting the BIRDiE model to block-level BISG probabilities we used tract-level random intercepts, we do not present block-level estimates.

We evaluate the small-area estimates by calculating the mean total variation distance between the estimated and true conditional distributions of party registration by race (averaging across geographic areas). Figure 9 summarizes our results, which qualitatively track the patterns found overall in Figure 6. The two BIRDiE models exhibit substantially lower error than the weighting, thresholding, and OLS estimators. Across all methods, the error is lower for White voters, who make up the bulk of the sample. Somewhat surprisingly, the amount of error does not appear to vary much for the BIRDiE models across different levels of geography—tract-level estimates are roughly as accurate as county-level estimates, on average.

Between the BIRDiE models, the mixed-effects model slightly outperforms the saturated model. This reflects the value in partially pooling estimates through the random effect structure.

We further evaluate the small-area estimates with two additional measures. First, we calculate the root-mean-square error (RMSE) of the estimated conditional probabilities by race within each geographic area, and then average this across all geographic areas. This captures the overall accuracy of the estimates. Second, to measure how well each method captures relative differences between geographic areas, we calculate the correlation between the estimated and true conditional

probabilities across all geographic areas by race. As in the main text, we remove area-race cells with fewer than 5 voters. A set of estimates which uniformly underestimates the proportion of Black voters which are registered Democrats, but which otherwise correctly orders geographic areas according to their proportion of Black Democrats, will score high on the correlation measure but also higher in RMSE.

Figure 10 summarizes our results, which closely track the findings of Section C.5. The BIRDiE models are more accurate at all geographic levels and for both Black and White voters. The weighting estimator performs the worst of all the methods.

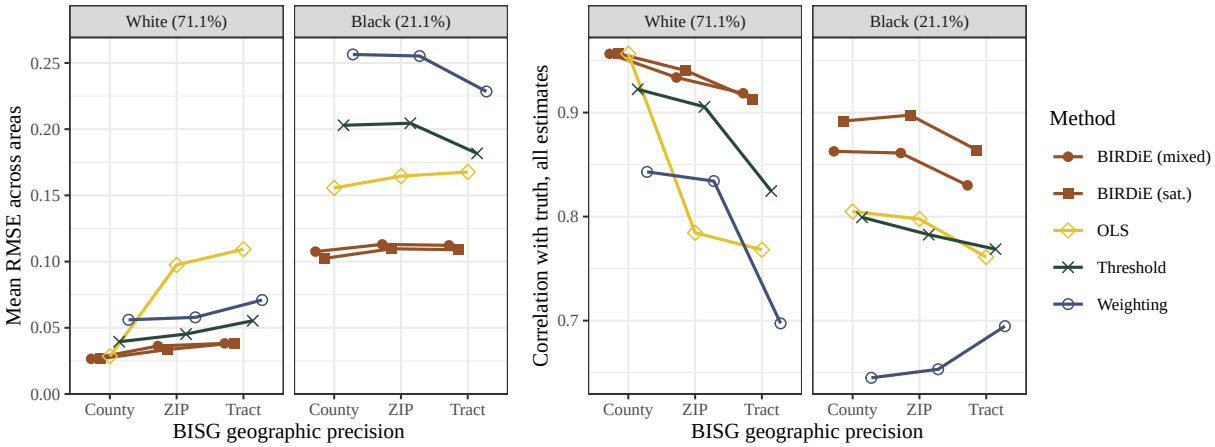

Figure 10: Accuracy of small-area estimates by race, as measured by root-mean-square error (RMSE; lower is better) and the correlation between the estimates and ground truth (higher is better).

C.6 Improved Individual Race Probabilities

As discussed in Section 4.3, we can use the conditional distribution $\mathbb{P}(Y \mid R, X, G)$ estimated with a BIRDiE model to create model-updated BISG probabilities $\tilde{\mathbf{P}}_{|Y} = \pi(\mathbf{R} \mid \hat{\Theta}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ by applying Bayes' rule. These updated probabilities may be more accurate than the original BISG probabilities.

For example, using the estimates produced by the mixed BIRDiE model applied to party registration with block-level BISG estimates, the MAP prediction accuracy increases from 79.6% with the input probabilities to 82.9% with the updated probabilities. These increases are signifi-

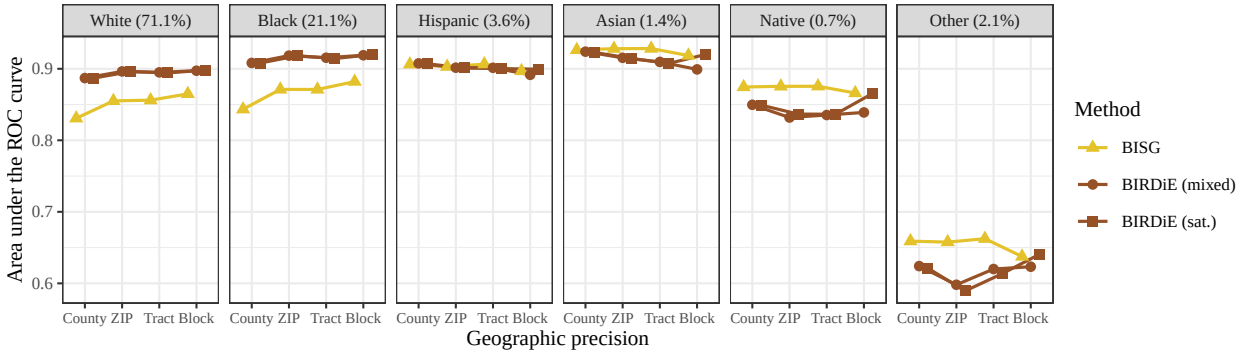

Figure 11: Race probability predictive accuracy, as measured by the area under the receiver operating characteristic (ROC) curve, for the input BISG probabilities as well as the BIRDiE-updated probabilities, by race and level of geographic precision. Larger values indicate more precise predictions.

cantly larger than the differences in accuracy between BISG probabilities using different levels of geographic precision.

The improvements are reflected in other accuracy measures as well. Figure 11 shows the accuracy of the predictions by race, as measured by the area under the receiver operating characteristic (ROC) curve. The updated probabilities are noticeably more accurate than the input probabilities for White and Black voters, about as accurate for Hispanic and Asian voters, and slightly less accurate for Native and "Other" voters for some geographic levels.

C.7 Estimates Conditional on an Additional Variable

The North Carolina voter file also provides an opportunity to demonstrate the methodology described in Section 4.4 to produce estimates conditional on another predictor variable that is not used in the BISG probabilities. We will estimate party registration rates by race among voters and nonvoters in the 2020 election. Following the discussion in Section 4.4, we will compute these estimates two ways: (1) by estimating party registration and 2020 turnout jointly by race, and (2) by first estimating 2020 turnout by race, then estimating party registration by race and 2020 turnout.

We will use a multinomial mixed-effects BIRDiE model applied to the block-level BISG probabilities, with random intercepts by tracts, for all the estimation. Fitting this model to a combined

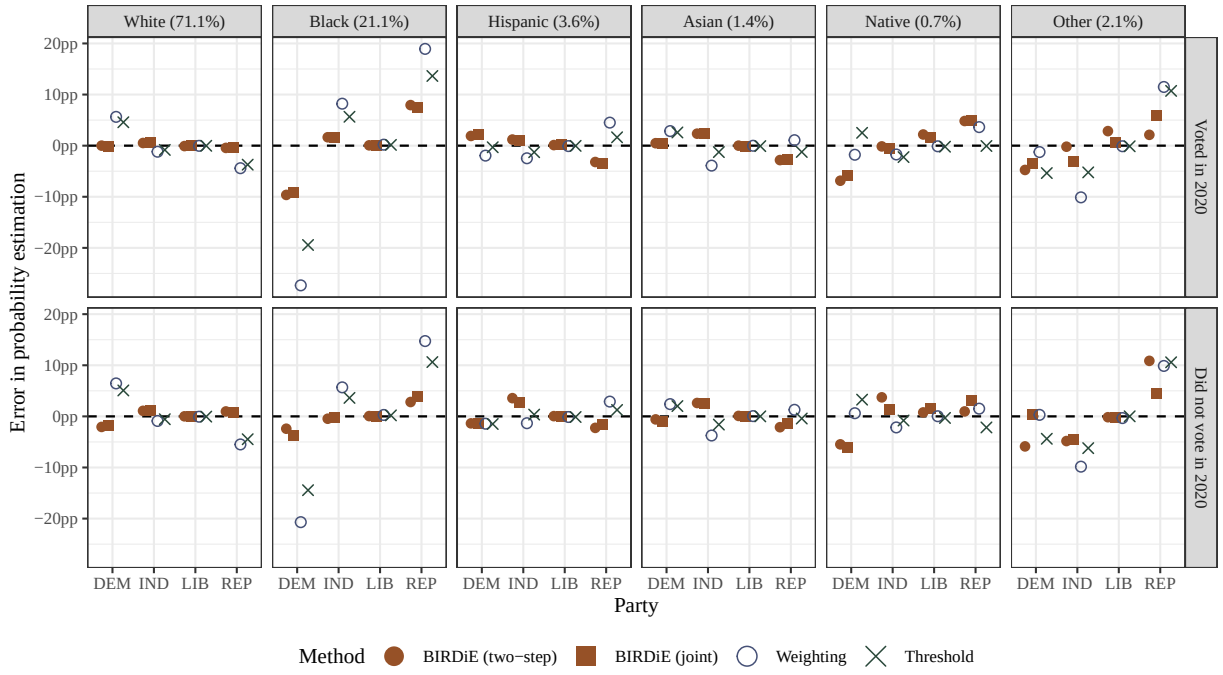

Figure 12: Error in the conditional expectation estimates for party registration by turnout and race, by estimation method. All methods used block-level data for this figure. Estimation uncertainty is minimal and hence suppressed from the figure for clarity.

2020 turnout/party variable (i.e., one with eight levels: no/DEM, yes/DEM, and so on) produces estimates of the joint distribution of party registration and turnout in 2020 by race. Normalizing these probabilities within turnout and race groups produces estimates of party registration by race and turnout. The total variation distance between these estimates and the true distribution is 0.051, which indicates close agreement. The accuracy of the most-likely race predictions from the BISG probabilities updated with both party and turnout is about the same as the party-alone accuracy, indicating that turnout is perhaps less correlated with race after controlling for location and party registration.

Figure 12 presents the errors in the estimates for this method (solid square), the two-step method (solid circle), and for the weighting (open circle) and thresholding (cross) estimators. The TV distance between these two-stage estimates and the true distribution is 0.052, very similar to the error in the joint-estimation approach. The two-step approach produces similar estimates to the joint-estimation approach, though the latter performs better in the "Other" category. As discussed

in Section 4.4, while both approaches produce highly accurate estimates in this example, we would expect the two-stage approach to be superior when one or both of the variables has more levels.

In contrast to the BIRDiE models, the weighting and thresholding estimates of party registration by 2020 turnout and race include large errors, especially for Black voters with all the errors exceeding 10 percentage points. The TV distance for the weighting estimator is 0.196, and the distance for the thresholding estimator is 0.147—around 3–4 times higher than for the estimates based on the BIRDiE models.

D Sensitivity Analysis

D.1 Local sensitivity analysis

In this section, we develop a sensitivity analysis that assesses how the bias in BISG race probabilities affect the estimates of racial disparities. In particular, we consider a setting where Assumptions CI-SG and ACC may be violated but Assumption CI-YS still holds. For example, consider the existence of an unobserved confounder that affects some or all of the variables except the outcome, i.e., (R, S, G, X). This leads to the violation of Assumption CI-SG, but Assumption CI-YS continues to be satisfied so long as such unobserved confounder does not affect the outcome. Unfortunately, even inaccurate BISG predictions can still lead to biased estimates of racial disparities.

Specifically, if either the Census data are inaccurate, or the conditional independence relation does not hold $S \not\perp\!\!\!\perp G, X \mid R$, then the BISG predictions $\widehat{\mathbf{P}}$ will differ from the "true" individual race probabilities $\mathbf{P}^* = \mathbb{P}(R \mid G, X, S)$. Our goal is to quantify how an error in these probabilities $\mathbf{P}^* - \widehat{\mathbf{P}}$ shifts the posterior and hence the estimates of racial disparities.

Denote by π_δ the posterior constructed using the error-corrected BISG race probabilities $\widehat{\mathbf{P}}_i + \delta_i$ as the input probabilities for the model (see Equation (3)), where π_{δ^*} is the true posterior with $\delta_i^* := \mathbf{P}_i^* - \widehat{\mathbf{P}}_i$. Estimating how π and π_δ differ in general is difficult, but we focus on the settings

where δ is small enough to make a linear approximation appropriate. In sum, we aim to quantify how the small error in BISG probabilities can alter the estimates of racial disparities.

For clarity, in this section we will use $\theta_{rG_iX_iY_i}$ to denote the model parameter or function thereof that represents $\pi(Y_i \mid R_i = r, G_i, X_i)$. This mirrors the notation of most of the specific models discussed in Section 4.2 above. Then define the following perturbation weight, which represents the ratio of posterior based on the biased and error-corrected BISG race probabilities:

$$w(\Theta, \delta^*) := \prod_{i=1}^{N} \left(1 + \frac{\theta_{\cdot G_i X_i Y_i}^{\top} \delta_i^*}{\theta_{\cdot G_i X_i Y_i}^{\top} \hat{\mathbf{P}}_i}\right) \propto \frac{\pi_{\delta^*}(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})}{\pi(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})},$$

Then, using a local linear approximation, we write the bias for a particular quantity of interest $g(\Theta)$ as

$$\begin{aligned}
\mathbb{E}_{\pi_{\delta^*}}[g(\Theta)] - \mathbb{E}_{\pi}[g(\Theta)] &= \left.\frac{d\, \mathbb{E}_{\pi_\delta}[g(\Theta)]}{d\delta}\right|_{\delta=0}^{\top} \delta^* + o(\|\delta^*\|) \\
&= \mathrm{Cov}_{\pi}\left(g(\Theta), \left.\frac{d \log w(\Theta, \delta)}{d\delta}\right|_{\delta=0}\right)^{\top} \delta^* + o(\|\delta^*\|), \qquad (6)
\end{aligned}$$

where the second equality is due to Theorem 2.1 of (Giordano et al. 2018; see also the idea of *local sensitivity* from Gustafson 1996).

With this representation, we can bound the total error in $\mathbb{E}_{\pi}[g(\Theta)]$ for sufficiently small δ as the following theorem shows.

Theorem D.1 (Bias Bound)**.** *Define* $\tilde{\vartheta}_{ir} := \frac{\theta_{rG_iX_iY_i}}{\theta_{\cdot G_iX_iY_i}^{\top}\hat{\mathbf{P}}_i}$. *Then for any input probabilities with total error* $\|\delta^*\|^2 = \sum_{i=1}^{n} \|\delta_i^*\|^2 \leq \Delta^2$,

$$|\,\mathbb{E}_{\pi^*}[g(\Theta)] - \mathbb{E}_{\pi}[g(\Theta)]| \lesssim \Delta \left\|\mathrm{Cov}_{\pi}(g(\Theta), \tilde{\vartheta})\right\|, \qquad (7)$$

as $\Delta \to 0$.

Proof. This is immediate from (6) once we compute

$$\begin{aligned}
\frac{d \log w(\Theta, \delta)}{d\delta_{ir}} &= \frac{d}{d\delta_{ir}} \sum_{i=1}^{N} \log\left(1 + \frac{\theta_{\cdot G_iX_iY_i}^{\top}\delta_i}{\theta_{\cdot G_iX_iY_i}^{\top}\hat{\mathbf{P}}_i}\right) \\
&= \frac{d}{d\delta_{ir}} \log\left(1 + \frac{\theta_{\cdot G_iX_iY_i}^{\top}\delta_i}{\theta_{\cdot G_iX_iY_i}^{\top}\hat{\mathbf{P}}_i}\right)
\end{aligned}$$

$$= \frac{1}{1 + \frac{\theta_{\cdot G_i X_i Y_i}^\top \delta_i}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}} \times \frac{\theta_{r G_i X_i Y_i}}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}$$
$$= \frac{\theta_{r G_i X_i Y_i}}{\theta_{\cdot G_i X_i Y_i}^\top (\hat{\mathbf{P}}_i + \delta_i)}$$

and evaluate at $\delta = 0$, since the worst-case bias for a fixed total error can be obtained by having the maximum allowable δ point in the direction of the gradient of $\log w(\Theta, \delta)$. □

The theorem shows that once researchers choose the amount of total error Δ, then the bound on the shift in a quantity of interest can be computed readily from posterior draws. It is important to note that both δ^* and $\text{Cov}_\pi(g(\Theta), \tilde{\vartheta})$ are vectors whose dimension depends on the sample size N. Thus, all else being equal, their norms will each grow as $\sqrt{N}$. However, each entry $\text{Cov}_\pi(g(\Theta), \tilde{\vartheta}_{ir})$ will tend to shrink as N increases, since each observation exerts less leverage on the overall posterior. Thus the overall impact of the sample size on the bound in Equation (7) may depend on specific features of the data. In particular, and as should be expected, the error is not guaranteed to vanish as $N \to \infty$. Practitioners should evaluate Equation (7) under a range of plausible Δ to understand how robust their findings are to worst-case linear violations of Assumptions $CI - SG$ and ACC.

D.2 OLS sensitivity analysis

For a different understanding of the effect of a particular δ, we can derive a result on the error in conditional probability estimates under the OLS estimator and particular configurations of δ. Unlike Theorem D.1, this result holds across all sizes of δ, and not just asymptotically as $\|\delta\| \to 0$. However, it applies to the OLS estimator, which, while unbiased, we do not recommend in practice. Despite this difference, we expect many of the qualitative conclusions to hold for BIRDiE models. The effect of any particular δ can of course be calculated directly by re-fitting the model to new race probabilities.

Here, we will work with a fixed $y \in \mathcal{Y}$ and among the subset of individuals with a particular $g \in \mathcal{G}$ and $x \in \mathcal{X}$. Then for notational simplicity we let $\hat{\mu}^{(\text{ols})}$ be the vector of estimates of

$\mathbb{P}(Y = y \mid R, G = g, X = x)$, and μ the corresponding true probabilities. Similarly, we write $\hat{\mathbf{P}}$ for the matrix of individual race probability estimates for the subset of individuals with $g \in \mathcal{G}$ and $x \in \mathcal{X}$; elsewhere in the text this would be notated $\hat{\mathbf{P}}_{\mathcal{J}(xg)}$

Proposition D.2 (OLS Bias from incorrect $\hat{\mathbf{P}}$)**.** *Let Assumption CI-YS hold. If the OLS estimator $\hat{\mu}^{(ols)}$ is calculated using race probabilities $\hat{\mathbf{P}}$ which differ from the true probabilities $\mathbf{P}^* = \hat{\mathbf{P}} + \delta$, then its bias satisfies*

$$\mathbb{E}[\hat{\mu}^{(ols)}] - \mu = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \delta \mu$$

Proof. We can write the OLS estimate as

$$\hat{\mu}^{(\text{ols})} = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \mathbf{1}\{\mathbf{Y} = y\}.$$

As shown in Theorem 4.2, under Assumption CI-YS, $\mathbb{P}(Y = y \mid S = s, G = g, X = x)$ is linear in the true $\mathbf{P}^*$. Thus letting $\varepsilon = \mathbf{1}\{\mathbf{Y} = y\} - \mathbb{P}(Y = y \mid S = s, G = g, X = x)$, we can substitute and find

$$\begin{aligned}\hat{\mu}^{(\text{ols})} &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\mathbf{P}^* \mu + \varepsilon) \\ &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \big((\hat{\mathbf{P}} + \delta)\mu + \varepsilon\big).\end{aligned}$$

Taking an expectation, since $\mathbb{E}[\varepsilon] = 0$ we find

$$\begin{aligned}\mathbb{E}[\hat{\mu}^{(\text{ols})}] &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \big((\hat{\mathbf{P}} + \delta)\mu\big) \\ &= \mu + (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\delta \mu);\end{aligned}$$

rearrangement yields the result. □

Informally, for BISG error δ to cause problems with the OLS estimate, two things must happen. First, within individuals it must be "correlated" (i.e., have nonzero inner product) with the true conditional probabilities μ. Since δ_i must always sum to zero, practically, this means that positive BISG errors must tend to occur in racial groups which have a relatively high occurrence of outcome y: $\mathbb{P}(Y = y \mid R = r, G = g, X - x) > \mathbb{P}(Y = y \mid G = g, X - x)$. Second, the vector $\delta\mu$ (where each entry measures this "correlation" between errors and relative frequencies)

must be correlated with the BISG probabilities themselves $\hat{\mathbf{P}}$. For example, if $\delta_i \mu$ is positive and tends to be larger for individuals with a high BISG probability of being Hispanic, then the overall OLS estimator the conditional probability of $Y = y$ among Hispanics will be biased upwards.

While Proposition D.2 applies within a (G, X) cell, if the same conditions hold across all (G, X) combinations, then the overall poststratified estimator will be similarly biased.

E Surname Groupings for North Carolina Robustness Analysis

We classify every surname in the voter file into one of nine groups, each containing surnames from one or more of 22 surname groups that we provide in the replication data and software. These groups are organized mainly around different regions of the world and different waves of immigration to the United States. To create the surname groups, each individual in the 1930 Census data was classified into one of the 22 groups. Then among the set of individuals with each surname, the group with the highest number of individuals relative to the whole population was assigned to that surname. For example, while most people named "Smith" fall into the Anglosphere group (containing 3rd or more generation White U.S. residents as of 1930, as well as immigrants from the U.K., Canada, Australia, etc.), there are relatively more Smiths among Black people than any other of the 22 groups. Thus "Smith" is assigned to the Black surname group. The full code for creating the 22 surname groupings from the 1930 Census data is available in the replication materials.

Because of the demographics of the United States, as well as limitations of the source data there is more geographic specificity in the surname groupings for some regions (e.g., Europe) than for others (e.g., South America and Africa). We collapse the 22 surname groups to nine for the robustness analysis in Section 5 based on the demographics of North Carolina specifically and to minimize the computational burden of performing the robust analysis.

The 50 roughly most frequent surnames in each group, along with a brief description of the group, are listed below. We stress that for the purposes of sensitivity analysis, the surname groups need only be correlated with countries of origin and racial subgroups. Perfect alignment is neither possible nor necessary.

Anlglosphere and Black surname group. Surnames which are relatively more prevalent among 3rd-or-more generation White U.S. residents and Black U.S. residents in 1930.

1. SMITH	11. LEWIS	21. HALL	31. COLLINS	41. COX
2. WILLIAMS	12. ROBINSON	22. CAMPBELL	32. STEWART	42. WARD
3. BROWN	13. WALKER	23. MITCHELL	33. MORRIS	43. RICHARDSON
4. JONES	14. ALLEN	24. CARTER	34. COOK	44. WATSON
5. DAVIS	15. WRIGHT	25. ROBERTS	35. ROGERS	45. BROOKS
6. TAYLOR	16. SCOTT	26. PHILLIPS	36. MORGAN	46. WOOD
7. MOORE	17. HILL	27. EVANS	37. COOPER	47. JAMES
8. JACKSON	18. GREEN	28. TURNER	38. BAILEY	48. BENNETT
9. WHITE	19. ADAMS	29. PARKER	39. REED	49. GRAY
10. CLARK	20. BAKER	30. EDWARDS	40. HOWARD	50. HUGHES

First wave European immigration surname group. Surnames associated with German, Nordic, and Irish immigrants.

1. JOHNSON	11. BURNS	21. CARROLL	31. SCHULTZ	41. HIGGINS
2. ANDERSON	12. OLSON	22. RILEY	32. PEARSON	42. OCONNOR
3. NELSON	13. WAGNER	23. BURKE	33. BARRETT	43. QUINN
4. MURPHY	14. MEYER	24. LARSON	34. BECK	44. SWANSON
5. PETERSON	15. SCHMIDT	25. CARLSON	35. POWERS	45. FITZGERALD
6. KELLY	16. RYAN	26. OBRIEN	36. LEONARD	46. CHRISTENSEN
7. SULLIVAN	17. DUNN	27. LYNCH	37. BENSON	47. MANNING
8. MURRAY	18. KELLEY	28. HANSON	38. LYONS	48. MCLAUGHLIN
9. MCDONALD	19. HANSEN	29. WEBER	39. MCCARTHY	49. DOYLE
10. KENNEDY	20. CUNNINGHAM	30. WALSH	40. ERICKSON	50. BRADY

Second wave European immigration surname group. Surnames associated with Eastern European, Italian, Jewish, Russian, Greek, and other Southern European immigrants.

1. FOX	11. ZIMMERMAN	21. KLINE	31. KATZ	41. NICHOLAS
2. NICHOLS	12. KLEIN	22. BERGER	32. MARINO	42. ROSENBERG
3. HOFFMAN	13. GROSS	23. STEIN	33. BRUNO	43. ROSSI
4. NEWMAN	14. GOODMAN	24. RAYMOND	34. MOSER	44. SINGER
5. SCHNEIDER	15. SHERMAN	25. FRIEDMAN	35. GOLDSTEIN	45. ABRAMS
6. KELLER	16. WOLF	26. LEVY	36. GOLDBERG	46. ACKERMAN
7. GREGORY	17. KRAMER	27. NOVAK	37. KAPLAN	47. HELLER
8. SCHWARTZ	18. NICHOLSON	28. KAUFMAN	38. KESSLER	48. STERN
9. COHEN	19. WEISS	29. LEVINE	39. ROMANO	49. SCHAFER
10. BECKER	20. RUSSO	30. LEHMAN	40. FINK	50. SHAPIRO

East Asian surname group. Surnames associated with Chinese, Japanese, and Korean immigrants.

1. LEE	11. BOWEN	21. HORNE	31. HAN	41. LIANG
2. YOUNG	12. LIU	22. XIONG	32. LAU	42. SUN
3. WONG	13. PAUL	23. LIM	33. MA	43. JUNG
4. WANG	14. CHAN	24. TANG	34. PUCKETT	44. ZHOU
5. PARK	15. TODD	25. CHO	35. CHIN	45. GEE
6. MAY	16. ZHANG	26. CHENG	36. GIL	46. ZHAO
7. JOSEPH	17. LANG	27. KANG	37. XU	47. SHIN
8. LOWE	18. YU	28. LAW	38. SONG	48. OHARA
9. CHANG	19. CHOI	29. CRAFT	39. KAY	49. ZHU
10. LIN	20. MOON	30. NG	40. STROUD	50. YEE

South Asian surname group. Surnames associated with Indian and Southwest Asian immigrants.

1. WILSON	11. CARR	21. DAVID	31. MOHAMED	41. SAMUEL
2. THOMAS	12. SINGH	22. HOWE	32. BOGGS	42. SEWELL
3. PATEL	13. BISHOP	23. HAHN	33. KUMAR	43. HASSAN
4. STEVENS	14. MANN	24. GOOD	34. WESTON	44. SADLER
5. WOODS	15. FRANCIS	25. JOHN	35. BEATTY	45. PINTO
6. SHAW	16. GILL	26. OSBORN	36. SWAIN	46. MAJOR
7. FERGUSON	17. YATES	27. ABRAHAM	37. GOMES	47. BARNHART
8. RAY	18. MARSH	28. RODRIGUES	38. JACOB	48. CARMICHAEL
9. WILLIS	19. ROY	29. PEREIRA	39. TOLBERT	49. MUHAMMAD
10. GEORGE	20. KAUR	30. SHARMA	40. PAYTON	50. GUPTA

Southeast Asian and Pacific surname group. Surnames associated with Southeast Asian and Pacific Islander immigrants, including Vietnamese and Filipino immigrants.

1. MILLER	11. SILVA	21. HUANG	31. PHAN	41. HOANG
2. MARTIN	12. SANTOS	22. WU	32. VO	42. CASH
3. KING	13. GREENE	23. ROWE	33. VU	43. BUI
4. NGUYEN	14. LI	24. BAUTISTA	34. LU	44. CHU
5. KIM	15. LE	25. HOUSTON	35. NGO	45. SINCLAIR
6. LONG	16. YANG	26. LAM	36. TAN	46. SORIANO
7. TRAN	17. LITTLE	27. HUYNH	37. HONG	47. ZHENG
8. CHEN	18. MORAN	28. HO	38. DANG	48. LESLIE
9. WEBB	19. PHAM	29. CHUNG	39. DO	49. ANGEL
10. GORDON	20. RAMSEY	30. TRUONG	40. LY	50. DUONG

Non-Cuban Hispanic surname group. Surnames associated with Mexican and Latin American immigrants, not including Cuban immigrants, and Puerto Rican residents.

1. GARCIA	11. RIVERA	21. MENDOZA	31. CASTRO	41. ALVARADO
2. RODRIGUEZ	12. GOMEZ	22. RUIZ	32. FERNANDEZ	42. DELGADO
3. MARTINEZ	13. DIAZ	23. CASTILLO	33. VARGAS	43. PENA
4. HERNANDEZ	14. CRUZ	24. GONZALES	34. GUZMAN	44. CONTRERAS
5. LOPEZ	15. REYES	25. VASQUEZ	35. MENDEZ	45. SANDOVAL
6. PEREZ	16. MORALES	26. ROMERO	36. MUNOZ	46. GUERRERO
7. SANCHEZ	17. GUTIERREZ	27. MORENO	37. SALAZAR	47. RIOS
8. RAMIREZ	18. ORTIZ	28. HERRERA	38. GARZA	48. ESTRADA
9. TORRES	19. RAMOS	29. MEDINA	39. SOTO	49. ORTEGA
10. FLORES	20. CHAVEZ	30. AGUILAR	40. VAZQUEZ	50. NUNEZ

Cuban surname group. Surnames associated with Cuban immigrants.

1. GONZALEZ	11. SUAREZ	21. CRANE	31. SARGENT	41. MARRERO
2. ALVAREZ	12. CONNER	22. FRYE	32. GORE	42. VALDES
3. JIMENEZ	13. SANTANA	23. PARRA	33. ZIEGLER	43. OLIVA
4. BOWMAN	14. DECKER	24. MAYO	34. TOMLINSON	44. MCCLENDON
5. DAVIDSON	15. SKINNER	25. DAVIES	35. LOWRY	45. QUEEN
6. ACOSTA	16. ABBOTT	26. BLANCO	36. PAGAN	46. MCCORD
7. MOLINA	17. GARRISON	27. WITT	37. LORD	47. CRESPO
8. MIRANDA	18. PONCE	28. CARRASCO	38. CARBAJAL	48. CORNEJO
9. CASTANEDA	19. PALACIOS	29. ALONSO	39. BETANCOURT	49. DUMAS
10. BALL	20. SLOAN	30. HAINES	40. PATINO	50. BUENO

"Other" surname group. Surnames not associated with one of the other categories, including those associated with later Western European immigration, Middle Eastern & North African-associated surnames, Native-associated surnames and Afro-Caribbean-associated surnames.

1. PERRY	11. WELCH	21. SIMON	31. FRANCO	41. MCKENZIE
2. HENRY	12. DAY	22. CUMMINGS	32. HAMMOND	42. BEIL
3. HUNT	13. STANLEY	23. CHANDLER	33. CLARKE	43. COCHRAN
4. ROSE	14. HOPKINS	24. SHARP	34. WATERS	44. NASH
5. PIERCE	15. LAMBERT	25. BARBER	35. FRANK	45. BRYAN
6. PETERS	16. NORRIS	26. GRIFFITH	36. ANDRADE	46. MEYERS
7. KNIGHT	17. WALTERS	27. PACHECO	37. LLOYD	47. CARSON
8. RICHARDS	18. STEELE	28. CROSS	38. FRENCH	48. WILKINSON
9. MORRISON	19. BUSH	29. GOODWIN	39. OWEN	49. ATKINSON
10. JACOBS	20. WOLFE	30. MULLINS	40. CHARLES	50. VINCENT

F Further details on tax study

F.1 Modeling details

Using a larger geography like PUMAs rather than ZCTAs is necessary to ensure a reasonable amount of data in each outcome-race-geography cell of the outcome model. PUMAs partition each state into areas containing roughly 100,000 people. Compared to alternative units of analysis like states or counties, PUMAs are therefore adaptive to population density; a large city might be contained entirely in a single county, but with a dozen or more distinct PUMAs, while the surrounding rural areas might be spread over many counties but only a few PUMAs. Since we expect more geographic heterogeneity in and around cities, this feature of PUMAs lends itself well to our analysis. Where PUMA was not available for the 3.4 million missing geocodes, we used an indicator for state of residence instead, which is available for every record in the sample.

There are 1,961 distinct PUMAs (or states) in the sample, compared to 28,880 ZCTAs. With 10 non-zero outcome levels and 6 racial groups, the typical cell in the outcome model is expected to have around 14 observations. Smaller racial groups would have fewer observations than this, on average. Therefore, to help regularize the PUMA-level model estimates, we impose a weak empirical Bayes prior based on a global estimate of HMID by race from the simple weighted estimator. The effective data size of the prior is just 0.1 observations per racial group. Thus the prior will have a meaningful impact only in those areas where there are close to zero expected members of a racial group, according to the BISG probabilities.

F.2 Estimating mortgage rates by race

The decennial census reports the number of homeowners by race, and the number of homeowners with a mortgage, but does not report the number of mortgages by race. Thus we are left to infer the mortgage-race distribution from this marginal information. Fortunately, both ownership-race and mortgage-ownership marginals are reported at fine geographic levels.

Table 2

Race	Total households	Fraction owner-occupied	Fraction with mortgage
White	82,343,859	72.2%	49.9%
Black	13,797,354	44.6%	31.7%
Hispanic	14,822,017	49.6%	33.7%
Asian	4,580,883	58.1%	44.4%
Native	758,975	57.2%	32.0%
Other	1,788,360	49.3%	35.6%

Census-reported number of households, and fraction that own their home, by race, and estimated fraction that have a mortgage, by race.

We therefore produce estimates of mortgage rates by race at the ZCTA level, then aggregate these estimates to the nation. Stratifying by ZCTA means that any variation in mortgage rates by race that is explained by geographic variation will be captured.

Within each ZCTA we estimate the number of mortgages by race by taking the fraction of homeowners with a mortgage and multiplying by the number of homeowners of each racial group. Nationwide, we find no association between the racial composition of a ZCTA and the fraction of homeowners with a mortgage. While not dispositive because of possible aggregation bias, this finding nevertheless suggests that after controlling for geography, little variation by racial group in the fraction of homeowners with a mortgage remains.

The table below reports our nationwide estimates of the fraction of each racial group with a mortgage.

G Additional discussion

G.1 Recommendations for Practitioners

Given the large amount of missing data inherent to the type of racial disparity estimation studied here, choices about data selection, processing, and modeling can have a significant impact on estimates and substantive conclusions. We collect here several recommendations for practitioners using BISG and BIRDiE methodology in their research.

- **Choose estimation methodology and input data based on the specifics of the research question.** As we have stressed, the causal structure of the research setting determines whether BISG should be used with the weighting or BIRDiE estimators. Research on populations which are very different from the general U.S. population may also benefit from the use of alternative or additional data on race and geography specific to that population. The choice of whether to use state-, county-, ZIP-, tract-, or block-level data likewise depends on aspects of the data under study, and the scale of geographic variation in the outcome variable and quantities of interest.

- **Focus on the calibration, not the predictive accuracy, of BISG probabilities.** Traditionally, BISG-type methods have been evaluated by their predictive accuracy, measured by the mean agreement between the thresholded racial categories and ground truth, or the AUROC of the BISG probabilities. We find that, especially in cases with abundant individual-level data, far more important than maximizing predictive accuracy is ensuring that the BISG probabilities are properly calibrated. Accurate but biased racial prediction will lead to bias in downstream estimates, regardless of which disparity estimator is used. To evaluate probabilistic calibration in validation settings, we recommend visual diagnostics like binned residual plots, as well as numerical summaries like the logarithmic score.

- **Decide whether additional covariates need to be collected.** Additional covariates can

make the BISG and BIRDiE assumptions more plausible, and can increase the accuracy of BISG probabilities and the precision of BIRDiE estimates. We recommend prioritizing covariates which are highly predictive of both outcome and race. However, including all available covariates without considering their effect on the various BISG and BIRDiE assumptions is likely to cause problems. In particular, if a covariate's distribution against both race and geography is not known, avoid making unrealistic independence assumptions that are required for the inclusion of the covariate in the BISG model. Rather, follow the approach outlined in Section 4.4. Covariate choice should be driven by substantive considerations, not convenience.

- **Unless computational limits are severe, use the mixed-effects BIRDiE model with group-level covariates.** The mixed-effects model uses the data itself to determine how much to pool disparity estimates across different geographies. This avoids the dual pitfalls of Simpson's paradox (a risk if no geographical information is used) and over-regularization (caused by the prior if no pooling is performed). When using the mixed-effects model, group-level covariates are likely to improve both overall and especially small-area accuracy, and they are easy to collect from public sources. For example, a good default for a ZIP-code level model applied to a political outcome variable would be to include the percentages of the major racial groups in the ZIP code, as well as the ZIP code's income level, population density, and partisanship, in the model.

- **Perform sensitivity analyses.** Both the BISG and BIRDiE models use priors, which should be perturbed to examine the sensitivity of the results to prior selection. The sensitivity of BIRDiE to its key identifying assumption should also be assessed, at minimum using the auxiliary covariate approach demonstrated in Section 5.4.

- **Consider validating estimates with a small-scale survey.** Even with administrative microdata observations in the millions, there is no substitute for a high-quality random sample

of individuals for whom race can be observed. Such a sample can be used to validate the various assumptions made by BISG and BIRDiE, as well as providing a sanity check against BIRDiE disparity estimates. Future improvements to BIRDiE could also directly integrate survey data into the workflow.

G.2 Ethical Considerations

As researchers have increasingly applied racial prediction and imputation methods to administrative records, including many publicly available datasets, there has been growing concern about ethical and privacy considerations around performing such predictions. While some scholars do not view racial prediction methods as privacy risks (Bun et al. 2021, "Statistical Inference is Not a Privacy Violation"), we believe it is important for researchers to consider the implications of their use of racial prediction methods (Kenny et al. 2023).

Recently, Lee & Velez (2023) studied public perception of these ethical considerations through a large factorial survey experiment that asked participants if they viewed a hypothetical study as ethically permissible based on three study factors. They find that studies which focused on accurate estimation of racial disparities were viewed as more ethically permissible than studies that overestimated or underestimated the size of disparities. This tracks with arguments made by many scholars of progressive tax policy (Brown 2022, Bearer-Friend 2019).

Compared to previous approaches to racial disparity estimation, which emphasized maximizing accuracy of individual racial predictions to minimize measurement error, BIRDiE focuses on the accuracy of estimating racial disparities. In fact, as our validation study demonstrated, different sets of individual race predictions of varying individual accuracies (county-based versus block-based BISG), when used with BIRDiE, produced similar estimates of racial disparities. To the extent that it allows researchers to focus on calibration of racial prediction rather than maximizing individual predictive accuracy, BIRDiE may alleviate some privacy concerns and reduce incentives to collect and link more personal data in an attempt to further increase accuracy. We

view this as a welcome change, consistent with the public's preference for focusing on accurate disparity information.

However, BIRDiE does allow for the creation of improved BISG probabilities that incorporate the outcome variable and thus can be more accurate as well as better calibrated. While this accuracy gain is a purely statistical phenomenon based on variables already present in the individual dataset under study, researchers should be cautious, for example, in releasing these improved racial predictions publicly. We urge practitioners to view racial prediction tools as a means to the end of accurate disparity estimation, and treat the intermediate probabilistic predictions with appropriate care.

G.3 Future Research

Much work remains to be done in accurately and reliably measuring racial disparities. First, the BISG probabilities themselves can be improved. Approaches to doing so include Imai et al. (2022), which accounts for some Census measurement error while remaining computationally tractable, and Greengard & Gelman (2023), which rakes BISG margins to improve calibration. More work on identifying and producing data sources which can be used as BISG inputs, rather than relying solely on Census tabulations, will also pay dividends for BISG quality.

Beyond the BISG probabilities, further empirical analyses could determine useful additional variables to condition on, which could allow analysts to weaken the required assumption. Additional study, possibly combined with qualitative research, could identify causal pathways that might threaten the assumptions that BISG and BIRDiE rely on, and develop data sources, like our auxiliary 1930 Census data, that could be used to evaluate the plausibility of those assumptions in real-world analyses, and their effect on numerical conclusions. The BIRDiE model could also be extended to directly model more complex types of outcome variables.

Finally, in cases where it is possible to obtain the individual race for a random subset of records, methods developed by Selén (1986), Fong & Tyler (2021), Egami et al. (2024), and Angelopoulos

et al. (2024) might be fruitfully applied.

References

Angelopoulos, A. N., Duchi, J. C. & Zrnic, T. (2024), 'PPI++: Efficient prediction-powered inference', *arXiv preprint arXiv:2311.01453* .

Bearer-Friend, J. (2019), 'Should the IRS know your race? The challenge of colorblind tax data', *Tax L. Rev.* **73**, 1.

Brown, D. A. (2022), *The whiteness of wealth: How the tax system impoverishes Black Americans–and how we can fix it*, Crown.

Bun, M., Desfontaines, D., Dwork, C., Naor, M., Nissim, K., Roth, A., Smith, A., Steinke, T., Ullmanand, J. & Vadhan, S. (2021), 'Statistical inference is not a privacy violation'.
URL: *https://differentialprivacy.org/inference-is-not-a-privacy-violation*

Buttice, M. K. & Highton, B. (2013), 'How does multilevel regression and poststratification perform with conventional national surveys?', *Political Analysis* **21**(4).

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Egami, N., Hinck, M., Stewart, B. M. & Wei, H. (2024), 'Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses.'.

Fong, C. & Tyler, M. (2021), 'Machine learning predictions as regression covariates', *Political Analysis* **29**(4), 467–484.

Giordano, R., Broderick, T. & Jordan, M. I. (2018), 'Covariances, robustness and variational Bayes', *Journal of machine learning research* **19**(51).

Greengard, P. & Gelman, A. (2023), 'BISG: When inferring race or ethnicity, does it matter that people often live near their relatives?'.

Gustafson, P. (1996), 'Local sensitivity of posterior expectations', *The Annals of Statistics* **24**(1), 174–195.

Imai, K., Olivella, S. & Rosenman, E. T. (2022), 'Addressing census data problems in race imputation via fully Bayesian improved surname geocoding and name supplements', *Science Advances* **8**(49), 1–10.

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T. & Imai, K. (2023), 'Comment: The essential role of policy evaluation for the 2020 census disclosure avoidance system', *Harvard Data Science Review* **Special Issue 2: Dierential Privacy for the 2020 U.S. Census**, 1–16.

Laird, N. (1993), The EM algorithm, *in* 'Computational Statistics', Vol. 9 of *Handbook of Statistics*, Elsevier, pp. 509–520.
URL: *https://www.sciencedirect.com/science/article/pii/S0169716105801385*

Lee, D. D. I. & Velez, Y. (2023), 'Ethical use of administrative data in inequality research'.

Rosenman, E. T., Olivella, S. & Imai, K. (2023), 'Race and ethnicity data for first, middle, and last names', *Scientific Data* **10**(299), 1–11.

Selén, J. (1986), 'Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data', *Journal of the American Statistical Association* **81**(393), 75–81.

Tzioumis, K. (2018), 'Demographic aspects of first names', *Scientific Data* **5**(1), 1–9.

Varadhan, R. & Roland, C. (2008), 'Simple and globally convergent methods for accelerating the convergence of any EM algorithm', *Scandinavian Journal of Statistics* **35**(2), 335–353.