Supplementary Material for “Estimating Heterogeneous Causal Effects of High-Dimensional Treatments: Application to Conjoint Analysis”

Max Goplerud Kosuke Imai Nicole E. Pashley

A The Details of the Immigration Conjoint Experiment

Attribute	# of Levels	Levels
Education	7	No formal education; Equivalent to completing fourth grade in the U.S.; Equivalent to completing eighth grade in the U.S.; Equivalent to completing high school in the U.S.; Equivalent to completing two years at college in the U.S.; Equivalent to completing a college degree in the U.S.; Equivalent to completing a graduate degree in the U.S.
Gender	2	Female; Male
Country of origin	10	Germany; France; Mexico; Philippines; Poland; India; China; Sudan; Somalia; Iraq
Language	4	During admission interview, this applicant spoke fluent English; During admission interview, this applicant spoke broken English; During admission interview, this applicant tried to speak English but was unable; During admission interview, this applicant spoke through an interpreter
Reason for Application	3	Reunite with family members already in U.S.; Seek better job in U.S.; Escape political/religious persecution
Profession	11	Gardener; Waiter; Nurse; Teacher; Child care provider; Janitor; Construction worker; Financial analyst; Research scientist; Doctor; Computer programmer
Job experience	4	No job training or prior experience; One to two years; Three to five years
Employment Plans	4	Has a contract with a U.S. employer; Does not have a contract with a U.S. employer, but has done job interviews; Will look for work after arriving in the U.S.; Has no plans to look for work at this time
Prior Trips to the U.S.	5	Never been to the U.S.; Entered the U.S. once before on a tourist visa; Entered the U.S. once before without legal authorization; Has visited the U.S. many times before on tourist visas; Spent six months with family members in the U.S.

Table A1: Table 1 in Hainmueller and Hopkins (2015). All attributes for immigrants and their levels.

B Additional Results for Comparison with `cjbart`

We compare the performance of our method with that of Robinson and Duch (2024) whose method is implemented using an open-source software package, `cjbart` (Robinson and Duch, 2023). We use the same set of moderators and factors considered in our earlier analyses. Figure A1 compares the estimated CAMCEs for country with Germany set as the reference category, calculated across all individual covariate vectors in the sample. Our method discovers a considerable degree of

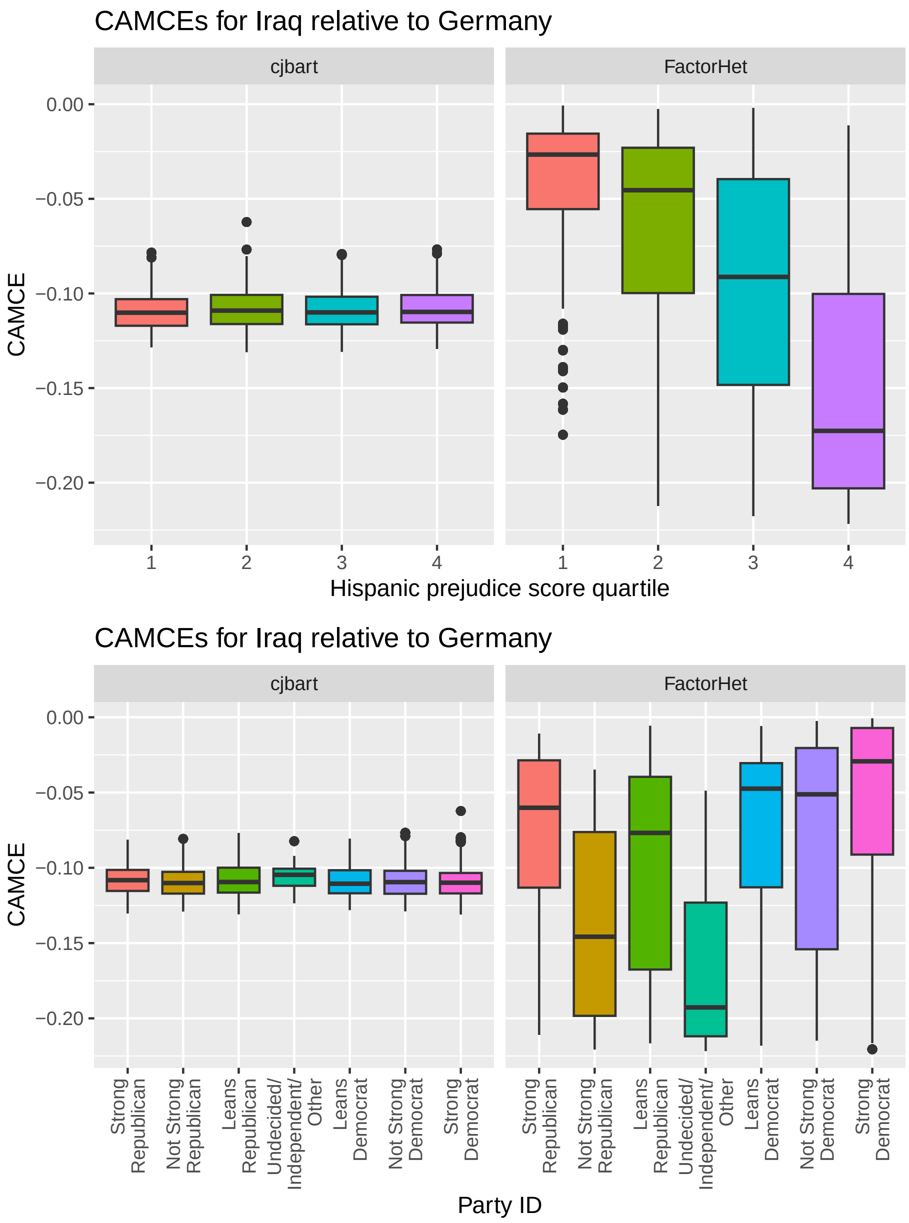


Figure A1: Comparison of discovered heterogeneous effects (the conditional average marginal component effects or CAMCEs) between the proposed method and the BART-based method `cjbart`. In both plots the y-axis corresponds to values estimated, either by our method (right) or by `cjbart` (Robinson and Duch, 2024) (left). The plots show the estimated effect of Iraq as compared to the baseline of Germany. In the top figure, the x-axis and color corresponds to the categories of individuals based on the quartile of their Hispanic prejudice score. In the bottom figure, the x-axis and color corresponds to party ID.

heterogeneity in the CAMCEs whereas `cjbart` shows limited treatment effect variation for most countries. Under our model, the estimated heterogeneous effects are more strongly associated with predictors than `cjbart`. For example, our method finds a clear association, on average, between the estimated CAMCEs and prejudice or party identification, whereas `cjbart` does not.

Figure A2 shows the distribution of CAMCEs for all countries. To simplify the visualization, we subset party ID to strong Republicans, strong Democrats, and Independent/other.

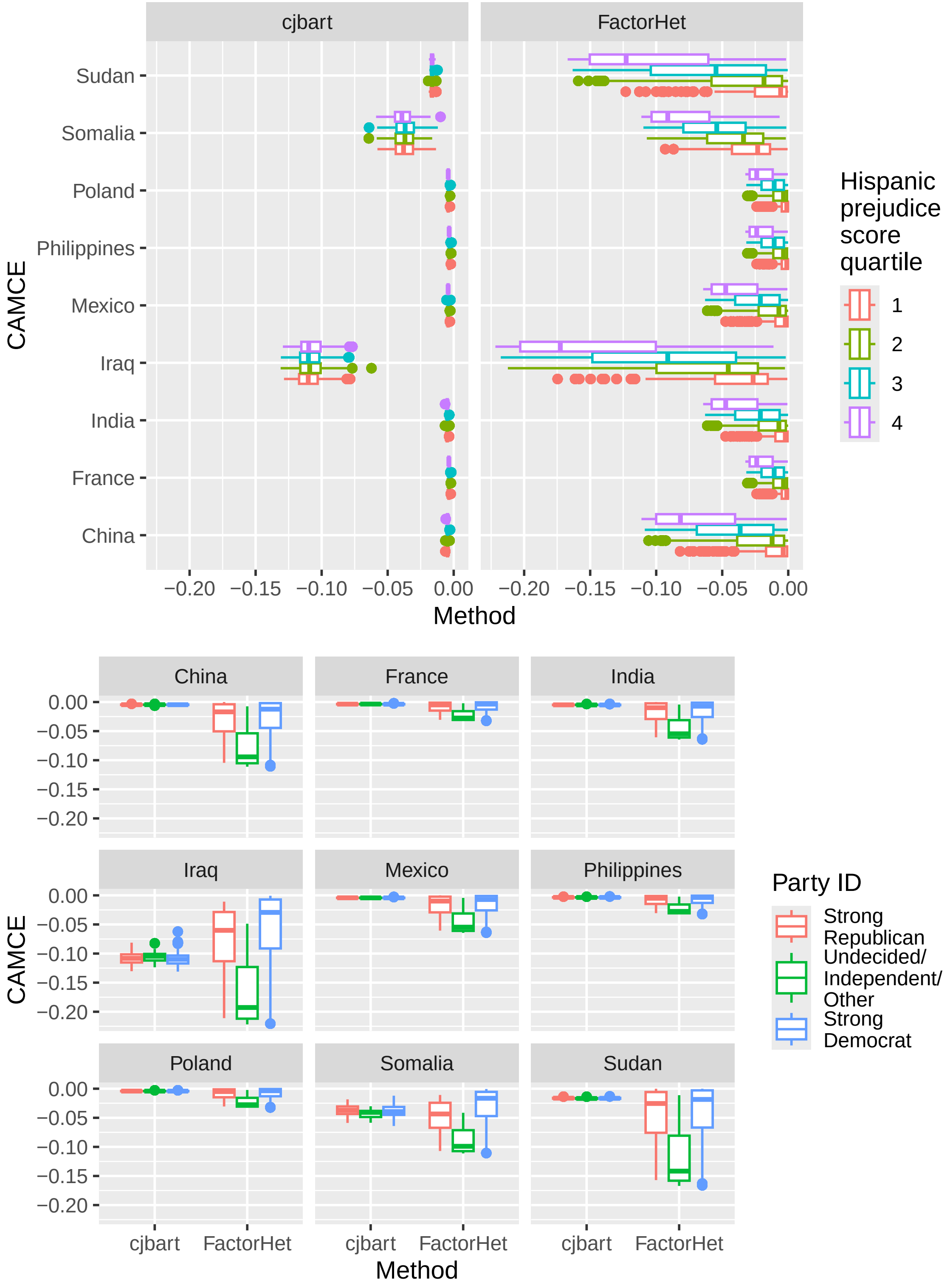


Figure A2: In both plots, the y-axis corresponds to the estimated values, either based on our method (right) or based on the method of Robinson and Duch (2024) (left), for the effect of a given country relative to the baseline of Germany. In the top figure, we color code based on quartile for the Hispanic prejudice score. In the bottom figure, we reduce the sample to those who identify as "Strong Republican", "Strong Democrat", or "Independent/Other" and color code by party ID.

C Proof of Proposition 1

To prove Proposition 1, we provide a more general result for one-parameter exponential family distributions, which include the specific Bernoulli result in the main text as a special case. We consider a random variable Y that is assumed to follow a single-parameter exponential family distribution with canonical parameter θ and $\mu = d\psi(\theta)/d\theta = \psi'(\theta)$. Since μ is monotone in θ, we index the density f using μ:

$$f_\mu(y) = c(y)\exp\left(y\theta - \psi(\theta)\right).$$

The maximum likelihood estimate of the mean $\hat{\mu}$ given N observations $\{y_i\}_{i=1}^N$ from Y is the sample average, $\frac{1}{N}\sum_{i=1}^N y_i = \hat{\mu}$, and the corresponding estimate of the canonical parameter is $\hat{\theta}$.

Proposition C.1 states that maximally heterogeneous groups in terms of Kullback-Leibler (KL) divergence of potential outcomes is equivalent to maximizing the log-likelihood over groups and their centroids for any choice of single parameter exponential family f.

Proposition C.1. *Assume a partition of N observations, indexed by $i \in \{1, \cdots, N\}$, into K groups whose memberships $Z_i \in \{1, \cdots, K\}$ are denoted by $\mathcal{Z}$. Define the estimated within-group average outcome under treatment $\boldsymbol{t}$ for group k and the estimated overall average outcome as $\hat{\zeta}_k(\boldsymbol{t}; \mathcal{Z}) = \sum_{i=1}^N I\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\}Y_i / \sum_{i=1}^N I\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\}$ and $\widehat{\widehat{Y}}(\boldsymbol{t}) = \sum_{i=1}^N I\{\boldsymbol{T}_i = \boldsymbol{t}\}Y_i / \sum_{i=1}^N I\{\boldsymbol{T}_i = \boldsymbol{t}\}$, respectively.*

Then, maximally heterogeneous groups in the terms of the Kullback-Leibler (KL) divergence of potential outcomes can be found by maximizing the log-likelihood function over the group membership and the centroids of groups, i.e.,

$$\underset{\mathcal{Z}}{\operatorname{argmax}}\left\{\sum_{k=1}^K\sum_{i=1}^N \mathbf{1}\{Z_i = k\}\mathrm{KL}\left(\hat{\zeta}_k(\boldsymbol{T}_i; \mathcal{Z}) \| \widehat{\widehat{Y}}(\boldsymbol{T}_i)\right)\right\} = \underset{\mathcal{Z}}{\operatorname{argmax}}\sum_{k=1}^K \sup_{\zeta_k}\sum_{i=1}^N \mathbf{1}\{Z_i = k\}\log f_{\zeta_k}(Y_i)$$

where $\mathrm{KL}(\mu_1, \mu_2)$ indicates the KL divergence between two single-parameter exponential family distributions with means μ_1 and μ_2 is defined as (Hastie, 1987):

$$\mathrm{KL}(\mu_1, \mu_2) = \mathbb{E}_{f_{\mu_1}(Y)}\left[\log f_{\mu_1}(Y) - \log f_{\mu_2}(Y)\right] = (\theta_1 - \theta_2)\,\mu_1 - \left[\psi(\theta_1) - \psi(\theta_2)\right].$$

To prove this proposition, we use Lemma C.1 which decomposes the total deviance of the observed data into the between and within components as in k-means (Everitt et al., 2011, ch. 5). This generalizes the standard Gaussian result (see Chi, Chi and Baraniuk, 2016).

Lemma C.1 (Deviance Decomposition for Exponential Family)**.** *Define the deviance for a single observation y as follows:*

$$D(y, \mu) = 2\left[\log f_y(y) - \log f_\mu(y)\right]$$

and the total deviance of the observed data when evaluated at the maximum likelihood estimate for each treatment $\boldsymbol{t}$—the sample average $\widehat{\widehat{Y}}(\boldsymbol{t})$ given randomization of $\boldsymbol{T}_i$—as follows

$$D_{\text{Total}} = \sum_{i=1}^N\sum_{\boldsymbol{t}\in\mathcal{T}} D\left(Y_i(\boldsymbol{t}), \widehat{\widehat{Y}}(\boldsymbol{t})\right)\mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}\} = \sum_{i=1}^N D\left(Y_i, \widehat{\widehat{Y}}(\boldsymbol{T}_i)\right),$$

where $\widehat{\widehat{Y}}(\boldsymbol{t}) = \sum_{i=1}^N \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}\}Y_i / \sum_{i=1}^N \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}\}$. Then, for any partition $\mathcal{Z}$ of the observations into K groups, D_{Total} can be decomposed as follows:

$$D_{\text{Total}} = \underbrace{\sum_{k=1}^K\sum_{i=1}^N \mathbf{1}\{Z_i = k\}\cdot 2\ \mathrm{KL}\left(\hat{\zeta}_k(\boldsymbol{T}_i; \mathcal{Z}), \widehat{\widehat{Y}}(\boldsymbol{T}_i)\right)}_{=D_{\text{Between}}} + \underbrace{\sum_{k=1}^K\sum_{i=1}^N \mathbf{1}\{Z_i = k\}D\left(Y_i(\boldsymbol{T}_i), \hat{\zeta}_k(\boldsymbol{T}_i; \mathcal{Z})\right)}_{=D_{\text{Within}}}$$

where $\hat{\zeta}_k(\boldsymbol{t};\mathcal{Z}) = \sum_{i=1}^N \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} Y_i / N_k(\boldsymbol{t};\mathcal{Z})$ *and* $N_k(\boldsymbol{t};\mathcal{Z}) = \sum_{i=1}^N \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\}$.

Proof. Define $\hat{\hat{\theta}}(\boldsymbol{t})$, $\hat{\theta}_k(\boldsymbol{t};\mathcal{Z})$ and $\theta_i(\boldsymbol{t})$ as the canonical parameters associated with, respectively, means $\widehat{\overline{Y}}(\boldsymbol{t})$, $\hat{\zeta}_k(\boldsymbol{t};\mathcal{Z})$, Y_i where $\theta_i(\boldsymbol{t})$ is used to define a saturated model for $Y_i(\boldsymbol{t})$. The result is proved below by re-arranging D_{Total}.

$$\begin{aligned}
D_{\text{Total}} =& \sum_{i=1}^N \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}\} \cdot 2 \left[\left(\theta_i(\boldsymbol{t}) - \hat{\hat{\theta}}(\boldsymbol{t})\right) Y_i(\boldsymbol{t}) - \left(\psi(\theta_i(\boldsymbol{t})) - \psi(\hat{\hat{\theta}}(\boldsymbol{t}))\right)\right] \\
=& \sum_{k=1}^K \sum_{i=1}^N \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} \cdot 2 \left(\theta_i(\boldsymbol{t}) - \hat{\hat{\theta}}(\boldsymbol{t}) + \hat{\theta}_k(\boldsymbol{t};\mathcal{Z}) - \hat{\theta}_k(\boldsymbol{t};\mathcal{Z})\right) Y_i(\boldsymbol{t}) \\
& - \sum_{k=1}^K \sum_{i=1}^N \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} \cdot 2 \left(\psi(\theta_i(\boldsymbol{t})) - \psi(\hat{\hat{\theta}}(\boldsymbol{t})) + \psi(\hat{\theta}_k(\boldsymbol{t};\mathcal{Z})) - \psi(\hat{\theta}_k(\boldsymbol{t};\mathcal{Z}))\right) \\
=& \sum_{k=1}^K \sum_{i=1}^N \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} \cdot 2\hat{\zeta}_k(\boldsymbol{t};\mathcal{Z}) \left[\left(\hat{\theta}_k(\boldsymbol{t};\mathcal{Z}) - \hat{\hat{\theta}}(\boldsymbol{t})\right) - \left(\psi(\hat{\theta}_k(\boldsymbol{t};\mathcal{Z})) - \psi(\hat{\hat{\theta}}(\boldsymbol{t}))\right)\right] \\
& + \sum_{k=1}^K \sum_{i=1}^N \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} D\left(Y_i(\boldsymbol{t}), \hat{\zeta}_k(\boldsymbol{t};\mathcal{Z})\right) \\
=& \sum_{k=1}^K \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \cdot 2 \text{ KL}\left(\hat{\zeta}_k(\boldsymbol{T}_i;\mathcal{Z}), \widehat{\overline{Y}}(\boldsymbol{T}_i)\right) + \sum_{k=1}^K \sum_{i=1}^N \mathbf{1}\{Z_i = k\} D\left(Y_i(\boldsymbol{T}_i), \hat{\zeta}_k(\boldsymbol{T}_i;\mathcal{Z})\right)
\end{aligned}$$

where the simplification of D_{Between} follows from noting that $\sum_{i=1}^N Y_i(\boldsymbol{t}) \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}, Z_i = k\} = N_k(\boldsymbol{t};\mathcal{Z})\hat{\zeta}_k(\boldsymbol{t}) = \sum_{i=1}^N \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}, Z_i = k\}\hat{\zeta}_k(\boldsymbol{t})$ by definition. □

Proof of Proposition C.1. Given Lemma C.1, maximizing D_{Between} over $\mathcal{Z}$ is equivalent to minimizing D_{Within} over $\mathcal{Z}$. Then, D_{Between} can be divided by two to obtain the left-hand side of the proposition. The right-hand side of the proposition is derived as follows. Minimizing the deviance is equivalent to maximizing the log-likelihood, i.e.,

$$\underset{\mathcal{Z}}{\operatorname{argmin}}\, D_{\text{Within}} = \underset{\mathcal{Z}}{\operatorname{argmax}} \left\{ \sum_{k=1}^K \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \log f_{\hat{\zeta}_k(\boldsymbol{T}_i;\mathcal{Z})}(Y_i(\boldsymbol{T}_i)) \right\}.$$

This can be written as a two-level optimization problem, noting that $Y_i = Y_i(\boldsymbol{T}_i)$ by the consistency assumption and that for fixed $\mathcal{Z}$, the maximum likelihood estimate of $\zeta_k(\boldsymbol{t})$ is $\hat{\zeta}_k(\boldsymbol{t};\mathcal{Z})$, i.e., the within-group observed average.

$$\underset{\mathcal{Z}}{\operatorname{argmin}}\, D_{\text{Within}} = \underset{\mathcal{Z}}{\operatorname{argmax}} \left\{ \sum_{k=1}^K \sup_{\zeta_k} \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \log f_{\zeta_k(\boldsymbol{T}_i)}(Y_i) \right\}$$

□

Finally, Proposition 1 in the main text uses the Bernoulli likelihood for f and is shown below.

$$\underset{\mathcal{Z}}{\operatorname{argmin}}\, D_{\text{Within}} = \underset{\mathcal{Z}}{\operatorname{argmax}} \left\{ \sum_{k=1}^K \sup_{\{\zeta_k\}} \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \left[Y_i \log \zeta_k(\boldsymbol{T}_i) + \{1 - Y_i\} \log\{1 - \zeta_k(\boldsymbol{T}_i)\}\right] \right\}$$

D Proof of Proposition 2

As before, we prove a more general result using the one-parameter exponential family distributions.

Proposition D.1 (Finding maximally heterogeneous groups with moderators)**.** *Suppose we extend the setting of Proposition C.1 and additionally model the conditional probability of each individual's group membership given categorical moderators $\{\pi_k(\boldsymbol{X}_i)\}_{k=1}^K$. Then, maximally heterogeneous groups in terms of KL divergence of potential outcomes with the entropy of group membership probabilities as a penalty term can be found by maximizing the log-likelihood function of the extended model,*

$$\begin{aligned} &\operatorname*{argmax}_{\mathscr{Z}} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \mathrm{KL}\left(\hat{\zeta}_k(\boldsymbol{T}_i; \mathscr{Z}) \| \widehat{\overline{Y}}(\boldsymbol{T}_i) \right) - \sum_{i=1}^{N} H(\{\hat{\pi}_k(\boldsymbol{X}_i; \mathscr{Z})\}_{k=1}^K) \right\} \\ = \; &\operatorname*{argmax}_{\mathscr{Z}} \sum_{k=1}^{K} \sup_{\zeta_k, \pi_k} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \left[\log f_{\zeta_k}(Y_i) + \log \pi_k(\boldsymbol{X}_i) \right] \end{aligned}$$

where $H(\{p_k\}_{k=1}^K) = -\sum_{k=1}^K p_k \log p_k$ (by convention, if $p_k = 0$, then $p_k \log p_k = 0$) is the entropy, and $\hat{\pi}_k(\boldsymbol{x}; \mathscr{Z}) = \sum_{i=1}^N \mathbf{1}\{Z_i = k, \boldsymbol{X}_i = \boldsymbol{x}\} / \sum_{i=1}^N \mathbf{1}\{\boldsymbol{X}_i = \boldsymbol{x}\}$ and $\hat{\zeta}_k(\boldsymbol{t}; \mathscr{Z})$ are the maximizers of the log-likelihood function of the right hand side of the above equation given $\mathscr{Z}$.

To prove this proposition, we use Lemma D.1.

Lemma D.1 (Entropy of Groups with Respect to Moderators)**.** *Define the set of observed categorical moderator values as $\mathcal{X}$ with $N(\boldsymbol{x}) = \sum_{i=1}^N \mathbf{1}\{\boldsymbol{X}_i = \boldsymbol{x}\}$ and $N_k(\boldsymbol{x}; \mathscr{Z}) = \sum_{i=1}^N \mathbf{1}\{Z_i = k, \boldsymbol{X}_i = \boldsymbol{x}\}$. Given $\mathscr{Z}$, the entropy of group membership probabilities given moderators, weighted by the frequency of the moderators, is defined as follows:*

$$H(\mathscr{Z}) = \sum_{\boldsymbol{x} \in \mathcal{X}} N(\boldsymbol{x}) H(\{\hat{\pi}_k(\boldsymbol{x}; \mathscr{Z})\}_{k=1}^K).$$

Then, $H(\mathscr{Z})$ can be expressed in the following two equivalent ways:

$$H(\mathscr{Z}) = \sum_{i=1}^{N} H(\{\hat{\pi}_k(\boldsymbol{X}_i; \mathscr{Z})\}_{k=1}^K) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{Z_i = k\} \log \hat{\pi}_k(\boldsymbol{X}_i; \mathscr{Z}).$$

Proof. The first expression follows by noting that the summation merely counts the number of times each $\boldsymbol{x}$ appears. The second expression is derived below by re-arranging $H(\mathscr{Z})$,

$$-H(\mathscr{Z}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{k=1}^{K} N(\boldsymbol{x}) \hat{\pi}_k(\boldsymbol{x}; \mathscr{Z}) \log \hat{\pi}_k(\boldsymbol{x}; \mathscr{Z}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{Z_i = k\} \log \hat{\pi}_k(\boldsymbol{X}_i; \mathscr{Z}),$$

where the last equality follows because $N(\boldsymbol{x}) \hat{\pi}_k(\boldsymbol{x}; \mathscr{Z}) = N_k(\boldsymbol{x}; \mathscr{Z})$ and it counts the number of times each combination of $(k, \boldsymbol{x})$ appears. □

Next, to prove Proposition D.1, we note that for any $\mathscr{Z}$, the KL divergence is equal to the log-likelihood evaluated at the maximum likelihood estimates plus a constant that does not depend on $\mathscr{Z}$ (see the definition of D_{Total}):

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \mathrm{KL}\left(\hat{\zeta}_k(\boldsymbol{T}_i; \mathscr{Z}) \| \widehat{\overline{Y}}(\boldsymbol{T}_i) \right) \; = \; \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \log_{\hat{\zeta}_k(\boldsymbol{T}_i; \mathscr{Z})}(Y_i(\boldsymbol{T}_i)) + \text{const.}$$

Adding the negative of group-moderator entropy $H(\mathcal{Z})$ to both sides and taking the maximum over $\mathcal{Z}$ gives the left-hand side of Proposition D.1. The equivalent right-hand side, using Lemma D.1 can be expressed as:

$$\underset{\mathcal{Z}}{\operatorname{argmax}} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \left[\log f_{\hat{\zeta}_k(\boldsymbol{T}_i;\mathcal{Z})}(Y_i(\boldsymbol{T}_i)) + \log \hat{\pi}_k(\boldsymbol{X}_i;\mathcal{Z}) \right] \right\}.$$

As in the proof of Proposition C.1, observing that $Y_i = Y_i(\boldsymbol{T}_i)$ by the consistency assumption and writing the above equation as two-level optimization problem over ζ_k and π_k establishes Proposition D.1. This follows by noting that for a fixed $\mathcal{Z}$, the maximum likelihood estimate of $\pi_k(\boldsymbol{x})$ is $\hat{\pi}_k(\boldsymbol{x})$ and the estimate of $\zeta_k(\boldsymbol{t})$ as $\hat{\zeta}_k(\boldsymbol{t})$ is unchanged as the optimization problem is separable. In addition, using the Bernoulli likelihood for f gives Proposition 2 in the main text. □

E Inclusion of Higher Order Interactions

Here we illustrate how the model and regularization penalties in Section 3.3 can be extended to include higher order interactions in a straightforward manner. We show below the model including all higher order interactions, and including only a subset is direct.

Let $\mathcal{J} = \{1, \dots, J\}$ be the set of J factors and let $\mathcal{T}$ be the set of all possible assignments on the $\mathcal{J}$ factors. Then our model for $\psi_k(\boldsymbol{T}_i)$ with all interactions among factors is

$$\begin{aligned} \psi_k(\boldsymbol{T}_i) &= \mu + \sum_{j=1}^{J} \sum_{l=0}^{L_j-1} \mathbf{1}\{T_{ij} = l\}\beta_{kl}^{j} + \sum_{j=1}^{J-1} \sum_{j'>j} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_{j'}-1} \mathbf{1}\{T_{ij} = l, T_{ij'} = l'\}\beta_{kll'}^{jj'} \\ &\quad + \cdots + \sum_{\boldsymbol{t}\in\mathcal{T}} \mathbf{1}\{\boldsymbol{T}_i = \boldsymbol{t}\}\beta_{k\boldsymbol{t}}^{12\cdots K} \\ &= \mu + \tilde{\boldsymbol{T}}_i^{\top}\boldsymbol{\beta}_k. \end{aligned}$$

In the above formulation, $\beta_{k\boldsymbol{t}}^{12\cdots K}$ is the K-way interaction coefficient in cluster k for assignment $\boldsymbol{t}$.

Let $\mathcal{T}_{-j}$ be the set of all possible assignments on the $\mathcal{J}$ factors except for factor j. With some slight notation abuse by letting $\beta_{kl\boldsymbol{t}_{-j}}^{12\cdots K}$ be the K-way interaction coefficient in cluster k for assignment l for j and $\boldsymbol{t}_j$ for the other $J-1$ factors, the ANOVA-type sum-to-zero constraints extend as follows:

$$\sum_{l=0}^{L_j-1} \beta_{kl}^{j} = 0, \sum_{l=0}^{L_j-1} \beta_{kll'}^{jj'} = \sum_{l'=0}^{L_{j'}-1} \beta_{kll'}^{jj'} = 0, \dots, \sum_{l=0}^{L_j-1} \beta_{kl\boldsymbol{t}_{-j}}^{12\cdots K} = 0 \tag{A1}$$

for $j, j' = 1, 2, \dots, J$ with $j' > j$ and for all $\boldsymbol{t}_{-j} \in \mathcal{T}_{-j}$. We write them compactly as,

$$\boldsymbol{C}^{\top}\boldsymbol{\beta}_k \ = \ \mathbf{0}, \tag{A2}$$

where each row of $\boldsymbol{C}^{\top}\boldsymbol{\beta}_k$ corresponds to one of the constraints given in Equation (A1).

For the structured sparsity, we have penalties of the form

$$\sum_{j=1}^{J} \sum_{l_j=1}^{L_j} \sum_{l'_j>l_j}^{L_j} \sqrt{(\beta_{l_j}^{j} - \beta_{l'_j}^{j})^2 + \sum_{j'\neq j} \sum_{l_{j'}=1}^{L_{j'}} (\beta_{l_j l_{j'}}^{jj'} - \beta_{l'_j l_{j'}}^{jj'})^2 + \cdots + \sum_{\boldsymbol{t}_{-j}\in\mathcal{T}_{-j}} (\beta_{l_j\boldsymbol{t}_{-j}}^{12\cdots K} - \beta_{l'_j\boldsymbol{t}_{-j}}^{12\cdots K})^2}$$

This will have $\sum_{j=1}^{J} L_j(L_j-1)/2$ terms, $L_j(L_j-1)/2$ terms for the jth factor.

For illustration, consider a simple example with one group and three factors—factor one has three levels, factor two has two levels, and factor three has two levels. In this case, our penalty contains 5 terms,

$$\sum_{l_1=1}^{L_1}\sum_{l_1'>l_1}^{L_1}\sqrt{(\beta^{1}_{l_1}-\beta^{1}_{l_1'})^2+\sum_{l_2=1}^{L_2}(\beta^{12}_{l_1l_2}-\beta^{12}_{l_1'l_2})^2+\sum_{l_3=1}^{L_3}(\beta^{13}_{l_1l_3}-\beta^{13}_{l_1'l_3})^2+\sum_{l_2=1}^{L_2}\sum_{l_3=1}^{L_3}(\beta^{123}_{l_1l_2l_3}-\beta^{123}_{l_1'l_2l_3})^2}$$
$$+\sum_{l_2=1}^{L_2}\sum_{l_2'>l_2}^{L_2}\sqrt{(\beta^{1}_{l_2}-\beta^{2}_{l_2'})^2+\sum_{l_2=1}^{L_2}(\beta^{12}_{l_1l_2}-\beta^{12}_{l_1l_2'})^2+\sum_{l_3=1}^{L_3}(\beta^{23}_{l_2l_3}-\beta^{23}_{l_2'l_3})^2+\sum_{l_1=1}^{L_1}\sum_{l_3=1}^{L_3}(\beta^{123}_{l_1l_2l_3}-\beta^{123}_{l_1l_2'l_3})^2}$$

$$+\sum_{l_3=1}^{L_3}\sum_{l_3'>l_3}^{L_3}\sqrt{(\beta^{1}_{l_3}-\beta^{3}_{l_3'})^2+\sum_{l_3=1}^{L_2}(\beta^{13}_{l_1l_3}-\beta^{12}_{l_1l_3'})^2+\sum_{l_1=1}^{L_3}(\beta^{13}_{l_1l_3}-\beta^{23}_{l_1l_3'})^2+\sum_{l_1=1}^{L_1}\sum_{l_2=1}^{L_2}(\beta^{123}_{l_1l_2l_3}-\beta^{123}_{l_1l_2l_3'})^2}$$

The first three terms encourages the pairwise fusion of the levels of factor one whereas the fourth encourages the fusion of the two levels of factor two and the fifth encourages the fusion of the two levels of factor three.

Using the sum of Euclidean norms of quadratic forms, we can write the penalty as

$$||\boldsymbol{\beta}^\top \boldsymbol{F}_1\boldsymbol{\beta}||_2+\ ||\boldsymbol{\beta}^\top \boldsymbol{F}_2\boldsymbol{\beta}||_2+\ ||\boldsymbol{\beta}^\top \boldsymbol{F}_3\boldsymbol{\beta}||_2+\ ||\boldsymbol{\beta}^\top \boldsymbol{F}_4\boldsymbol{\beta}||_2+\ ||\boldsymbol{\beta}^\top \boldsymbol{F}_5\boldsymbol{\beta}||_2,$$

where $\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3$ are appropriate positive semi-definite matrices to encourage the fusion of the pairs of levels in factor one, $\boldsymbol{F}_4$ encourages the fusion of the two levels in factor two, $\boldsymbol{F}_5$ encourages the fusion of the two levels in factor three, and $\boldsymbol{\beta}=[\beta^1_0\ \beta^1_1\ \beta^1_2\ \beta^2_0\ \beta^2_1\ \beta^{12}_{00}\ \beta^{12}_{10}\ \beta^{12}_{20}\ \beta^{12}_{01}\ \beta^{12}_{11}\ \beta^{12}_{21}\cdots\beta^{123}_{211}]^\top$.

More generally, for a fully interacted model we will have $\sum_{j=1}^{J} L_j(L_j-1)/2=G$ terms,

$$\sum_{g=1}^{G}||\boldsymbol{\beta}^\top \boldsymbol{F}_g\boldsymbol{\beta}||_2.$$

F Propriety of the Structured Sparse Prior

The proof of propriety for the structured sparse prior used in our paper is an application of Theorem 1 established in Goplerud (2021) and is reproduced here.

Theorem F.1 (Goplerud (2021))**.** *Consider the following structured sparse prior on $\boldsymbol{\beta}\in\mathbb{R}^p$ with regularization strength $\lambda>0$ penalizes K linear constraints $\boldsymbol{d}_k$ and L quadratic constraints $\boldsymbol{F}_\ell$ on the parameters where $\boldsymbol{F}_\ell$ is symmetric and positive semi-definite. The kernel of the prior is shown below.*

$$p(\boldsymbol{\beta})\propto\exp\left(-\lambda\left[\sum_{k=1}^{K}|\boldsymbol{d}_k^\top\boldsymbol{\beta}|+\sum_{\ell=1}^{L}\sqrt{\boldsymbol{\beta}^\top\boldsymbol{F}_\ell\boldsymbol{\beta}}\right]\right)$$

Further define $\boldsymbol{D}^\top=[d_1,\cdots,d_K]^\top$ and $\bar{\boldsymbol{D}}^\top=[\boldsymbol{D}^\top,\boldsymbol{F}_1,\cdots,\boldsymbol{F}_L]$. Then, for $\lambda>0$, the prior above is proper if and only if $\bar{\boldsymbol{D}}$ is full column rank.

In our specific case, we note that $K = 0$, $L = G$, and $\lambda = \lambda\bar{\pi}_k^{\gamma}$. Prior propriety of $p(\boldsymbol{\beta}_k \mid \{\boldsymbol{\phi}_k\}_{k=2}^K, \lambda)$, therefore, can be determined by empirically investigating whether $\bar{\boldsymbol{D}}$, i.e. the vertically stacked $\boldsymbol{F}_\ell$, is full column rank.

It is also possible to analytically show the propriety of the prior distribution in all cases considered in this paper. We focus on the case of $K = 1$ and arbitrary $\lambda > 0$ as the result follows automatically for the case in our paper.

Result F.1. *Assume a structured sparse prior for a factorial or conjoint design with J factors each with L_j levels where all pairwise interactions are included and levels of each factor are encouraged to be fused together (i.e. the model in the main text). The kernel of the prior is shown below where $\boldsymbol{F}_g$ are as defined in the main text.*

$$k(\boldsymbol{\beta}) = \exp\left(-\lambda \sum_{g=1}^{G} \sqrt{\boldsymbol{\beta}^\top \boldsymbol{F}_g \boldsymbol{\beta}}\right)$$

Assume that the linear sum-to-zero constraints $\boldsymbol{C}^\top\boldsymbol{\beta} = \boldsymbol{0}$ hold. Then, the structured sparse prior on the unconstrained $\tilde{\boldsymbol{\beta}}$ such that $\tilde{\boldsymbol{\beta}} \in \mathcal{N}(\boldsymbol{C}^\top)$ is proper. Or, equivalently, the following result holds.

$$\int_{\boldsymbol{\beta}:\boldsymbol{C}^\top\boldsymbol{\beta}=\boldsymbol{0}} k(\boldsymbol{\beta}) d\boldsymbol{\beta} < \infty.$$

Proof. Let $\mathcal{B}_{\boldsymbol{C}^\top}$ represent a basis for the linear constraints $\boldsymbol{C}^\top$. The integral for evaluating propriety can be written as,

$$\int_{\tilde{\boldsymbol{\beta}}} \tilde{k}(\tilde{\boldsymbol{\beta}}) d\tilde{\boldsymbol{\beta}} \quad \text{where} \quad \tilde{k}(\tilde{\boldsymbol{\beta}}) = \exp\left(-\lambda \sum_{g=1}^{G} \sqrt{\tilde{\boldsymbol{\beta}}^\top \mathcal{B}_{\boldsymbol{C}^\top}^\top \boldsymbol{F}_g \mathcal{B}_{\boldsymbol{C}^\top} \tilde{\boldsymbol{\beta}}}\right).$$

Note that $\boldsymbol{F}_g$ can be expressed as a sum of N_g outer products of $|\boldsymbol{\beta}|$-length vectors of the form $\boldsymbol{l}_i \in \{-1, 0, 1\}$ where -1 and 1 correspond to the two terms that are fused together and all other elements are 0, i.e., $\boldsymbol{F}_g = \sum_{g'=1}^{N_g} \boldsymbol{l}_{g'} \boldsymbol{l}_{g'}^\top$. Thus, one can define a matrix $\boldsymbol{Q}_g^\top = \left[\boldsymbol{l}_1, \cdots, \boldsymbol{l}_{N_g}\right]$ such that $\boldsymbol{Q}_g^\top \boldsymbol{Q}_g = \boldsymbol{F}_g$, which allows us to rewrite $\tilde{k}(\tilde{\boldsymbol{\beta}})$ as:

$$\tilde{k}(\tilde{\boldsymbol{\beta}}) = \exp\left(-\lambda \sum_{g=1}^{G} \sqrt{\tilde{\boldsymbol{\beta}}^\top \left[\mathcal{B}_{\boldsymbol{C}^\top}\right]^\top \boldsymbol{Q}_g^\top \boldsymbol{Q}_g \mathcal{B}_{\boldsymbol{C}^\top} \tilde{\boldsymbol{\beta}}}\right).$$

By applying Theorem F.1 and noting that the null spaces of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}$ are identical, the integral of $\tilde{k}(\tilde{\boldsymbol{\beta}})$ is finite if and only if $\boldsymbol{Q}\mathcal{B}_{\boldsymbol{C}^\top}$ is full column rank, where $\boldsymbol{Q}^\top = [\boldsymbol{Q}_1^\top, \cdots, \boldsymbol{Q}_G^\top]$. We demonstrate this fact in two steps. First, there exists a permutation matrix $\boldsymbol{P}_Q$ such that $\boldsymbol{P}_Q\boldsymbol{Q}$ has a block diagonal structure with $J + 1$ diagonal blocks. The first J blocks corresponding to the main terms for each factor j and the last block corresponds to all interaction terms. The null space of each block is spanned by the vector $\boldsymbol{1}$ as the corresponding block of $\boldsymbol{P}_Q\boldsymbol{Q}$ is a (transposed) orientated incidence matrix of a fully connected graph. Thus, the null space of $\boldsymbol{P}_Q\boldsymbol{Q}$, and hence $\boldsymbol{Q}$, is spanned by the $J + 1$ columns of a block diagonal matrix with $\boldsymbol{1}$ on each block. Second, consider the linear constraints $\boldsymbol{C}^\top\boldsymbol{\beta} = \boldsymbol{0}$. The only vector to satisfy this constraint and lie in the null space of $\boldsymbol{Q}$ must be $\boldsymbol{0}$ as, for each block, the only vector proportional to $\boldsymbol{1}$ and satisfying the corresponding sum-to-zero constraints must be $\boldsymbol{0}$. Thus, $\boldsymbol{Q}\mathcal{B}_{\boldsymbol{C}^\top}$ is full column rank and the prior is proper. □

G Derivations for the Basic Model

This section derives a number of results for the basic model. It first restates the main results concerning the elimination of the linear constraints $\boldsymbol{C}^\top \boldsymbol{\beta}_k = \boldsymbol{0}$. Then, it derives the Expectation Maximization algorithm, our measure of degrees of freedom, and some additional computational improvements used to accelerate estimation. In the following, we use $\tilde{\boldsymbol{T}}_i$ to denote the corresponding vector of indicators for whether certain treatments or interactions are present (i.e. stacking all $\mathbf{1}\{T_{ij} = l\}$, etc. from Equation A17). In addition, we use ψ_{ik} to indicate the linear predictor for observation i and group k.

G.1 Removing the Linear Constraints

The inference problem in the main text is presented as an optimization problem subject to linear constraints on the coefficients $\boldsymbol{\beta}_k$. Inference is noticeably easier if these are eliminated via a transformation of the problem to a lower-dimensional one by noting that $\boldsymbol{\beta}_k$ must lie in the null space of the constraint matrix $\boldsymbol{C}^\top$ (see, e.g., Lawson and Hanson 1974, ch. 20). Define $\tilde{\boldsymbol{\beta}}_k = \left(\mathcal{B}_{\boldsymbol{C}^\top}^\top \mathcal{B}_{\boldsymbol{C}^\top}\right)^{-1} \mathcal{B}_{\boldsymbol{C}^\top}^\top \boldsymbol{\beta}_k$ where $\mathcal{B}_{\boldsymbol{C}^\top}$ is a basis for the null space of $\boldsymbol{C}^\top$. The problem can thus be solved in terms of the unconstrained $\tilde{\boldsymbol{\beta}}_k \in \mathbb{R}^{p-\text{rank}(\boldsymbol{C}^\top)}$ given appropriate adjustment of the treatment design vectors, $\tilde{\tilde{\boldsymbol{T}}}_i = \mathcal{B}_{\boldsymbol{C}^\top} \tilde{\boldsymbol{T}}_i$, penalty matrices, $\tilde{\boldsymbol{F}}_g = \mathcal{B}_{\boldsymbol{C}^\top}^\top \boldsymbol{F}_g \mathcal{B}_{\boldsymbol{C}^\top}$, and linear predictor, $\psi_{i,k} = \left[\tilde{\tilde{\boldsymbol{T}}}_i\right]^\top \tilde{\boldsymbol{\beta}}_{Z_i} + \mu$. Once the algorithm convergences, the constrained parameters can be recovered by noting $\boldsymbol{\beta}_k = \mathcal{B}_{\boldsymbol{C}^\top} \tilde{\boldsymbol{\beta}}_k$.

Given the similarity of the unconstrained and constrained problems and for notational simplicity, we present all results herein dropping the second "tilde" notation on $\tilde{\boldsymbol{T}}_i$ and the "tilde" on $\boldsymbol{\beta}_k$ and note that, once estimated, $\tilde{\boldsymbol{\beta}}_k$ is projected back into the original space for the reported coefficients, average marginal component effects, etc. The results of Appendix J on approximating $\tilde{\boldsymbol{\beta}}_k$ as multivariate Gaussian imply that $\boldsymbol{\beta}_k$ will have a (singular) multivariate Gaussian distribution.

G.2 Expectation Maximization Algorithm

This section considers inference after removing the linear constraints as discussed in the prior subsection. Algorithm A1 summarizes our approach to maximizing Equation (9). Each iteration of our AECM algorithm involves two cycles where the data augmentation scheme enables iterative updating of the treatment effect parameters $\boldsymbol{\beta}$ and moderators $\boldsymbol{\phi}$. $\boldsymbol{\theta}$ collects both sets of parameters.

G.2.1 Updating Treatment Effect Parameters

We begin with the cycle of the AECM algorithm for updating $\{\boldsymbol{\beta}_k\}_{k=1}^K$ and μ given $\{\boldsymbol{\phi}_{k=2}^K\}$. To update $\boldsymbol{\beta}, \mu$, our data augmentation strategy requires three types of missing data. First, we use the standard group memberships of each unit i for inference in finite mixtures, i.e., $Z_i \in \{1, \cdots, K\}$. We also include two other types of data augmentation that result in a closed-form update. We use Polya-Gamma augmentation (ω_i; Polson, Scott and Windle 2013) for the logistic likelihood and data augmentation on the sparsity-inducing penalty (τ_{gk}^2; see, e.g., Figueiredo 2003; Polson and Scott 2011; Ratkovic and Tingley 2017; Goplerud 2021) yielding

Algorithm A1 AECM Algorithm for Estimating $\boldsymbol{\theta}$

```
Set Hyper-Parameters: $K$ (groups), $\lambda$, $\sigma_\phi^2$, $\gamma$ (prior strength), $\epsilon_1, \epsilon_2$ (convergence criteria), $T$ (number of iterations)
Initialize Parameters: $\boldsymbol{\theta}^{(0)}$, i.e. $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\phi}^{(0)}$; Appendix G.5 provides details.
For iteration $t \in \{0, \cdots, T-1\}$
    Cycle 1: Update $\boldsymbol{\beta}$
    1a. $E$-Step: Find the conditional distributions of $\{Z_i, \omega_i\}_{i=1}^N$ and $\{\{\tau_{gk}^2\}_{g=1}^G\}_{k=1}^K$ given $\{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}$ and $\boldsymbol{\theta}^{(t)}$ (Eq. (A3)). Derive $Q_\beta(\boldsymbol{\beta}, \boldsymbol{\theta}^{(t)})$ (Eq. (A4)).
    1b. $M$-Step: Set $\boldsymbol{\beta}^{(t+1)}$ such that $Q_\beta(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq Q_\beta(\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$
    Cycle 2: Update $\boldsymbol{\phi}$
    2a. $E$-Step: Find $p(Z_i = k \mid Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)})$. Derive $Q_\phi(\boldsymbol{\phi}, \{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)}\})$ (Eq. (A7)).
    2b. $M$-Step: Set $\boldsymbol{\phi}^{(t+1)}$ such that
     $Q_\phi(\boldsymbol{\phi}^{(t+1)}, \{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)}\}) \geq Q_\phi(\boldsymbol{\phi}^{(t)}, \{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)}\})$
    Check Convergence
    3. Stop if $\log p\left(\boldsymbol{\theta}^{(t+1)} | \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}_{i=1}^N\right) - \log p\left(\boldsymbol{\theta}^{(t)} | \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}_{i=1}^N\right) < \epsilon_1$ (Eq. (9)) or
    $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||_\infty < \epsilon_2$.
```

$$
\begin{aligned}
p(Y_i, \omega_i \mid Z_i, \boldsymbol{X}_i, \boldsymbol{T}_i) &\propto \frac{1}{2} \exp\left\{ \left(Y_i - \frac{1}{2}\right) \psi_{Z_i}(\boldsymbol{T}_i) - \frac{\omega_i}{2} \left[\psi_{Z_i}(\boldsymbol{T}_i)\right]^2 \right\} f_{PG}(\omega_i \mid 1, 0), && \text{(A3a)} \\
p(\boldsymbol{\beta}_k, \{\tau_{gk}^2\}_{g=1}^G \mid \lambda, \{\boldsymbol{\phi}_k\}) &\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_k^\top \left(\sum_{g=1}^G \frac{\boldsymbol{F}_g}{\tau_{gk}^2} \right) \boldsymbol{\beta}_k \right\} \prod_{g=1}^G \tau_{gk}^{-1} \exp\left\{ -\frac{(\lambda \bar{\pi}_k)^2}{2} \cdot \tau_{gk}^2 \right\}, && \text{(A3b)}
\end{aligned}
$$

where $f_{PG}(\cdot \mid b, c)$ represents the Polya-Gamma distribution with parameters (b, c) and $Z_i \sim \text{Multinomial}(1, \boldsymbol{\pi}_i)$ with the kth element of $\boldsymbol{\pi}$ equal to $\pi_k(\boldsymbol{X}_i)$. Note that $\boldsymbol{\beta}$ only enters Equation (A3) via a quadratic form. The first cycle of the AECM algorithm involves, therefore, maximizing the following function with respect to $\boldsymbol{\beta}$ given $\boldsymbol{\theta}^{(t)}$.

$$
\begin{aligned}
Q_\beta\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(t)}\right) &= \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}[\mathbf{1}\{Z_i = k\}] \left\{ \left(Y_i - \frac{1}{2}\right) \psi_k(\boldsymbol{T}_i) - \mathbb{E}[\omega_i \mid Z_i = k] \frac{[\psi_k(\boldsymbol{T}_i)]^2}{2} \right\} \\
&\quad + \sum_{k=1}^K -\frac{1}{2} \boldsymbol{\beta}_k^\top \left[\sum_{g=1}^K \boldsymbol{F}_g \cdot \mathbb{E}[1/\tau_{gk}^2] \right] \boldsymbol{\beta}_k + \text{const.} && \text{(A4)}
\end{aligned}
$$

where all expectations are taken over the conditional distribution of the missing data given the current parameter estimates. We note that the E-Step involves computing $p(\{\omega_i, Z_i\}, \{1/\tau_{gk}^2\} \mid \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}, \boldsymbol{\theta}^{(t)})$ which factorizes into, respectively, a collection of Polya-Gamma (PG), categorical, and Inverse-Gaussian random variables. Their conditional distributions are shown below,

$$p(\tau_{gk}^{-2} \mid \boldsymbol{\theta}) \sim \text{InverseGaussian}\left(\frac{\lambda}{\sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k}}, \quad \lambda^2\right), \tag{A5a}$$

$$p(Z_i = k \mid Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{\theta}) \propto p_{ik}^{Y_i}(1-p_{ik})^{1-Y_i}\pi_{ik}; \quad p_{ik} = \frac{\exp(\psi_{ik})}{1+\exp(\psi_{ik})}, \tag{A5b}$$

$$p(\omega_i \mid Z_i = k, \boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{\theta}) \sim \text{PG}\left(1, \psi_{ik}\right), \tag{A5c}$$

as well as the relevant expectations needed in $Q_\beta(\boldsymbol{\beta}, \boldsymbol{\theta})$,

$$\mathbb{E}\left[\tau_{gk}^{-2}\right] = \frac{\lambda}{\sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k}}, \tag{A6a}$$

$$\mathbb{E}[z_{ik}] = \mathbb{E}\left[\mathbf{1}\{Z_i = k\}\right] = \frac{p_{ik}^{Y_i}(1-p_{ik})^{1-Y_i}\pi_{ik}}{\sum_{\ell=1}^K p_{i\ell}^{Y_i}(1-p_{i\ell})^{1-Y_i}\pi_{i\ell}}, \tag{A6b}$$

$$\mathbb{E}[\omega_i \mid Z_i = k] = \frac{1}{2\psi_{ik}}\tanh\left(\frac{\psi_{ik}}{2}\right). \tag{A6c}$$

Note that as $\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k$ approaches zero, $\mathbb{E}[\tau_{gk}^{-2}]$ approaches infinity. To prevent numerical instability, we rely on the strategy in Goplerud (2021) (inspired by Polson and Scott 2011) where once it is sufficiently small, e.g. below 10^{-4}, and thus the restriction is almost binding, we ensure that restriction holds in all future iterations. We do so by adding a quadratic constraint $\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k = 0$. This implies that $\boldsymbol{\beta}_k$ lies in the null space of $\boldsymbol{F}_g$ and thus with an additional transformation, it can be removed and the problem be solved in an unconstrained space with a modified design.

To compute the update for $\boldsymbol{\beta}$, define $\check{\boldsymbol{\beta}}^\top = [\mu, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K]^\top$. We can create a corresponding design matrix $\check{\boldsymbol{T}} = [\mathbf{1}_N, \boldsymbol{I}_K \otimes \boldsymbol{T}]$ where $\tilde{\boldsymbol{T}}^\top = [\tilde{\boldsymbol{T}}_1, \cdots, \tilde{\boldsymbol{T}}_N]$ and diagonal weight matrix $\check{\boldsymbol{\Omega}} = \text{diag}\left(\{\{\mathbb{E}[z_{ik}]\mathbb{E}[\omega_i \mid Z_i = k]\}_{i=1}^N\}_{k=1}^K\right)$. Further, we can create the combined ridge penalty $\boldsymbol{\mathcal{R}} = \text{blockdiag}\left(\{0, \{\boldsymbol{R}_k\}_{k=1}^K\}\right)$ where $\boldsymbol{R}_k = \sum_g \boldsymbol{F}_g \mathbb{E}[\tau_{gk}^{-2}]$ and augmented outcome $\check{\boldsymbol{Y}} = \{\{\mathbb{E}[z_{ik}](Y_i - 1/2)\}_{i=1}^N\}_{k=1}^K$. The Q_β function is thus proportional to the following ridge regression problem and yields the update for the M-Step,

$$Q_\beta\left(\boldsymbol{\beta}; \boldsymbol{\theta}^{(t)}\right) = \check{\boldsymbol{Y}}^\top\left(\check{\boldsymbol{T}}\check{\boldsymbol{\beta}}\right) - \frac{1}{2}\check{\boldsymbol{\beta}}^\top \check{\boldsymbol{T}}^\top \check{\boldsymbol{\Omega}} \check{\boldsymbol{T}} \check{\boldsymbol{\beta}} - \frac{1}{2}\check{\boldsymbol{\beta}}^\top \boldsymbol{\mathcal{R}} \check{\boldsymbol{\beta}} + \text{const.},$$
$$\check{\boldsymbol{\beta}}^{(t+1)} = \left(\check{\boldsymbol{T}}^\top \check{\boldsymbol{\Omega}} \check{\boldsymbol{T}} + \boldsymbol{\mathcal{R}}\right)^{-1} \check{\boldsymbol{T}}^\top \check{\boldsymbol{Y}}.$$

One could reply on a generalized EM algorithm where Q_β is improved versus maximized for computational reasons, e.g. by using a conjugate gradient solver initialized at $\check{\boldsymbol{\beta}}^{(t)}$.

G.2.2 Updating Moderator Parameters

To update the moderator parameters $\boldsymbol{\phi}$, we use the second cycle of the AECM algorithm where only the Z_i are treated as missing data. The E-step involves recomputing the group membership probabilities, i.e., $p(Z_i \mid Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)})$, given the updates in the first cycle. The implied

Q-function is shown below,

$$\begin{aligned} Q_\phi(\boldsymbol{\phi}, \{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}^{(t)}\}) \;=\; & \sum_{k=1}^{K} \left[\sum_{i=1}^{N} \mathbb{E}[\mathbf{1}\{Z_i = k\}] \log \pi_k(\boldsymbol{X}_i) \right] \\ & + \sum_{k=1}^{K} \left[m\gamma \log \bar{\pi}_k - \lambda \bar{\pi}_k^{\gamma} \sum_{g=1}^{G} \sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k} \right] + \log p(\{\boldsymbol{\phi}_k\}_{k=2}^{K}), \end{aligned} \tag{A7}$$

where $\pi_k(\boldsymbol{X}_i)$ and $\bar{\pi}_k = \sum_{i=1}^{N} \pi_k(\boldsymbol{X}_i)/N$ are functions of $\boldsymbol{\phi}_k$. Note that if $\gamma = 0$, this simplifies to a multinomial logistic regression with $\{\mathbb{E}[\mathbf{1}\{Z_i = k\}]\}_{k=1}^{K}$ as the outcome. We perform the M-Step using a standard optimizer (e.g., L-BFGS) to optimize Q_ϕ and thus obtain $\boldsymbol{\phi}^{(t+1)}$.

G.3 Classification Maximum Likelihood

If classification maximum likelihood approach is desired, despite statistical concerns about this procedure's asymptotic bias (e.g., Bryant and Williamson 1978), it can be easily implemented by adapting the preceeding EM algorithm. Celeux and Govaert (1992) propose the "classification EM" algorithm in the spirit of how k-means classification is commonly implemented.

The adjustment proceeds as follows (Celeux and Govaert, 1992, p. 319): after conducting an E-step and obtaining $\tilde{\pi}_k(\boldsymbol{X}_i, Y_i, \boldsymbol{T}_i; \boldsymbol{\theta}) = \tilde{\pi}_{ik} = p(Z_i = k \mid Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{\theta})$ for use in evaluating Q_β and Q_ϕ, perform a classification or "hard assignment". That is, find $k_i^* = \text{argmax}_k\, \tilde{\pi}_{ik}$, i.e., the most probable cluster for observation i given its observed data and $\boldsymbol{\theta}$. In the subsequent M-step, use a modified weight $c_{ik} = 1$ if $k = k_i^*$ and otherwise $c_{ik} = 0$ in lieu of $\tilde{\pi}_{ik}$.

G.4 Degrees of Freedom

Our procedure for estimating $\check{\boldsymbol{\beta}}^{(t)}$ appears similar to the results in Oelker and Tutz (2017) where complex regularization and non-linear models can be recast as a (weighted) ridge regression. Using that logic, we take the trace of the "hat matrix" implied by our algorithm at stationarity to estimate our degrees of freedom. We also adjust upwards the degrees of freedom by the number of moderator coefficients (e.g., Khalili 2010; Chamroukhi and Huynh 2019).

Equation (A8) shows our procedure where $\boldsymbol{\mathcal{R}}$ and $\check{\boldsymbol{\Omega}}$ contain expectations calculated at convergence. p_x denotes the number of moderators, i.e. the dimensionality of $\boldsymbol{\phi}_k$. Before evaluating Equation (A8), for any two factor levels that are sufficiently close (e.g., $\sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k} < 10^{-4}$), we assume they are fused together and consider it as an additional linear constraint on the parameter vector $\boldsymbol{\beta}_k$.

$$\text{df} = \text{tr}\left[\left(\check{\boldsymbol{T}}^\top \check{\boldsymbol{\Omega}} \check{\boldsymbol{T}} + \boldsymbol{\mathcal{R}} \right)^{-1} \check{\boldsymbol{T}}^\top \check{\boldsymbol{\Omega}} \check{\boldsymbol{T}} \right] + p_x\,(K-1) \tag{A8}$$

From this, we can calculate a BIC criterion. We seek to find the regularization parameter λ that minimizes this criterion. To avoid the problems of a naive grid-search, we use Bayesian model-based optimization that attempts to minimize the number of function evaluations while searching for the value of λ that minimizes the BIC (`mlrMBO`; Bischl et al. 2018). We find that with around fifteen model evaluations, the optimizer can usually find a near optimal value of λ.

G.5 Computational Improvements

While the algorithm above provides a valid way to locate a posterior mode, our estimation problem is complex and high-dimensional. Furthermore, given the complex posterior implied by mixture of experts models, we derived a number of computational strategies to improve convergence. We use the SQUAREM algorithm (Varadhan and Roland 2008). Our software provides the option to use a generalized EM algorithm to update $\boldsymbol{\beta}$ using a conjugate gradient approach and $\boldsymbol{\phi}$ using a few steps of L-BFGS.

We also outline a way to deterministically initialize the model to provide stability and, again, speed up estimation on large problems. To do this, we adapt the procedure from Murphy and Murphy (2020) for initializing mixture of experts: (i) initialize the groups using some (deterministic) procedure (e.g. spectral clustering on the moderators); (ii) using only the main effects, estimate an EM algorithm—possibly with hard assignment at the E-Step (CEM; Celeux and Govaert 1992); (iii) iterate until the memberships have stabilized. Use those memberships to initialize the model. This has the benefit of having a deterministic initialization procedure where the group membership is based on the moderators but guided by which grouping seem to have sensible treatment effects, at least for the main effects. Given the memberships, update $\boldsymbol{\beta}$ using a ridge regression and $\boldsymbol{\phi}$ using a ridge regression and take those values as $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\phi}^{(0)}$.

H Extensions to the Basic Model

As noted in the main text, there are five major extensions to the basic model that applied users might wish to include:

1. Repeated tasks (observations) for a single individual
2. A forced-choice conjoint experiment
3. Survey to weight the sample estimate to the broader population
4. Adaptive weights for each penalty
5. Latent overlapping groups

All can be easily incorporated into the proposed framework above. This section outlines the changes to the underlying model.

H.1 Repeated Observations

This modification notes that in factorial and conjoint experiments it is common for individuals to perform multiple tasks. Typically, the number of tasks N_i is similar across individuals. The updated likelihood for a single observation i is shown below; we show both the observed and complete case. y_{im} represents the choice of person i on task $m \in \{1, \cdots, N_i\}$; p_{imk} is the probability of $Y_{im} = 1$ if person i was in group k, and $\tilde{\boldsymbol{T}}_{im}$ is the vector of treatment indicators for person i on task m.

$$L\left(\{Y_{im}\}_{m=1}^{N_i}\right) = \sum_{k=1}^{K} \pi_{ik} \left[\prod_{m=1}^{N_i} p_{imk}^{Y_{im}} (1-p_{imk})^{1-Y_{im}}\right]; \quad p_{imk} = \frac{\exp(\psi_{imk})}{1+\exp(\psi_{imk})}; \quad \psi_{imk} = \tilde{\boldsymbol{T}}_{im}^{\top} \boldsymbol{\beta}_k + \mu \tag{A9}$$

$$L^c(\{y_{im}, \omega_{im}\} \mid Z_i) = \prod_{t=1}^{N_i} \left[\frac{1}{2} \exp\left\{\left(Y_{im} - \frac{1}{2}\right)\psi_{i,Z_i} - \omega_{im} \frac{\psi_{im,Z_i}^2}{2}\right\} f_{PG}(\omega_{im} \mid 1, 0)\right] \tag{A10}$$

Note that because of the conditional independence of (y_{it}, ω_{it}) given Z_i and the parameters, the major modifications to the EM algorithm is that the E-Step must account for all t observations, i.e. the terms summed in Equation (A9). Some additional book-keeping is required in the code as the design of the treatments has $\sum_{i=1}^{N} N_i$ rows whereas the design of the moderators has N rows. Repeated observations can be easily integrated into the uncertainty estimation procedure outlined below.

H.2 Forced Choice Conjoint Design

A popular design of a conjoint experiment is the forced choice design where the respondents are required to choose between two profiles. Therefore, the researcher does not observe an outcome for each profile separately, but rather a single outcome is observed for each pair indicating which is preferred. Egami and Imai (2019) show that this can be easily fit into the above framework with some adjustment. Specifically, the model is modified to difference the indicators of the treatment levels for the pair of profiles (subtracting, e.g., the levels of the profile presented on the left from those of the profile presented on the right). The intercept for this model can be interpreted as a preference for picking a profile presented in a particular location. With this modification, estimation proceeds as before.

H.3 Standardization Weights

An additional modification to the problem is to weight the penalty. This could be done for two reasons. First, there is an issue of the columns having different variances/Euclidean norms because of the different number of factor levels L_j. Second, it is popular to weight the penalty based on some consistent estimator (e.g. ridge regression) to improve performance and, in simpler models, can be shown to imply various oracle properties (e.g. Zou 2006). We leave the latter to future exploration.

Define ξ_{gk} as a positive weight for the g-th penalty and the k-th group. The kernel of the penalty is modified to include them.

$$\log p(\boldsymbol{\beta}_k \mid \lambda, \gamma, \{\boldsymbol{\phi}_k\}) \propto -\lambda \bar{\pi}_k^{\gamma} \sum_{g=1}^{G} \xi_{gk} \sqrt{\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k} \tag{A11}$$

This has no implication on the rank of the stacked $\boldsymbol{F}_g$ (and thus the results in Appendix F) as they are all positive and thus only slightly modify the E-Step.

We employ weights in all of our analyses to account for the fact that different factors j may have different number of levels L_j. We use a generalization of the weights in Bondell and Reich (2009) to the case of penalized *differences*. Specifically, consider the over-parameterized model in Appendix H.4 where the penalty can be written entirely on the differences $\boldsymbol{\delta}_{\text{Main}}$, $\boldsymbol{\delta}_{\text{Int}}$, $\boldsymbol{\delta}_{\text{Main}-\text{Copy}}$. Note that each of those penalties has a simple (group) LASSO form and thus we adopt the approach in Lim and Hastie (2015) of weighting by the Frobenius norm of the associated columns in $\boldsymbol{T}_{\text{LOG}}$, i.e. the over-parameterized design matrix. At slight abuse of notation, define $[\boldsymbol{T}_{\text{LOG}}]_g$ as the columns of $\boldsymbol{T}_{\text{LOG}}$ corresponding to the differences penalized in the (group) lasso g, the weight can be expressed as follows:

$$\xi_{gk} = \frac{1}{\sqrt{N}} || \ [\boldsymbol{T}_{\text{LOG}}]_g \, ||_F$$

Ignoring the factor of $\sqrt{N}$, this exactly recovers the weight proposed in Bondell and Reich (2009) in the non-latent-overlapping non-interactive model of $(L_j + 1)^{-1}\sqrt{N_l^j + N_{l'}^j}$ where N_l^j, $N_{l'}^j$ are the number of observations for factor j in level l and level l' that are being encouraged to fuse together by the penalty in group g.

H.4 Latent Overlapping Groups

One feature of the above approach is that our groups are highly overlapping. Yan and Bien (2017) suggest that, in this setting, a different formulation of the problem may result in superior performance (see also Lim and Hastie 2015). Existing work on the topic has focused on group LASSO penalties (e.g. $\boldsymbol{F}_g = \boldsymbol{I}$) and thus some modifications are needed for our purposes. To address this, we note that we can again recast our model in an equivalent fashion. Instead of penalizing $\sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k}$, we can penalize the vector of differences between levels as long as we also impose linear constraints to ensure that the original model is maintained.

Consider a simple example with two factors each with two levels $\{1, 2\}$ and $\{A, B\}$. The relevant differences are defined such that $\delta^j_{1-2} = \beta^j_1 - \beta^j_2$ and $\delta^{jj'}_{(lm)-(l'm')} = \beta^j_{l,m} - \beta^{j'}_{l',m'}$. The equivalent penalty can be imposed as follows:

$$\begin{gathered}\sqrt{\left(\delta^j_{1-2}\right)^2 + \left(\delta^{jj'}_{(1A)-(2A)}\right)^2 + \left(\delta^{jj'}_{(1B)-(2B)}\right)^2} = \sqrt{\boldsymbol{\delta}^\top \boldsymbol{\delta}}; \quad \boldsymbol{\delta} = \begin{pmatrix} \delta^j_{1-2} \\ \delta^{jj'}_{(1A)-(2A)} \\ \delta^{jj'}_{(1B)-(2B)} \end{pmatrix} \\ \text{such that} \begin{bmatrix} \delta^j_{1-2} \\ \delta^{jj'}_{(1A)-(2A)} \\ \delta^{jj'}_{(1B)-(2B)} \end{bmatrix} = \begin{bmatrix} \beta^j_1 - \beta^j_2 \\ \beta^{jj'}_{1A} - \beta^{jj'}_{2A} \\ \beta^{jj'}_{1B} - \beta^{jj'}_{2B} \end{bmatrix}\end{gathered} \tag{A12}$$

The latent overlapping group suggests a slight modification. In addition to the above penalization of the ℓ_2 norm of the main and interactive differences,[1] it duplicates the main effect and penalizes it separately while ensuring that all effects maintain the accounting identities between the "latent" groups and the overall effect. Specifically, it modifies the above penalty to duplicate the column corresponding to δ^j_{1-2} and adds a new parameter $\delta^j_{(1-2)-\text{Copy}}$.

$$\sqrt{\boldsymbol{\delta}^\top \boldsymbol{\delta}} + |\delta^j_{(1-2)-\text{Copy}}| \quad \text{such that} \begin{bmatrix} \delta^j_{1-2} \\ \delta^{jj'}_{(1A)-(2A)} \\ \delta^{jj'}_{(1B)-(2B)} \end{bmatrix} + \begin{bmatrix} \delta^j_{(1-2)-\text{Copy}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \beta^j_1 - \beta^j_2 \\ \beta^{jj'}_{1A} - \beta^{jj'}_{2A} \\ \beta^{jj'}_{1B} - \beta^{jj'}_{2B} \end{bmatrix} \tag{A13}$$

Scoping out to the full problem, define $\boldsymbol{\delta}_{\text{Main}}$ as the main effect differences, e.g. δ^j_{1-2}, and $\boldsymbol{\delta}_{\text{Int}}$ as the interaction differences and $\boldsymbol{D}_{\text{Main}}$ as the matrix such that $\boldsymbol{D}_{\text{Main}}\boldsymbol{\beta} = \boldsymbol{\delta}_{\text{Main}}$, and $\boldsymbol{D}_{\text{Int}}$ as the corresponding matrix to create the vector of interactions. Define $\boldsymbol{\delta}_{\text{Main}-g}$ as the sub-vector of $\boldsymbol{\delta}_{\text{Main}-g}$ that corresponds to the (main) effect differences between levels l and l' of factor j penalized by $\boldsymbol{F}_g$ in the original notation. Similarly define $\boldsymbol{\delta}_{\text{Int}-g}$ and $\boldsymbol{\delta}_{\text{Main}-\text{Copy}-g}$.

[1] Note the related "hierarchical group LASSO" would add separate individual penalties for each of the interactions. It is easy to include that in our approach.

$$p(\boldsymbol{\beta}, \boldsymbol{\delta}_{\text{Main}}, \boldsymbol{\delta}_{\text{Int}}, \boldsymbol{\delta}_{\text{Main}-\text{Copy}}) = \sum_{g=1}^{G} \sqrt{\boldsymbol{\delta}_{\text{Main}-g}^{T}\boldsymbol{\delta}_{\text{Main}-g} + \boldsymbol{\delta}_{\text{Int}-g}^{T}\boldsymbol{\delta}_{\text{Int}-g}} + \sum_{g'=1}^{G} \sqrt{[\boldsymbol{\delta}_{\text{Main}-\text{Copy}-g}]^2}$$

$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{C}^{\top} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{D}_{\text{Main}} & -\boldsymbol{I} & \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{D}_{\text{Int}} & \mathbf{0} & -\boldsymbol{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta}_{\text{Main}} \\ \boldsymbol{\delta}_{\text{Int}} \\ \boldsymbol{\delta}_{\text{Main}-\text{Copy}} \end{bmatrix} = \mathbf{0} \tag{A14}$$

This also requires a modification of the design matrix $\tilde{\boldsymbol{T}}$ to ensure that (i) its dimensionality conforms with the expanded parameter vector and (ii) that for any choice of the expanded parameter that satisfies the constraints, the linear predictor for all observation (and thus the likelihood) is unchanged. Consider first the simple case without latent-overlapping groups. In this case, following Bondell and Reich (2009), note that the expanded design can be expressed as $\tilde{\boldsymbol{T}}^{\dagger} = \boldsymbol{T}\tilde{\boldsymbol{M}}^{\dagger}$ where $\tilde{\boldsymbol{M}}^{\top} = [\boldsymbol{I}, \boldsymbol{D}_{\text{Main}}^{\top}, \boldsymbol{D}_{\text{Int}}^{\top}]$ and $\tilde{\boldsymbol{M}}^{\dagger}$ is a left-inverse of $\tilde{\boldsymbol{M}}$. The latent-overlapping group formulation is a simple extension; we copy the columns of $\tilde{\boldsymbol{T}}^{\dagger}$ that correspond to $\boldsymbol{\delta}_{\text{Main}}$ and append them to get $\boldsymbol{T}_{\text{LOG}}$.

With this new design and parameterization in hand, we can again use the above results on projecting out the linear constraints to turn the problem into inference on an unconstrained vector $\boldsymbol{\beta}_k$ with a set of positive semi-definite constraints $\{\boldsymbol{F}_g\}_{g=1}^{2G}$ and inference proceeds identically to before.

I Estimators

Here we provide further details on the estimators. In particular, we discuss estimation of Average Marginal Component Effects (AMCEs) and Average Marginal Interaction Effects (AMIEs) based on our model. We consider a traditional factorial design, where each unit receives one treatment (profile), and a conjoint design in which each unit compares two treatments (profiles). We also discuss the impact of randomization restrictions on estimators and implied changes in interpretation of estimands.

I.1 Factorial designs

I.1.1 Without restrictions on randomization

For a unit in group k we have

$$\Pr(Y_i = 1 \mid \boldsymbol{T}_i, \boldsymbol{X}_i) = \zeta_k(\boldsymbol{T}_i) \tag{A15}$$

where $i = 1, 2, \ldots, N$ and for $k = 1, 2, \ldots, K$,

$$\zeta_k(\boldsymbol{T}_i) \;=\; \frac{\exp(\psi_k(\boldsymbol{T}_i))}{1 + \exp(\psi_k(\boldsymbol{T}_i))}. \tag{A16}$$

We model $\psi_k(\boldsymbol{T}_i)$ as

$$\psi_k(\boldsymbol{T}_i) \;=\; \mu + \sum_{j=1}^{J} \sum_{l=0}^{L_j - 1} \mathbf{1}\{T_{ij} = l\}\beta_{kl}^{j} + \sum_{j=1}^{J-1} \sum_{j'>j} \sum_{l=0}^{L_j - 1} \sum_{l'=0}^{L_{j'} - 1} \mathbf{1}\{T_{ij} = l, T_{ij'} = l'\}\beta_{kll'}^{jj'}, \tag{A17}$$

for each $k = 1, 2, \ldots, K$, with constraints

$$\boldsymbol{C}^\top \boldsymbol{\beta}_k \;=\; \boldsymbol{0} \tag{A18}$$

where $\boldsymbol{\beta}_k$ is a stacked column vector containing all coefficients for group k.

We can rewrite this to aid in the interpretation of $\boldsymbol{\beta}_k$ as follows:

$$\operatorname{logit}(\zeta_k(\boldsymbol{T}_i)) = \mu + \sum_{j=1}^{J} \sum_{l=0}^{L_j-1} \mathbf{1}\{T_{ij} = l\}\beta^j_{kl} + \sum_{j=1}^{J-1} \sum_{j'>j} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_{j'}-1} \mathbf{1}\{T_{ij} = l, T_{ij'} = l'\}\beta^{jj'}_{kll'}.$$

Thus, $\beta^j_{kl} - \beta^j_{kf}$ is the AMCE going from level f to level l of factor j on the logit probability of $Y_i = 1$ scale.

Let $\boldsymbol{t}$ be some combination of the J factors, where $\boldsymbol{t}_j$ is the jth factor's level and $\boldsymbol{t}_{-j}$ is the levels for all factors except j. This allows us to easily write, taking expectation over units in group k,

$$\begin{aligned}\mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right) &= \Pr\left(Y_i = 1 | Z_i = k, \boldsymbol{T}_{i,j} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right) \\ &= \frac{\exp(\zeta_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))}{1 + \exp(\zeta_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))},\end{aligned}$$

where $T_{ij} = l$ indicates for unit i forcing factor j to be assigned level l and $\boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}$ indicates forcing the assignment on all factors except for j to be assigned levels as in $\boldsymbol{t}_{-j}$.

The causal effects of interest (on the original Y scale) are defined as contrasts of these expectations. Without additional weighting (i.e., using traditional uniform weights for marginalization), the AMCE for level l vs f of factor j in group k is,

$$\begin{aligned}\delta^*_{jk}(l, f) =& \frac{1}{M} \sum_{\boldsymbol{t}_{-j}} \mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right) - \mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right) \\ =& \frac{1}{M} \sum_{\boldsymbol{t}_{-j}} \frac{\exp(\zeta_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))}{1 + \exp(\zeta_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))} - \frac{\exp(\zeta_k(T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))}{1 + \exp(\zeta_k(T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))},\end{aligned}$$

where M is the number of possible combinations of the other $J - 1$ factors (e.g., if we had J 2-level factors, $M = 2^{J-1}$). We can estimate this by plugging in the coefficients directly. Note that, because of the nonlinear nature of the estimator, this approach is consistent (under model assumptions) but not unbiased.

Alternatively, instead of summing over all *possible* $\boldsymbol{t}_{-j}$, we can use the empirical distribution of $\boldsymbol{t}_{-j}$ in the sample. This potentially changes the estimand. Define estimators

$$\widehat{\psi}_k(\boldsymbol{t}) \;=\; \mu + \sum_{j=1}^{J} \sum_{l=0}^{L_j-1} \mathbf{1}\{t_j = l\}\widehat{\beta}^j_{kl} + \sum_{j=1}^{J-1} \sum_{j'>j} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_{j'}-1} \mathbf{1}\{t_j = l, t_{j'} = l'\}\widehat{\beta}^{jj'}_{kll'}$$

and

$$\widehat{y}_k(\boldsymbol{t}) = \frac{\exp(\widehat{\psi}_k(\boldsymbol{t}))}{1 + \exp(\widehat{\psi}_k(\boldsymbol{t}))}$$

Then we can use the following overall estimator for the AMCE:

$$\frac{1}{N} \sum_{b=1}^{N} \left(\widehat{Y}_k(T_{bj} = l, \boldsymbol{T}_{b,-j}) - \widehat{Y}_k(T_{bj} = f, \boldsymbol{T}_{b,-j}) \right).$$

This is a consistent estimator (under model assumptions) of

$$\begin{aligned}
&\frac{1}{N}\sum_{b=1}^{N}\mathbb{E}\left(Y_i \mid Z_i = k, \boldsymbol{T}_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{T}_{b,-j}\right) - \mathbb{E}\left(Y_i \mid Z_i = k, \boldsymbol{T}_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{T}_{b,-j}\right)\\
=&\frac{1}{N}\sum_{b=1}^{N}\frac{\exp(\psi_k(T_{bj} = l, \boldsymbol{T}_{b,-j}))}{1+\exp(\psi_k(T_{bj} = l, \boldsymbol{T}_{b,-j}))} - \frac{\exp(\psi_k(T_{bj} = f, \boldsymbol{T}_{b,-j}))}{1+\exp(\psi_k(T_{bj} = f, \boldsymbol{T}_{b,-j}))},
\end{aligned}$$

conditioning on the treatments we actually observed.

Now, we turn to examination of the AMIEs. Without additional weighting (i.e., using traditional uniform weights for marginalization), the AMIE for level l of factor j and level q of factor s vs f of factor j and level r of factor s in group k is

$$\text{AMIE}^*_{jsk}(l, f, q, r) = \text{ACE}^*(l, f, q, r) - \delta^*_{jk}(l, f) - \delta^*_{sk}(q, r)$$

where

$$\begin{aligned}
&\text{ACE}^*(l, f, q, r)\\
=&\frac{1}{M^*}\sum_{\boldsymbol{t}_{-(j,s)}}\mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = l, T_{is} = q, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)}\right) - \mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = f, T_{is} = r, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)}\right)\\
=&\frac{1}{M^*}\sum_{\boldsymbol{t}_{-(j,s)}}\frac{\exp(\psi_k(T_{ij} = l, T_{is} = q, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)}))}{1+\exp(\psi_k(T_{ij} = l, T_{is} = q, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)}))} - \frac{\exp(\psi_k(T_{ij} = f, T_{is} = r, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)})))}{1+\exp(\psi_k(T_{ij} = f, T_{is} = r, \boldsymbol{T}_{i,-(j,s)} = \boldsymbol{t}_{-(j,s)}))},
\end{aligned}$$

where M^* is the number of possible combinations of the other $J - 2$ factors (e.g., if we had J two-level factors, $M^* = 2^{J-2}$).

We can use the following overall estimator for the ACE:

$$\widehat{\text{ACE}}^*(l, f, q, r) = \frac{1}{N}\sum_{b=1}^{N}\widehat{Y}_k(T_{bj} = l, T_{bs} = q, \boldsymbol{T}_{b,-(j,s)}) - \widehat{Y}_k(T_{bj} = f, T_{bs} = r, \boldsymbol{T}_{b,-(j,s)}).$$

This is then combined with the estimators for the AMCEs to get

$$\widehat{\text{AMIE}}^*_{jsk}(l, f, q, r) = \widehat{\text{ACE}}^*(l, f, q, r) - \widehat{\delta}^*_{jk}(l, f) - \widehat{\delta}^*_{sk}(q, r).$$

I.1.2 With restrictions on randomization

In this section we consider restricted randomization conditions. Let us assume that factor j and factor h are such that some levels of j are not well defined and hence excluded in combination with some levels of factor h under the randomization set up. Let $\mathcal{S}(j, h) \subset \{1, \ldots, L_j\}$ be the set of levels of factor j that are not defined for some levels of factor h. Similarly, let $\mathcal{S}(h, j) \subset \{1, \ldots, L_h\}$ be the set of levels of factor h that are not defined for some levels of factor j. In our example, if j is education and h is profession, we have $\mathcal{S}(j, h) = \{\text{No formal, 4th grade, 8th grade, High school}\}$ and $\mathcal{S}(h, j) = \{\text{Financial analyst, Research scientist, Doctor, Computer programmer}\}$.

When estimating the AMCE for level l vs f of factor $J - 1$ in group k, using the model rather than the empirical distribution, we consider,

$$\begin{aligned}
&\frac{1}{M_{def(j,h)}}\sum_{\boldsymbol{t}_{-j}:\boldsymbol{t}_h \notin \mathcal{S}(h,j)}\mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right) - \mathbb{E}\left(Y_i \mid Z_i = k, T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}\right)\\
=&\frac{1}{M_{def(j,h)}}\sum_{\boldsymbol{t}_{-j}:\boldsymbol{t}_h \notin \mathcal{S}(h,j)}\frac{\exp(\psi_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))}{1+\exp(\psi_k(T_{ij} = l, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))} - \frac{\exp(\psi_k(T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))}{1+\exp(\psi_k(T_{ij} = f, \boldsymbol{T}_{i,-j} = \boldsymbol{t}_{-j}))},
\end{aligned}$$

where $M_{def(j,h)}$ is the number of possible combinations of the other factors, restricted such that $\boldsymbol{t}_h \notin \mathcal{S}(h,j)$ (e.g., if we had J 3-level factors, and some of the levels of factor j were not defined for one level of factor h, this would be $2 \times 3^{J-2}$).

To use empirical distribution, we need a way to deal with profiles that are not well defined. We can accomplish this by only aggregating over those profiles that are sensible for all levels of factor j. That is, we use the following estimator,

$$\frac{1}{\sum_{i=1}^{N} \mathbb{I}\{T_{ih} \notin \mathcal{S}(h,j)\}} \sum_{b=1}^{N} \mathbb{I}\{T_{bh} \notin \mathcal{S}(h,j)\} \left(\widehat{Y}_k(T_{bj} = l, \boldsymbol{T}_{b,-j}) - \widehat{Y}_k(T_{bj} = f, \boldsymbol{T}_{b,-j}) \right).$$

Consider the case where we are estimating the AMCE for “doctor” vs “gardener” for profession. Because of the randomization restriction between certain professions and level of education, we will remove any profiles that have “4th grade” as level of education. Although “gardener” with “4th grade” education is allowable under the randomization, we must remove such profiles to have an “apples-to-apples” comparison with profession of doctor, which is not allowed to have “4th grade” education. Note that we do this dropping of profiles even if we are comparing “waiter” vs “gardener” for profession, which are both allowed to have “4th grade” as level of education, to ensure that all AMCEs for profession comparable.

Similarly for the AMIEs, we restrict the profiles we marginalize over to be only those that are defined for both factors in the interactions. Let factor j be restricted by some other factor h and let factor s be restricted by some other factor w. Then we have the following estimator,

$$\begin{aligned}
&\widehat{\text{ACE}}^*(l,f,q,r) \\
=&\sum_{b=1}^{N} \frac{\mathbb{I}\{T_{bh} \notin \mathcal{S}(h,j), T_{bw} \notin \mathcal{S}(w,s)\}}{\sum_{i=1}^{N} \mathbb{I}\{T_{ih} \notin \mathcal{S}(h,j), T_{iw} \notin \mathcal{S}(w,s)\}} \left(\widehat{Y}_k(T_{bj} = l, T_{bs} = q, \boldsymbol{T}_{b,-(j,s)}) - \widehat{Y}_k(T_{bj} = f, T_{bs} = r, \boldsymbol{T}_{b,-(j,s)}) \right).
\end{aligned}$$

The relevant AMCEs should be similarly restricted within the AMIE estimator, with restrictions applied based on the restrictions for all levels both factors in the interaction.

I.2 Conjoint designs

I.2.1 Without restrictions on randomization

Consider a conjoint experiment in which each unit i only compares two profiles. The response Y_i indicates a choice between two profiles. Let $\boldsymbol{T}_i^L$ be the levels for the left profile and $\boldsymbol{T}_i^R$ be the levels for the right profile that unit i sees. Here, we modify how we model ψ_k to

$$\begin{aligned}
\psi_k(\boldsymbol{T}_i^L, \boldsymbol{T}_i^R) &= \mu + \sum_{j=1}^{J} \sum_{l \in L_j} \beta_{kl}^{j} \left(\mathbf{1}\left\{T_{ij}^L = l\right\} - \mathbf{1}\left\{T_{ij}^R = l\right\} \right) \\
&\quad + \sum_{j=1}^{J-1} \sum_{j' > j} \sum_{l \in L_j} \sum_{l' \in L_{j'}} \beta_{kll'}^{jj'} \left(\mathbf{1}\left\{T_{ij}^L = l, T_{ij'}^L = l'\right\} - \mathbf{1}\left\{T_{ij}^R = l, T_{ij'}^R = l'\right\} \right).
\end{aligned}$$

If we use $Y_i = 1$ to indicate that unit i picks the left profile, then we have,

$$\begin{aligned}
\mathbb{E}\left(Y_i \mid Z_i = k, \boldsymbol{T}_i^L = \boldsymbol{t}^L, \boldsymbol{T}_i^R = \boldsymbol{t}^R\right) &= \Pr\left(Y_i = 1 \mid Z_i = k, \boldsymbol{T}_i^L = \boldsymbol{t}^L, \boldsymbol{T}_i^R = \boldsymbol{t}^R\right) \\
&= \frac{\exp(\psi_k(\boldsymbol{T}_i^L = \boldsymbol{t}^L, \boldsymbol{T}_i^R = \boldsymbol{t}^R))}{1 + \exp(\psi_k(\boldsymbol{T}_i^L = \boldsymbol{t}^L, \boldsymbol{T}_i^R = \boldsymbol{t}^R))}.
\end{aligned}$$

We can use the symmetry assumption that choice order does not affect the appeal of individual attributes. That is, there may be some overall preference for left or right accounted for by μ, but this preference is not affected by profile attributes. Then, we can define our effects, on the original Y scale, as contrasts of these expectations. Without additional weighting, the AMCE for level l vs l' of factor j in group k is,

$$\begin{aligned}\delta_{jk}(l,l') \;=\; & \frac{1}{2}\mathbb{E}\left[\left\{\Pr\left(Y_i = 1 \mid Z_i = k, T^L_{ij} = l, \boldsymbol{T}^L_{i,-j}, \boldsymbol{T}^R_i\right) - \Pr\left(Y_i = 1 \mid Z_i = k, T^L_{ij} = l', \boldsymbol{T}^L_{i,-j}, \boldsymbol{T}^R_i\right)\right\}\right. \\ & \left. + \left\{\Pr\left(Y_i = 0 \mid Z_i = k, T^R_{ij} = l, \boldsymbol{T}^R_{i,-j}, \boldsymbol{T}^L_i\right) - \Pr\left(Y_i = 0 \mid Z_i = k, T^R_{ij} = l', \boldsymbol{T}^R_{i,-j}, \boldsymbol{T}^L_i\right)\right\}\right].\end{aligned}$$

To save space, the outer expectation is over the random assignment, which corresponds to the expectation over the $\widetilde{M}$ possible combinations of the two profiles on the other $J-1$ factors (e.g., if we had J two-level factors, this would be 4^{J-1}). We can again estimate this by plugging in our coefficient estimates directly.

Alternatively, instead of summing over all *possible* $\boldsymbol{t}^L_{-j}$ and $\boldsymbol{t}^R_{-j}$, we can use the empirical distribution of $\boldsymbol{t}^L_{-j}$ and $\boldsymbol{t}^R_{-j}$ in the sample. Define

$$\widehat{Y}_k(\boldsymbol{t}^L, \boldsymbol{t}^R) = \frac{\exp(\widehat{\psi}(\boldsymbol{t}^L, \boldsymbol{t}^R))}{1 + \exp(\widehat{\psi}(\boldsymbol{t}^L, \boldsymbol{t}^R))}.$$

Then we can use the estimator

$$\begin{aligned}\widehat{\delta}_{jk}(l,l') =& \frac{1}{2N}\sum_{i=1}^{N}\left[\left\{\widehat{Y}_k(T^L_{ij} = l, \boldsymbol{T}^L_{i,-j}, \boldsymbol{T}^R_i) - \widehat{Y}_k(T^L_{ij} = l', \boldsymbol{T}^L_{i,-j}, \boldsymbol{T}^R_i)\right\}\right. \\ & \left. - \left\{\widehat{Y}_k(T^R_{ij} = l, \boldsymbol{T}^R_{i,-j}, \boldsymbol{T}^L_i) - \widehat{Y}_k(T^R_{ij} = l', \boldsymbol{T}^R_{i,-j}, \boldsymbol{T}^L_i)\right\}\right].\end{aligned}$$

Now we turn to examination of the AMIEs. Without additional weighting (i.e., using traditional uniform weights for marginalization), the AMIE for level l of factor j and level q of factor s vs m of factor j and level r of factor s in group k is

$$\text{AMIE}_{jsk}(l, f, q, r) = \text{ACE}(l, f, q, r) - \delta_{jk}(l, f) - \delta_{sk}(q, r)$$

Here we can use the estimator

$$\begin{aligned}\widehat{\text{ACE}}(l, f, q, r) =& \frac{1}{2N}\sum_{i=1}^{N}\left[(\widehat{Y}_k(T^L_{ij} = l, T^L_{is} = q, \boldsymbol{T}^L_{i,-(j,s)}, \boldsymbol{T}^R_i) - \widehat{Y}_k(T^L_{ij} = f, T^L_{is} = r, \boldsymbol{T}^L_{i,-(j,s)}, \boldsymbol{T}^R_i)\right) \\ & - \frac{1}{2N}\sum_{i=1}^{N}\left(\widehat{Y}_k(T^R_{ij} = l, T^R_{is} = q, \boldsymbol{T}^R_{i,-(j,s)}, \boldsymbol{T}^L_i) - \widehat{Y}_k(T^R_{ij} = f, T^R_{is} = r, \boldsymbol{T}^R_{i,-(j,s)}, \boldsymbol{T}^L_i)\right).\end{aligned}$$

This gives us

$$\widehat{\text{AMIE}}_{jsk}(l, f, q, r) = \widehat{\text{ACE}}(l, f, q, r) - \widehat{\delta}_{jk}(l, f) - \widehat{\delta}_{sk}(q, r).$$

I.2.2 With restrictions on randomization

Similar to Appendix I.1.2, adjustments to estimation need to be made when we have restricted randomizations. We again will do this by dropping profiles that have levels of factors not allowable for all levels of the factor(s) whose effects we are estimating (e.g., profiles with “4th grade” for education when estimating an effect for profession). However, now we estimate the effect for the right profile and the effect for the left profile, and then average the two (they should be equal under symmetry). When estimating the effect for the right profile, therefore, we will only drop pairings if the *right* profile has a level that is not allowed for some level of the factor we are estimating an effect of. For example, dropping pairings where the right profile has “4th grade” as level of education when estimating main effects of profession because “doctor” cannot have level “4th grade.” Again, this will drop more profiles than those that are not allowed under randomization to ensure an “apples-to-apples” comparison across levels of profession.

In this calculation, we use the empirical distribution for the levels of the left profile (which represents the “opponent”). Thus, the distribution of other factors for the profile we are calculating the effect of may differ than that distribution for its opponents. Similarly, when estimating the effect for the left profile, we only drop pairings in which the left profile has a restricted level for some level of the factor of interest. Estimation for the AMIE under randomization restrictions follows similarly.

J Quantification of Uncertainty

We quantify uncertainty in our parameter estimates by inverting the negative Hessian of the log-posterior at the estimates $\hat{\boldsymbol{\theta}}$, i.e. $\left[-\frac{\partial}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^T} \log p(\boldsymbol{\theta}|Y_i)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ or $\mathcal{I}(\hat{\boldsymbol{\theta}})$. This can be stably and easily computed using terms from the AECM algorithm following Louis (1982)’s method. Specifically, consider the model from the main text augmented with Z_i, i.e. the group memberships. Recall that $z_{ik} = \mathbf{1}\{Z_i = k\}$ for notational simplicity.

$$\begin{aligned} L^c(\boldsymbol{\theta}) = & \sum_{i=1}^{N} \left[\sum_{k=1}^{K} z_{ik} \log(\pi_{ik}) + z_{ik} \log L(Y_i \mid \boldsymbol{\beta}_k) \right] + \\ & \sum_{k=1}^{K} m \log(\lambda) + m\gamma \log(\bar{\pi}_k) - \lambda \bar{\pi}_k^{\gamma} \left[\sum_{g=1}^{G} \xi_{gk} \sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k} \right] + \log p(\{\boldsymbol{\phi}_k\}). \end{aligned} \tag{A19}$$

Louis (1982) notes that equation can be used to compute $\mathcal{I}_L(\hat{\boldsymbol{\theta}})$, where the subscript L denotes its computation via this method.

$$\mathcal{I}_L(\hat{\boldsymbol{\theta}}) = E_{p(\{Z_i\}_{i=1}^N | \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}_{i=1}^N, \hat{\boldsymbol{\theta}})} \left[-\frac{\partial L^c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^\top} \right] - \text{Var}_{p(\{Z_i\}_{i=1}^N | \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}_{i=1}^N, \hat{\boldsymbol{\theta}})} \left[\frac{\partial L^c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \tag{A20}$$

To address the issue with the non-differentiability of the penalty on $\boldsymbol{\beta}$ (and thus $L^c(\boldsymbol{\theta})$), we follow the existing research in two ways. First, for restrictions that are sufficiently close to binding, we assume them to bind and estimate the uncertainty *given* those restrictions. That is, we identify the binding restrictions such that $\sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k}$ is sufficiently small (say 10^{-4}) and note that if these are binding, we can use the null space projection technique to transform $\boldsymbol{\beta}_k$ such that it lies in an unconstrained space.

To further ensure stability, we modify the penalty with a small positive constant $\epsilon \approx 10^{-4}$ to ensure that the entire objective is (twice) differentiable. For notational simplicity, we derive the results below assuming $\boldsymbol{\beta}_k$ represent the parameter vector after projecting into a space with no linear constraints. The approximated log-posterior is shown below and denoted with a tilde. We thus evaluate $\mathcal{I}_L(\hat{\boldsymbol{\theta}})$ using $\tilde{L}^c$ in place of L^c.

$$
\begin{aligned}
\tilde{L}^c(\boldsymbol{\theta}) =& \sum_{i=1}^{N} \left[\sum_{k=1}^{K} z_{ik} \log(\pi_{ik}) + z_{ik} \log L(y_i | \boldsymbol{\beta}_k) \right] + \\
& \sum_{k=1}^{K} m \log(\lambda) + m\gamma \log(\bar{\pi}_k) - \lambda \bar{\pi}_k^{\gamma} \left[\sum_{g=1}^{G} \xi_{gk} \sqrt{\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k + \epsilon} \right] + \log p(\{\boldsymbol{\phi}_k\})
\end{aligned}
\tag{A21}
$$

This procedure has some pleasing properties that mirror existing results on approximate standard errors after sparse estimation; consider a simple three-level case: $\beta_1^j, \beta_2^j, \beta_3^j$. If β_1^j and β_2^j are fused, then their approximate point estimates and standard errors will be identical but *crucially* not zero. This is because while their difference is zero and assumed to bind with no uncertainty, this does not imply that the effects, themselves, have no uncertainty: $\beta_1^j - \beta_2^j$ will have a standard error of zero in our method. This thus mirrors the results from Fan and Li (2001) where effects that are shrunken to zero by the LASSO are not estimated with any uncertainty. One might relax this with fully Bayesian approaches in future research.

Second, note that if all levels are fused together, i.e. $\beta_1^j = \beta_2^j = \beta_3^j$, then all point estimates must be zero by the ANOVA sum-to-zero constraint *and* all will have an uncertainty of zero. Thus, when an entire factor is removed from the model, the approximate standard errors return a result consist with existing research.

J.1 Derivation of Hessian

To calculate the above terms, the score and gradient of $\tilde{L}^c$ are required. They are reported below:

$$
\begin{aligned}
\tilde{S}^c(\mu) &= \sum_{i=1}^{N} \left[\sum_{k=1}^{K} z_{ik} (Y_i - p_{ik}) \right] \\
\tilde{S}^c(\boldsymbol{\beta}_k) &= \sum_{i=1}^{N} z_{ik} \cdot (Y_i - p_{ik}) \tilde{\boldsymbol{T}}_i - \lambda \bar{\pi}_k^{\gamma} \sum_{g=1}^{G} \xi_{gk} (\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k)^{-1/2} \cdot \boldsymbol{F}_g \boldsymbol{\beta}_k \\
\tilde{S}^c(\boldsymbol{\phi}_k) &= \sum_{i=1}^{N} \left[z_{ik} - \pi_{ik} \right] \boldsymbol{X}_i + \frac{\partial \log p(\{\boldsymbol{\phi}_k\})}{\partial \boldsymbol{\phi}_k} + \sum_{k'=1}^{K} m\gamma \frac{\partial \log(\bar{\pi}_{k'})}{\partial \boldsymbol{\phi}_k} - \lambda \gamma \bar{\pi}_{k'}^{\gamma - 1} \cdot \frac{\partial \bar{\pi}_{k'}}{\partial \boldsymbol{\phi}_k} \cdot \left[\sum_{g=1}^{G} \xi_{g,k'} \sqrt{\boldsymbol{\beta}_{k'}^{\top} \boldsymbol{F}_{g,k'} \boldsymbol{\beta}_{k'}} \right] \\
H^c(\mu, \mu) &= \sum_{i=1}^{N} \left[- \sum_{k=1}^{K} z_{ik} p_{ik} (1 - p_{ik}) \right] \\
H^c(\mu, \boldsymbol{\beta}_k) &= - \left[\sum_{i=1}^{N} z_{ik} p_{ik} (1 - p_{ik}) \tilde{\boldsymbol{T}}_i \right] \\
H^c(\boldsymbol{\beta}_k, \boldsymbol{\beta}_k) &= - \left[\sum_{i=1}^{N} z_{ik} \cdot p_{ik} (1 - p_{ik}) \tilde{\boldsymbol{T}}_i \tilde{\boldsymbol{T}}_i^{\top} \right] - \lambda \bar{\pi}_k^{\gamma} \sum_{g=1}^{G} \xi_{gk} \boldsymbol{D}_{gk}
\end{aligned}
$$

where $\left[\boldsymbol{D}_{gk}\right]_{a,b} = -\left(\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k\right)^{-3/2} \boldsymbol{\beta}_k^{\top} \left[\boldsymbol{F}_g\right]_a \boldsymbol{\beta}_k^{\top} \left[\boldsymbol{F}_g\right]_b + \left(\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k\right)^{-1/2} \left[\boldsymbol{F}_g\right]_{a,b}$.

$$
H^c(\left[\boldsymbol{\beta}_k\right]_i, \boldsymbol{\phi}_\ell) = -\lambda \gamma \bar{\pi}_k^{\gamma - 1} \left[\sum_{g=1}^{G} \xi_{gk} (\boldsymbol{\beta}_k^{\top} \boldsymbol{F}_g \boldsymbol{\beta}_k)^{-1/2} \cdot \boldsymbol{\beta}_k^{\top} \left[\boldsymbol{F}_g\right]_i \right] \frac{\partial \bar{\pi}_k}{\partial \boldsymbol{\phi}_\ell}
$$

$$\begin{aligned}H^c(\boldsymbol{\phi}_k, \boldsymbol{\phi}_\ell) = &\sum_{i=1}^N -\left[(I[k=\ell] - \pi_{ik})\,\pi_{i\ell}\right] \boldsymbol{X}_i\boldsymbol{X}_i^\top + \frac{\partial^2 \log p(\{\boldsymbol{\phi}_k\})}{\partial \boldsymbol{\phi}_k \boldsymbol{\phi}_\ell^\top} + \sum_{k'=1}^K m\gamma \frac{\partial \log(\bar\pi_{k'})}{\partial\boldsymbol{\phi}_k\boldsymbol{\phi}_\ell^\top} + \\ &+ \sum_{k'=1}^K -\lambda\gamma \left[\sum_{g=1}^G \xi_{g,k'}\sqrt{\boldsymbol{\beta}_{k'}^\top \boldsymbol{F}_{g,k'}\boldsymbol{\beta}_{k'}}\right]\left[I(\gamma\notin\{0,1\})\cdot(\gamma-1)\bar\pi_{k'}^{\gamma-2}\cdot\left[\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k}\right]\left[\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_\ell}\right]^\top + \bar\pi_{k'}^{\gamma-1}\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k\boldsymbol{\phi}_\ell^\top}\right]\end{aligned}$$

The above results use the following intermediate derivations:

$$\begin{aligned}\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k} &= \frac{1}{N}\sum_{i=1}^N \pi_{i,k'}\left[I(k=k') - \pi_{ik}\right]\boldsymbol{X}_i \\ \frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k\boldsymbol{\phi}_\ell^\top} &= \frac{1}{N}\sum_{i=1}^N \left[\pi_{i,k'}\left(I(k'=\ell) - \pi_{i\ell}\right)\left(I(k=k') - \pi_{ik}\right) - \pi_{i,k'}\pi_{ik}\left(I(k=\ell) - \pi_{i\ell}\right)\right]\boldsymbol{X}_i\boldsymbol{X}_i^\top \\ \frac{\partial\log(\bar\pi_{k'})}{\partial\boldsymbol{\phi}_k} &= \frac{1}{\bar\pi_{k'}}\cdot\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k} \\ \frac{\partial\log(\bar\pi_{k'})}{\partial\boldsymbol{\phi}_k\boldsymbol{\phi}_\ell^\top} &= -\frac{1}{\bar\pi_{k'}^2}\left[\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k}\right]\left[\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_\ell}\right]^\top + \frac{1}{\bar\pi_{k'}}\cdot\frac{\partial\bar\pi_{k'}}{\partial\boldsymbol{\phi}_k\boldsymbol{\phi}_\ell^\top}\end{aligned}$$

Second, the variance of $\tilde{S}^c(\boldsymbol{\theta})$ over $p(\{z_{ik}\} \mid \boldsymbol{\theta})$. This is derived blockwise below.

$$\begin{aligned}\text{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\boldsymbol{\beta}_\ell)\right] &= \sum_{i=1}^N (Y_i - p_{ik})\cdot(Y_i - p_{i\ell})\cdot\mathbb{E}(z_{ik})\left(I(k=\ell) - \mathbb{E}(z_{i\ell})\right)\tilde{\boldsymbol{T}}_i\tilde{\boldsymbol{T}}_i^\top \\ \text{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\boldsymbol{\phi}_\ell)\right] &= \sum_{i=1}^N (Y_i - p_{ik})\cdot\mathbb{E}(z_{ik})\left(I(k=\ell) - \mathbb{E}(z_{i\ell})\right)\tilde{\boldsymbol{T}}_i\boldsymbol{X}_i^\top \\ \text{Cov}\left[\tilde{S}^c(\boldsymbol{\phi}_k), \tilde{S}^c(\boldsymbol{\phi}_\ell)\right] &= \sum_{i=1}^N \mathbb{E}(z_{ik})\left(I(k=\ell) - \mathbb{E}(z_{i\ell})\right)\boldsymbol{X}_i\boldsymbol{X}_i^\top \\ \text{Cov}\left[\tilde{S}^c(\mu), \tilde{S}^c(\mu)\right] &= \sum_{i=1}^N\left[\sum_{k=1}^K\sum_{k'=1}^K \mathbb{E}(z_{ik})\left(I(k=k') - \mathbb{E}(z_{ik'})\right)(Y_i - p_{ik})(Y_i - p_{ik'})\right] \\ \text{Cov}\left[\tilde{S}^c(\boldsymbol{\phi}_k), \tilde{S}^c(\mu)\right] &= \sum_{i=1}^N\left[\sum_{k'=1}^K \mathbb{E}(z_{ik})\left(I(k=k') - \mathbb{E}(z_{ik'})\right)(Y_i - p_{ik'})\boldsymbol{X}_i\right] \\ \text{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\mu)\right] &= \sum_{i=1}^N\left[\sum_{k'=1}^K \mathbb{E}(z_{ik})\left(I(k=k') - \mathbb{E}(z_{ik'})\right)(Y_i - p_{ik})(Y_i - p_{ik'})\tilde{\boldsymbol{T}}_i\right]\end{aligned}$$

This provides all terms needed to compute $\mathcal{I}_L(\hat{\boldsymbol{\theta}})$.

J.2 Repeated Observations

Now consider the case of repeated observations per individual i. In this scenario, each individual i performs N_i tasks. Note, after augmentation, the score has exactly the same form and thus the complete Score $\tilde{S}^c$ and Hessian $\tilde{H}^c$ are identical where the sum merely now runs over $\sum_{i=1}^N\sum_{m=1}^{N_i}$. The average for $\bar\pi_k$ is similarly a weighted average by N_i, although note that often each respondent answers an identical number of tasks so it is, effectively, the same as before. The covariance of $\tilde{S}^c$ is adjusted as shown below.

$$\operatorname{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\boldsymbol{\beta}_\ell)\right] = \sum_{i=1}^{N} \mathbb{E}(z_{ik}) \left(I[k=\ell] - \mathbb{E}(z_{i\ell})\right) \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk}) \tilde{\boldsymbol{T}}_{im}\right] \left[\sum_{m'=1}^{N_i} (Y_{im} - p_{im\ell}) \tilde{\boldsymbol{T}}_{im'}^\top\right]$$

$$\operatorname{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\boldsymbol{\phi}_\ell)\right] = \sum_{i=1}^{N} \mathbb{E}(z_{ik}) \left(I[k=\ell] - \mathbb{E}(z_{i\ell})\right) \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk}) \tilde{\boldsymbol{T}}_{im}\right] \boldsymbol{X}_i^\top$$

$$\operatorname{Cov}\left[\tilde{S}^c(\mu), \tilde{S}^c(\mu)\right] = \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbb{E}(z_{ik}) \big(I[k=k'] - \mathbb{E}(z_{ik'})\big) \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk}) \tilde{\boldsymbol{T}}_{im}\right] \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk}) \tilde{\boldsymbol{T}}_{im}^\top\right]\right]$$

$$\operatorname{Cov}\left[\tilde{S}^c(\boldsymbol{\beta}_k), \tilde{S}^c(\mu)\right] = \sum_{i=1}^{N} \left[\sum_{k'=1}^{K} \mathbb{E}(z_{ik}) \big(I[k=k'] - \mathbb{E}(z_{ik'})\big) \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk}) \tilde{\boldsymbol{T}}_{im}\right] \left[\sum_{m=1}^{N_i} (Y_{im} - p_{imk'})\right]\right]$$

J.3 Standard Errors on Other Quantities of Interest

Given the above results, we derive an approximate covariance matrix on $\hat{\boldsymbol{\theta}}$. We calculate uncertainty on other quantities of interest, e.g. AMCE and marginal effects, using the multivariate delta method. As almost all of our quantities of interest can be expressed as (weighted) sums or averages over individuals $i \in \{1, \cdots, N\}$, calculating the requisite gradient for the multivariate delta method simply requires calculating the relevant derivative for each observation. For example, all derivatives needed in the AMCE are of the following form; see Appendix I for more details.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{\exp(\psi_{ik})}{1 + \exp(\psi_{ik})}\right]$$

K Simulations

We detail our simulations and provide additional results in this section.

K.1 Setup

We generate the $\boldsymbol{\beta}_k$ used in our simulations following Equation 4 and calibrating their implied AMCEs to be roughly comparable to the magnitude found in our empirical example, i.e. ranging between around -0.30 and 0.30. The $\boldsymbol{\beta}_k$ and $\{\boldsymbol{\phi}_k\}_{k=2}^{3}$ used in all simulations are determined using one draw from the following procedure:

Simulating $\boldsymbol{\beta}_k$:

1. For each factor j and group k, draw the number of unique levels u with equal probability from $\{1, 2, 3\}$.
2. Draw u normal random variables independently from $N(0, 1/3)$; call these b_{ku}^j.
3. For $u = 1$, set $\beta_{kl}^j = 0$
4. For $u = 3$, de-mean $\{b_{ku}^j\}_{u=1}^3$ drawn in (2) and set all β_{kl}^j equal to the corresponding value.
5. For $u = 2$, assign b_{k3}^j equal to one of the two b_{ku}^j with equal probability. De-mean the $\{b_{ku}^j\}_{u=1}^3$ and set β_{kl}^j equal to the corresponding values.

Simulating $\boldsymbol{\phi}_k$: $\{\boldsymbol{\phi}_k\}_{k=1}^K \sim N(\mathbf{0}, 2 \cdot \boldsymbol{I})$

To evaluate our method, we calculate the AMCEs in each group simulations using Monte Carlo simulation where we sample 1,000,000 pairs of treatment profiles for the other attributes to marginalize over the other factors. The distribution of the $\boldsymbol{\beta}_k$ and average marginal component effects (with a baseline level of '1') used in the simulations are shown below:

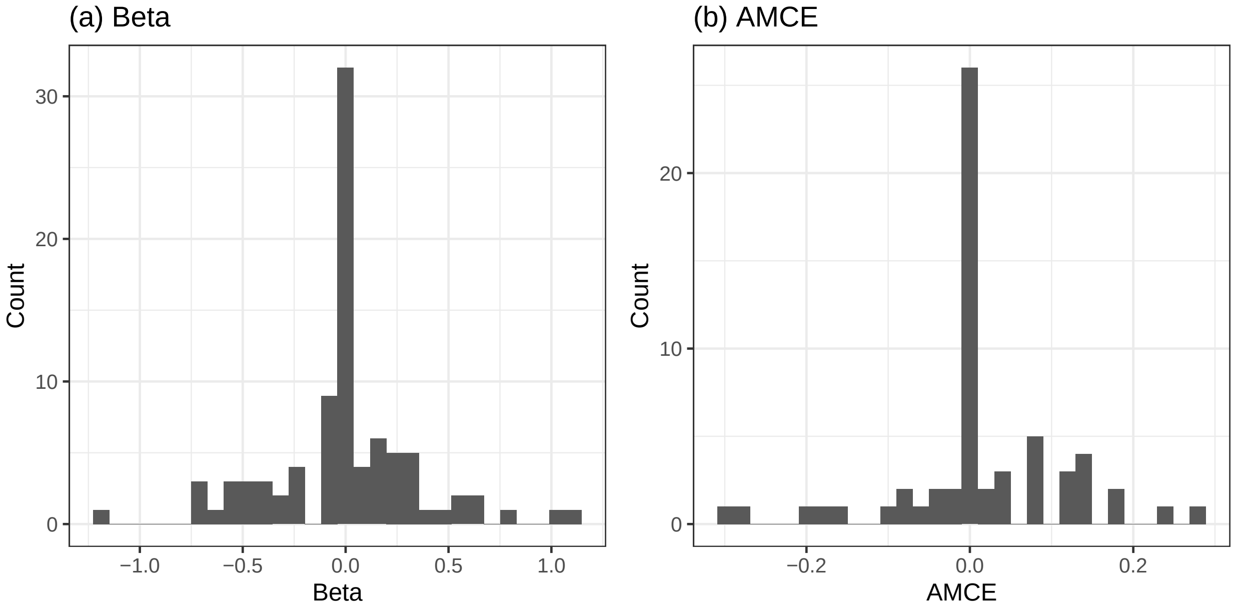


Figure A3: The distribution of parameters and AMCEs used in the simulation.

For each simulation, we draw N individuals who rate T profiles where $(N, T) \in \{(1000, 5), (2000, 10)\}$. For each individual i, we draw its moderators $\boldsymbol{x}_i$ from a correlated multivariate normal where $\boldsymbol{x}_i \sim N(\mathbf{0}_5, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = 0.25^{|i-j|}$ for $i, j \in \{1, \cdots, 5\}$. The distribution of group assignment probabilities π_{ik} is shown below from one million Monte Carlo simulation draws of $[1, \boldsymbol{x}_i^\top]$.

We see that the members are well-separated; the groups are somewhat unbalanced, i.e. $\bar{\boldsymbol{\pi}} = [0.217, 0.261, 0.522]$. If we consider the maximum probability for each person i, i.e. $\pi_i^* = \max_{k \in \{1,2,3\}} \pi_{ik}$, this distribution has a median of 0.93, a 25th percentile of 0.75 and a 75th percentile of 0.99.

In terms of simulating the treatment profiles and outcome, for each individual i, we draw a group membership Z_i using $\boldsymbol{\pi}_i$ generating using $\boldsymbol{X}_i$, $\boldsymbol{\phi}$ and Equation 4. For each task t, we then randomly draw a pair of treatments and then, given Z_i, draw the outcome Y_i given their observed treatments using the model in the main text.

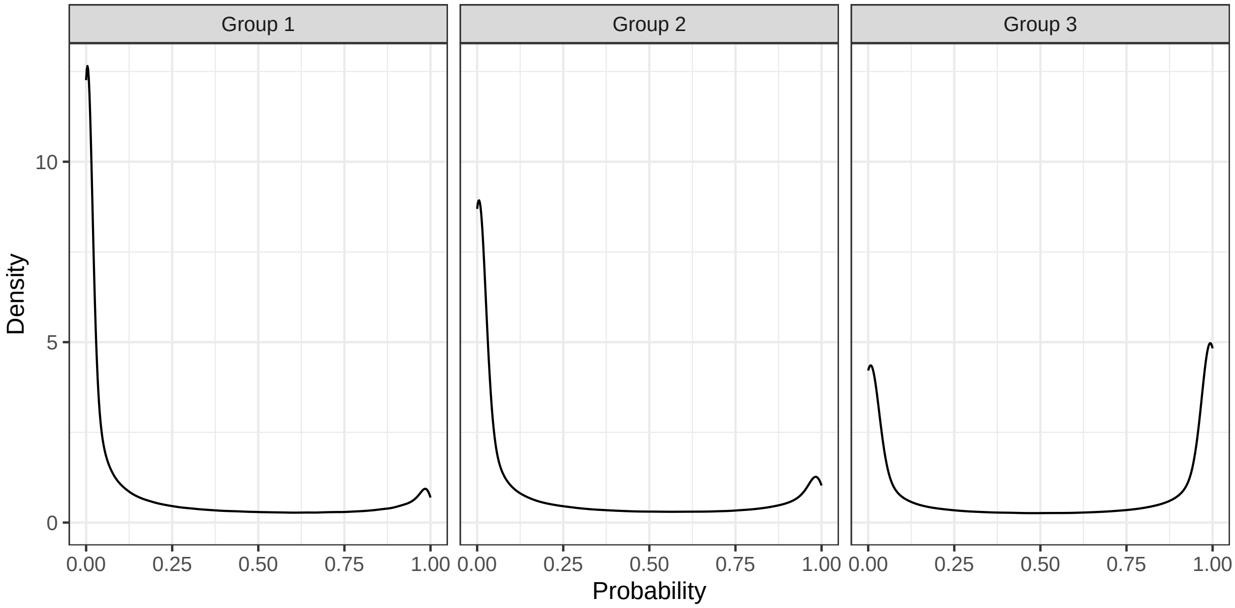
After estimating our model with $K = 3$, we resolve the problem of label switching by permuting our estimate group labels to minimize the absolute error between the estimated posterior membership probabilities $\{E[z_{ik}|\boldsymbol{\theta}]\}_{k=1}^K$ and $\boldsymbol{z}_i$ (the one-hot assignment of group membership).

K.2 Additional Results

We provide additional simulation results to complement those presented in the main text. Figure A5 presents the results for the simulations in the main text when considering the $\boldsymbol{\beta}_k$ (instead of the AMCE). It shows a similar pattern of some bias even at the larger sample size.

To address this issue, we consider an alternative procedure based on sample splitting. We fit the model using half of the data (selected at random) and then refit the model. To refit the model, we hold fixed the sparsity pattern estimated in the original estimation hold (i.e., which levels are

Figure A4: Group Membership Probabilities



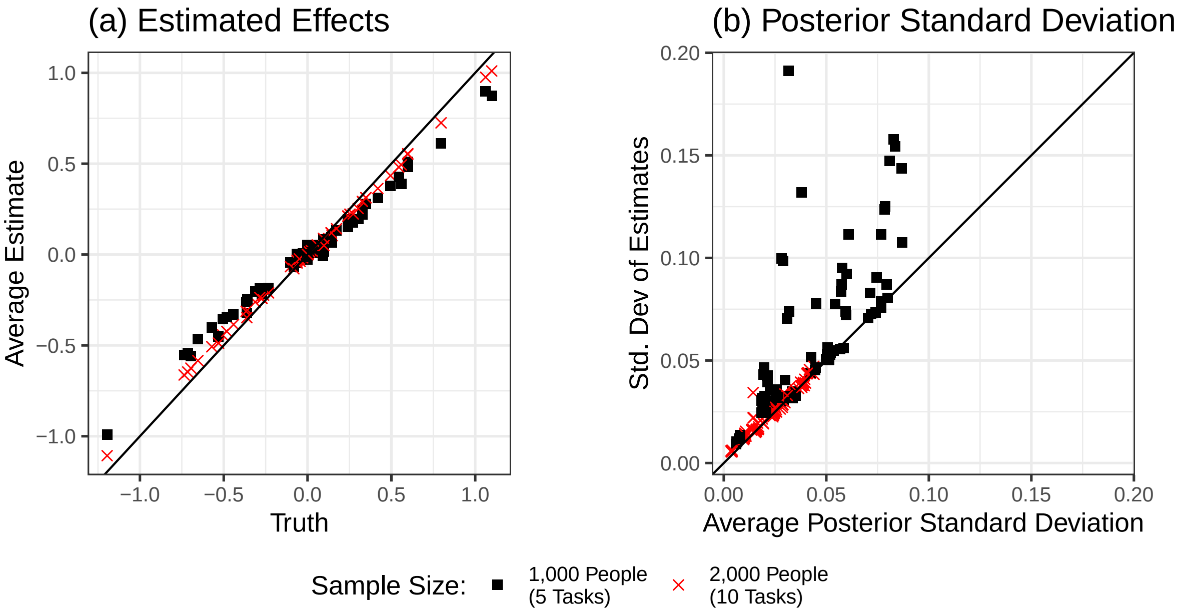


Figure A5: The empirical performance of the proposed estimator on simulated data. The black squares indicate the effects estimated in each group with the smaller sample size (1,000 people completing 5 tasks); the red crosses indicate effects estimated with the larger sample size (2,000 people completing 10 tasks).

fused together) using a tolerance of 10^{-3}. We also fix the estimated moderator relationship, i.e. $\pi_k(\boldsymbol{X}_i)$, and only estimate the treatment effect coefficients after fusion. Algorithm A2 states the procedure. To calculate the average marginal effects, as noted in Appendix I, we use the empirical distribution of treatments to marginalize over other factors. In this split version, we also use the distribution from the full dataset.

Figure A6 compares the estimators from the split sample and full data ("Full Sample", i.e. the

Algorithm A2 Refitting Procedure

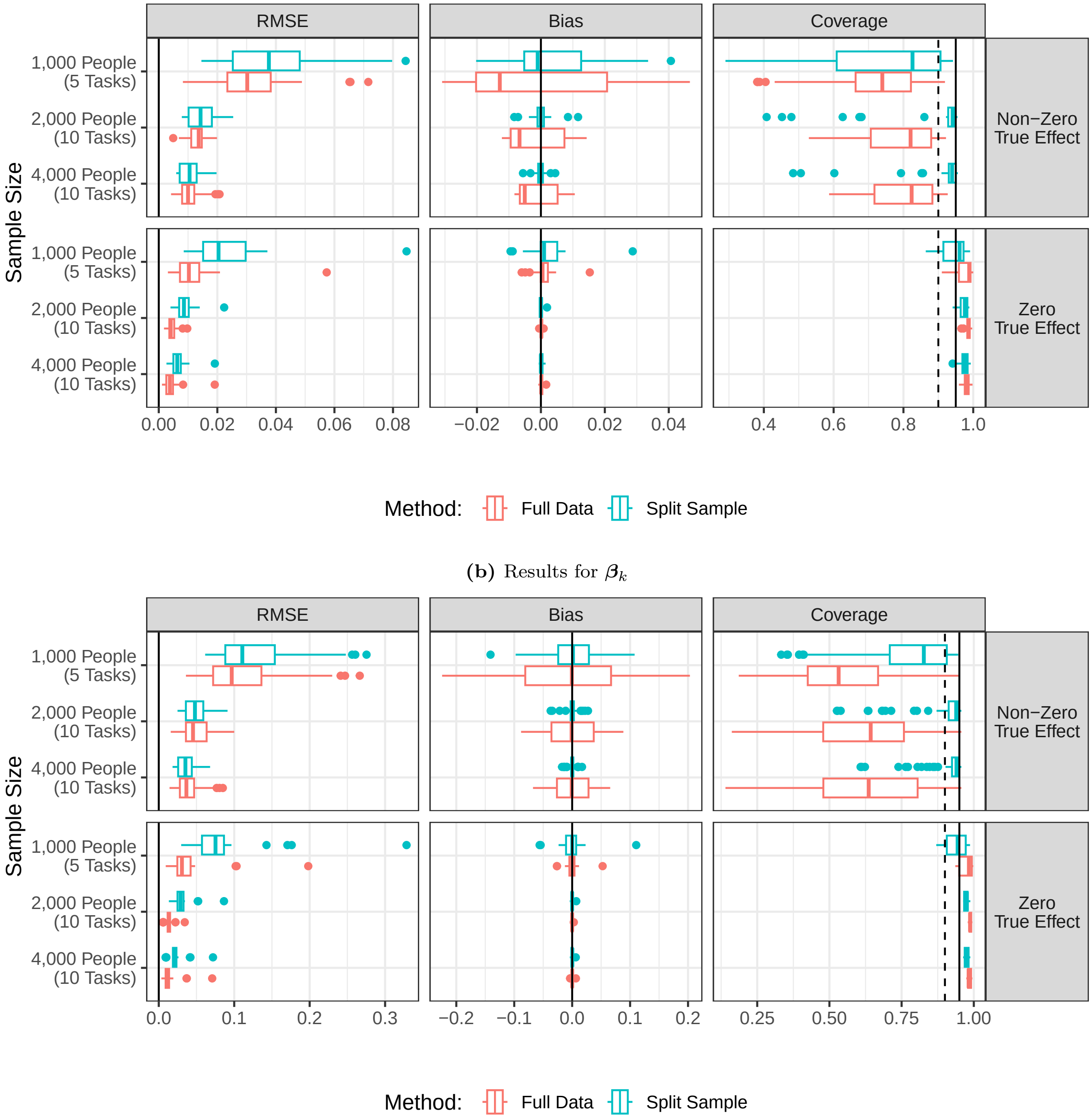
```
1. Randomly split the observations $i \in \{1, \cdots, N\}$ into two groups indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$
2. Using the data $i \in \mathcal{I}_1$, estimate the parameters of the model using Algorithm A1 in the main
   text. Define the resulting parameters from this as $\tilde{\boldsymbol{\theta}}$: $\{\tilde{\boldsymbol{\beta}}_k\}_{k=1}^K$, $\{\tilde{\boldsymbol{\phi}}_k\}_{k=2}^K$, $\tilde{\mu}$
3. Fuse levels $l$ and $l'$ of factor $j$ for group $k$ where the following condition holds for tolerance $\epsilon$

   $$\max\left\{\left|\tilde{\beta}^j_{kl} - \tilde{\beta}^j_{kl'}\right|\right\} \bigcup \left\{\bigcup_{j' \neq j} \bigcup_{m=0}^{L_{j'}-1} \left|\tilde{\beta}^{jj'}_{klm} - \tilde{\beta}^{jj'}_{kl'm}\right|\right\} \leq \epsilon$$

   For each combination where this is satisfied, construct matrices $\boldsymbol{R}_k$ that contain the required
   equality constraints, i.e. where $\boldsymbol{R}_k^T \tilde{\boldsymbol{\beta}}_k$ ensures that $\tilde{\beta}^j_{kl} = \tilde{\beta}^j_{kl'} = 0$ and/or $\tilde{\beta}^{jj'}_{klm} - \tilde{\beta}^{jj'}_{kl'm} = 0$.
   Define $\tilde{\pi}_k(\boldsymbol{X}_i)$ as follows:

   $$\tilde{\pi}_k(\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\phi}}_k)}{\sum_{k'=1}^K \exp(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\phi}}_{k'})}$$

4. Using the other half of the data $i \in \mathcal{I}_2$, estimate the refit parameters for the treatment effects,
   where $\boldsymbol{C}$ contains the original sum-to-zero constraints discussed in the main text.

   $$\{\hat{\boldsymbol{\beta}}_k^{\text{refit}}\}_{k=1}^K, \hat{\mu}^{\text{refit}} = \underset{\{\boldsymbol{\beta}_k\}_{k=1}^K,\ \mu}{\operatorname{argmax}} \sum_{i \in \mathcal{I}_2} \log\left(\sum_{k=1}^K \tilde{\pi}_k(\boldsymbol{X}_i) \zeta_k(\boldsymbol{T}_i)^{Y_i} \{1 - \zeta_k(\boldsymbol{T}_i)\}^{1-Y_i}\right) \quad \text{s.t.} \quad \boldsymbol{C}^T \boldsymbol{\beta}_k = \boldsymbol{0},\ \boldsymbol{R}_k^T \boldsymbol{\beta}_k = \boldsymbol{0}$$
```

methods shown in the main text) approaches. It shows the distribution of the root mean-squared error (RMSE), bias, and coverage across the estimated AMCE and coefficients. We split the results by whether the true underlying effect is zero to compare differences across those cases. We also consider one even larger sample size (4,000 respondents with 10 tasks) to examine a scenario where the split sample method has the same amount of data as the full sample method for the second step in the estimation process.

The figure corroborates the initial results. Specifically, the full data method has non-trivial bias that decreases slowly even at the largest sample sizes. By contrast, the bias is small in the split sample method. As the panel on coverage shows, this results in considerably better coverage—especially for quantities with a non-zero true effect. At the two larger sample sizes, the median frequentist coverage of the split sample method is close to the nominal 95%, with a few outliers that have low coverage. In terms of RMSE, the methods perform similarly.

K.3 Robustness to Misspecification

As noted in the main text, our methodology is not predicated on the assumption that the true data generating process is a mixture model. Rather, fitting a mixture model or a mixture of experts model is equivalent to finding maximally heterogeneous groups. Nevertheless, we consider a simulation setting in which the true data generating process is a mixture model. Under this assumption, we explore how the specification of different parts of the model (e.g., K and the choice of moderators) affects performance. Specifically, we explore different choices of K and misspecification of the moderator model $\pi_k(\boldsymbol{X}_i)$ from the ones used to generate the data.



(a) Results for AMCE

(b) Results for $\boldsymbol{\beta}_k$

Figure A6: The distribution of performance for each estimator across sample sizes. The top figure shows results for the AMCE; the lower figure shows results for the coefficients $\boldsymbol{\beta}_k$. Inside each figure, results are split by whether the true effect is zero ("Zero True Effect") or not ("Non-Zero True Effect"). The boxplot shows the distribution across all effects for each group. For the plots on RMSE and bias, the solid vertical line indicates zero. For coverage, the solid line indicates 95% coverage and the dashed line indicates 90%.

K.3.1 Data-Driven Choice of K

First, as noted in the main text, a common approach to choosing K can be information criterion. We use the BIC to calibrate our choice of λ, i.e. pick the λ that minimizes the BIC. In our simulations, we compare the BIC across $K \in \{1, 2, 3, 4\}$ to see which it would suggest choosing. Table A2 reports the probability of each K being chosen across 1,000 simulations. It shows that, even for the smallest data size, the BIC correctly identifies $K = 3$. The probability of correct selection rises as the sample size grows. However, as we note in the main text, this simulation example has relatively well separated clusters, and correctly specified likelihoods, and thus the information criterion approach is expected to perform well.

Sample Size	$K = 1$	$K = 2$	$K = 3$	$K = 4$
1,000 People (5 Tasks)	0	0.01	0.941	0.049
2,000 People (10 Tasks)	0	0.00	0.999	0.001
4,000 People (10 Tasks)	0	0.00	0.994	0.006

Table A2: Probability of K being chosen using smallest BIC

Other criterion based on cross-validation—e.g., splitting the sample and taking the model with the highest out-of-sample predictive likelihood or lowest RMSE—also show a high probability of choosing $K = 3$ (84% for the smallest sample size and 97-98% for the larger sample sizes).

K.3.2 Effect of Choice of K on Estimates

We first consider how different choices of K impact our results in the simulation study. To do this, we focus on the CAMCE discussed in the main text (Section 5.3) as this quantity is comparable across models with different K. For each individual i, we calculate our estimate of CAMCE using their moderators $\boldsymbol{X}_i$ and compare this against the true value, which can be calculated by plugging in the true values of $\pi_k(\boldsymbol{X}_i)$ and $\delta_{jk}(l, l')$ into Equation (11). We run models with $K \in \{2, 3, 4\}$ with both split-sample and full data methods discussed above.

Figure A7 shows a binned scatterplot of the true CAMCEs against the estimated CAMCEs for each individual i, i.e., for all true CAMCE in a bin, what is the average estimated CAMCE? As above, it shows that for the correct choice of $K = 3$, the estimates track the truth well. Interestingly, $K = 4$ also shows good performance but $K = 2$ shows some weaker performance, especially for certain ranges of the true CAMCE.

We also compute the marginalized error (i.e., the error in the estimated CAMCE vs the true CAMCE, averaged across all people and CAMCEs estimated in a simulation) and RMSE of the estimated CAMCEs. Figure A8 plots the distribution of RMSE and marginalized error across the 1000 simulations. Consistent with our earlier results, the figure shows that the full sample method for all choice of K has some non-vanishing bias while the split-sample method exhibits a considerably smaller error. Further, while the estimated error looks similar for $K \in \{2, 3, 4\}$, the correct choice ($K = 3$) has lower RMSE than either $K = 2$ or $K = 4$. The results for $K = 4$ are comparable to those for $K = 3$, but the case of $K = 2$ sees a considerably worse performance.

Next, we consider how different choices of K affect the ability to recover the average marginal effect. To do this, we average the CAMCE across all individuals used to fit the model and compare that AMCE in the population. Figure A9 plots the bias of the estimated AMCE by aggregating the individual-level effects; it is largely unaffected by the choice of K, corroborating Figure A8. As

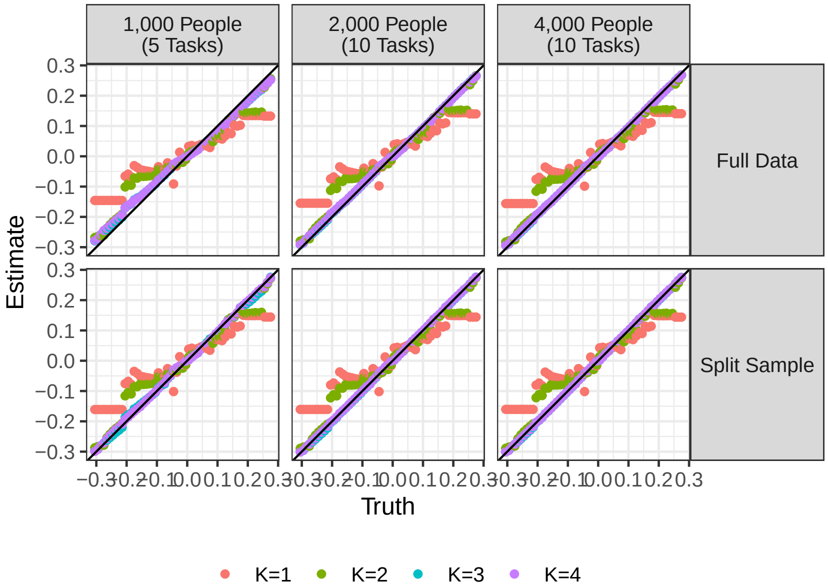


Figure A7: The binned scatterplot of the true CAMCEs versus the estimated CAMCEs. Results are shown for different sample sizes and estimation method (e.g., full data versus split sample). The color of the dot indicates the number of groups K.

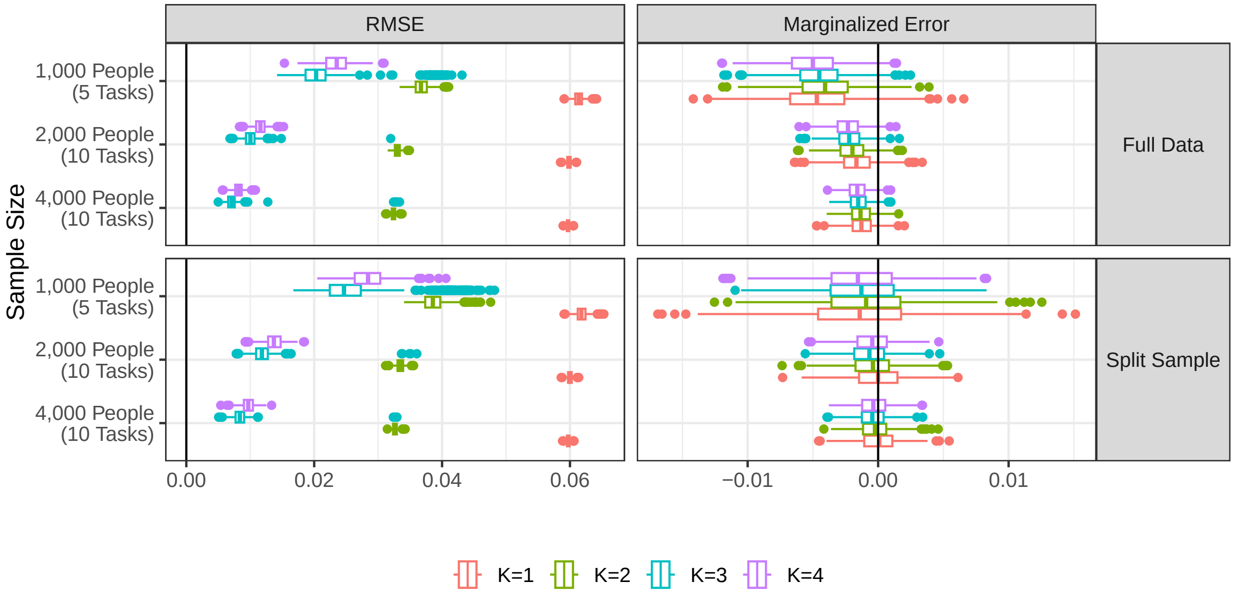


Figure A8: The distribution of performance across simulations. The top panel shows the performance in terms of RMSE and marginalized error, across all individuals and CAMCEs, for the model fit on the entire dataset. The bottom panel shows the results for a method estimated using the split sample method. The color of the boxplot indicates the number of groups K.

expected, there is regularization bias for the full data method that using the split sample approach eliminates.

As a final illustration on the choice of K, we also examine how much variability in the *true*

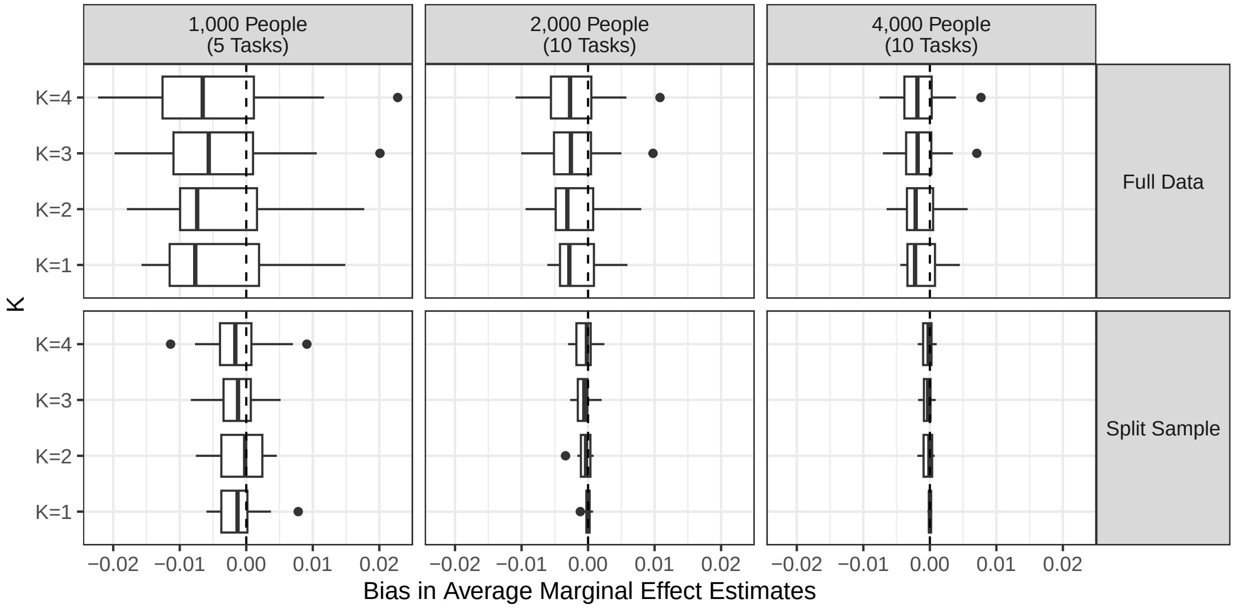


Figure A9: The distribution of bias in AMCEs by averaging CAMCEs by different K

CAMCE is explained by the estimated groups, inspired by how one might assess the quality of clustering in k-means. We compute this as follows: For each observation i, obtain its estimated group membership probabilities $\hat{\pi}_k(\boldsymbol{X}_i)$ for $k \in \{1, \cdots, K\}$. Using its true CAMCE, i.e. $\text{CAMCE}^*_j(l, l'; \boldsymbol{X}_i)$, compute the total variability in CAMCE across the N units and the between-group variability using $\hat{\pi}_k$ as group weights. Formally, we compute B_K and the total variability T.

$$B_K = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l'_j=1}^{L_j-1} N_k \left[\overline{\text{CAMCE}}^*_{k,j}(l_j, l'_j) - \overline{\text{CAMCE}}^*_j(l_j, l'_j)\right]^2; \quad N_k = \sum_{i=1}^{N} \hat{\pi}_k(\boldsymbol{X}_i);$$

$$T = \sum_{j=1}^{J} \sum_{l'_j=1}^{L_j-1} \sum_{i=1}^{N} \left[\text{CAMCE}^*_j(l_j, l'_j; \boldsymbol{X}_i) - \overline{\text{CAMCE}}^*_j(l_j, l'_j)\right]^2$$

$$\overline{\text{CAMCE}}^*_{k,j} = \frac{1}{N_k} \sum_{i=1}^{N} \hat{\pi}_k(\boldsymbol{X}_i) \cdot \text{CAMCE}^*_j(l_j, l'_j; \boldsymbol{X}_i); \quad \overline{\text{CAMCE}}^*_j = \frac{1}{N} \sum_{i=1}^{N} \text{CAMCE}^*_j(l_j, l'_j; \boldsymbol{X}_i)$$

Figure A10 reports the ratio of the between-group variability over the total variability across the 1,000 simulations for $K \in \{2, 3, 4\}$. With $K = 2$, we already able to explain around 50% of the variability in the data. As expected, $K = 2$ shows considerably lower B_K/T than higher K's, suggesting its groups are less distinct—or, equivalently, more internally heterogeneous—than $K \in \{3, 4\}$. There is limited improvement in quality with $K = 4$, which is consistent with the earlier results that the correct choice ($K = 3$) adequately summarizes the variability in the data.

K.3.3 Misspecified Moderators

We next consider how misspecifying the model for the moderators $\pi_k(\boldsymbol{X}_i)$ affects our simulated results. We show this in two ways; first, we fit a model with no moderators, that is, $\boldsymbol{X}_i = 1$. While this model has a number of limitations—e.g., for classifying and predicting heterogeneous effects

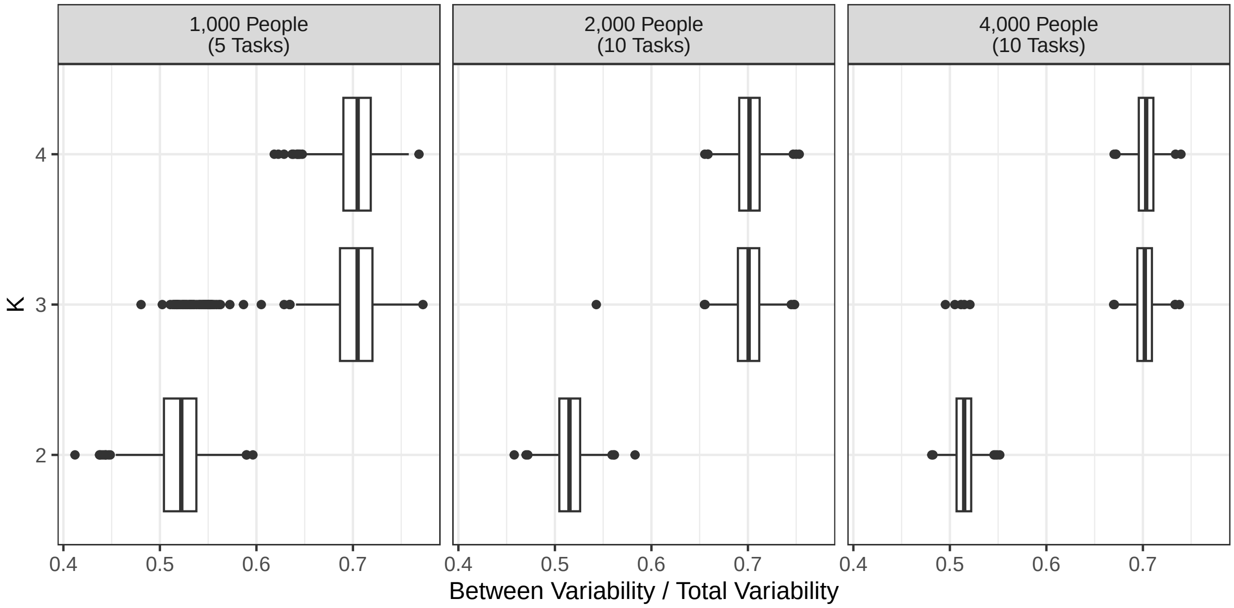


Figure A10: The distribution of B_K/T across simulations. The top panel shows results for the model fit on the entire dataset. The bottom panel shows the results for a method estimated using the split sample method.

for new individuals, it is a useful benchmark. Second, instead of using the true moderators (e.g., $\boldsymbol{X}_i$), we assume the researcher only has available the following non-linear transformations of the moderators (following Kang and Schafer 2007) and uses those instead:

$$\begin{aligned}
\boldsymbol{A}_{i,1} &= \sqrt{3}\exp(\boldsymbol{X}_{i,1}/2) - 2 \\
\boldsymbol{A}_{i,2} &= \sqrt{3}\boldsymbol{X}_{i,2}/\left[1 + \exp(\boldsymbol{X}_{i,1})\right] \\
\boldsymbol{A}_{i,3} &= 1/19\left[\boldsymbol{X}_{i,1} + \boldsymbol{X}_{i,3} + 0.6\right]^3 \\
\boldsymbol{A}_{i,4} &= 1/3\left[\boldsymbol{X}_{i,2} + \boldsymbol{X}_{i,4}\right]^2 - 1 \\
\boldsymbol{A}_{i,5} &= 2.5\sqrt{|\boldsymbol{X}_{i,5} + \boldsymbol{X}_{i,1}|} - 2.5.
\end{aligned}$$

We rescale the moderators $\{\boldsymbol{A}_i\}_{i=1}^N$ to have zero mean and unit variance in each simulated dataset.

Figure A11 replicates Figure A6 on the performance on estimating the AMCE where we show results with all moderators (i.e., in Figure A6) and with both types of mis-specification ("No Moderators" and "Non-Linear Transf." when $\boldsymbol{A}_i$ are used).

It shows that, for the smallest sample size, the no-moderator model incurs a penalty in terms of the RMSE of the estimated AMCEs, although it does not have considerably larger bias. At larger sample sizes, the difference between the moderator and no-moderator models decreases. With moderators that are included but mis-specified using some non-linear transformation, the performance is rather close to the one that uses the correct moderators.

To further illustrate the impact of excluding moderators, Figure A12 plots the estimated average posterior and posterior predictive probability (i.e., $\hat{\pi}_k(\boldsymbol{X}_i)$) in the group corresponding to the individual's sampled Z_i for all observations in the estimation data. It shows, as expected, that using the correctly specified moderators results in a considerably higher probability of each individual being assigned to group that corresponds to their sampled Z_i. The model with included but mis-

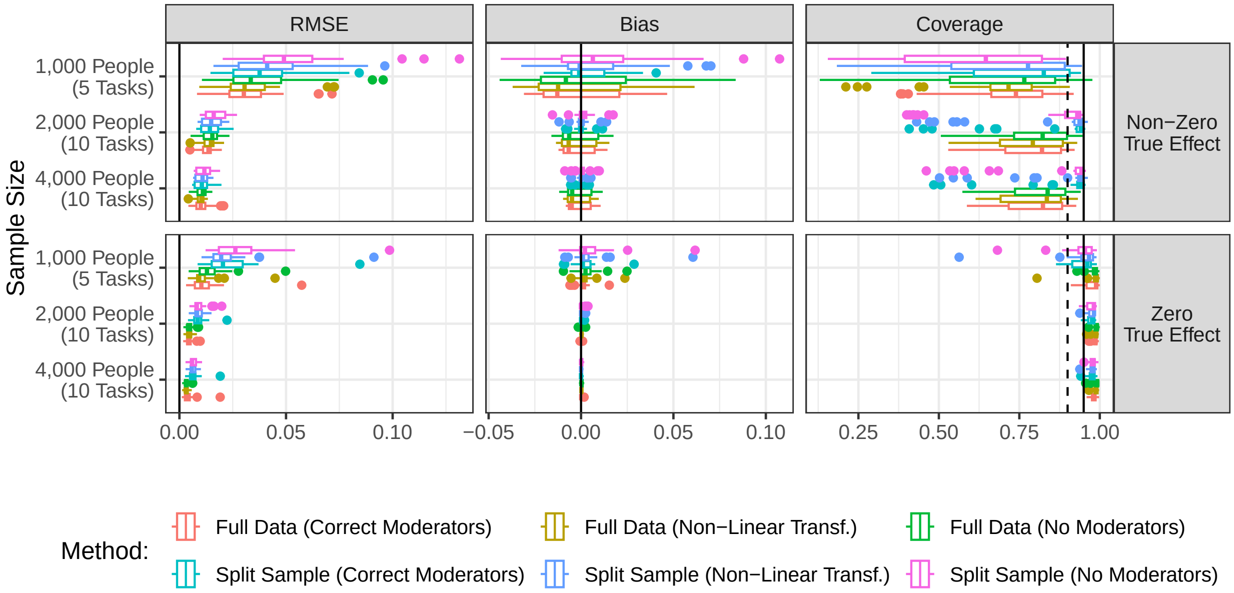


Figure A11: The distribution of performance for each estimator across sample sizes, with and without moderators. Inside each figure, results are split by whether the true effect is zero ("Zero True Effect") or not ("Non-Zero True Effect"). The boxplot shows the distribution across all effects for each group. For the plots on RMSE and bias, the solid vertical line indicates zero. For coverage, the solid line indicates 95% coverage and the dashed line indicates 90%.

specified moderators ("Non-Linear Transf.") is somewhere between the model without moderators and the correctly specified one.

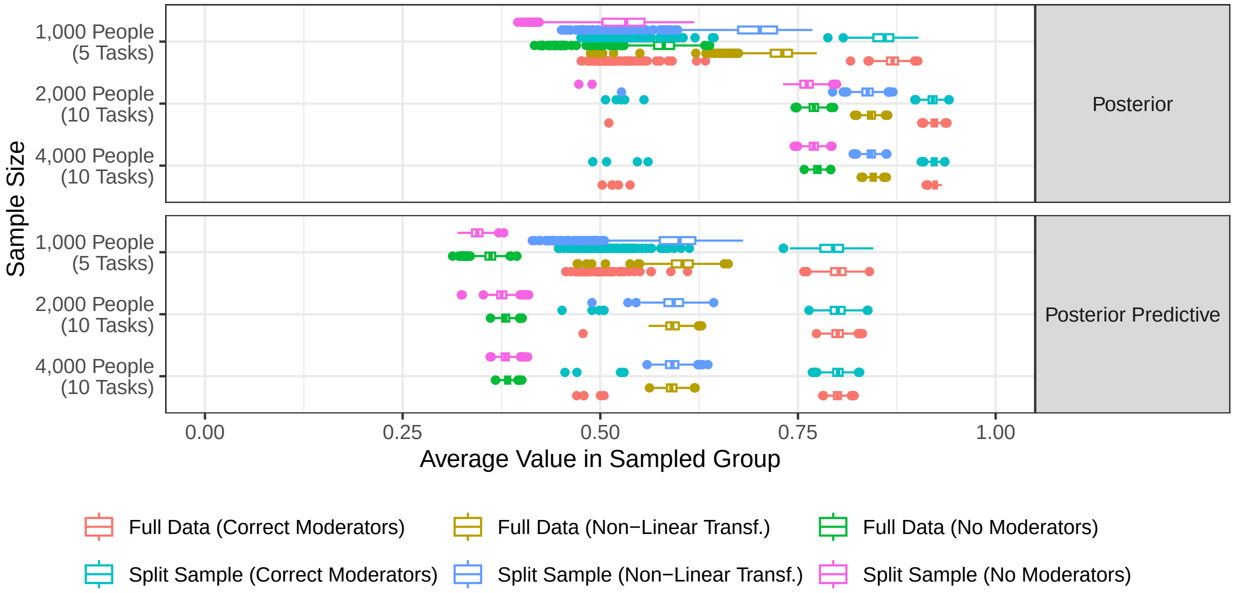


Figure A12: The average probability that is assigned to the group corresponding to an individual's sampled Z_i, showing the distributions across simulations.

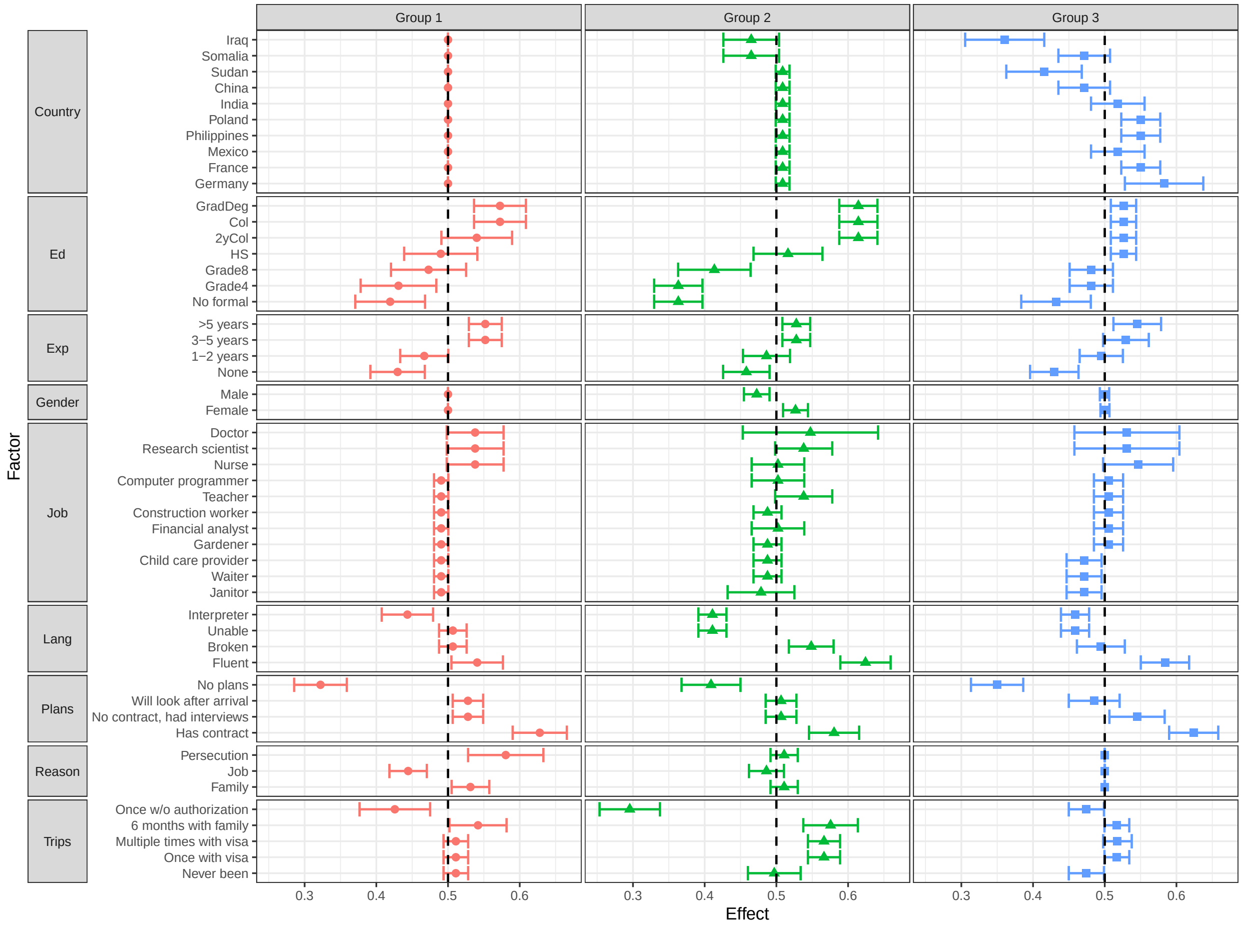


Figure A13: Estimated average marginal means using a three-group (right) analysis. The point estimates and 95% Bayesian credible intervals are shown.

L Additional Results for Immigration Conjoint Experiment

We provide some additional results for our main empirical analysis. First, focusing on the three-group model, we report a different quantity of interest. We use an analogue to the "marginal means" estimator in Leeper, Hobolt and Tilley (2020). We compute the probability of a profile being chosen *without* specifying a baseline category. The equation is shown below for the forced choice case; note it consists of two of the terms used for the AMCE.

$$\text{MM}_{jk}(l) \;=\; \frac{1}{2}\mathbb{E}\left[\left\{\Pr\left(Y_i = 1 \mid Z_i = k, T^L_{ij} = l, \boldsymbol{T}^L_{i,-j}, \boldsymbol{T}^R_i\right) + \Pr\left(Y_i = 0 \mid Z_i = k, T^R_{ij} = l, \boldsymbol{T}^R_{i,-j}, \boldsymbol{T}^L_i\right)\right\}\right]. \tag{A22}$$

The below plot ignores randomization restrictions when estimating this quantity to center the estimate around 0.50 as in Leeper, Hobolt and Tilley (2020). The results are substantively similar to the analysis in shown in the main paper using AMCEs.

Second, as noted in the main text, we found that sample splitting and refitting the model (see Appendix K.2) was somewhat unstable given different splits of the data. To illustrate this point, Figure A14 shows the 25th-75th percentile (and median) of the AMCEs estimated across twenty repetitions of splitting the data into halves and then using the refitting procedure described above.

We address the problem of label switching using a permutation of labels that minimizes the average mean absolute error between all pairs of estimates; we find a permutation by randomly permuting the labels for a randomly chosen set of estimates and repeat this repeatedly until the average mean absolute error stabilizes.

While Figure A14 shows instability in some of the estimated AMCE, it broadly shows a similar result to that in the main text. For example, one group (Group 2 when $K = 2$; Group 3 when $K = 3$) shows a clear effect of country across most splits whereas one group (Group 1 when $K = 2$ and Groups 1 and 2 when $K = 2$) generally shows a large penalty for immigrants who entered without legal authorization.

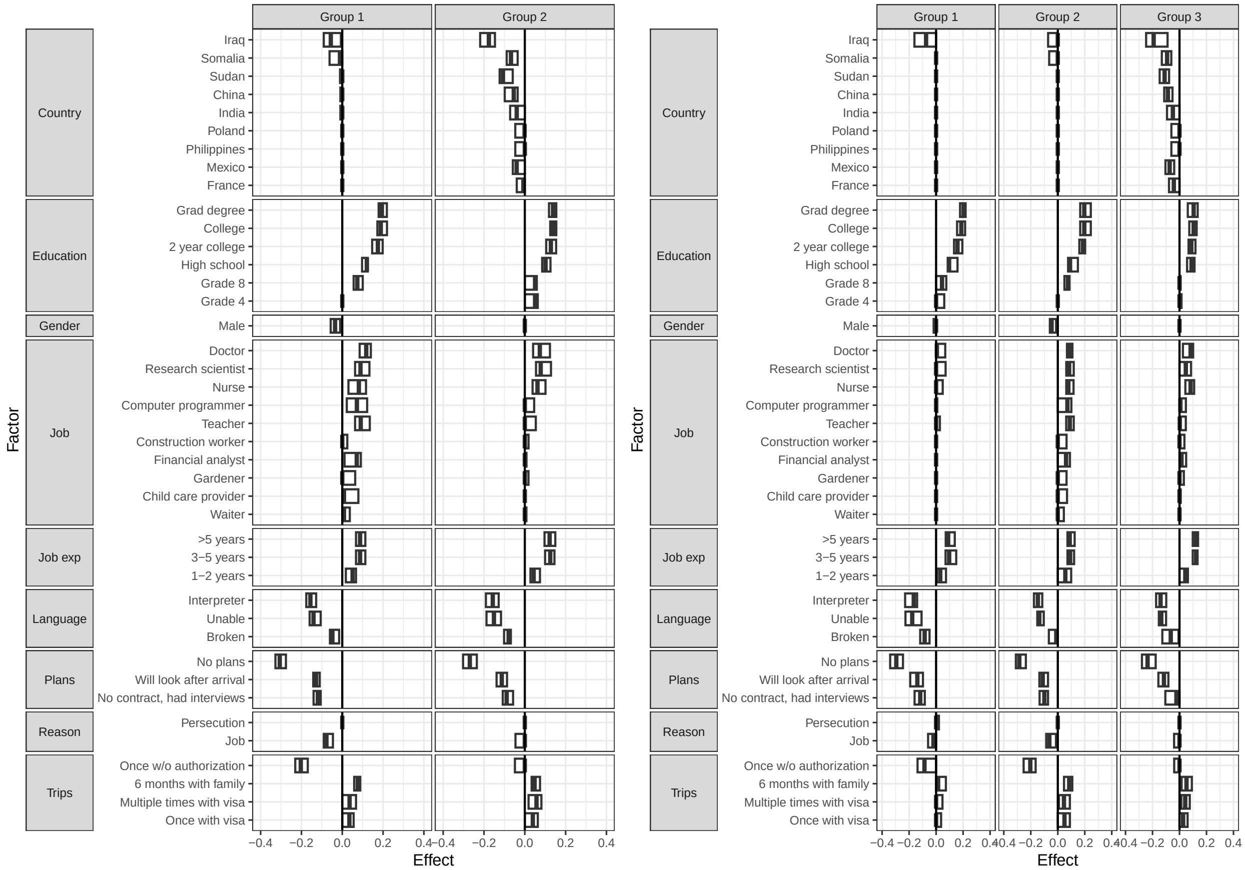


Figure A14: The distribution of AMCE from a two-group and three- model with twenty random splits of the data. The interquartile range and median are shown.

Third, Figure 5 in the main text reports the average effect of changing some moderator from x_0 to x_1 on π_k, i.e.,

$$\mathbb{E}\left[\pi_k(X_{ij} = x_1, \boldsymbol{X}_{i,-j}) - \pi_k(X_{ij} = x_0, \boldsymbol{X}_{i,-j})\right]. \tag{A23}$$

Figure A15 considers the impact on the average *absolute* distance, i.e.

$$\mathbb{E}\left[\left|\pi_k(X_{ij} = x_1, \boldsymbol{X}_{i,-j}) - \pi_k(X_{ij} = x_0, \boldsymbol{X}_{i,-j})\right|\right], \tag{A24}$$

to prevent positive and negative changes from canceling each other out. To interpret this quantity, Figure A15 also the absolute value of the difference reported in the main text, i.e., the absolute value of Equation A23 in a red $*$. Uncertainty is computed by drawing samples from the estimated asymptotic distribution of $\hat{\boldsymbol{\phi}}$, evaluating Equation A24 over those samples, and reporting the mean and $[0.025, 0.975]$ percentile interval. Figure A15 shows that, for certain groups, some covariates show a small average effect but a larger average of absolute effects (e.g., with $K = 3$, Group 2 and "Not Strong Republican" versus the baseline of "Strong Republican") .

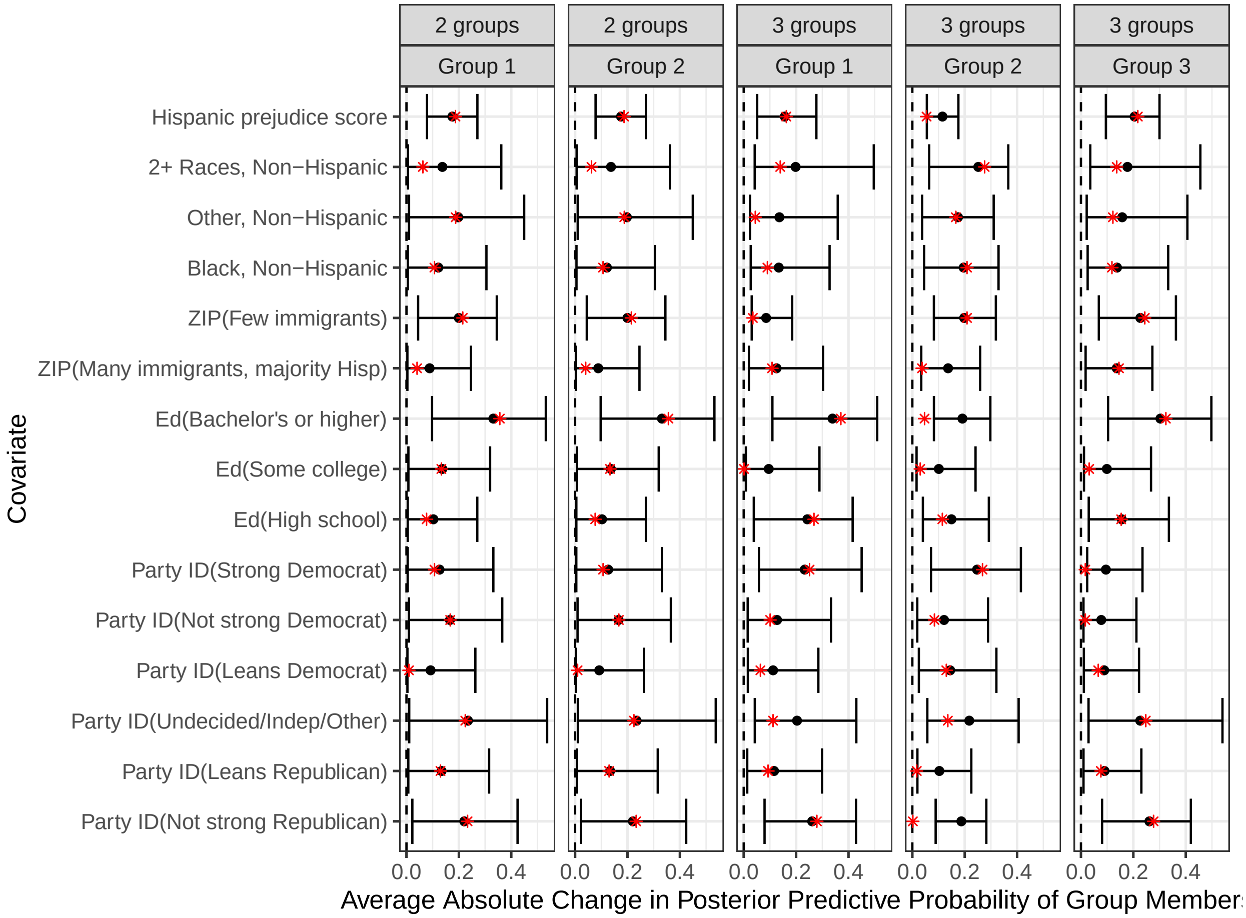


Figure A15: The average absolute effect of changing a moderator. The 2.5% to 97.5% percentile interval is shown.

Next, we discuss the two-factor interactions. The largest average marginal interaction effect (AMIE) was found between education and job in the three group analysis. This is visualized in Figure A16. The largest AMIE occurs between the levels of Teacher and High School and has magnitude of 0.0021.

Compared in magnitude to the AME, which for education was on average 0.111 and for job was on average 0.0237, this is clearly negligible. Given this, we have little hope of finding substantial higher-order interactions in this example.

If higher-order interactions were of interest, a pre-processing step to do some basic screening (see, e.g., Shi, Wang and Ding, 2023) might be implemented on the full dataset to a priori reduce the number of interactions considered. The sparsity inducing penalties of our method would then impose additional regularization.

Finally, we briefly remark upon choosing K using an information criterion. While this works well in the simulated example (see Appendix K.3.1), we find less clear results on the full data. Table A3 the results of optimizing the BIC over λ for $K \in \{1, 2, 3, 4\}$ as well as optimizing the AIC

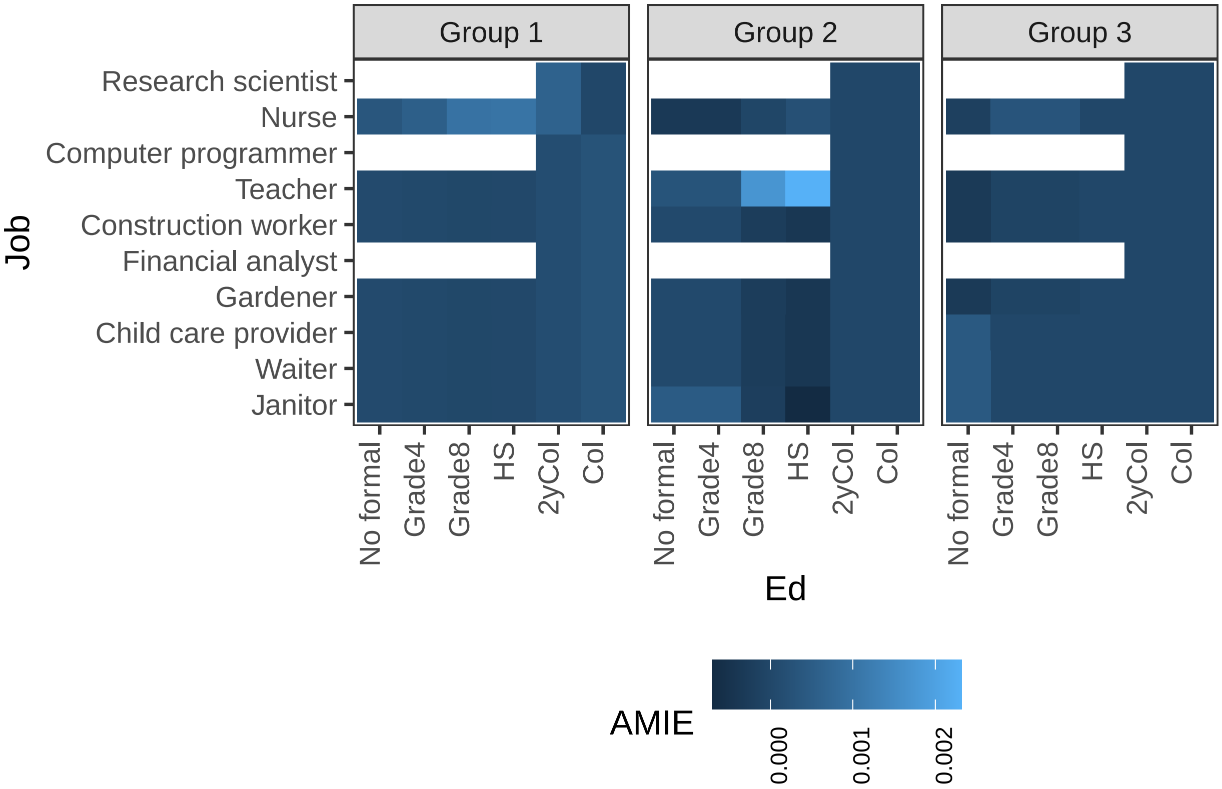


Figure A16: The average marginal interaction effect between education and job.

over λ. It shows that, if one uses the BIC, this suggests $K = 1$. However, if one uses the AIC, this suggests $K = 4$.

Optimizing BIC over λ			
$K = 1$	$K = 2$	$K = 3$	$K = 4$
6125	6270	6391	6529
Optimizing AIC over λ			
$K = 1$	$K = 2$	$K = 3$	$K = 4$
5968	5902	5871	5833

Table A3: Information criterion for different K

References

Bischl, Bernd, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas and Michel Lang. 2018. "mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions." *Working paper available at* `https://arxiv.org/pdf/1703.03373.pdf`.

Bondell, Howard D. and Brian J. Reich. 2009. "Simultaneous Factor Selection and Collapsing Levels in ANOVA." *Biometrics* 65:169–177.

Bryant, Peter and John A. Williamson. 1978. "Asymptotic Behaviour of Classification Maximum Likelihood Estimates." *Biometrika* 65:273–281.

Celeux, Gilles and Gérard Govaert. 1992. "A Classification EM Algorithm for Clustering and Two Stochastic Versions." *Computational Statistics & Data Analysis* 14:315–332.

Chamroukhi, Faicel and Bao-Tuyen Huynh. 2019. "Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models." *Journal de la Société Française de Statistique* 160:57–85.

Chi, Jocelyn T., Eric C. Chi and Richard G. Baraniuk. 2016. "k-POD: A Method for k-Means Clustering of Missing Data." *The American Statistician* 70:91–99.

Egami, Naoki and Kosuke Imai. 2019. "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis." *Journal of the American Statistical Association* 114:529–540.

Everitt, Brian S., Sabine Landau, Morven Leese and Daniel Stahl. 2011. *Cluster Analysis*. 5th edition ed. John Wiley & Sons.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96:1348–1360.

Figueiredo, Mário A.T. 2003. "Adaptive Sparseness for Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25:1150–1159.

Goplerud, Max. 2021. "Modelling Heterogeneity Using Bayesian Structured Sparsity." *Working paper available at* `https://arxiv.org/pdf/2103.15919.pdf`.

Hainmueller, Jens and Daniel J Hopkins. 2015. “The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants.” *American Journal of Political Science* 59:529–548.

Hastie, Trevor. 1987. “A Closer Look at the Deviance.” *The American Statistician* 41:16–20.

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.” *Statistical Science* 22:523–539.

Khalili, Abbas. 2010. “New Estimation and Feature Selection Methods in Mixture-of-Experts Models.” *Canadian Journal of Statistics* 38:519–539.

Lawson, Charles L. and Richard J. Hanson. 1974. *Solving Least Squares Problems*. Prentice-Hall.

Leeper, Thomas J, Sara B Hobolt and James Tilley. 2020. “Measuring Subgroup Preferences in Conjoint Experiments.” *Political Analysis* 28:207–221.

Lim, Michael and Trevor Hastie. 2015. “Learning Interactions via Hierarchical Group-Lasso Regularization.” *Journal of Computational and Graphical Statistics* 24:627–654.

Louis, Thomas A. 1982. “Finding the Observed Information Matrix When Using the EM Algorithm.” *Journal of the Royal Statistical Society, Series B, Methodological* 44:226–233.

Murphy, Keefe and Thomas Brendan Murphy. 2020. “Gaussian Parsimonious Clustering Models with Covariates and a Noise Component.” *Advances in Data Analysis and Classification* 14:293–325.

Oelker, Margret-Ruth and Gerhard Tutz. 2017. “A Uniform Framework for the Combination of Penalties in Generalized Structured Models.” *Advances in Data Analysis and Classification* 11:97–120.

Polson, Nicholas G., James G. Scott and Jesse Windle. 2013. “Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables.” *Journal of the American Statistical Association* 108:1339–1349.

Polson, Nicholas G. and Steve L. Scott. 2011. “Data Augmentation for Support Vector Machines.” *Bayesian Analysis* 6:1–24.

Ratkovic, Marc and Dustin Tingley. 2017. “Sparse Estimation and Uncertainty with Application to Subgroup Analysis.” *Political Analysis* 25:1–40.

Robinson, Thomas and Raymond Duch. 2023. *cjbart: Heterogeneous Effects Analysis of Conjoint Experiments*. R package version 0.3.2.
URL: *https://CRAN.R-project.org/package=cjbart*

Robinson, Thomas S and Raymond M Duch. 2024. “How to Detect Heterogeneity in Conjoint Experiments.” *The Journal of Politics* 86:412–427.

Shi, Lei, Jingshen Wang and Peng Ding. 2023. “Forward screening and post-screening inference in factorial designs.” *arXiv preprint arXiv:2301.12045*.

Varadhan, Ravi and Christophe Roland. 2008. “Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm.” *Scandinavian Journal of Statistics* 35:335–353.

Yan, Xiaohan and Jacob Bien. 2017. “Hierarchical Sparse Modeling: A Choice of Two Group Lasso Formulations.” *Statistical Science* 32:531–560.

Zou, Hui. 2006. “The Adaptive LASSO and Its Oracle Properties.” *Journal of the American Statistical Association* 101:1418–1429.