Online Appendix
“How to detect heterogeneity in conjoint experiments"

Thomas S. Robinson[*] Raymond M. Duch[†]

Appendix Sections

[*] Department of Methodology, London School of Economics and Political Science. Email: t.robinson7@lse.ac.uk

[†] Nuffield College, University of Oxford. Email: raymond.duch@nuffield.ox.ac.uk. Phone: +44 (0)1865 278515

A Further information on estimands and estimates

Table A1 shows how the estimands relate to the structure of conjoint datasets. Each estimand is a nested quantity that relates to the structure of the observed data collected via conjoint designs. As such, each estimand covers increasingly aggregate portions of the data.

Table A1. Nested causal quantities in a conjoint experiment

Subject	Round	Profile	Attribute	…	y	$y_{l'}$				
1	1	1	A	…	1	0	OMCE	RMCE	IMCE	AMCE
1	1	2	B	…	0	1				
1	2	1	A	…	0	0				
1	2	2	A	…	1	0				
⋮	⋮	⋮	⋮	⋱	⋮	⋮				
N	2	1	B	…	0	1				
N	2	2	A	…	1	1				

The above example reflects the structure of observations in the data collected from a conjoint experiment where the lth attribute has two possible levels (“A" and “B"). y is the observed forced-choice outcome in the experiment. $y_{l'}$ is the counterfactual *unobserved* outcome where the lth attribute is switched. The various causal estimands relate to different nested sets of observations within the data.

Relaxing the assumption of complete randomisation The main paper specifies the potential outcomes under the assumption of complete randomisation. Statistically, this assumption means that every possible combination of values across attributes is equally likely and there are no prohibited combinations. Not only is this assumption satisfied in many applications, but it also considerably simplifies the estimation. In some scenarios, however, researchers may impose restrictions to prevent implausible combinations of attributes. For example, if each profile is a political campaign, the average donation to a campaign could not exceed the total amount of donations.

In these cases, as shown by Hainmueller et al. (2014), the AMCE estimand must condition on the possibility that the remainder of the treated profile and the vector of other possible treatment options are in the intersection of the supports ($\mathcal{T}$) of $p(T_{ijk[-l]} = t, \boldsymbol{T}_{i[-j]k} = \boldsymbol{t}|T_{ikl} = l_1)$ and $p(T_{ijk[-l]} = t, \boldsymbol{T}_{i[-j]k} = \boldsymbol{t}|T_{ikl} = l_0)$, where t is the vector of all other attribute values for the jth profile in round k, and $\boldsymbol{t}$ is the set of possible vectors of all attributes in the other profile.

In our framework, by relaxing this assumption, the IMCE estimand becomes:[1]

$$\tau_{il} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, \cdots) - Y_{ijk}(t_l = l_0, \cdots)|(T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) \in \tilde{\mathcal{T}}, \mathcal{S}_i\big],$$

the RMCE becomes:

$$\tau_{ikl} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, \cdots) - Y_{ijk}(t_l = l_0, \cdots)|(T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) \in \tilde{\mathcal{T}}, \mathcal{S}_i, \mathcal{R}_{ik}\big],$$

and the OMCE becomes:

$$\tau_{ijkl} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, \cdots) - Y_{ijk}(t_l = l_0, \cdots)|(T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) \in \tilde{\mathcal{T}}, \mathcal{S}_i, \mathcal{R}_{ik}, \mathcal{P}_{ijk}\big].$$

This logic follows from the fact that these quantities are conditional variants of the AMCE, which itself is conditioned on the joint support of the probabilities of the two conditional potential outcomes.

B Further information on the BART estimation strategy

As we note in the main text, Bayesian Additive Regression Trees (BART) are a tree-based machine learning strategy for prediction and classification, developed by Chipman et al. (2010). In this section, we provide a more detailed explanation of the algorithm for interested readers.

The underlying principal of BART is that the outcome of interest y can be decomposed

[1] For the sake of notational simplicity, we replace $T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}$ in each of the potential outcomes with "$\cdots$".

into smaller parts. Therefore, an individual outcome y_i can be described as a function of covariates $\boldsymbol{x}_i$ such that,

$$y_i = f(\boldsymbol{x}_i) \approx \sum_{t=1}^{T} g_t(\boldsymbol{x}_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

where t indexes a set of functions g_t that in summation approximate the true data-generating function f.

In the BART model, each g_t is a tree-model, where the input data is recursively subset using a series of splitting criteria. We call each point where the data is split into two subsets a non-terminal node. Each non-terminal node has two child nodes, which may themselves either be non-terminal (i.e. they split the data again) or terminal. A terminal node represents a final subset of the data, determined by the conjunction of splitting rules of its ancestors.

The Bayesian aspect of these tree models comes from the fact the model assumes a prior over the structure of *each* tree (i.e. the number, position, and splitting criteria of non-terminal nodes), the terminal node parameters themselves, and an independent error variance prior. With regards to the tree structure, for example, whether any given node is non-terminal is determined by the prior probability,

$$\alpha(1 + d)^{-\beta}, \quad \alpha \in (0, 1), \beta \in [0, \infty),$$

where α and β are hyperparameters that can be specified by the researcher. The default values set by Chipman et al. (2010) ($\alpha = 0.95, \beta = 2$) are designed to heavily constrain each tree so they are small, which helps prevent the model from overfitting (Hill et al. 2020).

The terminal-node parameters differ substantially from regular tree-based methods. Unlike in conventional trees where the terminal node parameters of the tree are simply the conditional expectations of the observations in that partition, in a BART model these parameters are defined as random variables. In particular, the prior for each leaf node (i) in tree (j) is defined as:

$$\mu_{ij} = \mathcal{N}(0, \sigma_\mu^2), \quad \text{where} \sigma_\mu = 0.5/k\sqrt{m},$$

where m is the number of trees in the model and k is a hyperparameter choice of the researcher – Chipman et al. (2010) recommend a default value of 2, on the basis of cross-validation evidence.

Finally, the error variance prior is drawn from an inverse-gamma distribution, with a λ parameter set using the data, to give a 90% (default) chance that the model will yield a root mean squared error (RMSE) value lower than from an OLS regression.

There are, as a result of this prior specification, several hyperparameters that can be specified by the researcher. As several authors note, the cross-validation exercises and resultant default parameters provided by Chipman et al. (2010) are known to perform well across a variety of contexts (Kapelner and Bleich 2016; Carnegie and Wu 2019; Hill et al. 2020). That said, researchers can perform cross-validation of these parameters on their specific dataset to see if they can achieve better performance.[2]

Since we sum these individual models, we do not want the models to predict the same part of the variance of the outcome. Using the metaphor of a forest, we do not want the canopy of the trees to overlap. Instead, each tree should “develop" (by growing or shrinking) to cover only that part of the forest canopy not covered by the remaining trees in the forest. During training, therefore, the algorithm sequentially updates each individual tree model, conditional on the current performance of the rest of the trees. Specifically, for each tree t, the model first calculates the “residual variance" (R_t) or the portion of the

[2] The **cjbart** package allows users to pass specific hyperparameter arguments (see Sparapani et al. 2021) to the underlying BART algorithm via the `cjbart(...)` function.

variance in y that is not explained by the remaining $T-1$ trees:

$$R_t = y - \sum_{j \neq t} f_j(\boldsymbol{x}).$$

The algorithm then updates the structure of tree t in an attempt to improve performance over R_t. To do so, the algorithm probabilistically makes one of the following changes: splits a terminal node (p=0.25), removes the child nodes of a non-terminal node (p=0.25), swaps split criteria across two non-terminal nodes (p=0.1), or alters the splitting criteria for a single non-terminal node (p=0.4). Once a change has been made, the model decides whether to keep this change using the Metropolis Hastings MCMC algorithm.[3]

This process is then repeated for every other tree in the model, sequentially, and finally the model updates the error variance of the model as a whole (σ) (Kapelner and Bleich 2016). This entire process is repeated k times, as defined by the researcher. As Chipman et al. (2010) note, since BART only updates one tree at a time, and in sequence, it is only ever making small changes to the overall prediction, allowing it to fine tune its performance via small additions and subtractions.

Post-training, predictions are made by taking draws from the model posterior. In practise, a “draw" is simply the result of passing a covariate vector $\boldsymbol{x}_i$ down each tree in the BART model and summing the results. More formally, a single draw from the trained BART model can be denoted:

$$\hat{y}_i^{(b)} = \sum_{t=1}^{T} \hat{g}_t(\boldsymbol{x}_i),$$

where the superscript notation indicates the bth draw from the trained BART model, and $\hat{g}_t$ is the final tth tree-model optimised via the training algorithm discussed above.

As Chipman et al. (2010) show, with sufficient training, the BART model will converge

[3] Note that this acceptance decision is constrained by R_j but also by the prior state of the tree being updated, and hence is regularized by the initial priors over the tree structures.

on the posterior distribution of the true data-generating function. Recall that since the parameters of the model are random variables, repeated draws using the same covariate vector will yield different predicted values. Therefore, to generate the final prediction $\hat{y}_i$, we can repeat this process B times to get a posterior distribution of predictions (typically 1000) and then take the average:

$$\hat{y}_i = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_i^{(b)},$$

The set of posterior draws, moreover, can be used to quantify the uncertainty of the estimate, as discussed in Section 2.1 of the main paper.

C Simulation protocols and further details

C1 IMCE prediction

To test the accuracy of the IMCE predictions, we simulate datasets with two binary attributes where the IMCE is defined with respect to a series of covariates, and across simulations we vary the relationship between these covariates and the IMCE. Since we wish to benchmark the performance of the model against "known" IMCE values for an attribute, which crucially is not the change in probability of choosing one profile over *another* profile, in this simulation exercise we assume independence between all observations. This is very similar to the assumptions made in a conventional conjoint experiment, from which the AMCE (and as we argue IMCE) are recovered. Hard-coding this independence into the data-generating process allows for better control over the size and shape of heterogeneity.

To illustrate this strategy, suppose we observe two covariates – c_1 and c_2 – that are invariant at the individual-level, and randomly assign to each observation two dichotomous attributes. The first attribute X_1 takes values a or b, and the effect of being presented b over a is the difference between the two individual-level covariates (i.e. $\tau_{X_1} = c_1 - c_2$). In other words, the marginal component effect of b is heterogeneous, and dependent on individual-level characteristics. The second attribute X_2 takes values c or d, and the marginal effect of d over c is invariant across individuals. Taken together, we get the following schedule of IMCEs:

Table C1. Hypothetical correlation between IMCEs and two covariate values: c_1 and c_2 are randomly drawn from uniform distributions

i	c_1	c_2	τ_{X_1}	τ_{X_2}
1	0.1	0	0.1	0.1
2	0.25	0.05	0.2	0.1
3	0.15	0.15	0	0.1
⋮	⋮	⋮	⋮	⋮
I	0.05	0.25	-0.2	0.1

We can then generate an assignment schedule by sampling at random the attribute levels for $I \times J$ observations i.e. attribute-level assignments across J rounds of the experiment on I individuals. Note here that, since we pre-define the IMCEs, we do not sample two observations per round – since, the IMCE does not reflect the probability of choosing one profile over another.

Suppose the probability of choosing the profile is calculated as:

$$P(Y_{ijk} = 1) = 0.5 + \mathbb{I}(X_1 = b)\tau_{X_1} + \mathbb{I}(X_2 = d)\tau_{X_2}.$$

Given these probabilities, for each individual-round-profile, we have a separate predicted probability of that profile being "chosen", i.e. an observed outcome of 1. Table C2 presents an example of how these probabilities would be calculated given random assignment of attributes across rounds, and the pre-defined IMCEs in Table C1.

Given Tables C1 and C2, we train the BART model on the actual attribute-level assignments, the observed covariates, and the outcome:

The BART model then estimates the OMCEs (τ_{ijk}) by making predictions of Y when X_1

Table C2. Random attribute-level assignment, and calculation of probability

i	j	X_1	X_2	Calculation	Prob	Y
1	1	a	c	$0.5 + 0 + 0$	0.5	0
1	2	a	d	$0.5 + 0 + 0.1$	0.6	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
I	J	b	c	$0.5 + -0.2 + 0$	0.3	0

Table C3. Training data for the BART model

i	c_1	c_2	X_1	X_2	Y
1	0.1	0	a	c	0
1	0.1	0	b	c	1
1	0.1	0	a	d	1
⋮	⋮	⋮	⋮	⋮	⋮
I	0.25	0.05	b	c	0

is set to b *for all observations* and when it is set to a, and deducting these two values, as demonstrated in Table C4.

Table C4. Calculating the OMCE by deducting the predicted probabilities under the assumption of different attribute-levels

i	$\hat{Y}\|X_1 = b$	$\hat{Y}\|X_1 = a$	$\widehat{\tau}_{ijkl}$
1	0.63	0.5	0.13
1	0.71	0.6	0.11
⋮	⋮	⋮	⋮
I	0.29	0.5	-0.21

Finally, the IMCEs are recovered by averaging the predicted OMCE across observations for the same individual. For example, for $i = 1$ the predicted IMCE is:

$$\hat{\tau}_{il} = \frac{1}{J \times 2}(0.13 + 0.11 + ...) = 0.109...$$

Given we know the IMCE for this individual is 0.1, the prediction error for this specific subject is $\hat{\tau}_{il} - \tau_{il} \approx 0.109 - 0.1 \approx 0.009$. We use these prediction errors to assess the accuracy of the BART model and corresponding IMCE estimation strategy.

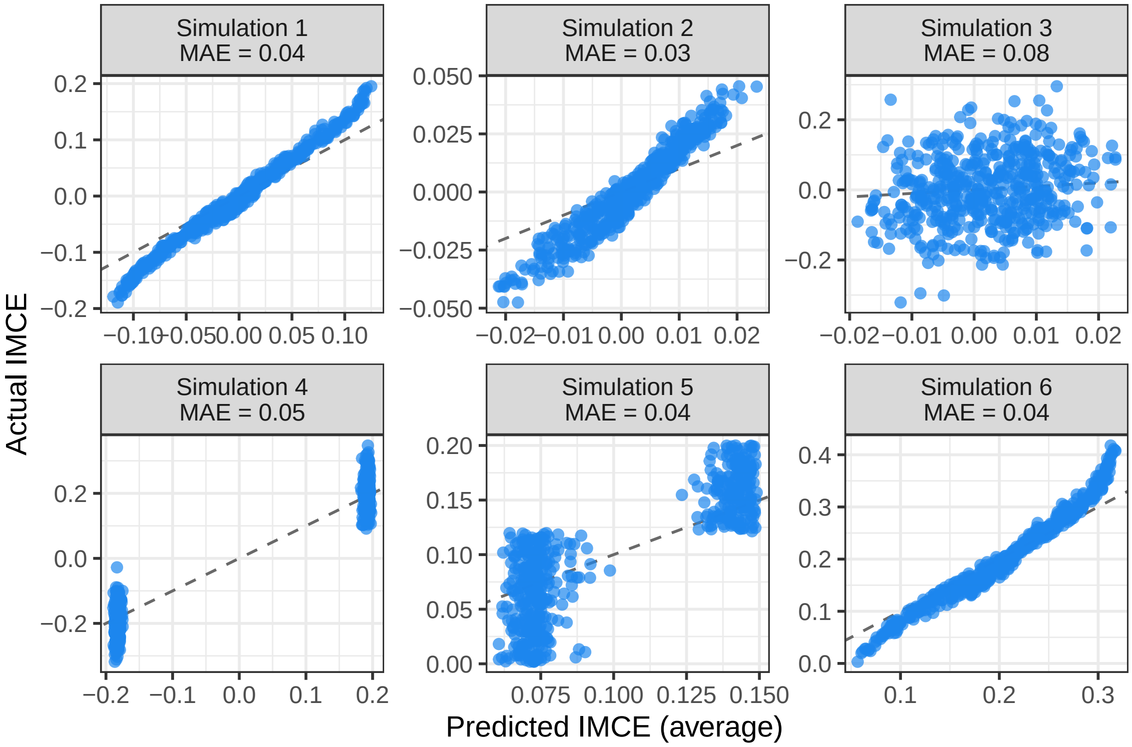
In our actual simulations, we complicate the DGP. We assume that each subject has three observed covariates: c_1 and c_2 are continuous covariates drawn from a random uniform distribution between 0 and some upper bound of heterogeneity (denoted h); c_3 is a binary variable generated from a binomial distribution with probability = 0.5. We also assume there is one *unobserved* covariate, c_4, which is normally distributed across subjects with mean 0 and standard deviation h. We randomly assign draws from each of these random variables to the 500 subjects.

Table C5 summarises the six scenarios we consider. In short, simulations 1 and 2 consider heterogeneity as a linear function of two observed covariates, varying the size of the heterogeneity parameter h. In simulation 3, treatment heterogeneity is largely random, although some small component (20%) is a linear function of the two covariates, and in simulation 4 heterogeneity is a function of a binary variable. In simulation 5 we simulate heterogeneity as a function of a missing covariate, and induce some correlation between an observed variable and this unobserved variable. Finally, in simulation 6, we consider an exponential function of heterogeneity (testing the predictive flexibility of the BART model).

For each of 100 iterations, we then generate the data by randomly assigning attribute levels to 500×5 observations, where each set of five observations correspond to the choices of a single subject. We calculate the predicted probability p of choosing each profile by multiplying the individuals' generated IMCEs by indicator variables for each of the two binary attributes plus a constant of 0.5 (such that, short of any attribute information, subjects are indifferent to the profile). We then draw binary outcomes from the binomial distribution using these predicted probabilities.

For each simulation and each iteration, we calculate the mean absolute error (MAE) between the BART models' IMCE prediction and the “true" IMCE. Figure C1 plots the average of each IMCE over 100 iterations, for each simulation specification. On average, we find that the MAE is low across heterogeneity specifications. Both linear, binary, and heteregeneity as a function of an unobserved covariate all have mean errors of approximately 0.04 to 0.05. When there is substantial random noise to the heterogeneity (simulation 3) we find greater error, but still quite low. What we do notice is at the tails of the IMCE distribution, the BART predicted effects are slightly conservative – as illustrated by the off-diagonal tails of the comparisons. This should be expected – the data is sparser at these points.

Figure C1. Average prediction error for each of 500 simulated IMCEs, varying the form of heterogeneity and its relationship to observed covariates.



Each panel depicts a separate Monte Carlo simulation, varying how heterogeneity in the IMCEs are defined. The individual points show the average error of the predicted IMCE across 500 iterations. The facet headings also report the mean absolute error (MAE) for each IMCE across these iterations.

C2 Coverage test

To test the uncertainty estimator we propose, we run Monte Carlo simulations in which we pre-define the IMCEs for each subject and assess the coverage of the resultant credible interval. As a naive comparison, we also estimate the variance of the IMCE as the simple mean of the OMCE variances for each subject i, i.e.

$$\widehat{\mathbb{V}(\tau_{il})} = \frac{1}{J \times K} \sum \widehat{\mathbb{V}(\tau_{ijkl})}$$

These IMCEs are themselves defined as normal distributions, where the mean for each subject is dependent on two subject-level covariates, and some standard deviation parameter σ_i:

$$\tau_{il} \sim \mathcal{N}([C_{1i} - C_{2i}], \sigma_i)$$

$$C_{1i}, C_{2i} \sim \text{Uniform}(0, c),$$

where c and σ_i are parameters set in the simulation.

In each iteration of the simulations, we take j draws from the IMCE distribution of each subject. These draws constitute the OMCEs for each subject in the experiment. We simultaneously generate a completely randomised treatment assignment schedule, for the IMCE attribute and one further dichotomous attribute where the IMCE is held fixed at 0.1 with zero variation. Given this assignment, we calculate the probability of picking each profile given the drawn OMCEs. We finally transform the outcome into a dichotomous measure by using the predicted probabilities to take draws from a binomial distribution.

After generating the simulated conjoint data, we calculate the **cjbart** predicted IMCEs and record whether or not the predicted interval contains the true IMCE mean, for each of the three variance estimation strategies. We repeat this process 500 times – generating new simulated data from the same (fixed) schedule of true IMCEs. We recover a single coverage rate for each measure by calculating the proportion of times the simulated IMCE contains the true population parameter for each hypothetical subject, and then take the average across these proportions.

To test the robustness of the coverage rate across contexts, we vary the number of subjects, rounds, the extent of IMCE heterogeneity, and the variance around the IMCE distributions. Table C6 details the parameter settings used for each of the seven separate simulation tests we run.

Table C7 reports the coverage rates for the two variance estimation methods. We find that, across different scenarios, the Bayesian interval produces near nominal simulated coverage rates. In general, coverage rates tend to be slightly conservative, estimating a slightly wider interval than necessary. We find, however, that in scenarios 4 and 5 where we increase the number of subjects, and where the naive estimator substantially underestimates the interval, the coverage of the Bayesian interval is closer to 0.95.

Table C5. Sources of heterogeneity in IMCEs, for each of 6 separate simulations

Sim.	f_{IMCE}	c	Details
1	$c_1 - c_2$	$c_x \sim$ Uniform$(0, h = 0.2)$	Effects are linearly heterogeneous between $-h$ and h
2	$c_1 - c_2$	$c_x \sim$ Uniform$(0, h = 0.05)$	As above, but the range is much smaller
3	$0.2(c_1 - c_2) + 0.8\mathcal{N}(0, 0.125)$	$c_x \sim$ Uniform$(0, h = 0.2)$	Covariates are a weak predictor of IMCE heterogeneity
4	If $c_3 = 1$, $\mathcal{N}(0.2, 0.05)$; else, $\mathcal{N}(-0.2, 0.05)$	$c_3 \sim$ Binomial$(1, 0.5)$	IMCE is either positive or negative dependent on observed binary variable
5	$c_4 \sim$ Uniform$(0, h = 0.2)$	$c_1 = 2 \times \mathbb{I}(c_4 > 0.6h) - \mathcal{N}(0, 0.25)$	IMCE is determined by unobserved covariate that also influences c_1.
6	$c_1 \times 2^{c_2} + c_2$	$c_x \sim$ Uniform$(0, h = 0.2)$	Exponential relationship between IMCE and covariates

Table C6. Simulation specifications testing the coverage rate of the confidence intervals

Sim.	Subjects	K	c	σ_i
1	500	5	0.25	0.05
2	500	5	0.05	0.02
3	500	10	0.05	0.02
4	1500	5	0.25	0.05
5	5000	5	0.25	0.05
6	500	5	0.25	Uniform$(0.001, 0.05)$
7	500	10	0.25	Uniform$(0.001, 0.05)$

Table C7. Comparison of coverage rates across the Bayesian and naive intervals.

Sim.	Naive Estimate	Bayesian
1	0.961	0.974
2	0.995	0.995
3	0.990	0.992
4	0.920	0.939
5	0.875	0.897
6	0.960	0.973
7	0.943	0.959

C3 RMCE simulation test

In Section 1 of the main paper we note that the RMCE, the marginal effect of an attribute-level within a specific round of the experiment, can be estimated as the average of the OMCEs within rounds of the experiment for each individual, rather than over all observations pertaining to that individual. This quantity can be useful to check whether the are any carryover or stability assumption violations that are necessary for valid conjoint analysis.

To check this assumption, we can train our first-stage model including a round-number indicator, allowing the model to learn any relationship between the outcome, effects, and rounds of the experiment. We then assess whether the estimated RMCEs correlate with the round indicator. If there are no carryover effects, in expectation the correlation should be zero.

To demonstrate this logic, we conducted a simulation where we repeatedly generated conjoint data where there either is or is not a serial correlation to the marginal effects of attribute-levels across rounds. Our simulated conjoint experiment contains three attributes (A, B, and C), each with two-levels (a1, a2, b1, etc.). Each experiment is run for 10 rounds and 250 subjects, with two profiles per round, and we simulate 100 separate experiments.

Within each round of each experiment, we define two sets of utility calculations to determine the forced-choice between profiles. In the "round-effect" scenario, the total utility of the subject i from profile j in round k is defined as:

$$\begin{aligned} U_{ijk}^{\text{Round-effect}} =& \mathcal{N}(0, 0.001) \\ & + 0.5r \times \mathbb{I}(A_{ijk} = a2) \\ & + (0.6 - 0.1r) \times \mathbb{I}(B_{ijk} = b2) \\ & + 0.5 \times \mathbb{I}(\mathcal{P}_{ijk} = c2), \end{aligned}$$

where r is the round of the experiment. In other words, the effect of level ‘a2’ increases over rounds, the effect of ‘b2’ decreases over rounds, and ‘c2’ has a constant effect.

The utility for the scenario in which there are no round effects, is calculated more simply as:

$$\begin{aligned}U_{ijk}^{\text{No round-effect}} =&\mathcal{N}(0, 0.001)\\&+ 1 \times \mathbb{I}(A_{ijk} = a2)\\&+ 0.2 \times \mathbb{I}(B_{ijk} = b2)\\&+ 0.5 \times \mathbb{I}(\mathcal{P}_{ijk} = c2).\end{aligned}$$

For each pair of profiles within the experiment, the profile that yields the higher utility gets assigned 1 and the other profile gets assigned 0. We calculate this separately for the round-effect and no round-effect utility calculations, yielding two experimental datasets.

We then estimate the OMCEs for each dataset, as detailed in Section 2, *including the round number indicator as a training variable*. This allows BART to flexibly use the round as an effect predictor if it helps refine predictions. In expectation, if there are no carryover or stability issues, then the round indicator variable should be uninformative. We then aggregate the OMCEs to the RMCE level by averaging the estimates within each round, for each hypothetical subject. Finally, we calculate the correlation between the estimated RMCEs and the round-number.

Figure C2 plots the distribution of these correlation coefficients by scenario and attribute, across the simulated experiments. For the no round-effects condition, each attribute’s distribution is centred on zero as expected – verifying that there is little information to be gleaned from the round indicator. For the round-effects scenario, however, there is a clear positive correlation for attribute A, and conversely a negative correlation for attribute B – clear evidence that the stability and no carryover assumption has been violated. Most interestingly, the relationship between round and attribute appears to have “leached" into the RMCE predictions for attribute C, despite the fact that in this scenario the marginal effect of C is unrelated to the round of the experiment. This clearly demonstrates why ensuring this assumption holds is so important – it may lead to biased estimates of attributes even if they are individually “well-behaved.

C4 OLS method comparison

In Figure 2 in the main paper, we demonstrate the ability of our BART method to effectively detect simulated heterogeneity. Table C8 reports the resulting correlations between the covariates and conjoint attributes, across 100 simulations.

Table C8. Average correlations between simulation covariates and conjoint attributes, over 100 simulations

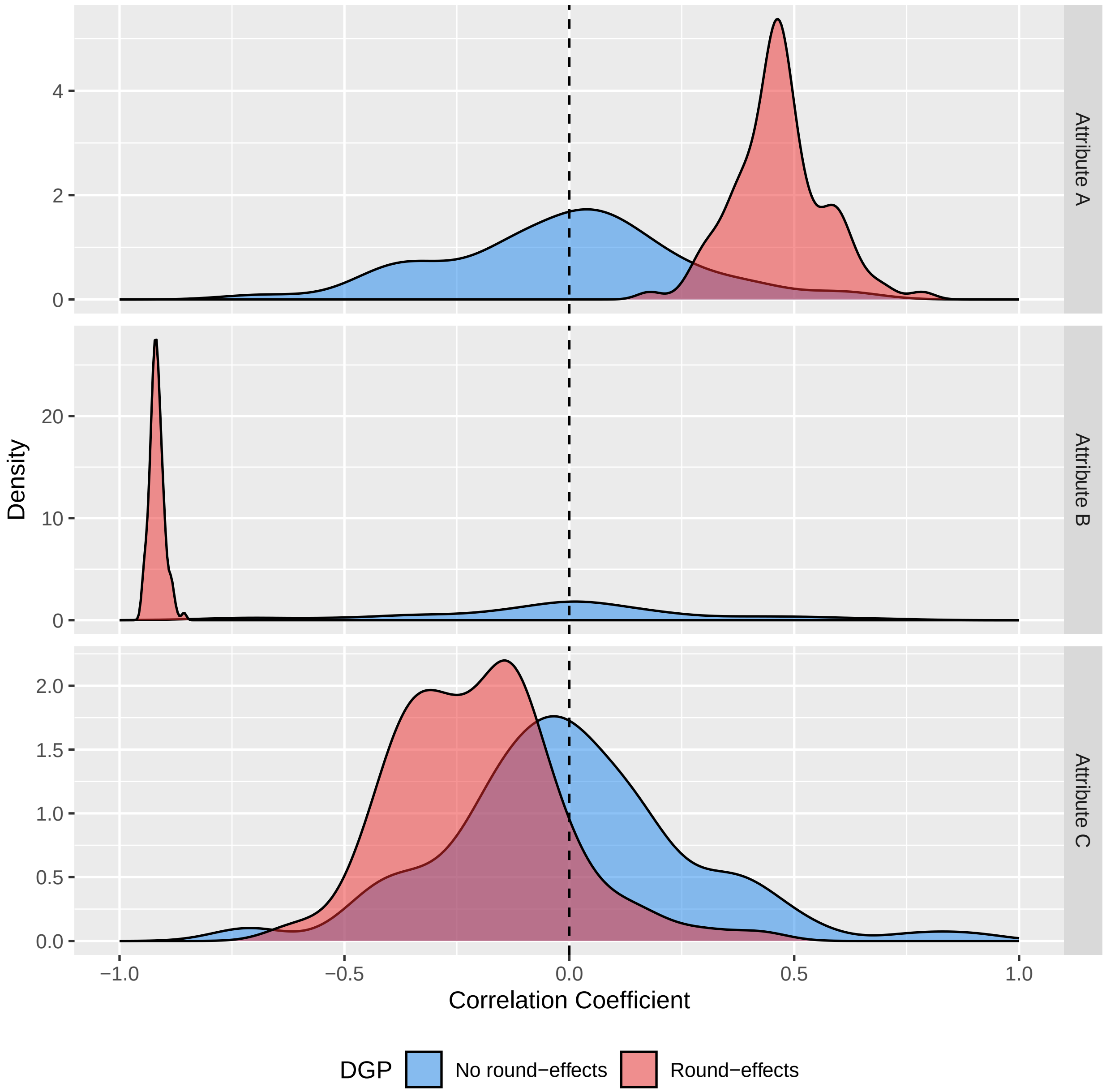
Attribute	c_1	c_2
A1	0.998	0.000
A2	0.004	-0.557
A3	-0.003	0.074

In this section, we replicate this exercise but with the OLS method proposed by Zhirkov (2022). Given the design requirements of this approach, we modify the simulation exercise in two ways. First, to ensure adequate power, we increase the number of conjoint rounds to 20 (with two profiles per round). Second, rather than force a choice between two profiles (using the defined utility function), we simply rescale the underlying utility to a 0-7 scale, and round the responses to the nearest integer – to mimic a rating-scale conjoint response. The underlying utility calculation and relationship between the binary (c_1) and interval (c_2) covariates are the same as in the main paper.

For each simulated subject we estimate a separate OLS regression model and record the coefficient for each of the three conjoint attributes (A1-3). Figures C3 and C4 plot the ordered distribution of the estimated IMCEs, colored by c_1 and c_2 values respectively.[4] As in

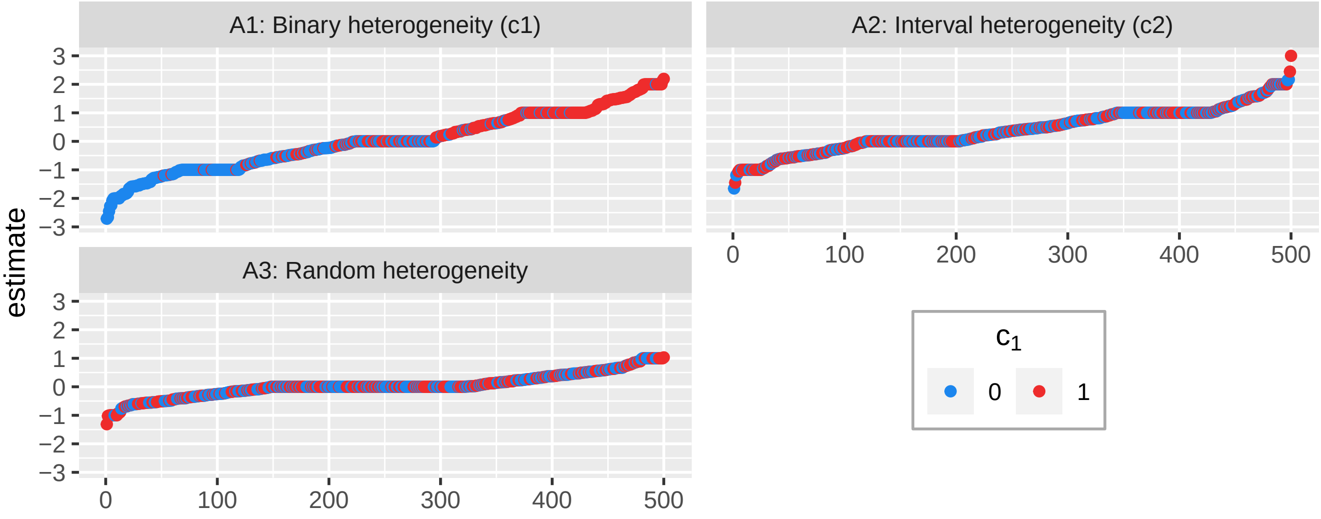
[4] The two figures do include 95 percent confidence intervals, but are very small and thus obscured by the plotted points. Moreover, in the flat regions, the model is performing poorly and returning essentially perfect fits.

Figure C2. Simulation evidence demonstrating how violations of the no carryover assumption can be detected by estimating the RMCE



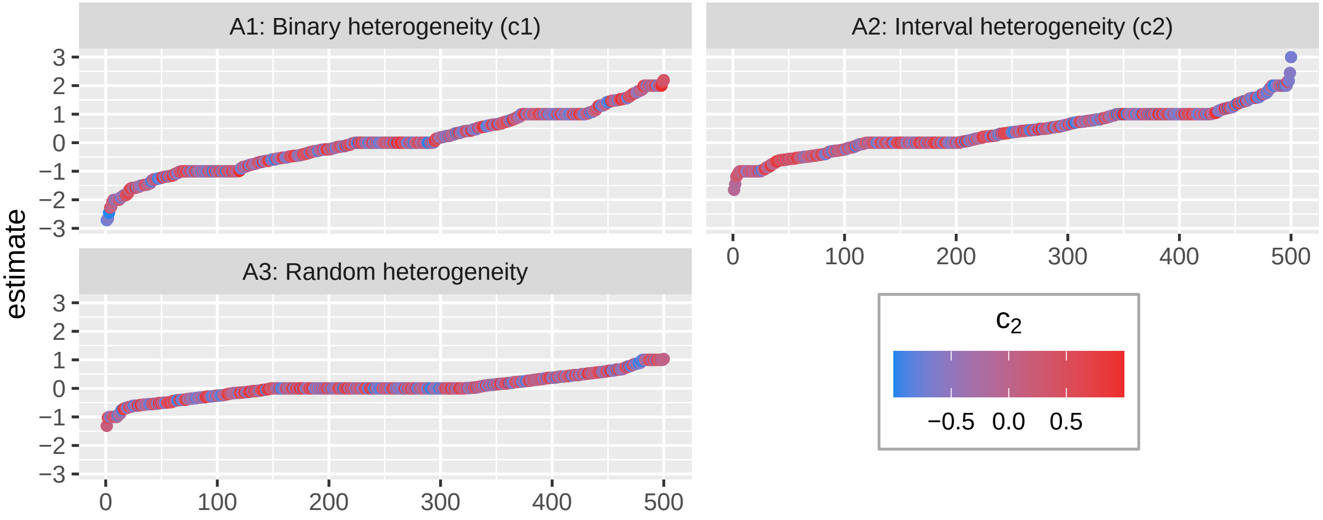
our proposed BART method, the OLS method does yield estimates that broadly align with the defined expectations of the simulation for the binary covariate c_1. In panel A1 of Figure C3, the IMCEs are largely sorted by the value of c_1. Note that the smooth continuity of this distribution, compared to the distribution in the main paper, can be attributed to using a rating scale outcome rather than the binary forced choice outcome. In Figure C4, although there is some slight suggestion of a negative correlation, the expected relationship is much harder to discern visually.

Figure C3. Detecting heterogeneity in IMCEs related to c_1 using simulated conjoint data derived from preferences over profiles, estimated with the OLS IMCE strategy



As in the main paper, we repeat this simulation exercise 100 times and record the correlation between each predicted IMCE and the two covariates in the design. Table C9 reports the average correlations between the covariates c_1 and c_2 and the three distributions of IMCEs respectively. There is a substantively large correlation between c_1 and A1, although this correlation is not as strong as observed using the BART strategy in the main paper. With respect to A2 and c_2, we see a much smaller (but nevertheless negative) correlation. Finally, as expected, we observe negligible correlations between c_1 and A2 and A3, and

Figure C4. Detecting heterogeneity in IMCEs related to c_2 using simulated conjoint data derived from preferences over profiles, estimated with the OLS IMCE strategy



between c_2 and A1 and A3.

It is noteworthy that across both binary and interval covariates, and compared to our BART approach[5], we observe relatively weaker correlations with the covariates, despite using exactly the same underlying utility function to simulate the hypothetical subjects' behavior. We suspect this is due to two factors. First, the OLS method cannot incorporate or model interactions between the individual-level covariates and the attributes (since there is no variation in these variables within each individual-level dataset). Second, using an an interval ratings outcome means smaller differences in utility lead to less stark differences in observed outcomes between profiles.[6] Researchers may want to consider these factors when deciding which outcomes to measure in their conjoint experiment, if analysing heterogeneity is a key part of the intended analysis.

[5] As well as the causal forest strategy detailed in Section E.

[6] This feature is in contrast to a forced-choice design, where even minuscule differences in utility between profiles result in one observation being assigned an outcome of 1 and the other an outcome of 0.

Table C9. Average correlations between simulation covariates and conjoint attributes, over 100 simulations

Attribute	c_1	c_2
A1	0.690	-0.002
A2	0.002	-0.156
A3	-0.003	0.000

D AMCE robustness check: further details and analysis

Hainmueller et al. (2014) conduct a conjoint experiment in which they consider the causal effects of immigrants' attributes on local individuals' attitudes towards these individuals. The study focuses on nine attributes of immigrants – including education, gender, country of origin – where the values of these attributes (the levels) are randomised over two profiles, and subjects pick which of the two immigrants they would prefer to 'give priority to come to the United States to live' (p.6).

To estimate the AMCEs parametrically, we run a linear probability model (LPM). We estimate the following model:

$$\begin{aligned}\text{ChosenImmigrant} = \alpha + \boldsymbol{\beta_1}\text{Education} + \boldsymbol{\beta_2}\text{Gender} + \boldsymbol{\beta_3}\text{CountryOfOrigin}\\ + \boldsymbol{\beta_4}\text{ReasonForApplication} + \boldsymbol{\beta_5}\text{Job} + \boldsymbol{\beta_6}\text{JobExperience} + \boldsymbol{\beta_7}\text{JobPlans}\\ + \boldsymbol{\beta_8}\text{PriorEntry} + \boldsymbol{\beta_9}\text{LanguageSkills},\end{aligned}$$

where $\boldsymbol{\beta_k}$ is the vector of coefficients for the $l-1$ levels within the kth attribute.

We then supply the same information to a BART model (including the ethnocentrism covariate embedded in the data) and recover the OMCE/IMCE estimates for each subject in the data. To aggregate the parameter estimates to the average marginal component effect, we simply take the average across the IMCEs.[7] We then plot these BART-estimated AMCEs against the parametric AMCEs as shown in Figure 3 in the main text. In Table D1 we present these same AMCE comparisons numerically, which further demonstrates the small divergence between parameter estimates for each attribute-level. Note that the 'Seek Better Job' parameter estimate failed to converge under the LPM specification.

Table D2 reports the 95 percent confidence interval and 95 percent credible interval for the AMCE estimates presented in Table D1. Overall, we find that the 95 percent credible intervals are slightly wider than the confidence intervals. Readers should note these two

[7] This can be computed automatically within the **cjbart** package by calling `summary()` on the IMCE object.

Table D1. Comparison of AMCE estimates for the Hainmueller et al. (2014) conjoint experiment using LPM and **cjbart** methods

		Coefficient		**Difference**
Attribute	**Level**	**LPM**	**cjbart**	**(% of LPM coefficient)**
Education	4th Grade	0.03	0.04	10.59
	8th Grade	0.06	0.06	-3.99
	High School	0.12	0.12	-2.26
	Two-Year College	0.15	0.16	1.23
	College Degree	0.18	0.18	0.54
	Graduate Degree	0.17	0.17	0.16
Gender	Male	-0.02	-0.02	-3.69
Country Of Origin	Germany	0.05	0.04	-15.70
	France	0.03	0.02	-14.79
	Mexico	0.01	0.01	-19.85
	Philippines	0.03	0.03	-18.62
	Poland	0.03	0.03	-11.79
	China	-0.02	-0.02	-11.27
	Sudan	-0.04	-0.04	-6.83
	Somalia	-0.05	-0.05	-6.37
	Iraq	-0.11	-0.11	-1.61
Reason For Application	Seek Better Job	-0.04	-0.04	0.03
	Escape Persecution	0.05	0.04	-14.14
Job	Waiter	-0.01	-0.01	-25.49
	Child Care Provider	0.01	0.01	-33.60
	Gardener	0.01	0.00	-37.07
	Financial Analyst	0.04	0.03	-34.89
	Construction Worker	0.04	0.03	-27.37
	Teacher	0.07	0.06	-8.51
	Computer Programmer	0.06	0.05	-20.68
	Nurse	0.08	0.07	-9.02
	Research Scientist	0.11	0.09	-11.97
	Doctor	0.14	0.13	-6.61
Job Experience	1-2 Years	0.06	0.06	-1.87
	3-5 Years	0.11	0.11	-0.46
	5+ Years	0.11	0.11	-1.24
Job Plans	Contract with Employer	0.12	0.12	-3.29
	Interviews with Employer	0.03	0.02	-23.21
	No Plans to Look for Work	-0.16	-0.16	1.43
Prior Entry	Once as Tourist	0.06	0.06	0.50
	Many Times as Tourist	0.05	0.05	1.41
	Six Months with Family	0.07	0.06	-13.29
	Once w/o Authorization	-0.11	-0.11	1.76
Language Skills	Broken English	-0.06	-0.06	-0.03
	Tried English but Unable	-0.13	-0.13	-0.66
	Used Interpreter	-0.16	-0.16	-0.69

uncertainty estimates are not directly comparable – the former being a frequentist statistic and the latter a Bayesian statistic.

Tables D3 and D4 replicate the same exercise for the Duch et al. (2021) data. The differences in parameter estimates across the two strategies are very small: typically less than a percentage point and at indistinguishable at two decimal places. These very small differences are most likely due to the large number of observations in this experiment. As before, while direct comparison is not possible, we find the credible interval is wider than the LPM confidence interval.

Table D2. AMCE uncertainty estimates for the Hainmueller et al. (2014) conjoint experiment using LPM and **cjbart** methods

		LPM	**cjbart**
Attribute	**Level**	**95% Conf. Interval**	**95% Cred. Interval**
Education	4th Grade	[0.00,0.06]	[-0.01,0.08]
	8th Grade	[0.03,0.09]	[0.02,0.10]
	High School	[0.09,0.15]	[0.03,0.17]
	Two-Year College	[0.12,0.19]	[0.08,0.24]
	College Degree	[0.15,0.21]	[0.08,0.25]
	Graduate Degree	[0.14,0.20]	[0.10,0.22]
Gender	Male	[-0.04,-0.01]	[-0.05,0.00]
Country Of Origin	Germany	[0.01,0.08]	[0.00,0.09]
	France	[-0.01,0.06]	[-0.05,0.08]
	Mexico	[-0.03,0.04]	[-0.03,0.06]
	Philippines	[0.00,0.07]	[-0.01,0.08]
	Poland	[0.00,0.07]	[-0.01,0.09]
	China	[-0.06,0.02]	[-0.07,0.03]
	Sudan	[-0.08,-0.01]	[-0.11,0.00]
	Somalia	[-0.09,-0.02]	[-0.11,0.00]
	Iraq	[-0.15,-0.07]	[-0.20,-0.04]
Reason For Application	Seek Better Job	[-0.06,-0.02]	[-0.07,-0.02]
	Escape Persecution	[0.02,0.08]	[-0.04,0.10]
Job	Waiter	[-0.04,0.02]	[-0.07,0.07]
	Child Care Provider	[-0.02,0.04]	[-0.04,0.09]
	Gardener	[-0.02,0.04]	[-0.05,0.10]
	Financial Analyst	[0.00,0.09]	[-0.02,0.13]
	Construction Worker	[0.00,0.07]	[-0.01,0.10]
	Teacher	[0.03,0.10]	[0.01,0.14]
	Computer Programmer	[0.01,0.11]	[-0.02,0.14]
	Nurse	[0.05,0.11]	[0.03,0.15]
	Research Scientist	[0.06,0.15]	[0.00,0.18]
	Doctor	[0.09,0.18]	[0.06,0.21]
Job Experience	1-2 Years	[0.04,0.09]	[0.02,0.10]
	3-5 Years	[0.09,0.13]	[0.04,0.15]
	5+ Years	[0.09,0.14]	[0.06,0.15]
Job Plans	Contract with Employer	[0.10,0.15]	[0.03,0.20]
	Interviews with Employer	[0.00,0.05]	[0.00,0.05]
	No Plans to Look for Work	[-0.18,-0.14]	[-0.22,-0.10]
Prior Entry	Once as Tourist	[0.03,0.08]	[0.00,0.09]
	Many Times as Tourist	[0.03,0.08]	[0.00,0.09]
	Six Months with Family	[0.05,0.10]	[0.00,0.10]
	Once w/o Authorization	[-0.14,-0.09]	[-0.18,-0.05]
Language Skills	Broken English	[-0.08,-0.03]	[-0.13,-0.01]
	Tried English but Unable	[-0.15,-0.11]	[-0.20,-0.06]
	Used Interpreter	[-0.18,-0.14]	[-0.24,-0.09]

Table D3. Comparison of AMCE estimates for the **?** conjoint experiment using LPM and **cjbart** methods

		Coefficient		Difference
Attribute	**Level**	**LPM**	**cjbart**	**(% of LPM coefficient)**
Vulnerability	Moderate (Twice the average risk of COVID-19 death)	0.05	0.05	0.37
	High (Five times the average risk of COVID-19 death)	0.18	0.18	0.07
Transmission	Moderate risk (Twice the average risk of catching and transmitting the COVID-19 virus)	0.04	0.04	-0.50
	High risk (Five times the average risk of catching and transmitting the COVID-19 virus)	0.16	0.16	0.08
Income	Lowest 20% income level	0.02	0.02	2.73
	Highest 20% income level	-0.03	-0.03	0.27
Occupation	Non-Key worker: Can work at home	-0.01	-0.01	2.92
	Non-Key worker: Cannot work at home	0.08	0.08	0.18
	Key worker: Education and childcare	0.21	0.21	0.07
	Key worker: Factory worker	0.14	0.14	0.10
	Key worker: Water and electricity service	0.15	0.15	0.17
	Key worker: Police and fire-fighting	0.21	0.21	0.09
	Key worker: Health and social care	0.26	0.26	0.13
Age	40 years old	0.05	0.05	0.66
	65 years old	0.09	0.09	-0.08
	79 years old	0.09	0.09	0.09

Table D4. AMCE uncertainty estimates for the **?** conjoint experiment using LPM and **cjbart** methods

		LPM	cjbart
Attribute	**Level**	**95% Conf. Int.**	**95% Cred. Int.**
Vulnerability	Moderate (Twice the average risk of COVID-19 death)	[0.05,0.06]	[-0.01,0.12]
	High (Five times the average risk of COVID-19 death)	[0.18,0.19]	[0.06,0.29]
Transmission	Moderate risk (Twice the average risk of catching and transmitting the COVID-19 virus)	[0.04,0.05]	[-0.01,0.09]
	High risk (Five times the average risk of catching and transmitting the COVID-19 virus)	[0.16,0.17]	[0.05,0.25]
Income	Lowest 20% income level	[0.01,0.02]	[-0.01,0.05]
	Highest 20% income level	[-0.03,-0.02]	[-0.09,0.00]
Occupation	Non-Key worker: Can work at home	[-0.01,0.00]	[-0.09,0.08]
	Non-Key worker: Cannot work at home	[0.08,0.09]	[0.00,0.15]
	Key worker: Education and childcare	[0.20,0.21]	[0.10,0.28]
	Key worker: Factory worker	[0.13,0.15]	[0.05,0.21]
	Key worker: Water and electricity service	[0.15,0.16]	[0.07,0.22]
	Key worker: Police and fire-fighting	[0.20,0.21]	[0.09,0.30]
	Key worker: Health and social care	[0.25,0.27]	[0.12,0.37]
Age	40 years old	[0.04,0.06]	[0.00,0.09]
	65 years old	[0.09,0.10]	[-0.04,0.15]
	79 years old	[0.08,0.09]	[-0.08,0.16]

E Causal Forest alternative estimation

As noted in the main text, our strategy can be generalised beyond the specific BART implementation that we pursue. One particularly interesting alternative is to use Causal Forests (Athey and Wager 2019). This strategy follows a similar logic to random forests where the final prediction is the average over many separate tree-models. Causal forests differ by using *causal* rather than decision trees: recursive partitions of the datas where splits are optimized to find treatment effect heterogeneity. In other words, each tree aims to partition the data such that the treatment effects *within* nodes are similar, but the conditional average treatment effects differ *across* nodes.

While this approach directly embeds intuitions about treatment effect heterogeneity into the estimation process, it nevertheless has some limitations compared to our proposed strategy in the main paper. Causal forests can currently only estimate treatment effects for binary treatment indicators. In the case of conjoint experiments, therefore, this has two important implications. First, where conjoint attributes have three or more levels, using causal forests requires running separate models for each binary comparison between the reference level and every other level. For example, a five-level attribute would require running four separate models. Moreover, since the treatment indicator must be binary, any experimental observations where the Lth attribute is neither the reference or current level of interest have to be dropped, resulting in fewer training examples.

E1 Simulation test

As in Section 2.2 of the main paper, and in Section C4 of the Appendix with respect to the Zhirkov (2022) OLS method, we test the causal forest estimation strategy using Monte Carlo simulation. We use use exactly the same utility specification and design as in the main paper, where 500 hypothetical subjects make a forced-choice between two profiles across 5 rounds of the experiment.

To estimate the IMCEs, and as noted as a limitation above, we run *separate* causal forest models for each of the three binary attributes in the simulated conjoint design. Prior to our main analysis, we also use the causal forests' inbuilt tuning algorithm to optimise all hyperparameters. We extract these optimal parameters once, and use them across all the models we estimate.

Figures E1 and E2 plot the estimated IMCEs by magnitude, colored by the values of the two covariates c_1 and c_2 respectively. The results are very similar to the BART analysis reported in the main paper: the models effectively distinguish both the binary relationship between c_1 and A1, as well as the more complex continuous relationship between c_2 and A2. Table E1 confirms this analysis via Monte Carlo simulation: the causal forest models correctly detects the designed correlations and otherwise finds negligible relationships, as we would expect.

Figure E1. Detecting heterogeneity in IMCEs related to c_1 using simulated conjoint data derived from preferences over profiles, estimated with the *causal forest* algorithm

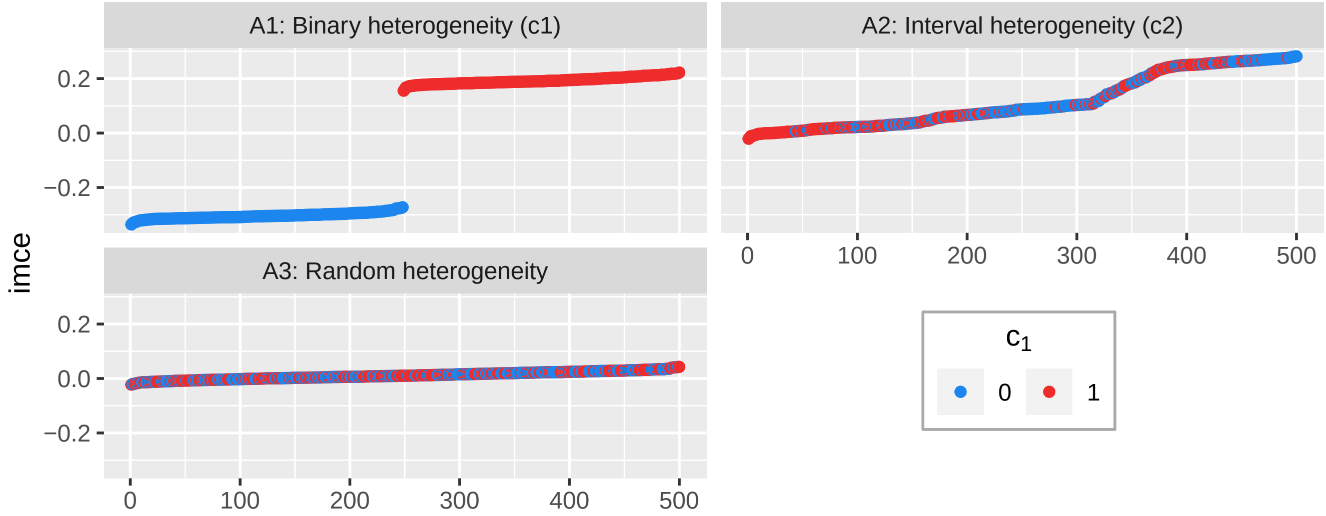


Figure E2. Detecting heterogeneity in IMCEs related to c_2 using simulated conjoint data derived from preferences over profiles, estimated with the *causal forest* algorithm

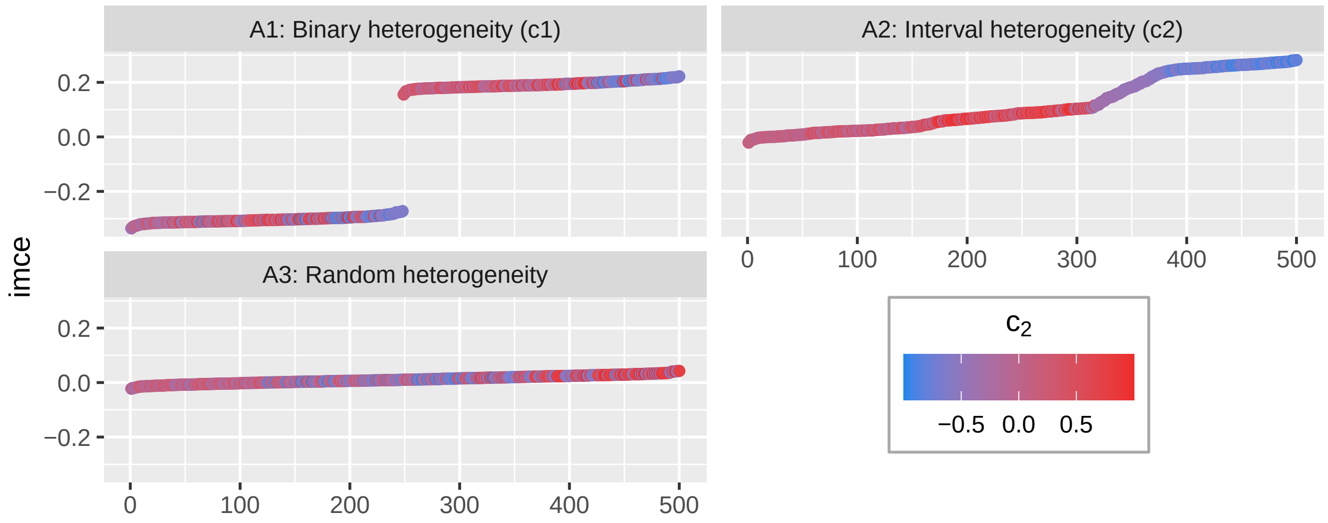


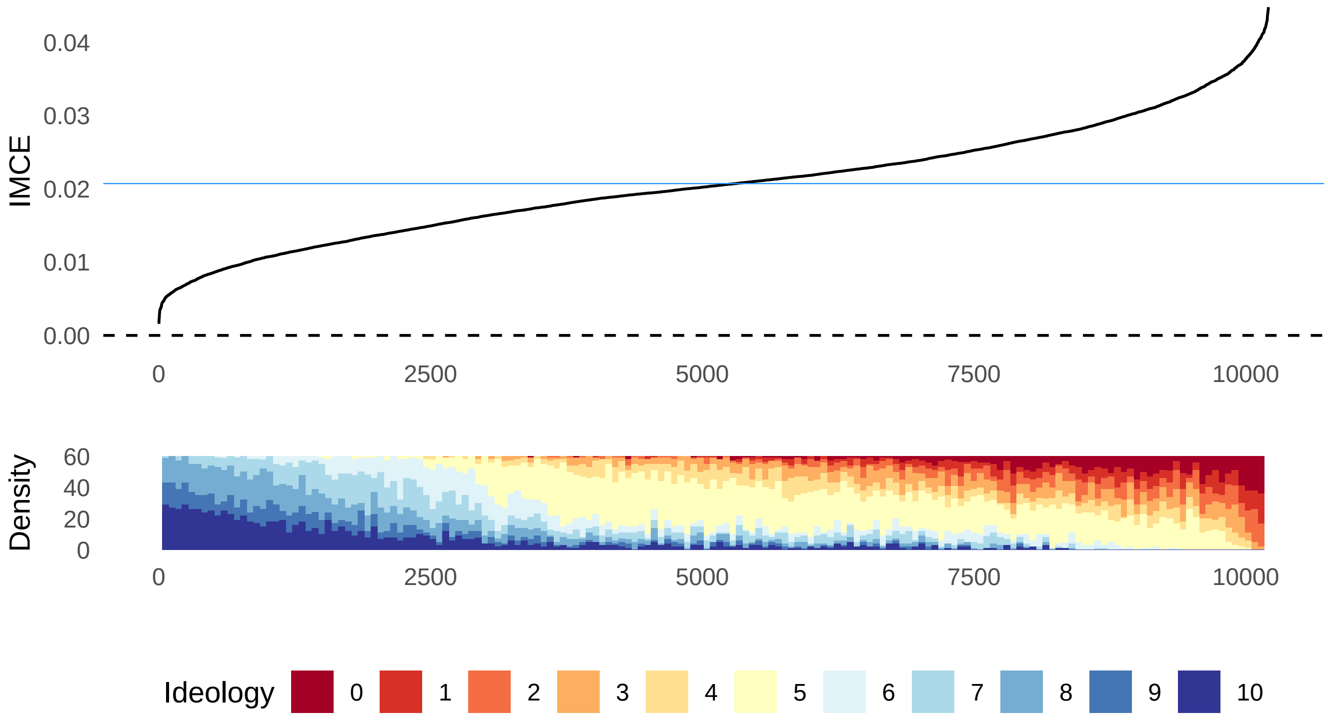
Table E1. Average correlations between simulation covariates and conjoint attributes, over 100 simulations estimated with the *causal forest* algorithm

Attribute	c_1	c_2
A1	0.997	0.002
A2	0.007	-0.582
A3	0.004	0.070

E2 Applied test

To explore whether our results in the main paper are robust across estimating strategies, we ran a causal forest model to estimate the IMCEs for the low income attribute-level of the conjoint experiment. We first tune and train a causal forest model using the **grf** package in R (Tibshirani et al. 2022), where the outcome is the binary choice variable, the treatment variable is a binary indicator for the income attribute-level, and we supply a training matrix of the other conjoint attributes *plus* the same covariates used in the BART models. All observations where subjects were assigned the “Highest 20% income level" were dropped prior to training, due to the limitations mentioned above.

Figure E3. Comparison of IMCEs for the “Lowest 20% income level" attribute-level ordered from smallest to largest and corresponding histogram of individuals’ self-reported ideology, using the **causal forest** estimation strategy



This causal forest model does not account for subject-level clustering of observations (see Figure E4).

Unlike in our BART strategy, the causal forest algorithm automatically returns predicted treatment effects rather than predicted outcomes. We therefore directly aggregate the output of the causal forest model (OMCEs) to the level of IMCEs by averaging these predictions for each individual separately.[8]

Figure E3 plots these IMCEs and the corresponding histogram of ideology values for every subject in the experiment. These results follow the same pattern as those presented in the main paper, with ideology clearly inversely related to the magnitude of the IMCEs: more right-leaning subjects have smaller (albeit positive) IMCEs.

[8] Since variance estimation in causal forest uses a bootstrap of little bags (Athey et al. 2019), aggregating the uncertainty estimates from the level of observation to the level of the individual is beyond the scope of this paper.

Causal forests also provide an in-built and simple variable importance measure (VIMP), by calculating a weighted sum of the number of times each covariate is used to split the data across all the trees in the forest. This measure is different from our BART strategy, since in the causal forest case (and like with random forests) one can rely on the independence of the estimates from each separate tree. In BART, since the trees are non-independent (they are trained to model the residual variance of the $T - 1$ other trees), interpreting split criteria directly is more challenging. Therefore, it is worth inspecting how this intrinsic model metric from the causal forests algorithm identifies important covariates.

Table E2 reports the VIMP scores for the covariate attributes in the model. For categorical variables, each dummy factor is assigned a separate score and so we sum these to get an importance measure for each covariate. Similar to our analysis in the main paper, subject ideology is identified as an important predictor. The causal forest importance measure diverges from our own in two ways. First, the causal forest does not identify subjects' country as an important predictor. We believe this difference is due to the fact that, for our random forest based measure in the main paper, the trees are able to split on multiple levels of the categorical variable at single decision nodes (as shown in Figure 5.) As a result, our variable importance measure regularises itself by collapsing levels of categorical variables. This is not possible in the causal forest measure since each node can only split on one level at a time. Second, and perhaps relatedly, the causal forest measure identifies subjects' age as an important feature. We are unsure precisely why this difference exists, but we note that theoretically the importance measures are quite different (see the discussion in Section 3.1), which may contribute to the divergence in scores.

Overall, these results help demonstrate two claims. First, that it is possible to substitute the BART-specific implementation we discuss in the main paper with alternative OMCE estimation strategies. Second, that our main substantive results appear robust to different

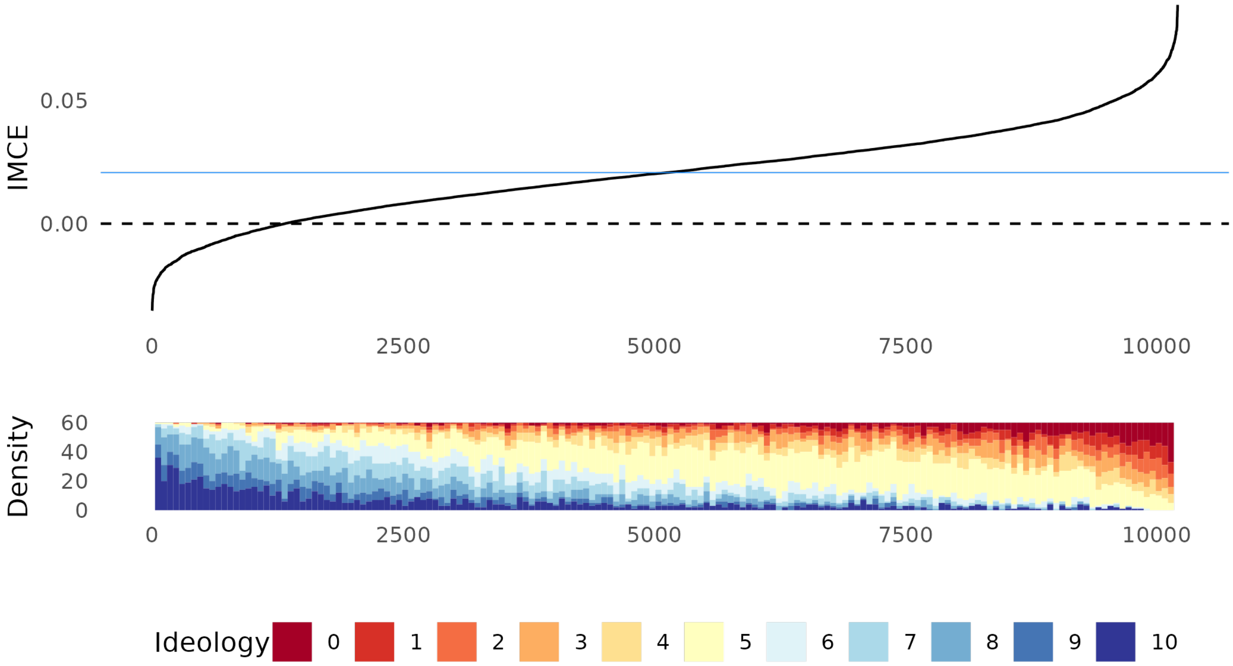
Table E2. Variable importance scores for the “Lowest 20% income level" attribute-level, using the intrinsic measure from the causal forest fitting algorithm

Covariate	Variable Importance
Country	0.010
Education	0.046
Gender	0.041
Hesitancy	0.067
Ideology	0.131
Income	0.027
Mandatory Vaccination	0.090
Age	0.130
WTP Access	0.047
WTP Private	0.052

ML estimating strategies, providing further evidence of their robustness.

Finally, one advantage of causal forest estimation is the ability to model the subject-clustering component of conjoint designs by supplying the subject identifiers to the algorithm (see Athey et al. 2019). We can therefore assess whether deliberately clustering affects our findings by comparing the results presented in Figure E3 with a clustered-variant (as shown in Figure E4). Substantively, the results are very similar. The distribution of right-leaning subjects stretches slightly further along the IMCE distribution, and the most extreme IMCEs are slightly larger, but not drastically so and do not affect our interpretation of the results.

Figure E4. Causal forest estimation of IMCEs for the “Lowest 20% income level" by subject ideology, accounting for subject-level clustering of observations



F Example of pIMCE estimation

In Section 2.4 of the main paper, we extend the logic of the IMCE to cases where we do not assume that possible profiles are distributed uniformly. In particular, we adapt the logic of de la Cuesta et al. (2022) by weighting the IMCE potential outcomes by the marginal distributions of the attributes in the population of interest.

To demonstrate this approach empirically, we consider a *hypothetical* case where we alter the marginal distributions of the age, income, and vulnerability attributes based on their distribution in US adult population. Table F1 summarises the population marginals we use. To make our hypothetical scenario more realistic, we approximate the distribution of age categories using the American Community Survey, by summing the proportion of US adults whose age is closest to each attribute-level in the Duch et al. (2021) design.[9] For the proportion of vulnerable adults, we use data provided by the Henry J Kaiser Family Foundation, which found that 37.6% of US adults had a higher risk of serious illness due to COVID-19.[10] We divide this percentage equally between the two higher vulnerability attribute-levels. For income, since the lower (upper) level refer to the 20% lowest (highest) income levels, we follow these distributions in the marginal distribution of age. For the remaining two attributes, we assume uniform distributions.

We first inspect the pIMCEs for the 65 year-old attribute-level, which our original analysis suggested was correlated with subjects' own age. Figure F1 plots a comparison of the pIMCE estimates against the original (unweighted) IMCEs generated from our standard strategy, for each US respondent in the Duch et al. (2021) data. While we do not see substantially different estimates using the pIMCE strategy, there is a notable compression of effect sizes into three distinct clusters. Figure F2 confirms this analysis: while the pIMCE

[9] The ACS categories do not perfectly align with the conjoint age levels, so these proportions are approximate. We also scale the proportions to consider only US subjects aged 15 years and older.

[10] https://www.kff.org/coronavirus-COVID-19/issue-brief/how-many-adults-are-at-risk-of-serious-illness-if-infected-with-coronavirus/ [Accessed 16th August 2022].

Table F1. Assumed marginal distributions of attribute-levels in the population

Attribute	**Level**	**Marginal Probability**
Vulnerability	Average	0.62
	Moderate	0.19
	High	0.19
Transmission	**All**	0.33
Income	Lowest 20%	0.20
	Average	0.60
	Highest 20%	0.20
Occupation	**All**	0.13
Age	25 years old	0.33
	40 years old	0.31
	65 years old	0.22
	79 years old	0.14

effects are slightly more extreme at either tail of the distribution, by and large they follow the same sort of pattern and magnitude. The jumps in the pIMCE line reflect the clustering of effect sizes seen in Figure F1. As shown in the bottom panels of Figure F2, however, the distribution of these estimates across both the IMCEs and pIMCEs correlate similarly with subjects' age: the strongest effects are for older respondents who are closer to the age of the attribute-level in question, consistent with our theoretical expectations.

Figures F3 and F4 repeat this exercise for the "High risk" transmission attribute-level in the conjoint experiment. We use the same marginal distributions as presented in Table F1. Here we see only minor differences between the IMCEs and pIMCEs for subjects. Similarly the effects distribution, while clearly heterogeneous, does not correlate anywhere near as strongly with subjects' age (compared to when analysing the age *attribute*).

Figure F1. Comparison of each US subjects' pIMCE estimate for the “65 years old" attribute-level, against the standard IMCE estimate

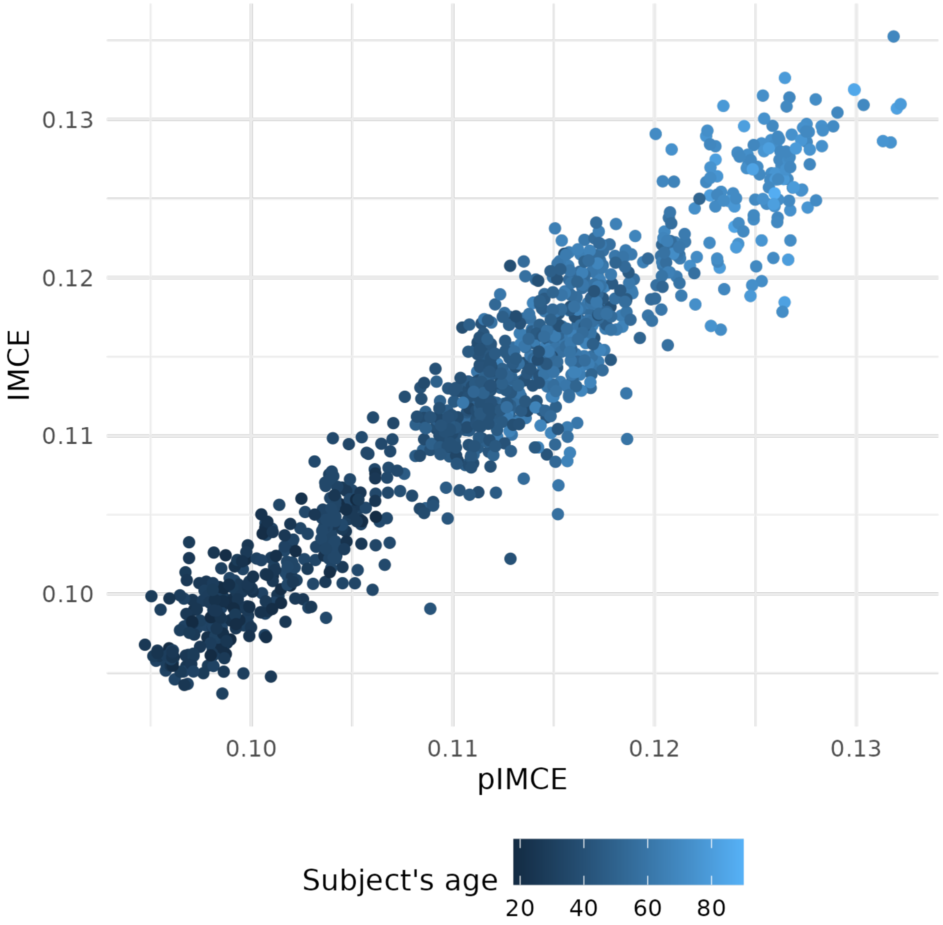
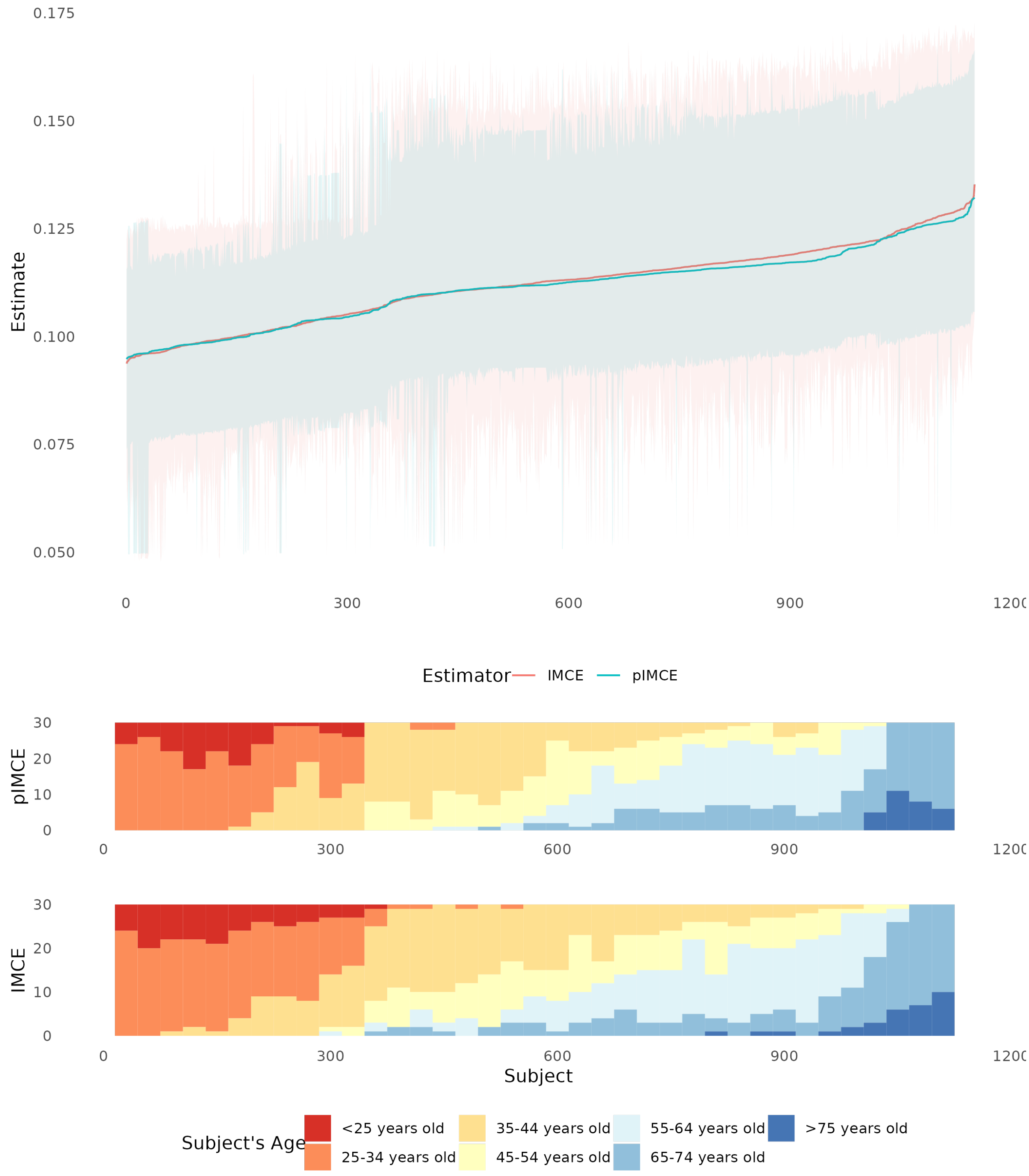


Figure F2. Comparison of the distribution of subjects' age against PIMCE and IMCE estimates for the "65 years old" attribute-level



Shaded areas around the IMCE/pIMCE lines indicate the respective 95% credible intervals.

Figure F3. Comparison of each US subjects' pIMCE estimate for the “high risk" transmission attribute-level, against the standard IMCE estimate

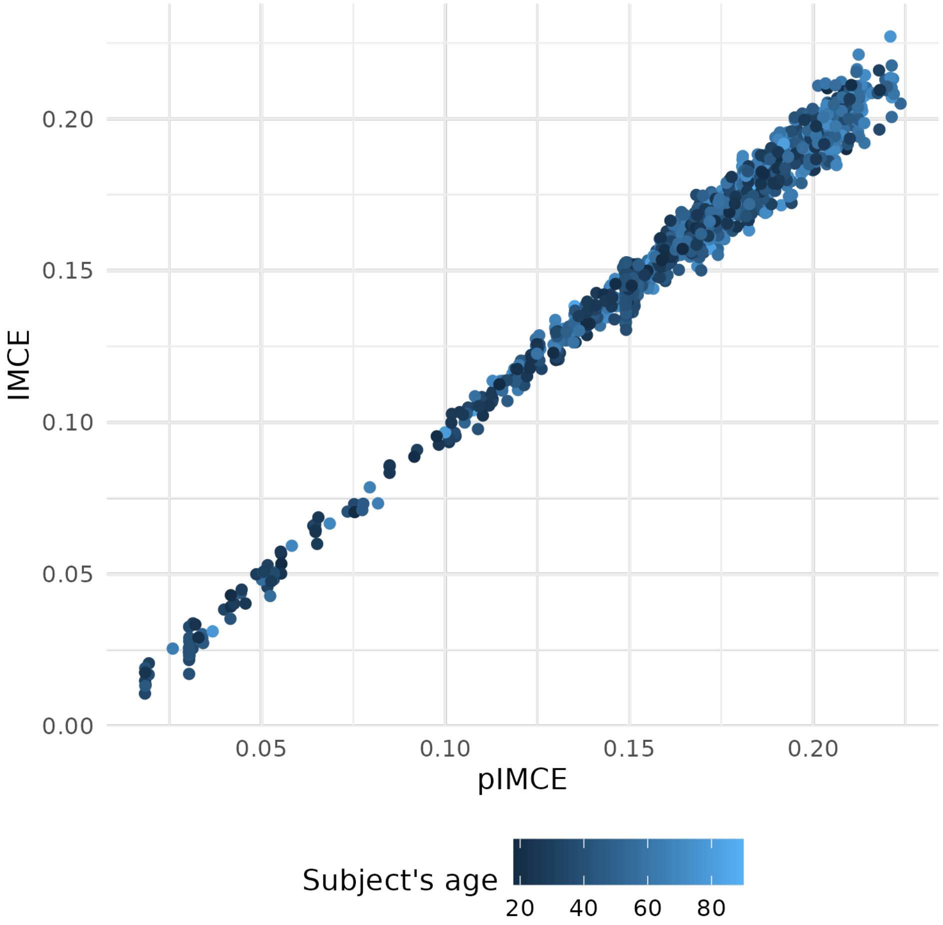
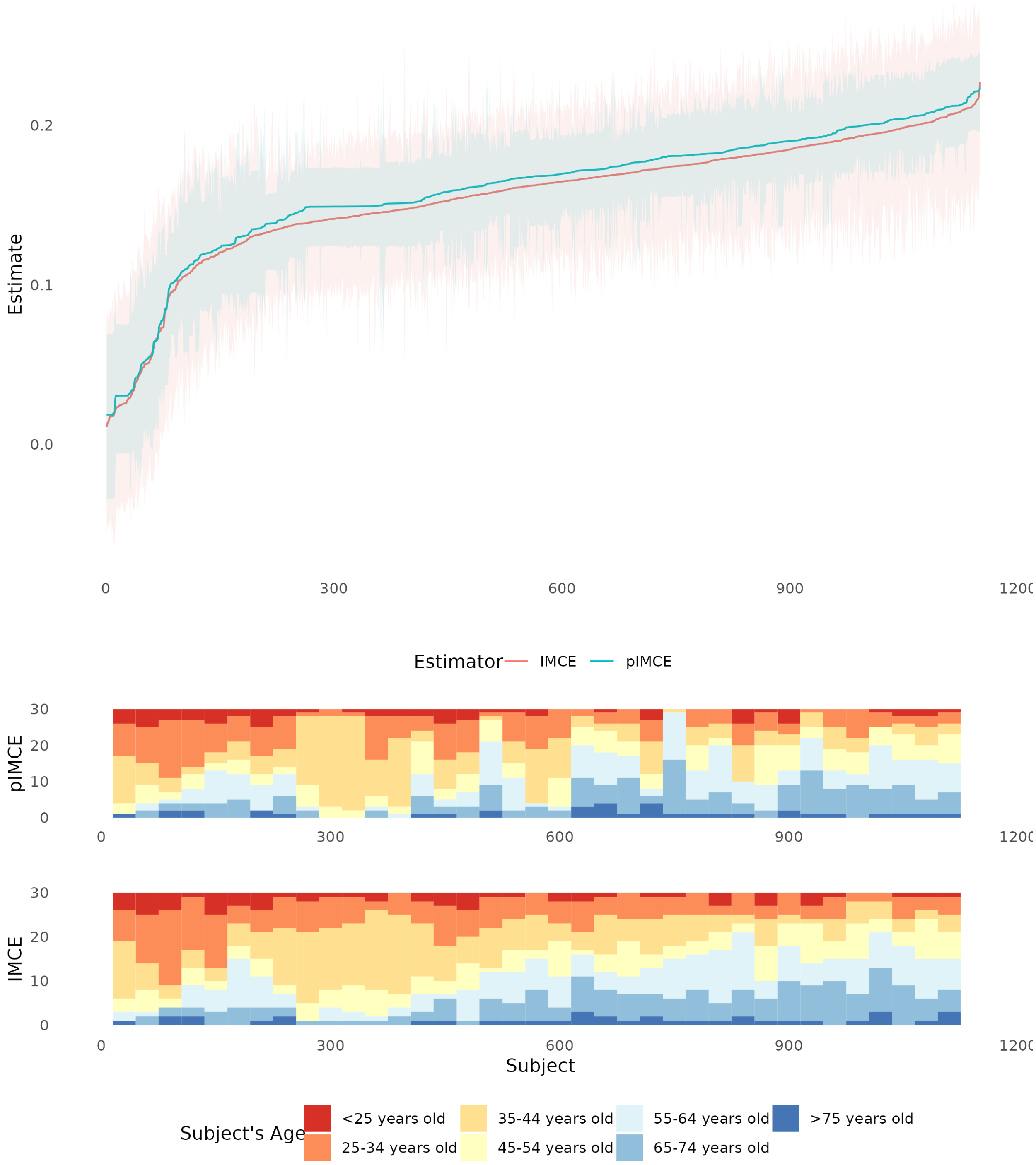


Figure F4. Comparison of the distribution of subjects' age against PIMCE and IMCE estimates for the "high risk" transmission attribute-level



Shaded areas around the IMCE/pIMCE lines indicate the respective 95% credible intervals.

G Additional figures

Figure G1. Detecting heterogeneity in IMCEs using simulated conjoint data derived from preferences over profiles (continuous covariate)

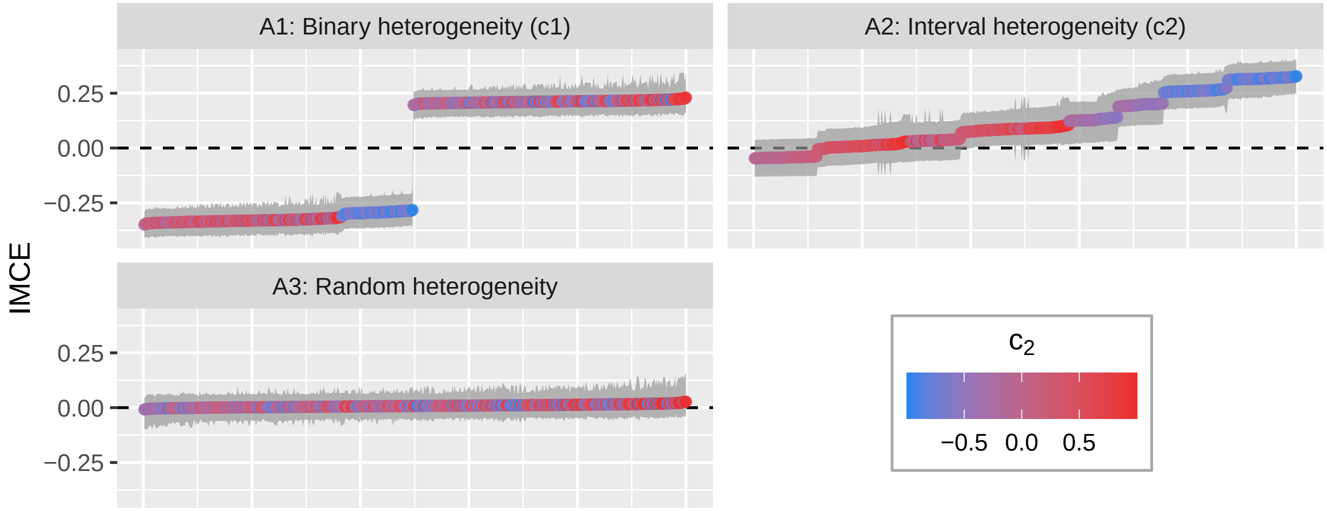


Figure G2. IMCE predictions by ideology values, using models on trained $k = 5$ random batches of the full experimental data

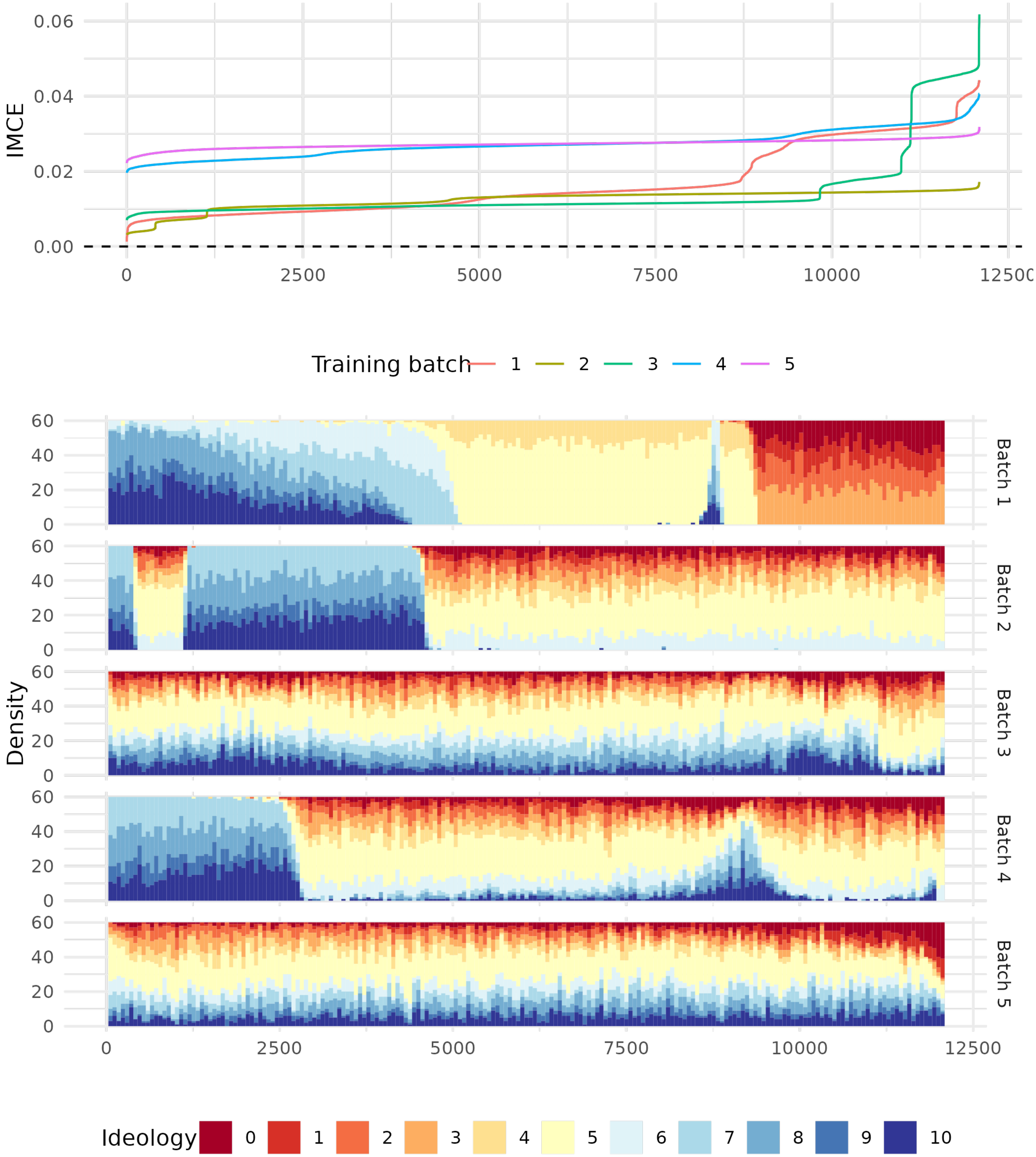
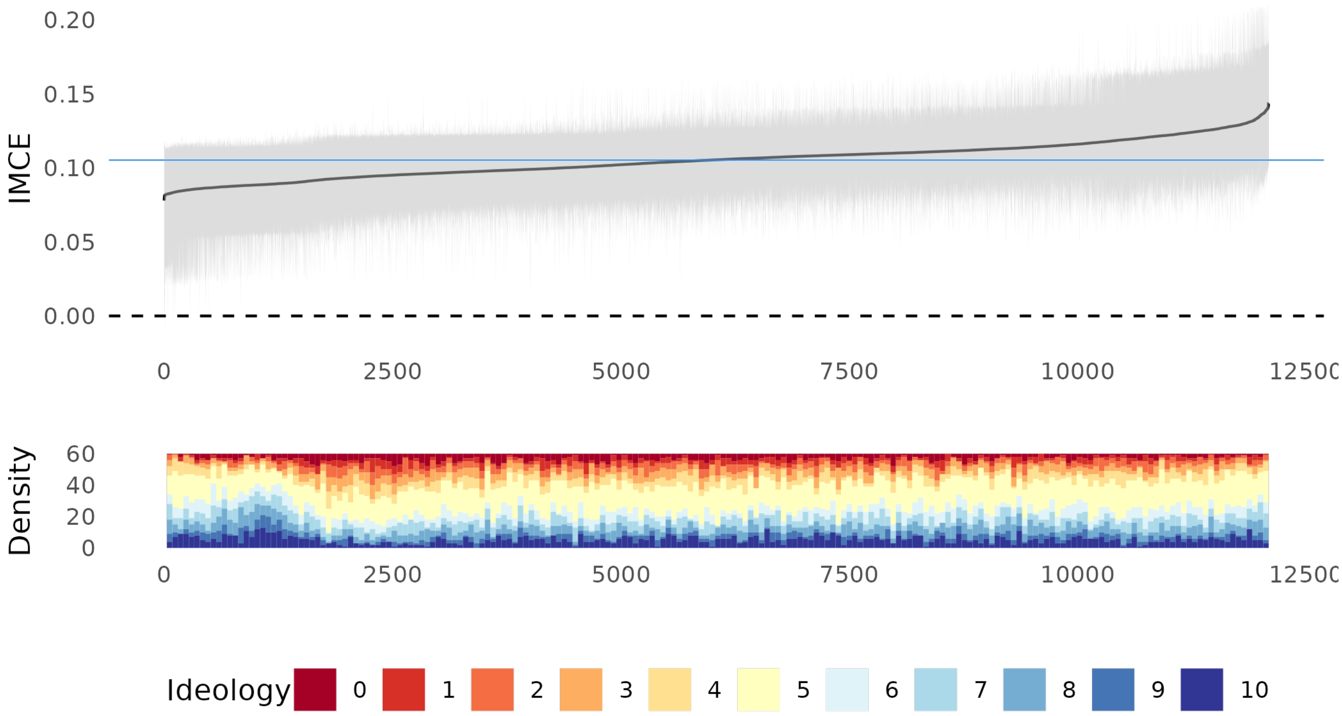


Figure G3. IMCE predictions by ideology values, for the “65 years old" attribute-level



References

Athey, Susan , Julie Tibshirani, and Stefan Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Athey, Susan and Stefan Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies 5*(2), 37–51.

Carnegie, Nicole Bohme and James Wu (2019). Variable selection and parameter tuning for bart modeling in the fragile families challenge. *Socius 5*, 2378023119825886.

Chipman, Hugh A. , Edward I. George, and Robert E. McCulloch (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics 4*(1), 266–298.

de la Cuesta, Brandon , Naoki Egami, and Kosuke Imai (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis 30*(1), 19–45.

Duch, Raymond , Laurence S. J. Roope, Mara Violato, Matias Fuentes Becerra, Thomas S. Robinson, Jean-Francois Bonnefon, Jorge Friedman, Peter John Loewen, Pavan Mamidi, Alessia Melegaro, Mariana Blanco, Juan Vargas, Julia Seither, Paolo Candio, Ana Gibertoni Cruz, Xinyang Hua, Adrian Barnett, and Philip M. Clarke (2021). Citizens from 13 countries share similar preferences for covid-19 vaccine allocation priorities. *Proceedings of the National Academy of Sciences 118*(38).

Hainmueller, Jens , Daniel J. Hopkins, and Teppei Yamamoto (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis 22*(1), 1–30.

Hill, Jennifer , Antonio Linero, and Jared Murray (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application 7*(1).

Kapelner, Adam and Justin Bleich (2016). bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software 70*(4), 1–40.

Sparapani, Rodney , Charles Spanbauer, and Robert McCulloch (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software 97*(1), 1–66.

Tibshirani, Julie , Susan Athey, Erik Sverdrup, and Stefan Wager (2022). *grf: Generalized Random Forests*. R package version 2.2.0.

Zhirkov, Kirill (2022). Estimating and using individual marginal component effects from conjoint experiments. *Political Analysis 30*(2), 236–249.