How to detect heterogeneity in conjoint experiments*

Thomas S. Robinson† Raymond M. Duch‡

Abstract

Conjoint experiments are fast becoming one of the dominant experimental methods within the social sciences. Despite recent efforts to model heterogeneity within this type of experiment, the relationship between the conjoint design and lower-level causal estimands is underdeveloped. In this paper, we clarify how conjoint heterogeneity can be construed as a set of nested, causal parameters that correspond to the levels of the conjoint design. We then use this framework to propose a new estimation strategy, using machine learning, that better allows researchers to evaluate treatment effect heterogeneity. We also provide novel tools for classifying and analysing heterogeneity post-estimation using partitioning algorithms. Replicating two conjoint experiments, we demonstrate our theoretical argument, and show how this method helps estimate and detect substantive patterns of heterogeneity. To accompany this paper, we provide new a `R` package, **cjbart**, that allows researchers to model heterogeneity in their experimental conjoint data.

*This version: March 17, 2022. This study was partially funded by the National Agency for Research and Development ANID, and we acknowledge the research support provided by FONDECYT 2020 grant number 1201397. We also acknowledge the Health Economics Research Centre at the University of Oxford Nuffield Department of Population Health for their support of the CANDOUR project and its Director, Professor Philip Clarke.

†Assistant Professor. School of Government and International Affairs, Durham University. Contact: Al-Qasimi Building, Elvet Hill Road, Durham, DH1 3TU, UK. Email: thomas.robinson@durham.ac.uk

‡Professor. Nuffield College, University of Oxford. Contact: Nuffield College, New Road, Oxford, OX1 1NF, UK. Email: raymond.duch@nuffield.ox.ac.uk. Phone: +44 (0)1865 278515

Conjoint experiments are an increasingly popular method of experimental research in discrete-choice contexts. In the last decade, the number of papers per year that mention "conjoint experiments" has risen sixfold, from 110 articles published in 2010 to 600 published in 2020.[1] Conjoint designs offer researchers an efficient means of recovering multiple causal parameters across a wide range of research areas, including radical right voting (Chou et al. 2021), tax preferences (Ballard-Rosa et al. 2017), asylum-seeking (Bansak et al. 2016), ethical principles guiding machine behavior (Awad et al. 2018), product design and pricing decisions (Orme 2010), and contemporary drivers of migration (Spilker et al. 2020; Duch et al. 2020).

The predominant causal quantity estimated in conjoint experiments is the average marginal component effect (AMCE; Ganter 2021), which is formally defined as "the effect of a particular attribute value of interest against another value of the same attribute while holding equal the joint distribution of the other attributes in the design, averaged over this distribution as well as the sampling distribution from the population" (Bansak et al. 2021, 29). While theoretically complex, in practise this quantity is easily estimated using conventional regression techniques, and allows researchers to isolate the average effect of attributes on the probability of choosing a profile.

This *average* quantity, by definition, may mask significant heterogeneity in subjects' behaviour. Researchers often want to know whether AMCEs differ depending on non-randomised characteristics of the subjects who take part in the study. To do so, studies typically present separate models for distinct sub-groups within the data.[2] Despite the simplicity of this approach, we know subgroup analysis is suboptimal. First, the strategy presumes that we have strong theoretical and empirical reasons to restrict our analysis to

[1] Based on a Google Scholar keyword searches for "conjoint experiment".

[2] Bansak et al. (2016), for example, run separate estimations for Left/Right, Young/Old, High/Low Education, and Above/Below Median Income respondents, Rehmert (2020) compares effects for veteran versus first-time delegates at party primaries, and Spilker et al. (2020) conduct subgroup analyses on age, income, education and location.

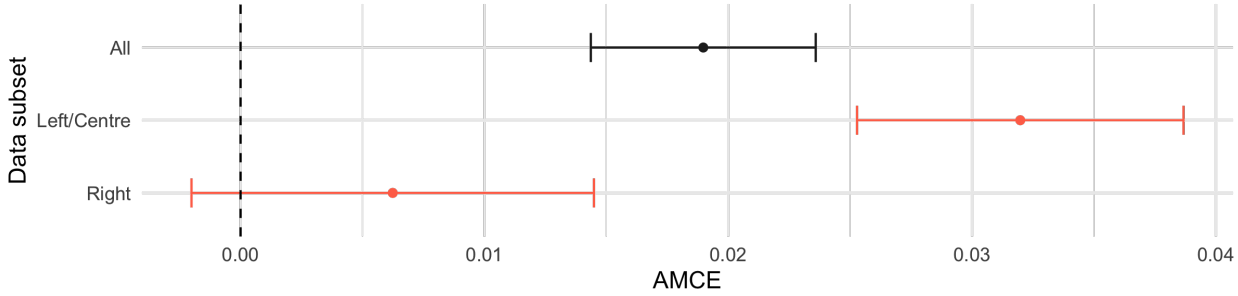
a set of sub-groups. Beyond convenient dichotomous splits in the data, subgroup analysis becomes unwieldy once researchers want to consider more complex groups of respondents. And typically we have under-developed notions about the functional form of the heterogeneous data generating process. Second, directly interpreting subgroup differences across models can be misleading if each subgroup's preference differs over the reference level (Leeper et al. 2020). Third, subgroup analyses reduce the number of observations in each model, increasing uncertainty by preventing the models from "borrowing" shared variation between subsets of the data.

We propose a strategy for detecting and characterizing heterogeneity in these marginal effects that overcomes these limitations by exploiting the richness of the data generated in conjoint experiments. We build on a growing corpus of work highlighting the utility of machine-learning methods in experimental settings (Hill 2011; Green and Kern 2012; Wager and Athey 2018; Künzel et al. 2019) to develop a new technique that allows researchers to estimate heterogeneous treatment effects in conjoint experiments. We also present strategies to *characterise* the potential drivers of this heterogeneity, without imposing *a priori* assumptions about the form it takes.

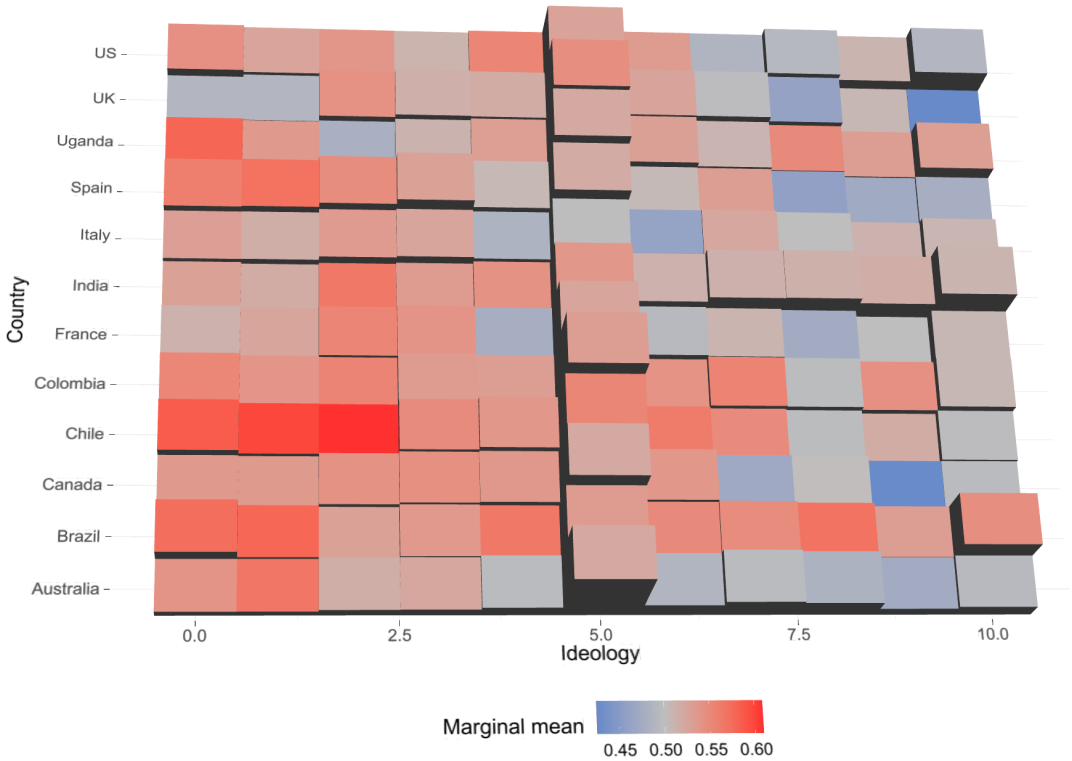
To illustrate the challenge that scholars face, consider a recent conjoint experiment conducted by Duch et al. (2021). This 13-country conjoint experiment asked subjects to choose amongst two profiles of potential prioritized recipients of a COVID-19 vaccine. In Figure 1a we replicate the AMCE and sub-group estimates for hypothetical profiles who had low incomes. On average, subjects were more likely to choose profiles that were labelled low-income relative to those on an average income, and subgroup analyses suggest the effect of this attribute-level is conditioned by subjects' own ideological stance.

However, Figure 1b suggests the narrative is not quite as simple as the subgroup analysis would suggest. Here we plot the selection probabilities (by colour) and densities (by height) for low-income profiles, conditioning on both ideological self-identification and the

Figure 1. Impact of respondents' ideology on choosing to prioritise vaccinating low-income profiles



(a) AMCE estimates for the "Lowest 20% income-level" attribute-level, estimated on the full data and subsets containing Left/Centre and Right-leaning subjects respectively.



(b) Proportion of profiles selected (marginal mean) that contain the "Lowest 20% income-level" attribute-level, by subjects' ideology and country. The height of the bars reflect the number of observations in each cell.

country of each subject in the Duch et al. (2021) study. The data comprising this plot covers 82,503 forced choices made by the 15,536 participants in the study. While we do see a general trend that right-leaning subjects are less likely to prioritise low-income profiles, there is quite clear heterogeneity in this relationship *across* contexts. For some countries – like Brazil, Uganda, and India – the relationship is far less pronounced.

The subgroup strategy demonstrated in Figure 1a fails to capture this cross-country variation in part because it involves a more complicated, *a priori* specification of subgroups. In turn, running separate models for left and right-identifying subjects would entail estimating 24 separate models, each powered by far fewer observations. Even if this were feasible, this strategy would still omit variation within the dichotomous ideological splits. For example, subjects in Canada and the UK both exhibit notable variation *within* right- and left-leaning subjects respectively.

In contrast to the subgroup approach, our strategy can identify heterogeneity in treatment effects without imposing any *a priori* functional form on the model. In other words, given the rich data that is generated by the conjoint design, we allow the model itself to find interactions between randomised conjoint attributes and subjects' characteristics. Moreover, by training a model on all available observations our strategy leverages the full density of the experimental data, improving the precision and power of the estimates.

In this study we make three novel contributions to the study of treatment effect heterogeneity in conjoint experiments. First, we address a recent interest in subject-specific or conditional treatment effects within conjoint analysis (Abramson et al. 2020; Zhirkov 2021). Our contribution is to clarify how these lower-level causal quantities are situated within the structure of conjoint designs: we present a simple derivation of nested causal effects that disaggregates the AMCE to the level of the individual, round, and observation within the experiment.ches

Second, we build on non-parametric approaches that leverage machine learning esti-

mators in order to estimate heterogeneity in experimental treatment effects (Hill 2011; Green and Kern 2012; Wager and Athey 2018; Künzel et al. 2019; Duch et al. 2020). We estimate heterogeneous treatment effects by predicting counterfactual treatment outcomes at the observation-level; these estimated effects are then aggregated to produce higher-level treatment effect estimates. To do so, we propose a non-parametric strategy based on Bayesian Additive Regression Trees (BART) – a highly-flexible, easy-to-implement algorithm that is suited to estimating heterogeneity in experimental treatment effects (Hill 2011; Green and Kern 2012; Duch et al. 2020). Unlike typical subgroup analyses and other recent approaches that focus on modelling each individual separately (Zhirkov 2021), this model is able to leverage the full support of the data rather than relying on much smaller subsets of observations. As part of this approach, we also provide variance estimators that exhibit good coverage, allowing researchers to quantify the uncertainty over these predicted effects.

Third, we provide a strategy for characterising the extent and types of heterogeneity once we have estimated the nested causal quantities. Our method builds on recent developments in the machine learning field that allow for bias-corrected estimates of the importance of variables within random forest models (Ishwaran and Lu 2019). We repurpose this strategy to measure how important different subject-level covariates are for partitioning the distribution of estimated individual-level marginal component effects, and thus identify which variables are driving treatment effect heterogeneity.

We demonstrate our approach with two different datasets: first, we replicate an archetypal conjoint study on preferences over immigrants conducted by Hainmueller et al. (2013), and second we use the recent (Duch et al. 2021) conjoint study of global preferences over Covid-19 vaccination policies to explore preference heterogeneity. To accompany the paper, we also provide a new `R` package – *cjbart* – that implements our proposed method, allowing researchers to estimate and analyse treatment effect heterogeneity within con-

joint experiments. This package is available on the Comprehensive R Archive Network (CRAN).

In the next section, we first deconstruct the typical parameter of interest in conjoint studies indicating its relation to lower-level quantities of interest. Section 2 then provides details on our estimation strategy. Section 3 provides new tools for characterising and analysing the resultant distribution of individual-level quantities of interest. Finally, Section 4 presents an applied demonstration of our approach.

1 Nested causal quantities in conjoint designs

Conjoint discrete choice experiment allow for efficient estimation of multiple causal parameters that affect preferences. Subjects are presented with profiles that randomly vary along a defined set of attributes. Since the value of each attribute is simultaneously randomised (we call the possible values “attribute-levels"), and subjects make a discrete choice over which profile they “prefer",[3] we can estimate the marginal effects of each attribute on subjects' choices through repeated observation.

In this section, we demonstrate how the recoverable parameters from this type of experiment can be conceived of as a series of nested causal quantities, related to the multi-level structure of the conjoint design. We restrict our focus to cases where there is complete randomisation of each attribute-level.[4] This assumption simplifies the analysis and estimation of the causal parameters, and is the typical design employed by researchers in practice.

Suppose N individuals (indexed by i) choose between J profiles across K rounds of the experiment. Within each round of the experiment, we randomly assign attribute-levels across L attributes for each profile (Hainmueller et al. 2013). Having run the experiment,

[3]Note that this preference may not strictly be in utility terms. For example, conjoint designs have asked subjects to consider which candidate subjects believe to be more corrupt (Spencer and Theodoridis 2020).
[4]In other words, where the probability of assigning each attribute-level is constant within each attribute and entirely independent of the values of other attributes.

the researcher faces a data structure with $N \quad J \quad K$ rows and $L + X$ columns (where X are any covariates observed for each subject), from which causal parameters of interest can be estimated.

The most common parameter estimated from this design is the **average marginal component effect** (AMCE). This estimand is theoretically complex. It reflects the overall effect of a specific attribute-level on the probability of choosing a profile (compared to some baseline reference level), after accounting for the possible effects of the other attributes in the design. To account for these other effects, the parameter is averaged over the effect variations caused by these other attributes. We defer describing the potential outcome estimand to Appendix Section A, and instead here focus on the conditional expectations from the data for ease of interpretation.[5]

With complete randomisation of these attributes, the AMCE for a single attribute-level can be estimated from the observed data as

$$\hat{\gamma}_l = \mathrm{E}[Y_{ijk} | T_{ijkl} = l_1] \quad \mathrm{E}[Y_{ijk} | T_{ijkl} = l_0];$$

where Y_{ijk} is a binary indicator of whether an individual (i) chose a given profile (j) in a round (k) of the conjoint experiment, T_{ijkl} is the value randomly assigned to the lth attribute for that specific profile, l_1 is the attribute-level of interest, l_0 is a reference category.[6] As a difference-in-means, and again assuming complete randomisation of each attribute's value, this quantity is easily recovered using linear probability models (LPMs).

By definition, the AMCE captures the central tendency of subjects' behavior with respect to each attribute of the design. Often, however, we are interested in whether these

[5] Appendix Section A describes the Hainmueller et al. (2013) potential outcome estimand of the AMCE and shows how the lower-level causal quantities defined in this paper relate to it. We keep our notation similar to Hainmueller et al. (2013), though we make some slight abbreviations to improve the succinctness of our specification.

[6] Strictly, the AMCE for the attribute-level l_1 should be denoted γ_{l_1} since there may be more than two possible values within attribute l. We drop the subscript in this paper for clarity.

effects differ dependent on subject characteristics or the context of the experiment. For example, Duch et al. (2021) run the same conjoint in multiple countries to assess whether individuals around the world exhibit the same behavior over prioritising Covid-19 vaccinations. Each country-model in their design reflects a subgroup analysis of the total sample data.

As others have also noted, the AMCE can be disaggregated into more granular causal quantities of interest (Abramson et al. 2020; Zhirkov 2021). Here we formalise this logic with respect to the structure of the data generating process itself. First, we can disaggregate the AMCE into the N individual-level effects. By the law of iterated expectations, we can state the AMCE $_{l} = \mathsf{E}[\ _{il} | i]$; where i indicates the respondents of the conjoint survey. In other words, the average marginal component effect is the average over the individual-level effects ($_{il}$). This lower-level parameter is the **individual-level marginal effect** (IMCE), and reflects the change in probability for a specific subject i of choosing a profile given an attribute-level (compared to some reference category) averaged over the effects of all other attributes.

By a similar logic to the AMCE, the IMCE can be estimated as:

$$\hat{}_{il} = \mathsf{E}[Y_{ijk} | T_{ijkl} = l_1; i] \quad \mathsf{E}[Y_{ijk} | T_{ijkl} = l_0; i]:$$

This specification is similar to a subgroup analysis of AMCEs, but rather than subsetting the data along a single covariate at some (arbitrary) split along that dimension, we consider the conditional effect that corresponds to all covariate dimensions associated with subject i (both observed and unobserved).

This individual-level marginal effect is substantively useful because it allows researchers to inspect heterogeneity in the treatment effects derived from conjoint experiments (Abramson et al. 2020), and is commensurate with more general heterogeneous effect estimation strategies (Künzel et al. 2019). By recovering a vector of individual-level estimates, we

can compare how non-randomised aspects of the data (i.e. subjects' characteristics) correspond to the magnitude and direction of the individual-level predicted effects.

The IMCE in turn can also be considered an average, this time over the repeated observations taken for that individual (i.e. the choices over profiles subjects make across multiple rounds of the conjoint experiment). This aggregation is more complicated because, in conventional conjoint designs, subjects see J profiles per round and make a forced choice *between* these profiles.[7] First, therefore, we can disaggregate the **round-level marginal component effect** (RMCE). This is the effect of a component within a specific round (k) of the experiment for a given individual, and is an average effect when $J \quad 2$:

$$\hat{}_{ikl} = \mathsf{E}[Y_{ijk} j T_{ijkl} = l_1; i; k] \quad \mathsf{E}[Y_{ijk} j T_{ijkl} = l_0; i; k]:$$

Finally, we can see that the RMCE is itself an average across the individual profiles of that round – a key characteristic of the forced choice conjoint design. We define the **observation-level marginal component effect** (OMCE) as

$$\hat{}_{ijkl} = [Y_{ijk} j T_{ijkl} = l_1; i; j; k] \quad [Y_{ijk} j T_{ijkl} = l_0; i; j; k]:$$

Unlike the IMCE and RMCE, the OMCE is by definition a counterfactual parameter. It relies on a wholly unobservable feature of the world: an alternate treatment schedule in which just a single attribute-level were changed for profile j presented to individual i in round k of the experiment. In other words, there is no averaging across observations. Consequently, we conceive of the OMCE as the most granular theoretical causal quantity of interest within the conjoint design.

While the OMCE is counterfactual, it serves a useful statistical purpose given its analytical relationship to the IMCE. A conjoint design randomises attributes across both rounds and profiles of the experiment and we assume that there is no serial correlation to sub-

[7] Typically, conjoint experiments have 2 profiles per round.

jects' behavior – in other words, the observed effects are stable and do not carryover across rounds (Hainmueller et al. 2013). As a consequence, each OMCE, denoted $_{ijkl}$, can be thought of as an independent draw from the individual-level distribution. We can leverage this feature when making predictions about these individual-level marginal effects (as shown in Section 2.1).

To summarise, notice that we derived these three quantities – the IMCE, RMCE and OMCE – as disaggregations of the AMCE. In reverse, we can see that, under the assumption of a completely randomised design, the AMCE itself is a three-step aggregation: averaging over observations, rounds of the experiment, and then subjects. Algebraically, we can express this relationship as:

$$
\begin{aligned}
{}_{l} &= \mathsf{E}_i\left[\mathsf{E}_k\left[\mathsf{E}_j\left[{}_{ijkl}\right]\right]\right] \\
&= \mathsf{E}_i\left[\mathsf{E}_k\left[\mathsf{E}_j\left[Y\,jT_{ijkl} = l_1; i; j; k\right] - \left[Y\,jT_{ijkl} = l_0; i; j; k\right]\right]\right];
\end{aligned}
$$

where subscripts under the expectation symbol indicate over what level the conditional means are taken.

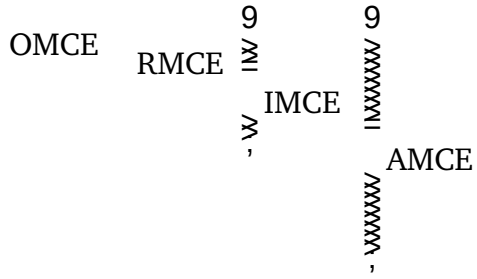
In short, conjoint designs allow researchers to estimate a series of nested causal quantities related to each other by the repeated application of the law of iterated expectations. In Table 1, we illustrate this relationship visually. Each respective estimand is a nested quantity that relates to the structure of the observed data collected via conjoint designs. As such, each estimand covers increasingly aggregate portions of the data.

2 Estimating the IMCE

Estimating lower-level marginal effects give us specific leverage over questions about the heterogeneity of these effects. We propose that the most effective level of analysis is the individual-level, since we can analyse how the IMCE varies dependent on characteristics

Table 1. Nested causal quantities in a conjoint experiment

Subject	Round	Profile	Attribute	⋮⋮⋮	y	$y_{l'}$
1	1	1	A	⋮⋮⋮	1	0
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.

of the subjects. In this section, therefore, we propose a three step strategy to recover estimates of the IMCEs.

First, we model the relationship between the forced choice outcome, conjoint attribute-levels, and subject-level covariates. This allows us to estimate some function that captures the potentially heterogeneous relationship between the conjoint attributes and subjects' characteristics when making choices in the experiment. Second, we use the trained model to predict counterfactual outcomes at the observation-level from which we can estimate OMCEs. Third, following the nested logic outlined in Section 1, we aggregate these OMCE estimates to the level of the individual in order to recover estimates of the IMCEs.

It is worth noting that researchers could use any number of possible estimators to model subject-level heterogeneity in the first step. We provide a specific implementation in this paper and accompanying software that uses Bayesian Additive Regression Trees (BART) (Chipman et al. 2010), but other researchers may wish to pursue alternative types of model. To that extent, the general approach detailed here can be considered a meta-strategy for estimating individual-level marginal effects in conjoint designs.

One key benefit of this meta-strategy is that *all* data is used when estimating the relationship between observed covariates, attribute-level assignments, and the conjoint outcome. This is in contrast to both subgroup analysis (where effects are modelled for only a smaller number of individuals who share a covariate value) and more recent approaches that recommend running separate models for each individual (Zhirkov 2021). Particularly, when modelling each individual separately, the cost constraints on experimental survey length may lead to large imprecision in the estimates. In our proposed method, the model leverages the full support of the data, across all observations, to discover covariate interactions that modify the causal effect at the individual-level. In Section 4 we demonstrate the comparative performance of our method compared to a subset-based strategy.

Moreover, by using machine learning, this method improves the analysis of potential heterogeneity in two ways. First, it reduces researcher degrees of freedom to arbitrarily run many subgroup analyses, which we would expect to inflate the chances of false positive discoveries. Second, it enables the identification of more complex relationships between variables. Common to many machine-learning methods, the model itself (rather than the researcher) determines the final functional form of the relationship between the supplied predictor variables and the outcome.

2.1 Parameter estimation

Step 1 In the first step, we use BART to model potential heterogeneity in the observed experimental data defined as:

$$P(Y_{ijk} = 1 | T_{ijk}; X_i) = f(T_{ijk}; X_i) \quad \hat{f}(T_{ijk}; X_i);$$

where Y_{ijk} is the observed binary outcome, T_{ijk} is the vector of treatment assignments across the L attributes, and X_i is the vector of covariate information for subject i considering profile j in round k of the experiment. f is some unknown true data generating

process, and $\hat{f}$ is an estimated model of that function.

BART is a tree-based supervised machine learning strategy which models the response surface by *summing* the predictions of many constrained individual tree models (Chipman et al. 2010). Unlike random forest models, the outcome is not the average across a set of complex trees. Instead each tree is a "weak learner" that seeks to explain only the *residual* variance in the outcome not explained by the T 1 other trees. BART models also have convenient Bayesian properties that allow us to recover variance estimates at the IMCE level, which we discuss below.

To estimate the BART model, we supply a matrix of "training" data at the observation-level. The training data are simply the results of the conjoint experiment. Each row reflects a profile within a round shown to a specific subject. The matrix columns comprise the observed individual decision (0 or 1) regarding that profile; the assigned attribute-levels for each of the L attributes in the vignette (which vary within individuals); and covariate columns that are invariant at the individual-level. During training, the BART algorithm iterates through the trees in the model, many times over, updating the tree parameters to minimize the error between a vector of predictions $\hat{Y}$ and the observed outcomes Y.[8]

Step 2 Using the final trained model ($\hat{f}$), we predict counterfactual outcomes (i.e. whether the profile was selected or not) changing the value of attribute-levels. Specifically, to recover a vector of OMCE estimates of attribute-level l_1, we take z draws from the predicted posterior using a "test" matrix which is identical to the training dataset, except each element in the column corresponding to attribute l is set to the value l_1.[9] We then repeat

[8] We use a probit-specific version of BART that better handles the binary outcome typical of this type of discrete-choice design. The probit outcomes are transformed back to probabilities prior to the computation of OMCEs.

[9] In our software implementation, $z = 1000$ – the default used in the Sparapani et al. (2021) implementation of BART in R. These draws are taken using a Gibbs Sampler, obtained through a Monte Carlo Markov Chain (MCMC) backfitting algorithm. Chipman et al. (2010) show that, with sufficient burn-in, these sequential draws converge to the posterior of the true data generating process (p.275).

this process, except the value of this column is now set to l_0, the reference category. This process yields two separate matrices of dimensions $z \quad N$, which approximate the posterior distribution for each observation for two separate attribute values respectively (l_1 and l_0). Subtracting these two matrices yields a single matrix of predicted OMCE estimates – z per observation. Finally, therefore, to recover a parameter estimate of the OMCE, we simply average these z predictions for each observation to yield a vector of observation-level effects:

$$\text{OMCE} = \hat{\tau}_{ijkl} = \frac{1}{z} \quad \hat{f}(T_{ijkl} = l_1; X_i) \quad \hat{f}(T_{ijkl} = l_0; X_i) \quad :$$

Step 3 Finally, consistent with the logic outlined in Section 1, the IMCE estimates can then be calculated by averaging the OMCEs for each individual i:

$$\text{IMCE} = \hat{\tau}_{il} = \frac{1}{J \quad K} \quad \hat{\tau}_{ijkl}.$$

2.2 Uncertainty estimation

We can also use the $z \quad N$ matrix of predicted OMCEs from the BART model to estimate the uncertainty both at the observation and individual level. We propose two separate approaches: a credible interval that allows for a Bayesian interpretation of the parameter uncertainty, and a confidence interval that allows for frequentist inference.[10]

Aggregated credible intervals. The Bayesian approach takes the 1 posterior interval of the OMCE-level predictions. To aggregate this interval to the IMCE level, we concatenate the posterior draws for each OMCE estimate, and take the =2 and (1)=2 quantiles of this combined vector. Given that the posterior distribution is a random variable, this credible interval indicates the central 1 proportion of the probability mass for the parameter's posterior. In other words, since the parameter itself is random in the Bayesian

[10]Both strategies are implemented in the accompanying statistical software **cjbart**.

framework, we are straightforwardly estimating the range that the parameter will likely fall in.

Combined parameter estimates. To make frequentist claims about the parameter estimate, we treat the uncertainty estimation as akin to a multiple imputation problem (Rubin 2009). The OMCEs reflect individual draws from the IMCE, each with an estimated variance. To calculate the total variance of the IMCE, we can "combine" these estimates as a weighted sum of the within and between variance components:

$$\hat{V}(\hat{\pi}_{il}) = V_W + \left(1 + \frac{1}{J \quad K}\right) V_B$$

$$\hat{V}_W = \frac{1}{J \quad K} \sum_{j} \sum_{k} V(\hat{\pi}_{ijkl})$$

$$\hat{V}_B = \frac{\sum (\hat{\pi}_{ijkl} \quad \hat{\pi}_{il})^2}{(J \quad K) \quad 1}.$$

The confidence intervals derived from this approach allow users to make claims about the likelihood of any given interval containing the true (non-random) parameter value.

2.3 Simulation tests of the estimation strategy

Using Monte Carlo simulations, we find that our method effectively detects IMCE heterogeneity caused by heterogeneous *preferences*. We simulate a full conjoint experiment in which subjects make choices between two profiles. Each profile contains three conjoint attributes that are randomly assigned one of two values: $A_1 = \{a, b\}; A_2 = \{c, d\}; A_3 = \{e, f\}$. To induce heterogeneity, we define subjects' *preferences* over attribute levels as a function of two individual-level covariates – varying this relationship across attributes. The first covariate c_1 is a binary variable drawn from a binomial distribution of size 1 with probability 0.5; the second covariate c_2 is a continuous variable drawn from a uniform distribution with bounds [-1,1].

We define the change in utility as a result of observing the second level for each attribute as follows:

$$U_{A_1} \begin{cases} N(\quad = 1; \quad = 1); & \text{if } c_1 = 1 \\ N(\quad = \quad 1; \quad = 1); & \text{otherwise.} \end{cases}$$

$$U_{A_2} \quad N(\quad = jc_2 \quad 0{:}2j; \quad = 1)$$

$$U_{A_3} \quad N(\quad = 0; \quad = 0{:}5)$$

We then simulate the conjoint experiment run on 500 subjects, for 5 rounds each, in which individuals choose between 2 profiles. For each observation, we calculate the utility for subject i given profile j in round k as:

$$U_{ijq} = \mathrm{I}(A_1 = b) \qquad U_{A_1} + \mathrm{I}(A_2 = d) \qquad U_{A_2} + \mathrm{I}(A_3 = f) \qquad U_{A_3} + \quad ;$$

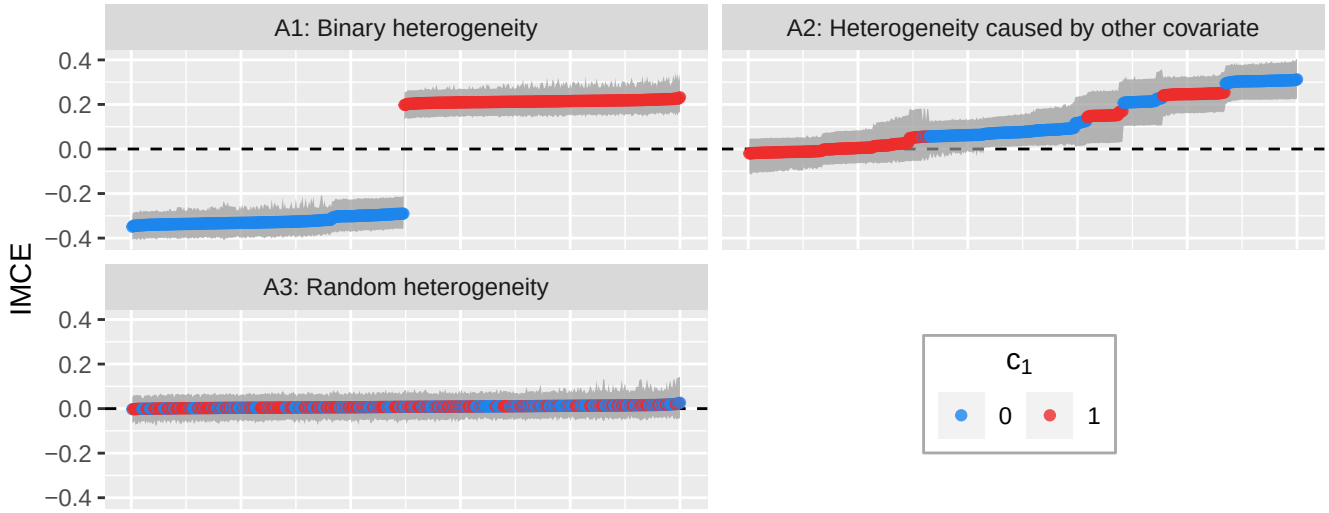
where $N(0; 0{:}0005)$ adds a small amount of noise to each utility calculation (to prevent exact draws).

For each round j that subject i sees, the profile that yields the higher change in utility is "chosen" ($Y = 1$), and the other is not ($Y = 0$). Therefore, we mimic the technical dependence between observations that forms the basis of the discrete choice design.

Given this specification, the BART estimation strategy should predict heterogeneous IMCEs for the first two attributes (A1 and A2) but not for the last attribute (A3). Moreover, when we inspect the correlation between IMCEs and the covariate c_1 we should find that the IMCEs for A1 are positive when $c_1 = 1$, but negative when $c_2 = 0$. We should not observe a correlation between c_1 and the continuous heterogeneity in A2.

Figure 2 demonstrates the results of this experiment. Our strategy effectively discovers heterogeneous IMCEs when the heterogeneity over preferences is a function of a binary variable – the positive and negative preferences perfectly correspond to the values of this covariate. Conversely, in the third facet, the completely random assignment of utility

Figure 2. Detecting heterogeneity in IMCEs using simulated conjoint data derived from preferences over profiles



Point estimates of the IMCEs for 500 subjects shown with 95% Bayesian intervals (as described in Section 2.2)

across individuals yields no sign of heterogeneity in IMCEs nor correlation between c_1 and the size of effects.

Importantly, the model also effectively discovers heterogeneity in the IMCE for A2 but does not exhibit correlation with c_1. This separation between heterogeneity detection and its correlation with covariates is important. Under a conventional, subsetting strategy, the analyst would likely also note that conditional AMCEs for A2 do not covary with c_1. However, they would not be able to tell, subsetting based on c_1, that there is substantial heterogeneity to the marginal component effect. We conjecture that as the complexity of the covariance between covariates and IMCEs increases it will become harder for the analyst to adequately pre-specify models that would be capable of detecting this heterogeneity.

Over 100 simulations of this exercise, we calculate the correlation between the covariate c_1 and the three distributions of IMCEs pertaining to the three attributes respectively. On average, the correlation between c_1 and A1 is 0.998. By comparison, the average cor-

relation between c_1 and A2 and A3, where *a priori* neither are related to c_1, are 0.002 and 0.004 respectively.

We extend this discussion of the simulated performance of our method in the Appendix. In Section B1 we demonstrate that the estimation method exhibits good predictive accuracy when IMCEs themselves are simulated across DGPs of varying form and complexity. We also find that our two variance estimation strategies both exhibit good coverage of the 95% interval (Section B2). Finally, we test whether RMCEs can be used to detect whether effects are serially correlated by round (a violation of a conjoint experiment's assumptions) in Section B3.

2.4 Applied test of BART-estimated AMCEs

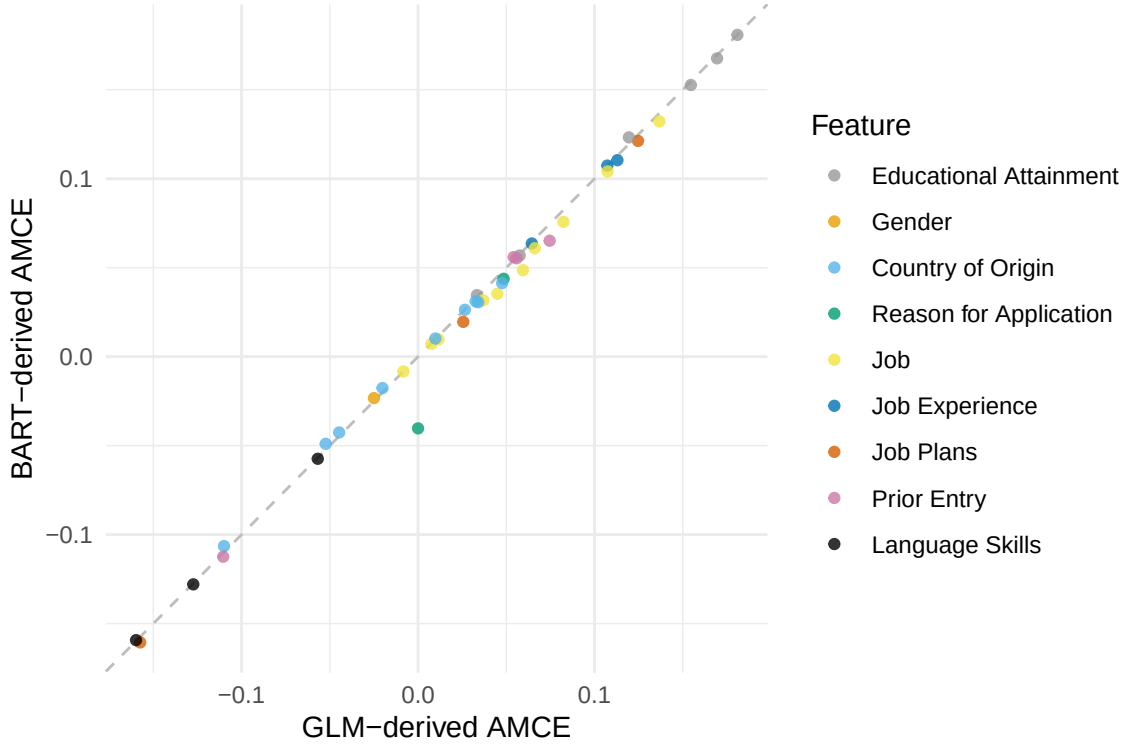
In Section 1 we note that the AMCE can be considered the double aggregation of the hypothetical observation-level effects. As an applied sense check of our method, we demonstrate this empirically using data from an archetypal conjoint experiment by Hainmueller et al. (2013). In the experiment U.S. subjects made a series of forced choices between two profiles describing potential immigrants, indicating which they would prefer to admit. The attributes presented in the profiles reflected traits hypothesized to matter in typical immigration decision making, including the migrant's profession, country of origin, and language skills.

We compare the *average* marginal component effects estimated by a linear probability model (the conventional estimating strategy) to the AMCEs recovered by averaging over the IMCEs estimated using our heterogeneous effects strategy. We expect to see similar estimates across both strategies. Full estimation details are provided in Appendix Section C.

Figure 3 compares the point estimates of the AMCE under our BART strategy and those of the conventional linear probability model approach. For all but one coefficient, we see

that the predicted effects are very similar. This is strong *prima facie* evidence that the BART model is appropriately estimating the response surface: the individual-level effects do, in aggregate, reflect the parametric estimation of the AMCE. The only coefficient for which there is disagreement between the two strategies is the 'seek better job' migration motivation level – this is because the linear probability model failed to recover an estimate due to co-linearity with the reference category. Interestingly, the flexibility of BART allows us to recover a point estimate for this parameter.

Figure 3. Comparison of conventional GLM-derived AMCE to AMCEs recovered from the BART estimated IMCEs



The off-diagonal point on this line is for the attribute-level 'seek better job' which could not be recovered from the parametric model due to colinearity with the reference category.

3 Comparing Sources of Heterogeneity

A particular attraction of heterogeneous effects estimation is that we are able to test whether treatment effects differ at the individual-level. To date, however, we have lacked principled methods of characterising any observed heterogeneity. In this section, we therefore propose two tools researchers can use to systematically recover indicators of *which* covariates are driving heterogeneity in the marginal treatment effects and the interactions between variables.[11] Both tools rely on tree-based learning methods to group the predicted IMCEs based on covariate information. In general, tree-based modelling approaches are well suited to this type of problem since they work by partitioning the outcome variable into clusters (or terminal nodes) where the differences in outcomes between members of the same cluster are as small as possible (Breiman et al. 1984).

We first introduce a standardised variable importance (VIMP) measure that summarises how well different covariates predict each distribution of IMCEs. This standardised measure can be used to explore the potential sources of heterogeneity in the marginal component effects systematically across all attributes in the experiment. Second, we show how single regression trees can subsequently be fit to better inspect the determinants of heterogeneity for specific attribute-levels of interest. This second step builds on the VIMP analysis by using the tree's decision rules to identify clusters, defined by subject covariates, that best define this heterogeneity. For each cluster, researchers can recover the conditional marginal component effect and thus also analyse the extent of heterogeneity in the treatment effects.

[11] Both tools are implemented in our R package **cjbart**.

3.1 Random forest variable importance

Our first tool summarises which covariates matter for predicting differences in the IMCE distributions for all attribute-levels in a conjoint experiment. We use random forests to estimate the relationship between the predicted IMCEs and subject-level covariates. Random forests operate by estimating many separate decision-trees, where the training data is bootstrapped across trees, and each tree considers only a random subset of variables. The result is an ensemble model that is less prone to bias (Breiman 2001). We then use variable importance metrics recovered from the trained random forest to identify variables that are particularly predictive of heterogeneity. In turn, these variables can drive subsequent analyses which we present in Section 3.2.

More formally, for each attribute-level, we train a random forest to model the heterogeneity in the predicted IMCE distribution. We use bold notation here to denote that we are estimating a vector of IMCEs, using a matrix of subjects' covariate information (X) as the predictor variables:

$$\hat{\tau}_{il} \quad \hat{f}^{\mathrm{RF}}(X).$$

Once this model has been trained, we then recover variable importance measures (VIMPs) – a common form of model analysis for tree-based methods – to understand which covariate dimensions are most useful for partitioning the data (Archer and Kimes 2008). In general, VIMP measures work by deliberately permuting a predictor variable and measuring how much this random noise degrades the performance of the model. A higher degradation of performance is indicative that the variable in question is more predictive of the outcome. For our purposes, we use VIMP scores to measure how well the included subject covariates predict each vector of IMCEs. Higher importance scores suggest that partitioning the IMCEs on these variables is informative.

Standard permutation-based variable importance measures are known to suffer from significant bias, particularly when variables are measured on different scales or, in the case of nominal variables, have different numbers of categories (Strobl et al. 2007). We therefore use a recent adaptation of VIMPs that permutes the *assignment* of observations down the tree rather than the value of the observations (Ishwaran and Lu 2019). Under this strategy, whenever a tree splits on the target variable the observation is randomly assigned to a child-node. Moreover, this VIMP measure allows one to recover bias-corrected variance estimates of VIMP scores, by using delete-d jackknife estimation (see Shao and Wu 1989).

The importance of different subject-level covariates may differ dependent on the specific attribute-level in question. We therefore recover separate VIMP scores for each IMCE separately. In practise, we recover a VIMP importance score for each combination of covariate and attribute-level, allowing us to plot a heatmap of variable importance across the design as a whole. In Section 4 we demonstrate how this schedule of VIMP scores can be analysed to understand what drives heterogeneity for each attribute-level in the conjoint experiment.

3.2 Single decision tree partitioning

The random forest VIMP tool compares how well subject-level covariates predict each IMCE distribution. Given its reliance on random forests, however, it is less useful for substantively interpreting the partitioned IMCE space. The final model contains many trees, where each individual tree only considers a random subset of variables and a bootstrap sample of the data. We therefore propose a complementary tool that fits a single decision tree on an attribute-level of interest. Like the random forest model, the single-tree model recursively partitions the vector of IMCEs using a matrix of covariate information. Unlike the random forest method, since only one model is fit the individual splitting rules from

this tree can be directly interpreted and used to inspect the heterogeneity in the IMCEs.

Single tree models typically fit many splits to the data, making interpretation difficult. This feature reflects the inherent trade-off in machine learning methods between the complexity of the fit model and the risk of mispredicting observations. In other words, a more complex tree may reduce prediction error (in training) but the incurred complexity reduces the variance of the model (leading to overfitting). Therefore, to ensure the tree is interpretable, we follow the convention of “pruning" the fit model. Since the partitioning is recursive and “greedy", we know that earlier splits in the tree are those that provide the greatest leverage over differentiating observations.[12] By removing later splits, pruning has the effect of paring back the cluster definitions (i.e. the combination of decision rules) to a more parsimonious level.

In practice, trees are pruned by setting a complexity parameter (cp). In the case of continuous outcomes, this determines the minimum increase in the overall R^2 of the model needed in order for a split to be kept in the model (Therneau et al. 1997). For the purpose of interpreting IMCE heterogeneity, we find that a complexity parameter of about 0.02 is sufficient to constrain the decision-tree to a depth that is substantively meaningful – yielding about 2 - 3 levels of partitioning.

Post-pruning, we can use the fit model to describe the underlying heterogeneity in the IMCE distribution. One very useful feature of decision trees is that, in this context, the terminal nodes reflect the conditional average marginal component effects defined by the splitting rules in the tree. This is similar to estimating marginal component effects for specific subgroups. Crucially, unlike manual subsetting approaches where subgroups have to be specified *a priori*, with a decision tree the clusters are *discovered* during model fitting

[12] In a recursive partitioning algorithm, the first split selects that variable (and cutting point) which minimizes the loss function associated with the resulting two partitions of the data. This process is then repeated for each child node, holding fixed the initial split. The fact that the parent split is not re-evaluated once the next layer of decision rules are determined means that the algorithm is “greedy".

itself. This is particularly useful since the tree, splitting sequentially on multiple variables, may define complex groups. For example, it may find a stronger effect for subjects aged under 25 years old *and* who are ideologically left-leaning compared to left-leaning but older respondents. We illustrate this approach in the next section.

4 Analysing heterogeneity in a multi-national conjoint experiment

In this final section, we consider an application of the framework and estimation strategy outlined in Sections 1 and 2. We analyse causal heterogeneity in a very large conjoint dataset that encompasses a diverse group of subjects surveyed from 13 countries, and then compare our approach to a recent alternative strategy proposed in Zhirkov (2021).

Detecting heterogeneous effects Our data is taken from the Duch et al. (2021) multi-national study on Covid-19 vaccine prioritization. This experiment asks subjects to choose which of two hypothetical individuals should be given priority for a Covid-19 vaccine. Each profile displays five attributes – the recipients' vulnerability to the virus, likely transmission of the virus, income, occupation, and age – and all values are totally randomly assigned. Subjects make a total of 8 choices in the experiment. The data also contains information on *subjects*' country of origin, age, gender, ideology, income, education, hesitancy over vaccination, and measures of their willingness to pay for a vaccine.

The original study finds consistent AMCEs across all the countries surveyed. Nevertheless, it is reasonable to suspect that these AMCEs may mask heterogeneity with respect to individual-level covariates. This experiment is particularly suited to a study of heterogeneous effects, since with approximately 250,000 observations in total and harmonised covariate information across countries, there is ample data to model complex relationships (at the cost of computational intensity). To take advantage of the diversity of our data,

we train a BART model on all five conjoint attributes and the set of covariate information for each profile using **cjbart**, using all observations from the 13 countries surveyed in the experiment. From this model, we automatically recover a schedule of IMCE estimates for each attribute-level.

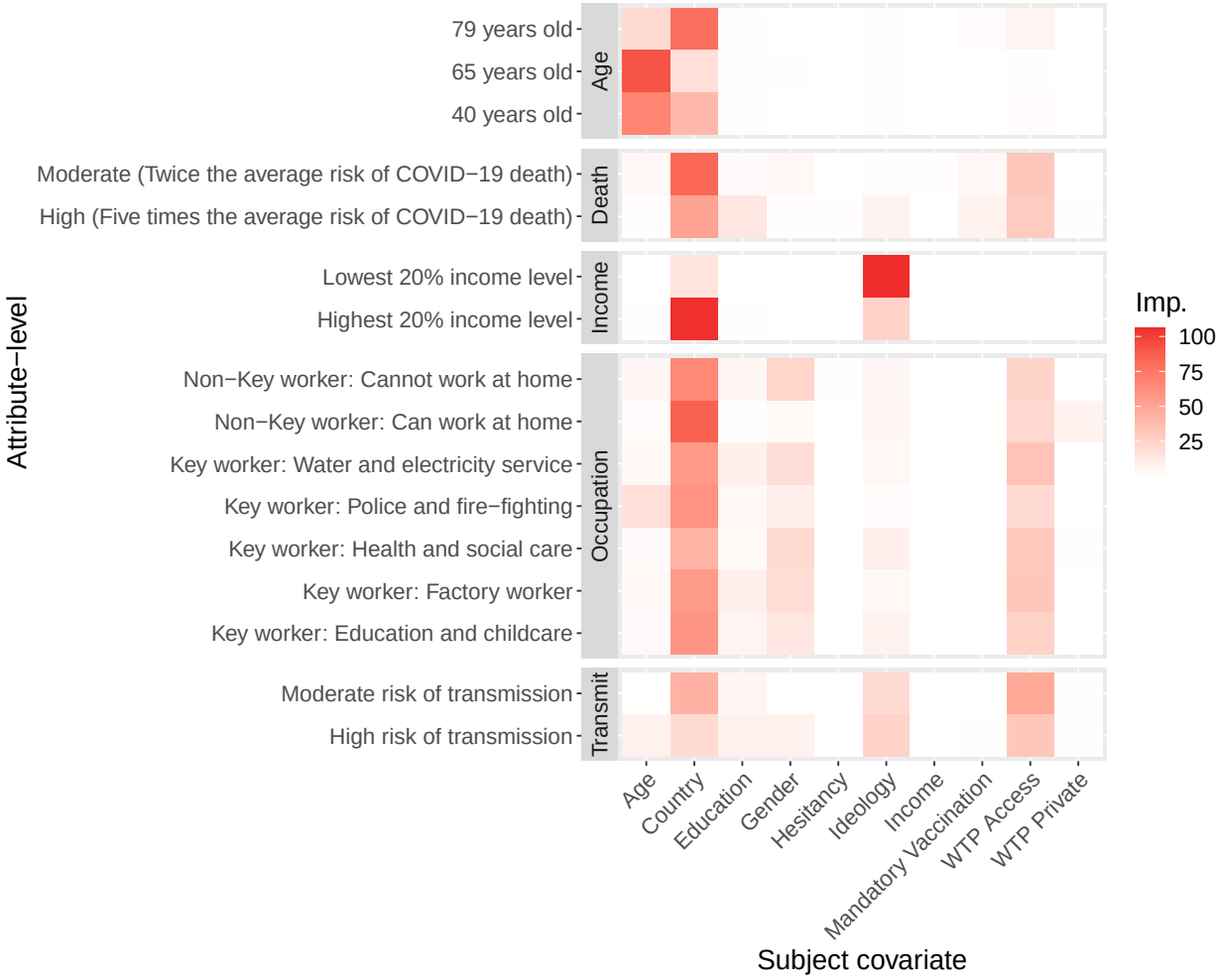
With multiple covariates, however, systematically identifying the drivers of heterogeneity is difficult. This is particularly acute in the case of conjoint experiments where we have separate IMCE vectors for each attribute-level, which means researchers are faced with a dense schedule of predicted effects. We address this challenge by using the tree-based measure of variable importance, as discussed in Section 3.1.

We use our proposed VIMP tool as the first step in identifying plausible sources of heterogeneity in the schedule of IMCEs estimated from the Duch et al. (2021). The method estimates a standardised importance score for each combination of the 10 covariates and 16 attribute-levels in the conjoint design. Figure 4 provides a graphical summary of how well each covariate predicts the attribute-levels in the Duch et al. (2021) conjoint. Clearly, the country of a respondent is a highly predictive factor across most attribute-levels in the model. This is perhaps unsurprising, given the diversity of contexts considered and differing levels of Covid-19 infections at the point the experiment was fielded.

Most interestingly, some subject-level variables appear to condition IMCEs for specific attributes. For example, while subjects' age is not a particularly important predictor of heterogeneity across most attributes, it is very predictive when considering the age of the potential vaccine recipient. In general, this suggests that whether one is willing to prioritise individuals based on age may well be driven by one's own age (which we explore in more detail below), and second that this is perhaps most important for the 65 year old label where the risks of Covid-19 begin to become more severe.

Similarly, ideology appears particularly important when partitioning the IMCEs related to the potential vaccine recipient's income. This finding accords with conventional expec-

Figure 4. Variable importance matrix having estimated separate random forest models on each attribute-level in the model. Higher values indicate variables that were more important in terms of predicting the estimated IMCE distribution



tations about the relationship between political ideology and service provision, but does usefully highlight that one's own ideological position does appear to predict how willing one is to prioritise those on low incomes.

Given the results from the VIMP summary measure, we can use our second proposed tool described in Section 3.2 – a single pruned decision tree – to inspect this heterogeneity in more detail. On the basis of the variable importance heatmap in Figure 4, for example, we would expect that subjects' age is used to partition the IMCE vectors for prioritising

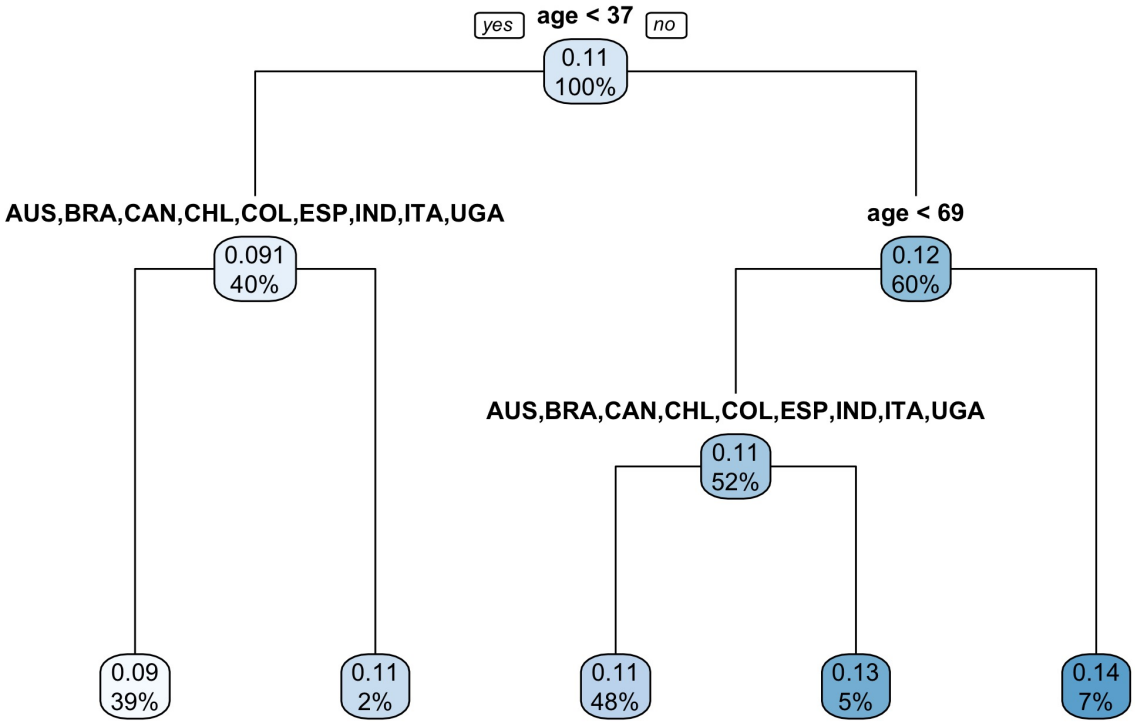
subjects of different ages.

Figure 5 presents a single decision tree for the IMCEs related to prioritising vaccines for “65 year olds". Note first that the split confirms the VIMP analysis results in Figure 4 that identify subject’s age as an important source of heterogeneity for the “65 years old" attribute-level. Here we can clearly see that the initial split is between those subjects aged younger than 37 or not. Notably, while all final clusters have average marginal component effects that are positive, older subjects (over the age of 37) exhibit a predicted average marginal effect (0.12) that is about 33 percent larger than younger subjects. Moreover, in the terminal nodes of the tree, the group with the highest predicted average effect is that defined solely by being older than 69 – suggesting some egocentric aspect to vaccine prioritisation beliefs.

These two complementary tools, the VIMP and single decision tree splitting, provide a comprehensive and robust way to identify sources of treatment effect heterogeneity in conjoint experiments. Finally, we show one further way of summarizing these results visually by plotting the full ordered distribution of IMCEs for a given variable against the corresponding distribution of a covariate. In Figure 4 we noted that subjects’ ideology was an important predictor of IMCEs for the income-related attribute-levels in the conjoint experiment. In Figure 6, therefore, we visualize this particular relationship by plotting the IMCEs against a histogram of subjects’ self-reported ideology.

As Figure 6 shows, there is quite clear and distinct heterogeneity. The IMCE effect line exhibits a marked jump, at which point the marginal component effect almost doubles. Those lower IMCEs (around the 0.01 mark) are individuals whose ideology is right-leaning (at or above 6 on a 0-10 scale). In contrast, after the jump in the effect, subjects are much more left-leaning. Clearly, ideology does not play a perfect role. Within these two portions of the distribution, varying degrees of ideology are more uniformly distributed, and at the very right of the IMCE distribution we see that something else appears to be driving a

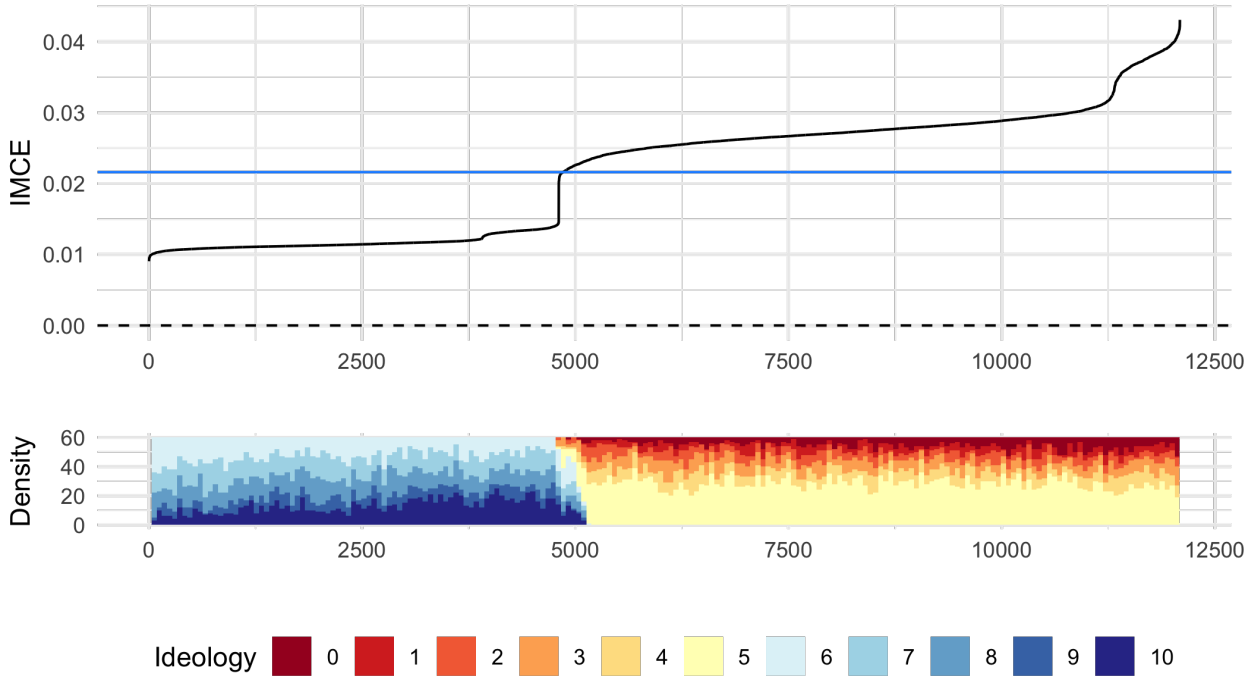
Figure 5. Pruned decision-tree of predicted IMCEs for prioritising vaccines for those “65 years old", using subject-level covariate information to partition the vector of individual-level effects.



further uptick in the predicted IMCE, to approximately four times the effect size of right-leaning subjects. In Appendix Section 4 we show that an alternate, OLS-based estimation strategy fails to detect this heterogeneity and proves difficult to use in conjoint experiments without very large numbers of observations per subject (as in this vaccination study).

Comparison to OLS-based approach To demonstrate the comparative performance of our approach, we also estimate IMCEs using an alternative strategy proposed recently by Zhirkov (2021). In short, this method estimates separate OLS regression models for each subject. The resultant coefficients are unbiased estimates of the same IMCE quantity we outline in Section 1.

Figure 6. 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 our method we find a strong correlation between individuals’ ideology and the predicted IMCEs for the low income attribute-level of the Duch et al. (2021) experiment. Under the Zhirkov (2021) OLS strategy, we would expect to see a similar result – both in terms of the distribution of IMCEs and its correlation with individuals’ self-reported ideology. To test this expectation, we estimate separate linear probability models for each individual in our data and, again, compare the ordered distribution of IMCEs to a corresponding histogram of subjects’ ideology.

Two practical features of the regression approach complicate this analysis using OLS. Since each subject completed eight rounds of the conjoint experiment (a number we think is quite typical for a conjoint design), each model has only 16 observations (2 profiles per round) and thus the individual models will be imprecise. Zhirkov (2021) directly acknowledges this limitation, and notes that the OLS approach ideally requires subjects to

rate closer to 30 profiles in total. We believe that, while this large number of activities may be feasible in principle, it goes beyond the budgetary constraints of most researchers.

Moreover, even when the number of observations approaches 30, Zhirkov (2021) recommends using interval rating scales rather than the binary, forced-choice outcome. While many conjoints implement both a rating and forced-choice scales of measurement, we believe the forced-choice outcome is the most interesting aspect. It allows us to think of the effects directly in terms of marginal probabilities, and thus to consider the behaviour of subjects (a choice of candidate) rather than just an attitude (the subjects' rating of two candidates).

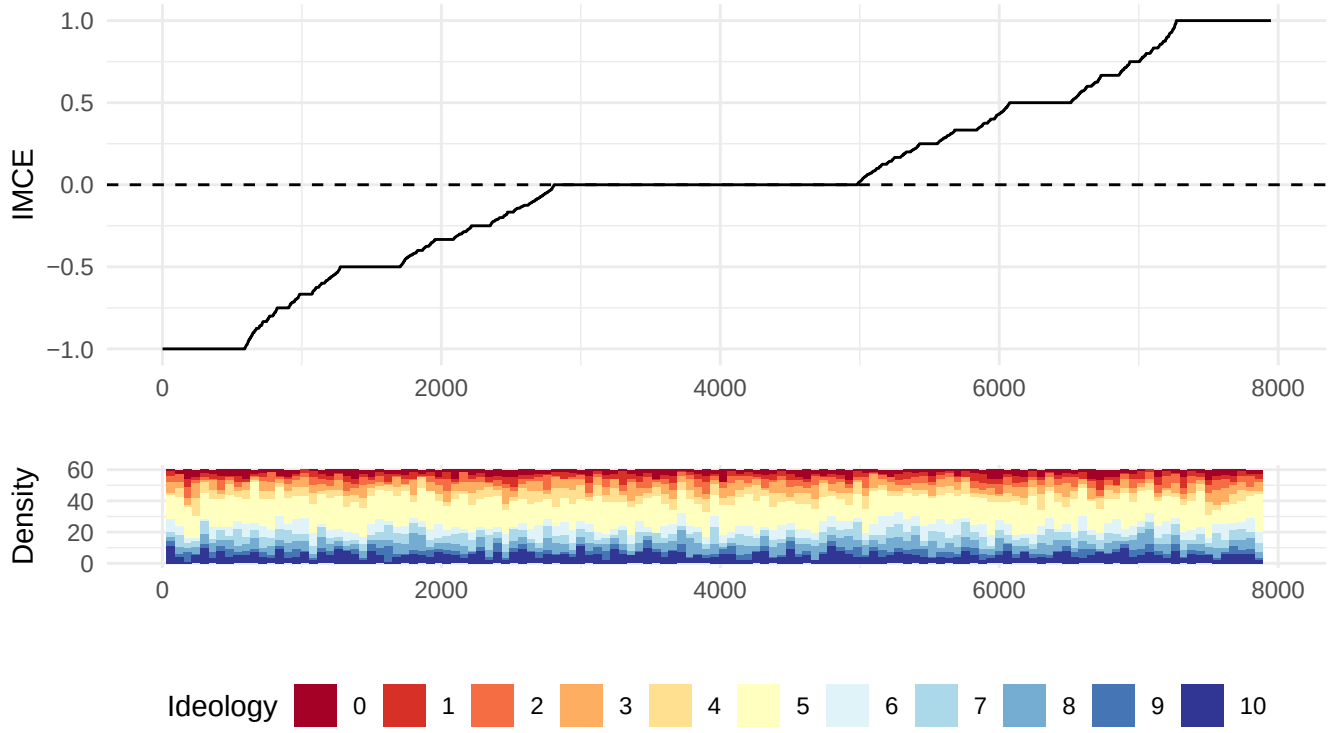
Using the Duch et al. (2021) data, the OLS approach yields 5369 IMCE estimates outside of the range that reflect possible changes in probability (-1 to 1). In our analysis below, we exclude all estimates that fall outside of this range. Excluding these estimates, however, leads to a 34 percent reduction in the number of IMCEs we can inspect.[13]

Figure 7 displays the ordered distribution of estimated IMCEs using this OLS strategy, plot against a histogram of individuals' self-reported ideology. We do not observe the same correlation as in our BART estimation. The correlation coefficient between the IMCEs and ideology in the OLS case is negligible and statistically insignificant ($r = \ 0{:}01; \ p = 0{:}20$) compared to a strong correlation with respect to BART ($r = \ 0{:}75; \ p < 0{:}001$). Looking at the distribution of IMCEs, moreover, the OLS strategy does not seem to have modelled the data well. The distribution is very symmetric, centred on zero and with tails that contain implausibly large effects.

While these are not the ideal conditions for the Zhirkov (2021) approach, we do believe the vaccine experiment resembles a typical conjoint design. As such, we believe our ML-based approach has some distinct advantages. In particular, since the ML heterogeneity model leverages all observations in the data, our estimation strategy is less reliant on

[13]Of these individuals, only 5935 uncertainty estimates were parametrically recoverable.

Figure 7. Comparison of estimated IMCEs using OLS method proposed in Zhirkov (2021), on the “Lowest 20% income level" attribute-level within Duch et al. (2021)



having many observations per experimental subject. We are able to estimate IMCEs for each of 12,000 subjects in the Duch et al. (2021) experiment compared to just over 7800 using the OLS method.[14] Relatedly, our ML-derived IMCE estimates are, on face value, more plausible in terms of the magnitude of effects.

Perhaps most importantly, our approach is able to detect and capture how subject covariate information modifies the size and direction of these marginal component effects. The OLS method rests on the fact that this heterogeneity is implicitly detected when the marginal effects are modelled for each individual separately. In our proposed method, since the trees in the BART model can identify interactive effects between the supplied covariates and the attribute-levels, it can explicitly model these effect modifiers. The result,

[14]These counts exclude those individuals who did not answer the ideology question in the survey.

in this case, is that our method identifies the correlation between subjects' ideology and their treatment of low-income vaccine recipients in a way that the OLS strategy does not.

5 Discussion and Conclusion

The attraction of conjoint experiments is a rich data generating process that allows us to tease out the choice characteristics that shape individuals' decision making. This type of experimental design is fast becoming one of the dominant experimental methods within the social sciences. A rich methodological literature is developing that explores how advances in conjoint estimation can enhance its informative value. Others, for example, have explored how to improve generalizability by weighting profile distributions to their actual occurrence in the populations of interest (de la Cuesta et al. 2021), and how to use eye-tracking software to better understand the decision-making process and the processing of conjoint vignettes (Jenke et al. 2021).

We make a small contribution to this wider development, by clarifying how the conjoint design relates to the structure of the data collected, and how we can leverage the nature of this data generation to estimate heterogeneous treatment effects across conjoint attributes. Heterogeneity can be characterized in terms of a set of nested, causal estimands that correspond to the repeated observations across individuals, rounds, and profiles of the conjoint design. Using machine learning tools, we show how to estimate heterogeneous treatment effects in the conjoint design using the potential outcomes framework. Our strategy allows researchers to assess treatment effect heterogeneity in a straightforward and flexible manner.

We suggest that machine learning is particularly useful given its ability to identify more complicated relationships between predictor variables without the need for researchers to specify these *a priori*. By reducing researcher degrees of freedom, our proposed general

method provides a more robust means of analysing heterogeneity compared to *ad hoc* subgroup analyses. Moreover, since our estimation strategy leverages all observations in the modelling stage, our method has greater statistical power than approaches that rely on estimating separate subset models.

Our specific estimation strategy relies on BART – a powerful Bayesian method for estimating heterogeneity that lends itself well to the potential outcomes logic of our framework. To accompany this paper, we provide a new `R` package, **cjbart**, that allows researchers to use our method on their experimental conjoint data. More generally, however, our proposed meta-strategy could be used with other forms of modelling. For example, researchers may wish to use random forests or neural networks instead.

Finally, estimating heterogeneity – that is, generating individual-level estimates of treatment effects – is only half the battle. Once researchers recover these individual-level estimates, the challenge is to identify which dimensions drive differences in the effects. We provide two complementary tools that help researchers make sense of the estimated distribution of individual-level effects. We demonstrate how random forests variable importance measures (VIMP) can be used to summarise which variables are most important for predicting heterogeneity in the IMCEs. We then show how single regression tree models can be used to partition IMCE distributions into clusters, where the decision rules provide information about which covariates define those clusters. This paper also shows how these results can be visualized to aid analysis.

References

Abramson, Scott F. , Korhan Kocak, Asya Magazinnik, and Anton Strezhnev (2020, July). Improving preference elicitation in conjoint designs using machine learning for heterogeneous effects.

Archer, Kellie J and Ryan V Kimes (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis 52*(4), 2249–2260.

Awad, Edmond , Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan (2018). The moral machine experiment. *Nature 563*(7729), 59–64.

Ballard-Rosa, Cameron , Lucy Martin, and Kenneth Scheve (2017). The structure of american income tax policy preferences. *The Journal of Politics 79*(1), 1–16.

Bansak, Kirk , Jens Hainmueller, and Dominik Hangartner (2016, 09). How economic, humanitarian, and religious concerns shape european attitudes toward asylum seekers. *Science 354*.

Bansak, Kirk , Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto (2021). *Conjoint Survey Experiments*. Cambridge University Press.

Breiman, Leo (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L. , J. Friedman, C.J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*. Taylor & Francis.

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

Chou, Winston , Rafaela Dancygier, Naoki Egami, and Amaney A. Jamal (2021). Competing for loyalists? how party positioning affects populist radical right voting. *Comparative Political Studies 54*(12), 2226–2260.

de la Cuesta, Brandon , Naoki Egami, and Kosuke Imai (2021). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis*, 1–27.

Duch, Raymond , Denise Laroze, Thomas Robinson, and Pablo Beramendi (2020). Multimodes for detecting experimental measurement error. *Political Analysis 28*(2), 263–283.

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).

Duch, Raymond M , Adrian Barnett, Maciej Filipek, Laurence Roope, Mara Violato, and Philip Clarke (2021). Cash versus lotteries: Covid-19 vaccine incentives experiment. *medRxiv*.

Duch, Raymond M , Denise Laroze, Constantin Reinprecht, and Thomas S Robinson (2020). Nativist policy: the comparative effects of trumpian politics on migration decisions. *Political Science Research and Methods*, 1–17.

Ganter, Flavien (2021). Identification of preferences in forced-choice conjoint experiments: Reassessing the quantity of interest. *Political Analysis*, 1–15.

Green, Donald P. and Holger L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly 76*(3), 491–511.

Hainmueller, Jens , Daniel J. Hopkins, and Teppei Yamamoto (2013). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis 531*, 1–30.

Hill, Jennifer L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*(1), 217–240.

Ishwaran, Hemant and Min Lu (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine 38*(4), 558–582.

Jenke, Libby , Kirk Bansak, Jens Hainmueller, and Dominik Hangartner (2021). Using eye-tracking to understand decision-making in conjoint experiments. *Political Analysis 29*(1), 75–101.

Künzel, Sören R. , Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*(10), 4156–4165.

Leeper, Thomas J. (2020). *cregg: Simple Conjoint Analyses and Visualization*. R package version 0.4.0.

Leeper, Thomas J. , Sara B. Hobolt, and James Tilley (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis 28*(2), 207–221.

Orme, B. (2010). *Getting started with conjoint analysis: Strategies for product design and pricing research second edition*. Research Publishers LLC.

Rehmert, Jochen (2020). Party elites'preferences in candidates: Evidence from a conjoint experiment. *Political Behavior*.

Rubin, D.B. (2009). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley.

Shao, Jun and CF Jeff Wu (1989). A general theory for jackknife variance estimation. *The annals of Statistics*, 1176–1197.

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.

Spencer, Douglas M and Alexander G Theodoridis (2020). "appearance of corruption": Linking public opinion and campaign finance reform. *Election Law Journal: Rules, Politics, and Policy 19*(4), 510–523.

Spilker, Gabriele , Vally Koubi, and Tobias Böhmelt (2020, 07). Attitudes of urban residents towards environmental migration in kenya and vietnam. *Nature Climate Change 10*.

Strobl, Carolin , Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*(1), 1–21.

Therneau, Terry M , Elizabeth J Atkinson, et al. (1997). An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation.

Wager, Stefan and Susan Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Zhirkov, Kirill (2021). Estimating and using individual marginal component effects from conjoint experiments. *Political Analysis*, 1–14.

Online Appendix for "How to detect heterogeneity in conjoint experiments"

A Potential outcome estimands

Borrowing heavily from Hainmueller et al.'s (2013) formulation, we can express the AMCE in potential outcomes as:

$$\text{AMCE} = {}_{l_1} = \mathsf{E}\ Y_i(l_1; T_{ijk[-l]}; T_{i[-j]k}) \quad Y_i(l_0; T_{ijk[-l]}; T_{i[-j]k})\ j\ (T_{ijk[-l]}; T_{i[-j]k})\ 2\ \tilde{T}\ ;$$

where $i\ 2\ N$ indexes subjects, j indexes profiles within each round k of the experiment, l_1 is some attribute-level of interest within attribute l, l_0 is a reference category within the same attribute, $T_{ijk[-l]}$ reflects the assigned treatment vector for all other attributes in the design, $T_{i[-j]k}$ is the set of treatment vectors excluding T_{ijk}, and $\tilde{T}$ reflects the intersection of the support of the probability of observing specific realizations of $T_{ijk[-l]}$ and $T_{i[-j]k}$ with $T_{ijkl} = l_1$ and $T_{ijkl} = l_0$ respectively.

Since the lower-level causal estimands are conditional versions of the AMCE, we can express them as follows:

$$\text{IMCE} = {}_{il_1} = \mathsf{E}\ Y_i(l_1; T_{ijk[-l]}; T_{i[-j]k}) \quad Y_i(l_0; T_{ijk[-l]}; T_{i[-j]k})\ j\ (T_{ijk[-l]}; T_{i[-j]k})\ 2\ \tilde{T}; n = i\ ;$$

$$\text{RMCE} = {}_{ikl_1} = \mathsf{E}\ Y_i(l_1; T_{ijk[-l]}; T_{i[-j]k}) \quad Y_i(l_0; T_{ijk[-l]}; T_{i[-j]k})\ j\ (T_{ijk[-l]}; T_{i[-j]k})\ 2\ \tilde{T}; n = i; K = k\ ;$$

$$\text{OMCE} = {}_{ijkl_1} = Y_i(l_1; T_{ijk[-l]}; T_{i[-j]k}) \quad Y_i(l_0; T_{ijk[-l]}; T_{i[-j]k})\ j\ (T_{ijk[-l]}; T_{i[-j]k})\ 2\ \tilde{T}; n = i; J = j; K = k:$$

B Simulation protocols

B1 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. $_{X_1} = c_1 \quad c_2$). 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:

We can then generate an assignment schedule by sampling at random the attribute levels for $I \quad J$ 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.

Table A1. Hypothetical correlation between IMCEs and two covariate values: c_1 and c_2 are randomly drawn from uniform distributions

i	c_1	c_2	x_1	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

Suppose the probability of choosing the profile is calculated as:

$$P(Y_{ijk} = 1) = 0.5 + \mathbb{I}(X_1 = b)\ _{x_1} + \mathbb{I}(X_2 = d)\ _{x_2}.$$

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 A2 presents an example of how these probabilities would be calculated given random assignment of attributes across rounds, and the pre-defined IMCEs in Table A1.

Table A2. 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

Given Tables A1 and A2, we train the BART model on the actual attribute-level assignments, the observed covariates, and the outcome:

The BART model then estimates the OMCEs ($_{ijk}$) by making predictions of Y when X_1 is set to b *for all observations* and when it is set to a, and deducting these two values, as demonstrated in Table A4.

Finally, the IMCEs are recovered by averaging the predicted OMCE across observations

Table A3. 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

Table A4. Calculating the OMCE by deducting the predicted probabilities under the assumption of different attribute-levels

i	$\hat{Y}jX_1 = b$	$\hat{Y}jX_1 = a$	b_{ijkl}
1	0.63	0.5	0.13
1	0.71	0.6	0.11
⋮	⋮	⋮	⋮
I	0.29	0.5	-0.21

for the same individual. For example, for $i = 1$ the predicted IMCE is:

$$\hat{}_{il} = \frac{1}{J \quad 2}(0{:}13 + 0{:}11 + {:}{:}{:}) = 0{:}109{:}{:}{:}$$

Given we know the IMCE for this individual is 0.1, the prediction error for this specific subject is $\hat{}_{il} \quad {}_{il} \quad 0{:}109 \quad 0{:}1 \quad 0{:}009$. 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.

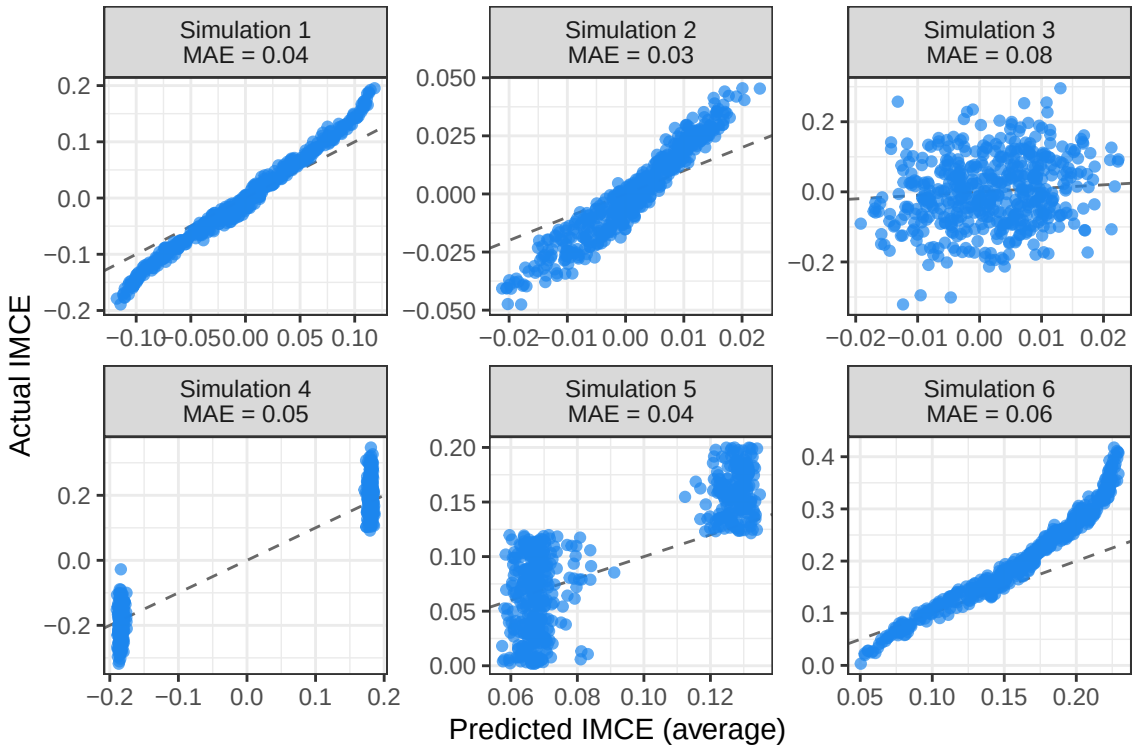
Table A5 summarises the six scenarios we consider. In short, simulations 1 and 2

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 A1 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 hetoregeneity 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 A1. 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.

B2 Coverage test

To test the frequentist and Bayesian variance estimators we propose, we run Monte Carlo simulations in which we pre-define the IMCEs for each subject and assess the coverage of the resultant confidence intervals derived from the Bayesian credible intervals and the Rubin's combination rule estimators. 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.

$$V(\backslash_{il}) = \frac{1}{J} \sum V(\backslash_{ijkl})$$

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 $_i$:

$$_{il} \quad N([C_{1i} \quad C_{2i}]; \ _i)$$

$$C_{1i}; C_{2i} \quad \text{Uniform}(0; c);$$

where c and $_i$ 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 A6 details the parameter settings used for each of the seven separate simulation tests we run.

Table A7 reports the coverage rates for the three variance estimation methods we consider. We find that, across different scenarios, both the Bayesian interval and confidence

interval strategies produce 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 both the Bayesian and frequentist intervals are closer to 0.95.

Table A5. Sources of heterogeneity in IMCEs, for each of 6 separate simulations

Sim.	f_{IMCE}	c		Details
1	$c_1 \quad c_2$	c_x	Uniform$(0; h = 0{:}2)$	Effects are linearly heterogeneous between h and h
2	$c_1 \quad c_2$	c_x	Uniform$(0; h = 0{:}05)$	As above, but the range is much smaller
3	$0{:}2(c_1 \quad c_2) + 0{:}8N\,(0; 0{:}125)$	c_x	Uniform$(0; h = 0{:}2)$	Covariates are a weak predictor of IMCE heterogeneity
4	If $c_3 = 1;\ N\,(0{:}2; 0{:}05)$; else, $N\,(\ 0{:}2; 0{:}05)$	c_3	Binomial$(1; 0{:}5)$	IMCE is either positive or negative dependent on observed binary variable
5	c_4 Uniform$(0; h = 0{:}2)$	$c_1 = 2 \quad \mathsf{I}(c_4 > 0{:}6h)$ $N\,(0; 0{:}25)$		IMCE is determined by unobserved covariate that also influences c_1.
6	$c_1 \quad 2^{c_2} + c_2$	c_x	Uniform$(0; h = 0{:}2)$	Exponential relationship between IMCE and covariates

Table A6. Simulation specifications testing the coverage rate of the confidence intervals

Sim.	Subjects	K	c	$_i$
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 A7. Comparison of coverage rates across the Bayesian and frequentist intervals.

Sim.	Naive Estimate	Parametric	Bayesian
1	0.965	0.973	0.977
2	0.996	0.997	0.996
3	0.990	0.994	0.992
4	0.938	0.948	0.954
5	0.919	0.928	0.933
6	0.962	0.971	0.975
7	0.950	0.958	0.965

We also compare our two proposed variance estimators to a naive variance estimator that is simply the average of the variance across OMCEs.

B3 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:

$$\begin{aligned} U_{ijk}^{\text{Round-effect}} &= N(0; 0{:}001) \\ &\quad + 0{:}5r \quad \mathbb{I}(A_{ijk} = a2) \\ &\quad + (0{:}6 \quad 0{:}1r) \quad \mathbb{I}(B_{ijk} = b2) \\ &\quad + 0{:}5 \quad \mathbb{I}(C_{ijk} = c2); \end{aligned}$$

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:

$$\begin{aligned} U_{ijk}^{\text{No round-effect}} &= N(0; 0{:}001) \\ &+ 1 \quad \mathbb{I}(A_{ijk} = a2) \\ &+ 0{:}2 \quad \mathbb{I}(B_{ijk} = b2) \\ &+ 0{:}5 \quad \mathbb{I}(C_{ijk} = c2){:} \end{aligned}$$

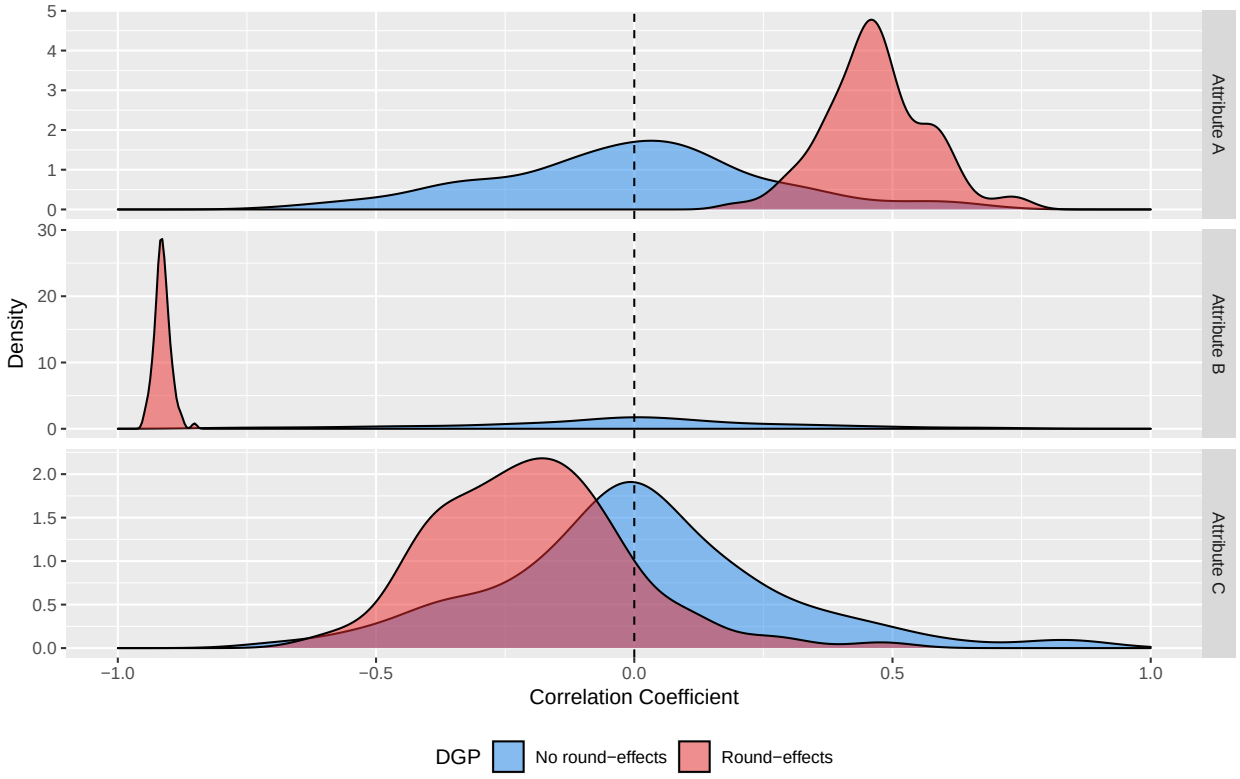
For each pair of profiles within the experiment, the profile that yields the higher utility gets assigned 1 and the other profile gets assigned 0. We calculate this separately for the round-effect and no round-effect utility calculations, yielding two experimental datasets.

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 A2 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 en-

suring this assumption holds is so important – it may lead to biased estimates of attributes even if they are individually "well-behaved.

Figure A2. Simulation evidence demonstrating how violations of the no carryover assumption can be detected by estimating the RMCE



C Hainmueller et al. (2013) estimation

Hainmueller et al. (2013) 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 using the **cregg** package within R (Leeper 2020). We estimate the following model:

$$\begin{aligned}\text{ChosenImmigrant} = \quad & + \ {}_1\text{Education} + {}_2\text{Gender} + {}_3\text{CountryOfOrigin} \\ & + \ {}_4\text{ReasonForApplication} + {}_5\text{Job} + {}_6\text{JobExperience} + {}_7\text{JobPlans} \\ & + \ {}_8\text{PriorEntry} + {}_9\text{LanguageSkills};\end{aligned}$$

where $_k$ 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.[15] We then plot these BART-estimated AMCEs against the parametric AMCEs as shown in Figure 3 in the main text. In Table A8 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.

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

Table A8. Comparison of AMCEs recovered using LPM and **cjbart** double-aggregation

		Coefficient		Difference
Attribute	**Level**	**LPM**	**cjbart**	**(% of LPM coefficient)**
Educational Attainment	4th Grade	0.03	0.03	3.31
	8th Grade	0.06	0.06	-1.63
	High School	0.12	0.12	2.75
	Two-Year College	0.15	0.15	-1.67
	College Degree	0.18	0.18	-0.43
	Graduate Degree	0.17	0.17	-1.44
Gender	Male	-0.02	-0.02	-7.00
Country of Origin	Germany	0.05	0.04	-13.54
	France	0.03	0.03	-1.24
	Mexico	0.01	0.01	3.07
	Philippines	0.03	0.03	-10.63
	Poland	0.03	0.03	-5.32
	China	-0.02	-0.02	-12.73
	Sudan	-0.04	-0.04	-5.09
	Somalia	-0.05	-0.05	-6.83
	Iraq	-0.11	-0.11	-3.53
Reason for Application	Seek Better Job	0.00	-0.04	-Inf
	Escape Persecution	0.05	0.04	-10.28
Job	Waiter	-0.01	-0.01	0.78
	Child Care Provider	0.01	0.01	-17.02
	Gardener	0.01	0.01	-5.74
	Financial Analyst	0.04	0.04	-21.72
	Construction Worker	0.04	0.03	-14.60
	Teacher	0.07	0.06	-8.23
	Computer Programmer	0.06	0.05	-18.54
	Nurse	0.08	0.08	-8.43
	Research Scientist	0.11	0.10	-3.60
	Doctor	0.14	0.13	-3.72
Job Experience	1-2 Years	0.06	0.06	-1.76
	3-5 Years	0.11	0.11	-0.32
	5+ Years	0.11	0.11	-2.66
Job Plans	Contract with Employer	0.12	0.12	-3.13
	Interviews with Employer	0.03	0.02	-23.96
	No Plans to Look for Work	-0.16	-0.16	1.60
Prior Entry	Once as Tourist	0.06	0.06	-1.30
	Many Times as Tourist	0.05	0.06	2.89
	Six Months with Family	0.07	0.06	-12.96
	Once w/o Authorization	-0.11	-0.11	1.38
Language Skills	Broken English	-0.06	-0.06	0.50
	Tried English but Unable	-0.13	-0.13	0.03
	Used Interpreter	-0.16	-0.16	-0.73