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 several scholars advancing novel ways to model heterogeneity within this type of design, the relationship between these new quantities and the conjoint design is underdeveloped. In this note, 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 that allows researchers to evaluate treatment effect heterogeneity and which exhibits good statistical properties. Replicating two conjoint experiments, we first demonstrate our theoretical argument, and then show how this method helps uncover interesting heterogeneity. To accompany this paper, we provide new a R package, **cjbart**, that allows researchers to model heterogeneity in their experimental conjoint data.

[*] 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. This design offers researchers an efficient means of recovering multiple causal parameters across a wide range of research areas, including tax preferences (Ballard-Rosa et al. 2017), asylum-seeking (Bansak et al. 2016), and ethical principles guiding machine behavior (Awad et al. 2018).

The causal quantity of interest typically reported in these experiments is the average marginal component effect (AMCE) – the relative effect of presenting a given attribute-level, holding all other attributes constant (Hainmueller et al. 2013). These treatment effects can differ depending on non-randomised characteristics of the subjects who take part in the study. We identify this treatment effect heterogeneity by deconstructing the conjoint data generating process into quantities that help indicate the presence of heterogeneity.

Two recent trends inform our approach. First, there is an interest in subject-specific or conditional treatment effects within conjoint analysis (Abramson et al. 2020; Zhirkov 2021). Our contribution is to situate these lower-level causal quantities within the structure of conjoint designs in a more general sense: we present a simple derivation of nested causal effects that disaggregate the AMCE to the level of the individual, round, and observation of the experiment.

Secondly, we build on non-parametric approaches that leverage machine learning estimators 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 are aggregated to produce higher-level estimates of treatment effects, including their uncertainty. 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).

We demonstrate our approach in two different contexts: a foundational conjoint study on preferences over immigrants conducted by Hainmueller et al. (2013), and a recent conjoint study exploring preferences over Covid-19 vaccination policies (Duch et al. 2021). 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 conjoint experiments. This package is available on the Comprehensive R Archive Network (CRAN).

We will first deconstruct the typical parameter of interest in conjoint studies indicating its relation to lower-level quantities of interest. In Section 2 we provide details on our estimation strategy. And in Section 3 we re-estimate the conjoint analysis conducted in Hainmueller et al. (2013) and in Duch et al. (2021).

1 Nested causal quantities in conjoint designs

In this section, we argue the recoverable parameters from conjoint experiments can be conceived of as a series of nested causal quantities, related to the multi-level structure of the conjoint design. The conjoint is a popular discrete choice experiment that allows for efficient estimation of multiple causal parameters that affect preferences. Suppose N individuals complete J rounds of the experiment choosing between Q profiles, within which we randomly assign attribute-levels across K attributes of each profile (Hainmueller et al. 2013). This $N \times J \times Q \times K$ structure, which is fully randomised, allows us to recover causal parameters over different levels of the data.

The most common parameter estimated from this design is the **average marginal component effect** (AMCE). This quantity reflects the effect of a specific attribute-level (compared to some baseline reference category) on the probability of choosing a given profile, on average across all N subjects, J rounds, and Q profiles. With a fully randomised design, the AMCE is

$$\tau_k = \mathbb{E}[Y|A_k = l, \mathbf{A}_{-k}] - \mathbb{E}[Y|A_k = l', \mathbf{A}_{-k}],$$

where Y is a binary indicator of whether the profile was chosen, A_k is the treatment assignment of the *k*th attribute in the conjoint design, l is the attribute-level of interest, l' is a corresponding reference category, and $\mathbf{A}_{-k}$ are the other $K-1$ conjoint attributes. Conveniently, assuming randomisation, this quantity is easily estimated using linear probability models (Hainmueller et al. 2013).

As other 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. First, we can think about disaggregating the AMCE effect into the N individual-level effects. By the law of iterated expectations, we can state the AMCE $\tau_k = \mathbb{E}[\tau_{ik}|i]$, 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. This lower-level parameter is the **individual-level marginal effect** (IMCE), which we will denote as τ_{ik} – it reflects the marginal effect of an attribute-level (compared to a reference category) conditional on subject i, and can be estimated as:

$$\tau_{ik} = \mathbb{E}[Y|A_k = l, \mathbf{A}_{-k}, i] - \mathbb{E}[Y|A_k = l', \mathbf{A}_{-k}, i].$$

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).

The 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). By recovering a vector of individual-level estimates, we can compare how non-randomised aspects of the data (i.e. subject' individual characteristics) correspond to the magnitude and direction of the predicted effect.

The IMCE in turn can also be considered an average, this time over the repeated observations taken for that individual (i.e. the individual choices over profiles within rounds of the conjoint experiment). This aggregation is more complicated because, in conventional conjoint designs, subjects see $Q = 2$ profiles per round and make a forced choice *between* these two profiles. The first level of disaggregation therefore is the **round-level marginal component effect**. This is the effect of a component within a specific round of the experiment for a given individual, and is an average effect when $Q \geq 2$:

$$\tau_{ijk} = \mathbb{E}[Y|A_k = l, \mathbf{A}_{-k}, i, j] - \mathbb{E}[Y|A_k = l', \mathbf{A}_{-k}, i, j].$$

Finally, we can further disaggregate the causal quantity of interest by expressing the *round*-level marginal component effect itself as an average across the individual observations of that round – a key characteristic of the forced choice conjoint design. We define the **observation-level marginal component effect** (OMCE) as

$$\tau_{ijkq} = [Y|A_k = l, \mathbf{A}_{-\mathbf{k}}, i, j, q] - [Y|A_k = l', \mathbf{A}_{-\mathbf{k}}, i, j, q].$$

Unlike the IMCE and RMCE, this 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 q presented to individual i in round j. Consequently, the OMCE reflects the most granular theoretical causal quantity within the conjoint experiment.

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 the AMCE itself is a three-step aggregation, averaging over observations, rounds of the experiment, and then

Table 1. Nested causal quantities in a conjoint experiment

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

subjects. Algebraically, this relationship can be expressed as follows:

$$\begin{aligned}\tau_k &= \mathbb{E}_i\Big[\mathbb{E}_j\big[\mathbb{E}_q[\tau_{ijqk}]\big]\Big]\\ &= \mathbb{E}_i\Bigg[\mathbb{E}_j\Big[\mathbb{E}_q\big[[Y|A_k = l, \mathbf{A}_{-\mathbf{k}}, i, j, q] - [Y|A_k = l', \mathbf{A}_{-\mathbf{k}}, i, j, q]\big]\Big]\Bigg],\end{aligned}$$

where subscripts under the expectation symbol indicate over what level the conditional means are taken. This same logic is also demonstrated visually in Table 1, where we demonstrate what observations are used in order to estimate each estimand.

The OMCE is useful because of its analytical relationship to the IMCE. Given randomisation across rounds of the conjoint experiment, we would expect there to be no serial correlation in these estimates – in other words, there is stability and no carryover in the rounds (Hainmueller et al. 2013). As a consequence, each OMCE, denoted τ_{ijk}, can be thought of as an independent draw from the individual-level distribution. We leverage this fact when making predictions about these individual-level marginal effects in Section 2.1., and show that in practise aggregating over observations then over individuals yields good estimates of the parametrically-derived AMCE. Moreover, in Section 2.4 we discuss how our method can be used to estimate the RMCE to test this stability assumption.

2 Estimating the IMCE

Our goal is to estimate the individual-level marginal component effect, so that we can then analyse how these vary dependent on characteristics of the subjects. We propose to recover estimates of the IMCEs in a two-stage process. First, we model the relationship between the forced choice outcome, conjoint attribute-levels, and subject-level covariates in order to estimate the OMCEs. The goal is to estimate some function $\hat{f}$ that models the potentially heterogeneous relationship, from which we can derive counterfactual predictions at the observation-level. To that extent, our approach is a meta-strategy whereby any prediction algorithm could be used. Second, we aggregate these OMCE estimates to the level of the individual in order to recover estimates of the IMCEs.

In the first stage, to estimate the OMCE, we use Bayesian Additive Regression Trees (BART) to fit a model on the experimental data. BART is a tree-based supervised machine learning strategy which attempts to model 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.

2.1 Parameter estimation

To estimate the underlying OMCE model, we supply a matrix of “training" data at the observation-level. This training data are simply the results of the conjoint experiment. Each row is a profile, within a round shown to a subject. The matrix columns comprise the observed individual decision (0 or 1) regarding that profile; the assigned attribute-levels for each of the K 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.[1]

The final trained model $\hat{f}$ can then be used to predict counterfactual outcomes (i.e. whether the profile was selected or not) given specific attribute values. Specifically, to recover a vector of OMCE estimates of attribute-level l, 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 k is set to the value l. We then repeat this process, except the value of this column is now set to l', the reference category.

Given the Bayesian nature of this estimation strategy, this yields two separate matrices of dimensions $z \times N$, where z is the number of *dependent* draws from the trained model using a Gibbs Sampler. Collectively these draws approximate the posterior distribution. 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:

$$\text{OMCE} = \hat{\tau}_{ijkq} = \frac{1}{z}\left(\hat{f}_{i,A_k=l,\mathbf{A_{-k}},j,q} - \hat{f}_{i,A_k=l',\mathbf{A_{-k}},j,q}\right).$$

As shown in Section 1, the IMCE estimates can then be calculated by averaging the J OMCEs for each individual i:

$$\text{IMCE} = \hat{\tau}_{ik} = \frac{1}{J}\sum_{j=1}^{J}\hat{\tau}_{ijqk}.$$

[1] 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.

2.2 Uncertainty estimation

We leverage the $z \times N$ matrix of predicted OMCEs from the BART model to estimate the uncertainty both at the OMCE and IMCE 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.[2]

Aggregated credible intervals. The Bayesian approach takes the $(1-\alpha)$ posterior interval of the OMCE-level predictions. To aggregate this interval to the IMCE level, we concatenate the posterior draws for the k observations to yield a $J \times z$ vector, and take the $\alpha/2$ and $(1-\alpha)/2$ quantiles of this vector. Given that the posterior distribution is a random variable, this credible interval should be interpreted as indicating the central $1-\alpha$ proportion of the probability mass for the parameter's posterior. In other words, since the parameter itself is random in the Bayesian 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:

$$\mathbb{V}(\hat{\tau}_{ik}) = \mathbb{V}_W + (1+\frac{1}{J})\mathbb{V}_B$$

$$\widehat{\mathbb{V}}_W = \frac{1}{J}\left(\sum_{i=1}^{J}\widehat{\mathbb{V}(\tau_{ijkq})}\right)$$

$$\widehat{\mathbb{V}}_B = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\tau}_{ijkq}-\hat{\tau}_{ik})^2}{J-1}}.$$

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

2.3 Simulation tests of the estimation strategy

For the sake of brevity, we defer a full discussion of the simulated performance of our method to the Appendix. Using Monte Carlo simulations, we find that our method effectively detects IMCE heterogeneity caused by heterogeneous *preferences* (Section A1), and exhibits good predictive accuracy when IMCEs themselves are simulated, across DGPs of varying form and complexity (Section A2). We also find that our two variance estimation strategies both exhibit good coverage of the 95% interval (Section A3).

2.4 RMCE estimation

To estimate the RMCE, we can average the OMCEs within rounds of the experiment for each individual, rather than over all observations pertaining to that individual. This step 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 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 can then assess whether this is true by correlating the aggregated RMCEs with the same round indicator. If there are no carryover effects, in expectation the correlation should be zero. In Appendix Section A4 we demonstrate this logic via simulation.

3 Analysing heterogeneity in two conjoint experiments

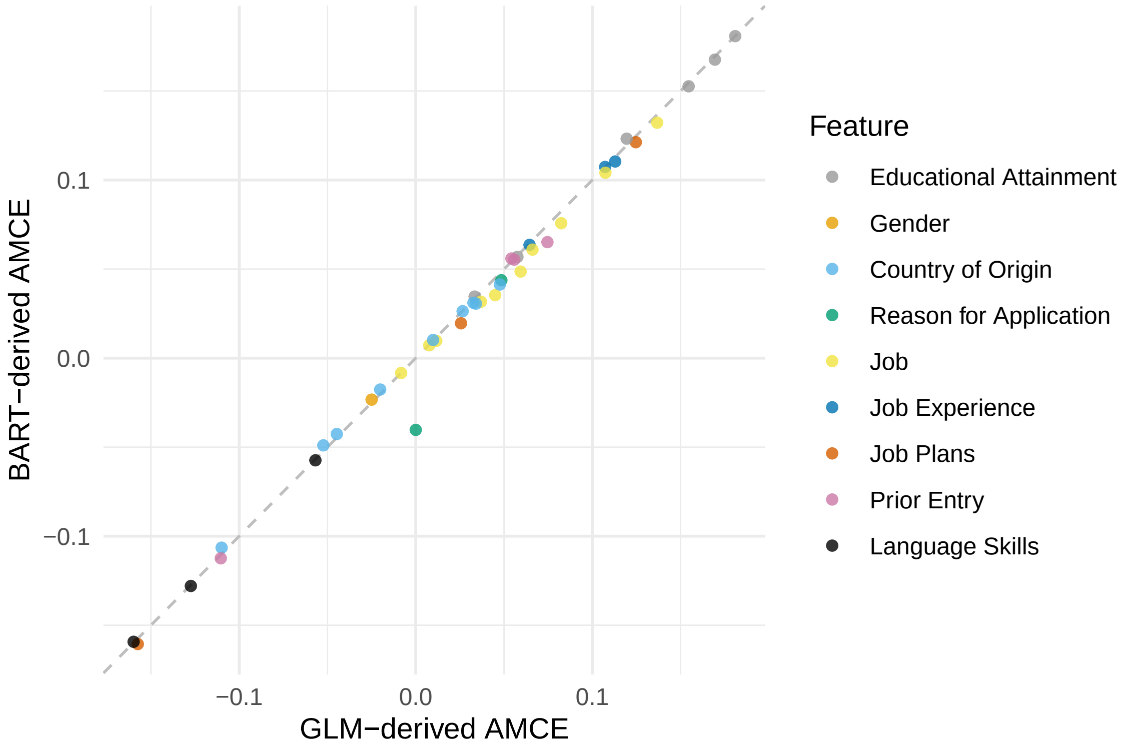
In this final section, we consider two applications of the framework and estimation strategy outlined in Sections 1 and 2. First, we demonstrate empirically the relationship between counterfactual observation-level effects, estimated with BART, and conventional AMCEs from OLS estimation. Second, we analyse causal heterogeneity in a very large conjoint dataset that encompasses a diverse group of subjects from 13 countries.

BART-estimated AMCEs In Section 1 we note that the AMCE can be considered the double aggregation of observation-level effects. We demonstrate this is the case empirically, using the immigration experiment from Hainmueller et al. (2013). We compare the *average* marginal component effects estimated by a linear probability model to the AMCEs recovered from our heterogeneous effects strategy. Our expectation is that the AMCE generated by averaging over all individual-level estimates from the BART estimation should be similar to those estimated by the parametric approach. Estimation details are provided in Appendix Section B.

Figure 1 compares the point estimates of the AMCE under our BART strategy and those of the parametric approach. For all but one coefficient, we see that the predicted effects are essentially the same. 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 parametric estimation approach fails to recover an estimate due to colinearity with the reference category. Interestingly, the flexibility of BART allows us to recover a point estimate for this parameter.

Detecting heterogeneous effects Next, we consider a recent conjoint experiment conducted across 13 countries over preferences for Covid-19 vaccine prioritisation (Duch et al. 2021). The study finds consistent AMCEs across all the countries surveyed. Nevertheless, 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). We demonstrate one way to leverage IMCE predictions in order to understand

Figure 1. 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.

how causal effects vary by subject-level characteristics.

The experiment asks subjects to choose which of two hypothetical individuals should be given priority for a Covid-19 vaccine. Each profile's vulnerability to the virus, likely transmission of the virus, income, occupation, and age are totally randomly assigned. Subjects makes 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. We then train a pooled model on all five conjoint attributes and the set of covariate information for each profile using **cjbart**.

We then recover a schedule of IMCE estimates for each attribute-level, and consider how these IMCEs vary with education level – a factor that is likely correlated with individuals' exposure and experience of the Covid-19 pandemic. In particular, here we compare occupation attributes from the original experiment, and how IMCEs differ by subjects' education level. Figure 2 presents the estimated IMCE distribution for the factory worker occupation (relative to someone not in work). In Duch et al. (2021) this attribute has a strong positive effect, confirmed by a **cjbart**-estimated AMCE of 0.14 ($\sigma = 0.04$). The IMCE distribution however suggests that this average masks considerable heterogeneity. In particular, the left-hand tail of the distribution drops away to zero, and most importantly this correlates with a high density of low-educated workers. In other words, those with a lower education level prioritise factory workers less than those who are highly educated, all else being equal. In fact, the conditional marginal component effect for low-educated workers is approximately 25% smaller than those who are either high or medium educated ($p < 0.001$ in both cases).[3]

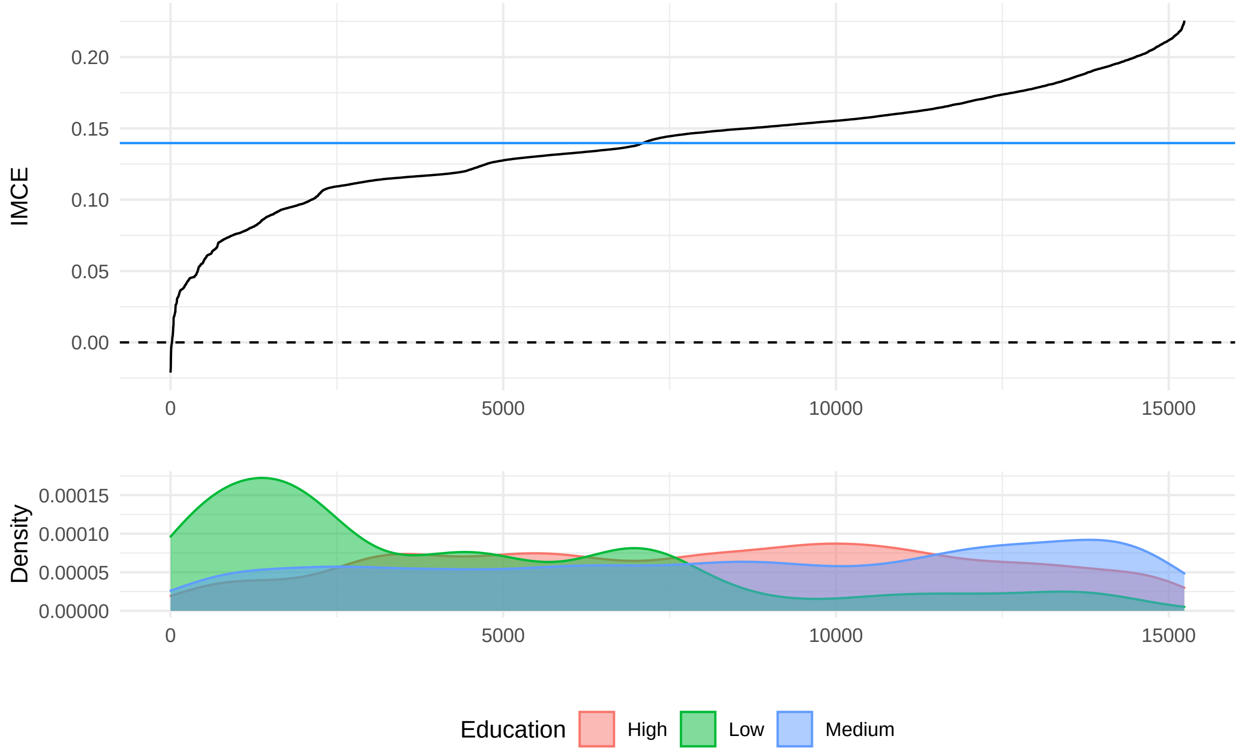
In Appendix Figures A4-A6 we find similar results related to other key-worker occupation attributes. We do not find this education heterogeneity for non-key workers able to work from home, as demonstrated in A7.

4 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 individual's decision making. We exploit the structure of this data, along with machine learning tools, to estimate heterogeneous conjoint treatment effects. Heterogeneity is characterized in terms of a set of nested, causal estimands that correspond to the levels of the conjoint design. This framework allows researchers to assess treatment effect heterogeneity in straightforward and flexible manner.

[3] For brevity, we focus explicitly on subjects' education. More generally, researchers could regress the IMCEs on the full set of covariates to explore effect heterogeneity across different dimensions.

Figure 2. Distribution of IMCEs for the “Key worker: Factory worker" occupation attribute within Duch et al. (2021)



To accompany this paper, we provide a new R package, **cjbart**, that allows researchers to model heterogeneity in their experimental conjoint data by using the powerful BART estimation strategy.

References

Abramson, S. F., K. Kocak, A. Magazinnik, and A. Strezhnev (2020, July). Improving preference elicitation in conjoint designs using machine learning for heterogeneous effects.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan (2018). The moral machine experiment. *Nature 563*(7729), 59–64.

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

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

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

Duch, R., D. Laroze, T. Robinson, and P. Beramendi (2020). Multi-modes for detecting experimental measurement error. *Political Analysis 28*(2), 263–283.

Duch, R., L. S. J. Roope, M. Violato, M. F. Becerra, T. Robinson, J.-F. Bonnefon, J. Friedman, P. Loewen, P. Mamidi, A. Melegaro, M. Blanco, J. Vargas, J. Seither, P. Candio, A. G. Cruz, X. Hua, A. Barnett, and P. M. Clarke (2021). Who should be first in line for the covid-19 vaccine? surveys in 13 countries of the public's preferences for prioritisation. *medRxiv*.

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

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

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

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

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

0.4.0.

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

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

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

A Simulation protocols

A1 Preference-derived IMCE

In this test, 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:

$$\begin{aligned}
\Delta U_{A_1} \sim & \begin{cases} \mathcal{N}(\mu = 1, \sigma = 1), & \text{if } c_1 = 1 \\ \mathcal{N}(\mu = -1, \sigma = 1), & \text{otherwise} \end{cases} \\
\Delta U_{A_2} \sim & \mathcal{N}(\mu = |c_2 - 0.2|, \sigma = 1) \\
\Delta U_{A_3} \sim & \mathcal{N}(\mu = 0, \sigma = 0.5)
\end{aligned}$$

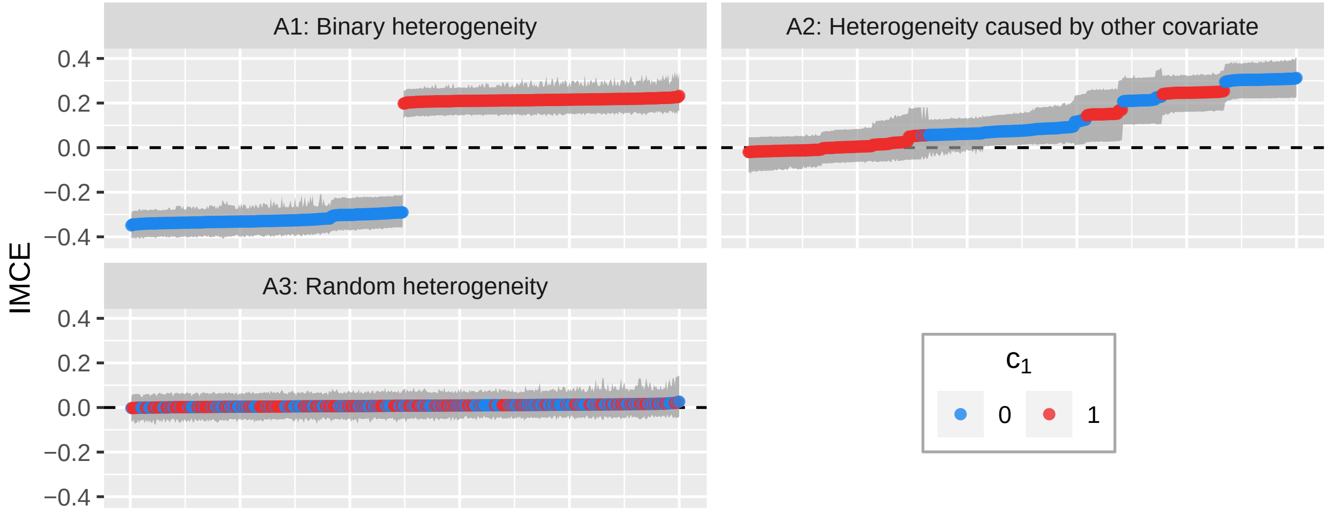
We then simulate the conjoint experiment run on 500 subjects, for 5 rounds each, in which individuals choose between 2 profiles. For each observation, we calculate the utility for subject i in round j given profile q as:

$$U_{ijq} = \mathbb{I}(A_1 = b) \times \Delta U_{A_1} + \mathbb{I}(A_2 = d) \times \Delta U_{A_2} + \mathbb{I}(A_3 = f) \times \Delta U_{A_3} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 0.0005)$ 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

Figure A1. 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)

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 A1 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 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 correlation between c_1 and A2 and A3, where *a priori* neither are related to c_1, are 0.002 and 0.004 respectively.

A2 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 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	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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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:

$$\text{Prob} = 0.5 + \mathbb{I}(X_1 = b)\tau_{X_1} + \mathbb{I}(X_2 = d)\tau_{X_2}.$$

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 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.

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

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

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

The BART model then estimates the OMCEs (τ_{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.

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

i	$\hat{Y} \vert X_1 = b$	$\hat{Y} \vert X_1 = a$	$\widehat{\tau}_{ijk}$
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:

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

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

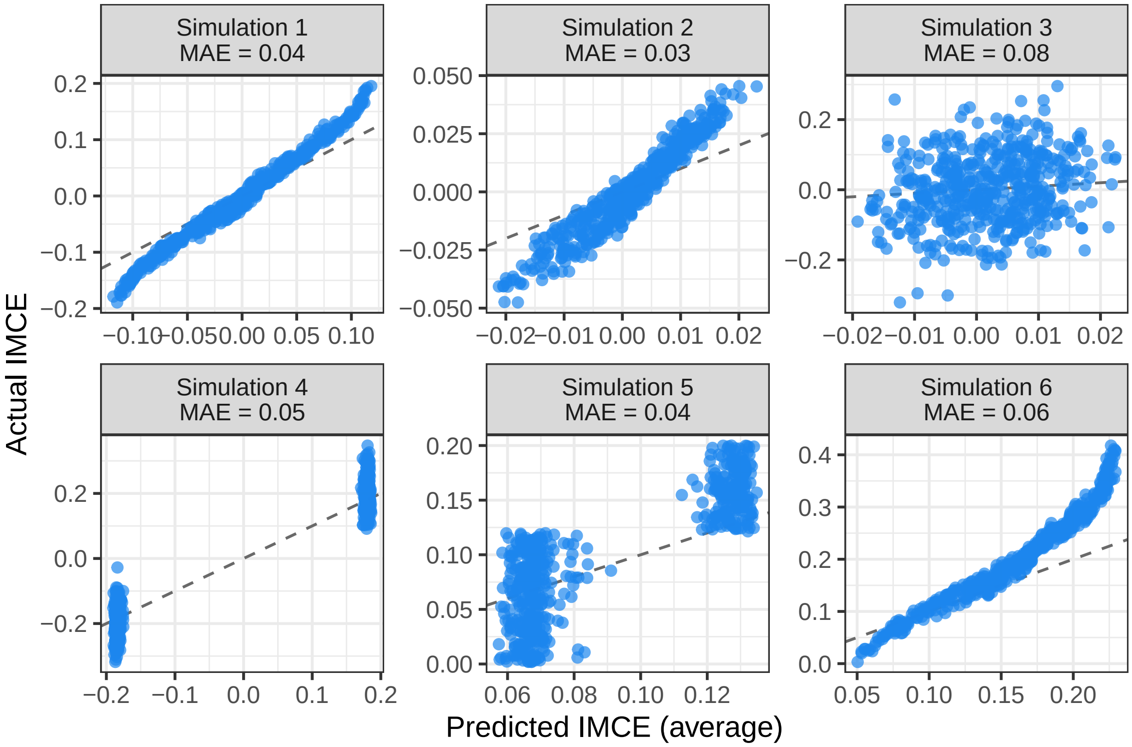
In our actual simulations, we complicate the DGP. We assume that each subject has three observed covariates: c_1 and c_2 are continuous covariates drawn from a random uniform distribution between 0 and some upper bound of heterogeneity (denoted h); c_3 is a binary variable generated from a binomial distribution 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 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 A2 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 A2. 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.

A3 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.

$$\widehat{\mathbb{V}(\tau_{ik})} = \frac{1}{J}\sum \widehat{\mathbb{V}(\tau_{ijk})}$$

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:

$$\tau_{ik} \sim \mathcal{N}([C_{1i} - C_{2i}], \sigma_i)$$
$$C_{1i}, C_{2i} \sim \text{Uniform}(0, c),$$

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 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	$c_1 - c_2$	$c_x \sim$ 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 A6. Simulation specifications testing the coverage rate of the confidence intervals

Sim.	Subjects	J	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 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.

A4 RMCE simulation test

To demonstrate how our approach can be used to check for carryover and stability assumption violations, we simulate conjoint experiment where there either is or is not a round-effect on the marginal effects of attribute-levels.

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 in round j from profile q is defined as:

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

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:

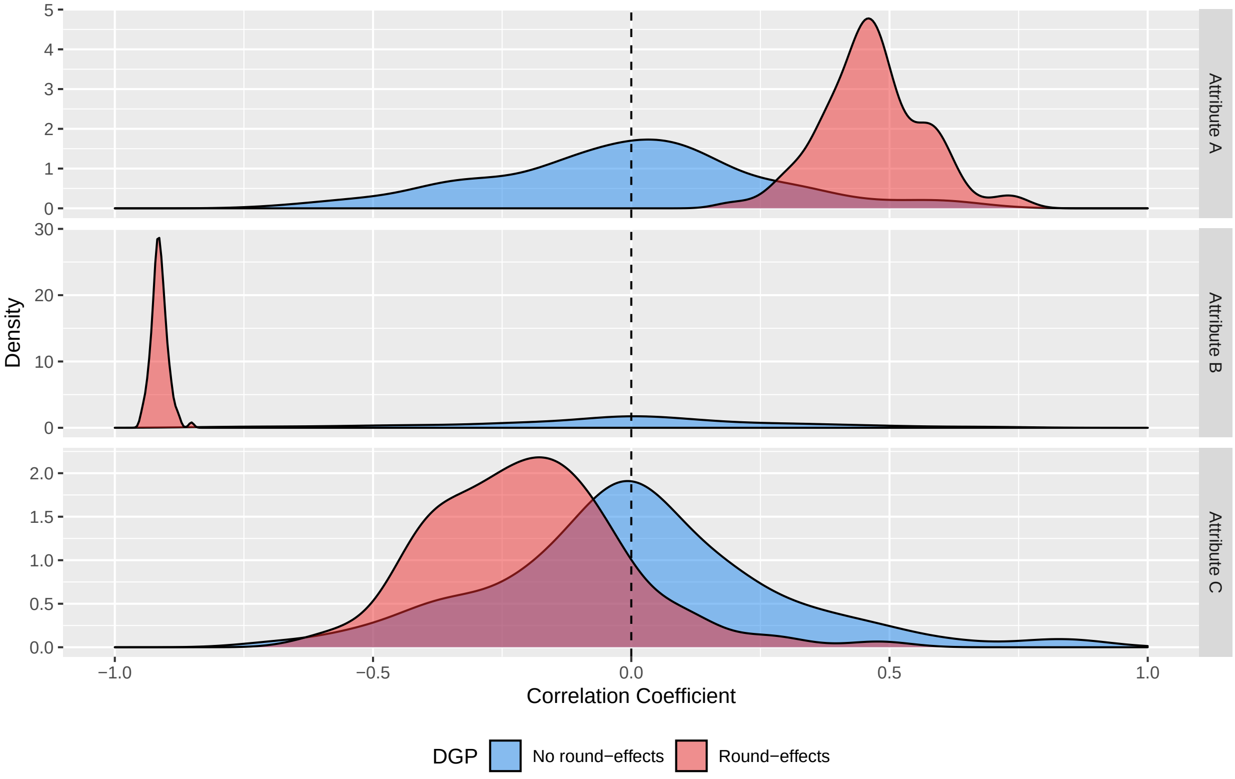
$$\begin{aligned} U_{ijq}^{\text{No round-effect}} =& \mathcal{N}(0, 0.001) \\ & + 1 \times \mathbb{I}(A_{ijq} = a2) \\ & + 0.2 \times \mathbb{I}(B_{ijq} = b2) \\ & + 0.5 \times \mathbb{I}(C_{ijq} = c2). \end{aligned}$$

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 A3 plots the distribution of these correlation coefficients by scenario and attribute, across the simulated experiments. For the no round-effects utility calculation, 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. Perhaps most interestingly, however, is that 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", particularly when using powerful predictive strategies like BART.

Figure A3. Simulation evidence demonstrating how violations of the no carryover assumption can be detected by estimating the RMCE



B 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:

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

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.[4] We then plot these BART-estimated AMCEs against the parametric AMCEs as shown in Figure 1 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.

[4] 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

C Duch et al. (2021) estimation

Figure A4. Distribution of IMCEs for the “Key worker: Water and electricity service" attribute within Duch et al. (2021)

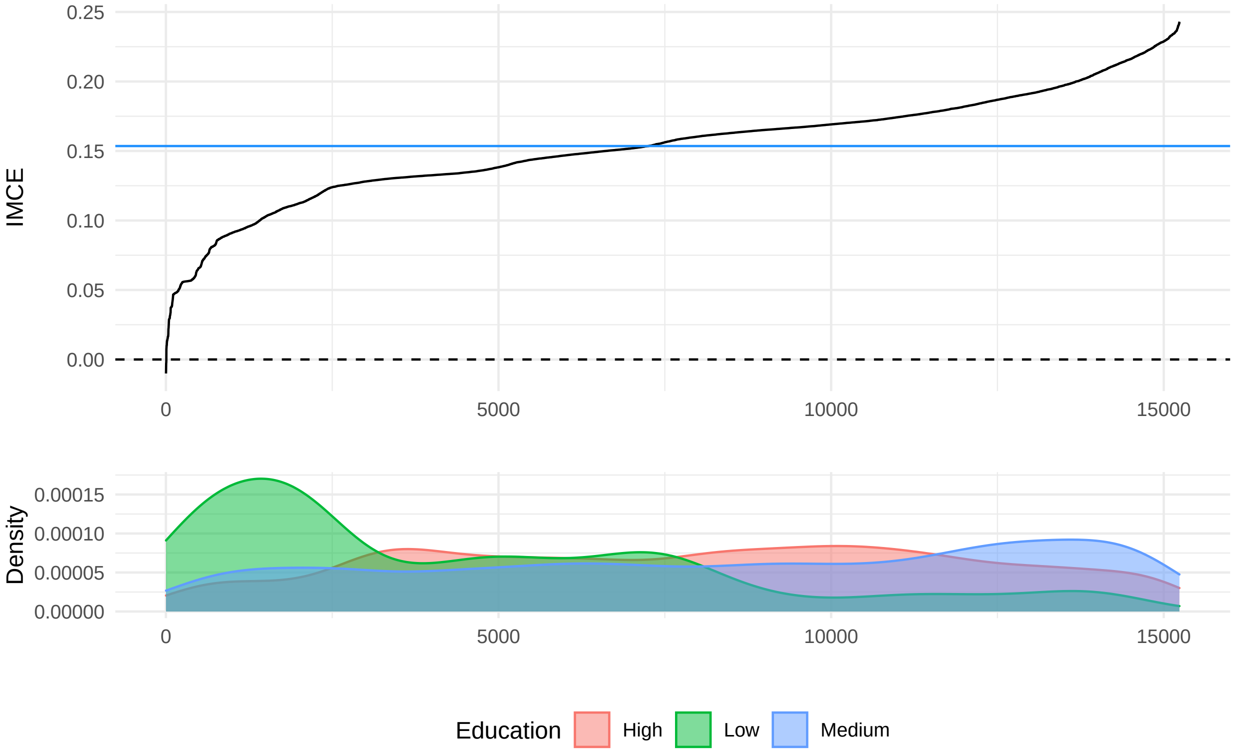


Figure A5. Distribution of IMCEs for the “Key worker: Health and social care" attribute within Duch et al. (2021)

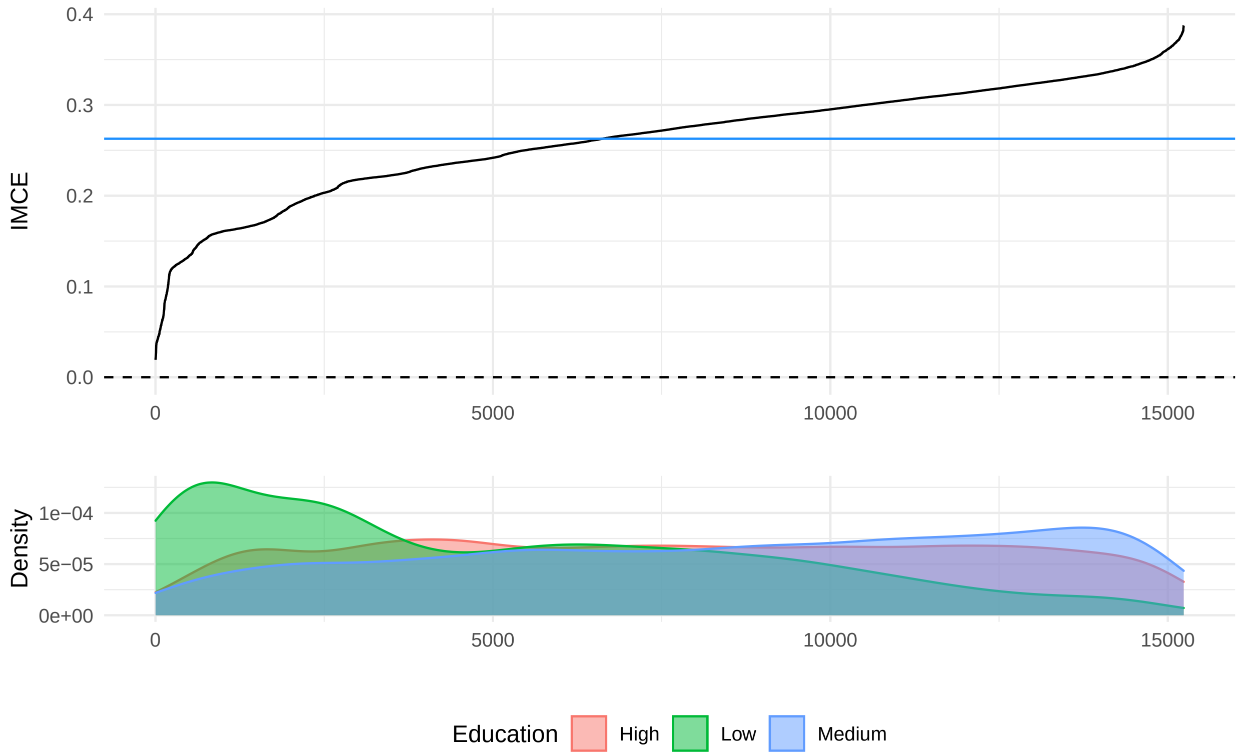


Figure A6. Distribution of IMCEs for the “Key worker: Police and fire-fighting" attribute within Duch et al. (2021)

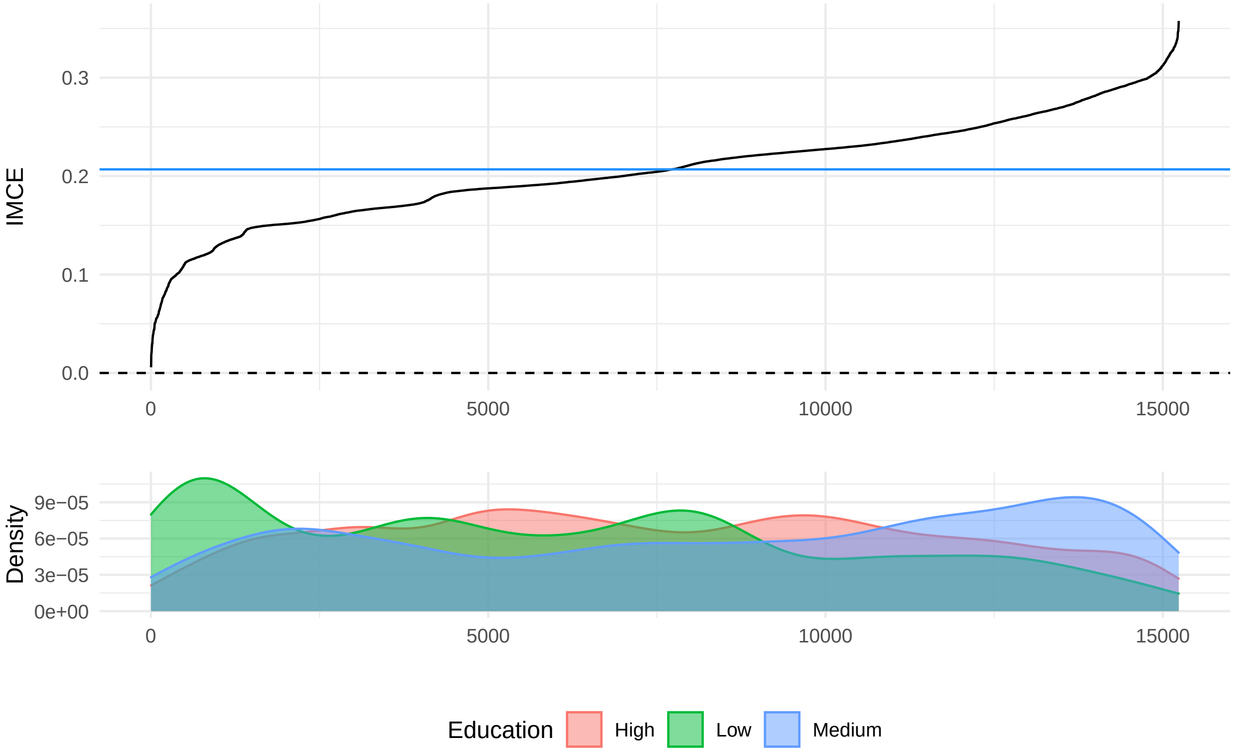
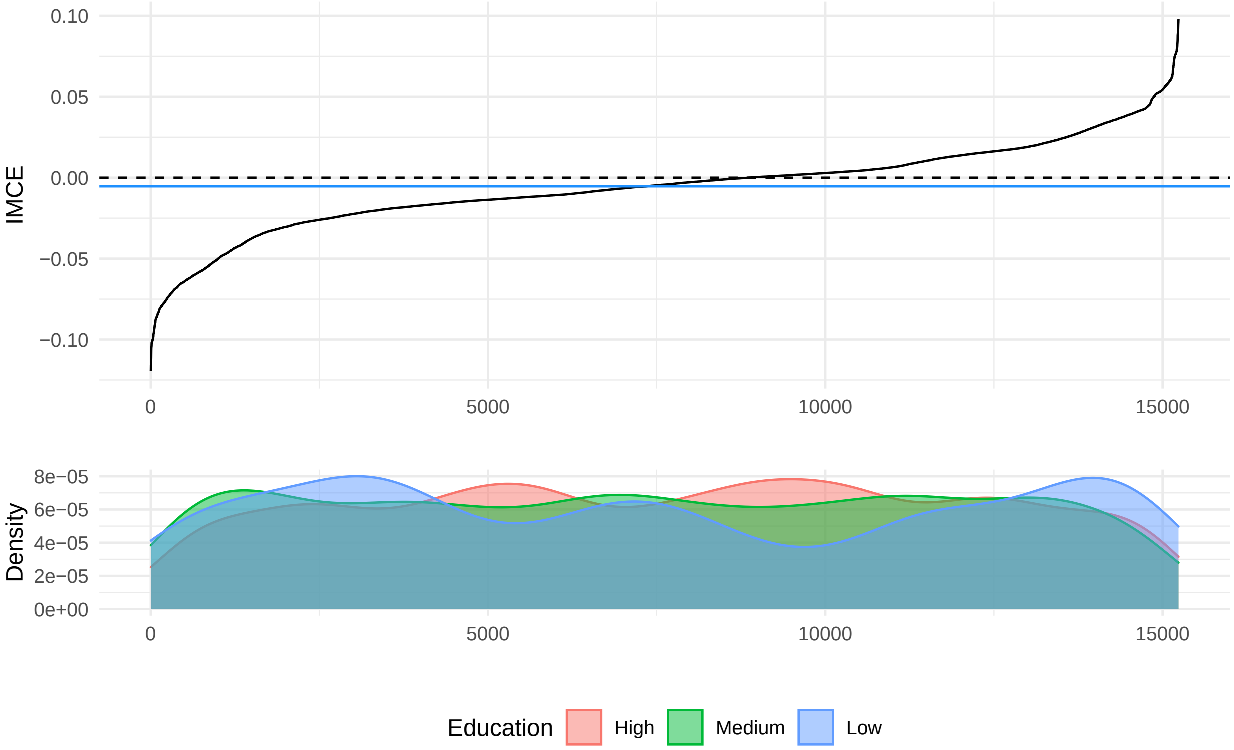


Figure A7. Distribution of IMCEs for the “Non-Key worker: Can work at home" attribute within Duch et al. (2021)


IMCE
Density
8e−05
6e−05
4e−05
2e−05
0e+00
Education
High
Medium
Low