Improving Preference Elicitation in Conjoint Designs using Machine Learning for Heterogeneous Effects

Scott F Abramson[*], Korhan Kocak[†], Asya Magazinnik[‡]& Anton Strezhnev[§]

July 13, 2020

Abstract

Conjoint analysis has become a standard tool for preference elicitation in political science. However the typical estimand, the Average Marginal Component Effect (AMCE), is only tangentially linked to theoretically relevant quantities. In this paper we clarify the necessary theoretical assumptions to interpret the AMCE in terms of individual preferences, explain how heterogeneity in marginal component effects can drive misleading conclusions about preferences, and provide a set of tools based on the causal/generalized random forest method (Athey et al., 2019; Wager & Athey, 2018) that allow applied researchers to detect effect heterogeneity between respondents and derive theoretically relevant quantities of interest from estimates of individual-level marginal component effects. We illustrate this method with an application to a recently conducted conjoint experiment on candidate preferences in the 2020 U.S. Democratic Presidential primary.

[*]Assistant Professor, Department of Political Science, University of Rochester, email: sabramso@ur.rochester.edu

[†]Postdoctoral Associate, Department of Politics, Princeton University, email: kkocak@princeton.edu

[‡]Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, email: asyam@mit.edu

[§]CDS-Moore Sloan Faculty Fellow, New York University Center for Data Science, email: as6672@nyu.edu

1 Introduction

Preference measurement is central to the study of political behavior (Achen, 1975; Ansolabehere et al., 2008; Campbell et al., 1960). To elicit voter preferences, scholars have increasingly turned to conjoint experimental designs (Bansak et al., 2019). In political science, applications of this technology have typically focused upon the Average Marginal Component Effect (AMCE) as their estimand of interest (De la Cuesta et al., 2019; Hainmueller et al., 2014). As shown by Abramson, Kocak, and Magazinnik (2020, henceforth AKM), the AMCE reflects an aggregation rule that does not map neatly onto many statements about electoral outcomes and restricts researchers' ability to make claims about aggregate preferences. In this paper we re-frame the issues identified by AKM as a heterogeneous effects problem and highlight how applied researchers can exploit advances in machine learning to make statements about electoral majorities and the distributions of preferences.

Our main results highlight that, to the degree it is feasible, political scientists should divorce the estimation of preferences from their aggregation. Broadly, this comports with standard practice in conjoint analysis conducted outside of the study of politics. In marketing, where conjoint experiments have the widest use, conjoint output typically consists of individual-level partworth utilities that are estimated from an explicitly stated model (Netzer et al., 2008; Scholz et al., 2010). Only after estimated individual preferences are obtained are quantities of interest—a predicted market share, for example—calculated. Contrast this with the typical conjoint experiment in political science whose target is an AMCE. Here, researchers are effectively conducting the estimation and aggregation of preferences simultaneously.

AKM highlight why this is problematic in two ways. First, they show a direct mapping from the AMCE to the Borda rule. In practice, very few elections are conducted via the Borda rule. Moreover, we are unaware of any study that uses a conjoint experiment to make explicit claims about this set of exceptional elections where there is a direct mapping from the AMCE to the electoral rule. Second, AKM show that the least-squares estimator proposed by Hainmueller et al. (2014) is isomorphic to the linear probability ideal point model of Heckman and Snyder Jr (1997) with observed choice characteristics. As such, the AMCE is equivalent to the average ideal point that would be obtained from this model. The average ideal point is rarely of interest outside of a class of probabilistic voting models that require strong additional assumptions for the mean

voter's preference to be relevant in characterizing equilibria. Again, we are unaware of a single study that attempts to use a conjoint or conjoint-like design to evaluate hypotheses generated by any probabilistic voting model.

In their response to AKM, Bansak et al. (2020) show that the AMCE can be interpreted as an expected vote share. To be clear, this expectation is taken over both the distribution of voter preferences and the distribution of candidate attributes. As a consequence, interpreting the AMCE as an expected vote share runs into the same fundamental problem of aggregating heterogeneous preferences that drives AKM's main result. Note that the AMCE is not reflective of a vote-share in any particular head-to-head election. Rather, it is the difference between the expected vote-shares of a candidate with feature a_1 and a candidate with feature a_0 when they run in each election against all other possible candidates defined by an arbitrary randomization distribution. As AKM show, by also averaging over voter preferences, identical expected vote shares can be generated from a preference distribution that results in a single landslide in favor of a_1 (and an overwhelming majority of losses) as well as from a preference distribution where candidates with feature a_1 win all elections. In other words, by averaging over the intensive and extensive margins of voter preferences, the expected vote share as defined by Bansak et al. (2020) can be generated by a wide range of voter preference distributions. Unfortunately, the AMCE alone does not allow researchers to discern which of the many preference distributions consistent with their estimate matches the true preference distribution of the target population.

While the AMCE is an easy and straightforward quantity to estimate – the conjoint analogue of the Average Treatment Effect – interpreting it as a measure of aggregate preferences requires strong additional assumptions. We clarify in this paper that the key implicit assumption necessary to use the AMCE to draw inferences about the direction of majority preference is a "constant effects" assumption with respect to individual marginal component effects. Essentially, that there exists no meaningful cross-unit heterogeneity in how a given feature affects respondents' choice behavior. This assumption is rarely justifiable, which suggests that researchers interested in making statements about preferences should be particularly interested in finding sources of heterogeneity in the marginal component effects. But researchers are justifiably concerned about the risks of specification searching and false positives when conducting sub-group analyses without strong a-priori knowledge of the sources of effect heterogeneity. We draw on recent advances in the application

of machine learning techniques to the estimation of heterogeneous treatment effects to propose a principled method for uncovering sub-groups that vary in their preferences for a given feature. Specifically, we apply the causal/generalized random forest method of Athey et al. (2019), Wager and Athey (2018), which extends the random forest algorithm of Breiman (2001) to optimize for differences in treatment effects among splits of the data.[1] At its core, our approach leverages observed individual-level covariates to relax the implicit assumption of constant treatment effects in order to make statements about the distribution of preferences in the sample.

In addition to providing a robust method of uncovering preference heterogeneity within a conjoint experiment, we also highlight some additional theoretical concerns with the marginal component effect as a measure of preferences at the individual level. Specifically, we emphasize that the marginal component effect does not directly measure of how an individual will tend to evaluate a pairwise comparison. We define a quantity of interest that reflects individual behavior in such pairwise contest, the feature choice probability (FCP), which is defined as the expected share of tasks in which an individual will choose a profile with an attribute value of t_1 when it is directly compared to a profile with attribute value of t_0 with the expectation being taken over all other profile features. When attributes are binary, we show that the marginal component effect is a linear transformation of the FCP. However, when attributes take on more than two levels, we show that the marginal component effect may diverge from the feature choice probability. Intuitively, this is because the marginal component effect averages over tasks involving direct pairwise comparisons between the two attribute levels of interest and all indirect comparisons of one of the two levels of interest against another possible level of the attribute. When preferences are not transitive, this use of indirect comparisons can lead to misleading marginal component effects. Moreover, even if individual preferences are transitive, we show that the average marginal component effect can suggest a stable social preference when, in fact, there are Condorcet-like cycles. We suggest a method for potentially uncovering such cycles by comparing the estimated AMCEs to non-parametric estimates of the average feature choice probability (AFCP) for all possible level pairs in a given attribute.

We highlight that one advantage of the average marginal component effect is that by aver-

[1] Our approach to recovering individual level heterogeneity using covariates can be viewed as an analogue of the common practice in marketing of estimating individual part-worth utilities via hierarchical models that exploit observed covariates. See e.g. Allenby and Rossi (1998), Lenk et al. (1996).

aging over indirect comparisons, estimates of the AMCE are more precise than estimates of the AFCP. The AFCP uses only those tasks that have a direct comparisons while the AMCE borrows information about preferences from indirect comparisons. Nevertheless, there is a clear trade-off between interpretability and precision in the choice of the quantity of interest. We caution that even if the extreme assumption of effect homogeneity holds, there are still additional behavioral assumptions required to interpret the AMCE, as many researchers wish to do, as a preference for a feature in a pairwise comparison.

The remainder of the paper is structured as follows. In Section 2 we review the theoretical framework for conjoint effects developed in Hainmueller et al. (2014) and illustrate the relationship between the marginal component effect and the pairwise preference for a feature, the FCP. We decompose the marginal component effect into its respective FCPs and explain the necessary additional assumptions for interpreting a marginal component effect in terms of preferred profiles in a pairwise comparison task. In Section 3 we focus on the problem of estimating *average* MCEs when individuals exhibit heterogeneity in their preferences. We then detail our approach to uncovering the distribution of individual MCEs using causal forests. In Section 4 we provide simulation evidence for the performance of the causal forest approach in a conjoint setting with heterogeneity driven by two unobserved subgroups with opposed preferences and differing preference intensities resulting in a divergence between mean and median preferences. The median of the causal forest predicted MCEs reliably recovers the true mean when observed covariates are strong predictors of subgroup membership, but bias grows as covariates become weaker in their predictive ability. In Section 5 we demonstrate our proposed method using data from a conjoint experiment conducted in June and July of 2019 by Data for Progress (Schaffner & Green, 2019) that surveys a sample of Democratic primary voters. Even in this sample, which is far more homogeneous than the broader American electorate, we uncover substantial heterogeneity in preferences for different attributes and demonstrate that attributes with comparable AMCEs can exhibit significantly different distributions of individual preferences. In Section 6 we conclude.

2 The AMCE & Pairwise Preferences

2.1 Setup

This section develops the first theoretical contribution of this paper: the connection between the AMCE and the *preference* for a particular attribute over another in a pairwise comparison. It clarifies that the AMCE is an average of differences in potential outcomes across two different *tasks* where one attribute in one profile is manipulated to take on a different level rather than a direct comparison of two different *profiles*. Despite this, the AMCE can be expressed in terms of a combination of these direct comparisons, providing a formal justification for using marginal component effects to draw inferences about how individuals would respond to pairwise comparisons.

However, when attributes of interest have more than two unique levels and individuals do not have transitive preferences, the sign of the marginal component effect and the pairwise preference for a given attribute level over another may not coincide. This occurs because the marginal component effect not only incorporates comparisons between profiles with an attribute level of interest a_1 and some alternative a_0, but *also* indirect pairwise comparisons between a_1 and each of the other attribute levels, a_a. On the one hand, this has the benefit of estimation efficiency, since the standard estimator of the AMCE will incorporate information from all tasks. On the other hand, it requires an additional assumption of single-peaked preferences in order to interpret the sign of the AMCE as indicative of whether the average pairwise preference for a feature is greater than or less than $\frac{1}{2}$.

We follow the potential outcomes framework for defining conjoint designs and estimands developed in Hainmueller et al. (2014). Consider a standard forced-choice conjoint design with a sample of N respondents indexed by i. Each respondent receives K choice tasks. For each task, they select a single preferred profile from among the J profiles. Each profile is composed of L discrete attributes, with each attribute l containing D_l unique levels. Let the complete set of treatments across all profiles and tasks assigned to individual i be defined as $\bar{\mathbf{T}}_i$. The treatment given to respondent i for the jth profile in the kth task is written as the L-dimensional vector T_{ijk}, with the lth level of that profile being T_{ijkl}. Let $\mathbf{T}_{ik}$ denote the set of attributes for all j profiles in choice task k.

The respondent's potential outcomes in each task can be written as the vector $Y_{ik}(\bar{\mathbf{T}}_i)$ with

profile-level components $Y_{ikj}(\bar{\mathbf{T}}_i)$. In other words, $Y_{ijk}(\bar{\mathbf{T}}_i)$ denotes the outcome that respondent i would assign—a 1 or 0 in a forced-choice conjoint—to choice j in task k if that respondent was assigned the treatment regimen $\bar{\mathbf{T}}_i$.

To simplify the analysis, Hainmueller et al. (2014) place a restriction on what treatments can affect which outcomes. This is analogous to a Stable Unit Treatment Value (SUTVA) assumption (Rubin, 1986) with respect to the tasks assigned to a particular respondent. We assume that there is no carryover in treatment assignments and that the profiles presented in previous tasks for a given respondent do not change the potential outcomes for subsequent tasks.

Assumption 1. *(Stability and No-Carryover).*

For each i and all possible pairs of treatments $\bar{\mathbf{T}}_i$ and $\bar{\mathbf{T}}'_i$

$$Y_{ijk}(\bar{\mathbf{T}}_i) = Y_{ijk'}(\bar{\mathbf{T}}'_i) \quad \text{if} \quad \mathbf{T}_{ik} = \mathbf{T}'_{ik'}$$

for any j, k, and k'

This allows us to pool across repeated choice tasks by assuming that responses in a single task depend only on the treatments assigned in that task. We can therefore write the potential outcomes in terms of only the treatment profiles assigned in the relevant task denoted $Y_{ik}(\mathbf{t})$. The observed outcome vector for a given task maps onto the potential outcomes as $Y_{ik} = Y_{ik}(\mathbf{T}_{ik})$ where $\mathbf{T}_{ik}$ is the observed treatment assignment for task k.

Following Hainmueller et al. (2014), we make a further assumption to allow us to pool across values assigned to each profile in an individual task.

Assumption 2. *(No profile order effects).*

$$Y_{ijk}(\mathbf{T}_{ik}) = Y_{ij'k}(\mathbf{T}'_{ik})$$

if $T_{ijk} = T'_{ijk}$, $T_{ij'k} = T'_{ijk}$, $T_{ijk} \neq T_{ij'k}$, $T'_{ijk} \neq T'_{ij'k}$ for any i, j, j', k

This assumption states that for two non-identical profiles, swapping the profile order swaps the potential outcomes.[2] What this allows us to do is establish that the potential outcome for an

[2]This formulation differs slightly from Hainmueller et al. (2014) in that it only assumes order invariance for profiles that are non-identical (where a respondent has a preference).

individual profile rating Y_{ijk} depends on the treatment assigned for that profile and the unordered set of treatments for the other comparison profiles.[3] We can refine the consistency assumption mapping observed to potential outcomes as $Y_{ijk} = Y_{ijk}(T_{ijk}, \mathbf{T}_{i[-j]k})$ where $\mathbf{T}_{i[-j]k}$ is defined as the unordered set of the non-jth profiles.

Next, we further split the T_{ijk} term in the potential outcome notation in terms of the level assigned to the lth attribute and the levels assigned to all other attributes $[-l]$. Define $Y_{ijk}(t_l, t, \mathbf{t})$ as the potential outcome we would observe for unit i in profile j in task k if that unit/task's treatment assignment were $T_{ijkl} = t_l$, $T_{ijk[-l]} = t$, $\mathbf{T}_{i[-j]k} = \mathbf{t}$. This yields the consistency assumption of $Y_{ijk} = Y_{ijk}(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$.

One minor notational issue is that we would like to define the potential outcomes in a way that does not depend on profile order or task order. Under Assumption 1, respondents' potential outcomes are the same across tasks if those tasks have identical profile sets. We can therefore suppress the k index when writing the potential outcomes. Hainmueller et al. (2014) further use Assumption 2 to suppress the j index, writing the individual potential outcomes and consistency assumption as $Y_{ijk} = Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$. Under order invariance, this is correct when the jth profile is unique among the profiles in the choice set. However, when some profiles in a task are identical and the design requires a forced choice, $Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$ is ill-defined. Consider the simple case where $J = 2$ and the two profiles are identical: $T_{ijk} = T_{i[-j]k}$. Under a forced choice conjoint design, $Y_{ijk} \neq Y_{i-jk}$ since only one profile can receive a 1.

Under the consistency assumption from Hainmueller et al. (2014) this implies

$$Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k}) \neq Y_i(t_{i[-j]kl}, T_{i[-j]k[-l]}, \mathbf{T}_{ijk})$$

However, $T_{ijk} = T_{i-jk}$ implies $t_{ijkl} = t_{i[-j]kl}$, $T_{ijk[-l]} = T_{i[-j]k[-l]}$, which yields

$$Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k}) \neq Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$$

which is by definition false.

While the presence of identical profiles is ultimately irrelevant to identification of the quantities

[3] Since in this paper we focus on the case where $J = 2$, this assumption primarily serves to state that a respondent will still select the same profile in a pair if that pair is swapped, unless that pair is identical.

of interest in Hainmueller et al. (2014) since these tasks cancel one another out (though the possibility of having tasks with identical profiles does affect the scale of the treatment effect via the randomization distribution), it is worth clarifying the consistency assumption and how the conjoint estimands of interest aggregate across profiles in a task. Under Assumption 1, we can make the consistency assumption $Y_{ijk} = Y_{ij}(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$. We define $Y_i()$ as the following function:

Definition 1. *Individual potential outcomes*

$$Y_i(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k}) = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$$

In a forced choice conjoint task where all profiles are different, this quantity is identical to the observed outcome (either 1 or 0), consistent with the notation used in Hainmueller et al. (2014). When all profiles are identical, this will equal $\frac{1}{J}$ (since the profile is selected once and not selected $J-1$ times). When only some profiles are identical, this quantity will lie between 0 and 1 inclusive, depending on how many duplicates of profile j exist in the remainder of the choice set. The treatment effects defined in Hainmueller et al. (2014) for conjoint experiments are described in terms of differences in potential outcomes when respondents are exposed to tasks with a different set of profiles. The most basic contrast is the effect of changing one attribute in one profile, holding all other attributes in that profile and all other profiles constant. This quantity, which we label the "component effect," is defined as

Definition 2. *Component effect*

$$CE_{il}(t_1, t_0, t, \mathbf{t}) = Y_i(t_1, t, \mathbf{t}) - Y_i(t_0, t, \mathbf{t})$$

The component effect represents the change in an individual's response if the lth attribute of the jth profile were set to t_1 versus t_0, holding the rest of the profile constant (at t) and the other profiles constant at $\mathbf{t}$. However, this quantity still depends on the entire vector of profiles, namely the other non-manipulated attributes t and the non-manipulated profiles $\mathbf{t}$. In a conjoint experiment, researchers are interested in summarizing the effect of manipulating the single attribute across a wide variety of possible profile combinations. Hainmueller et al. (2014) define

such a quantity by *marginalizing* over the distribution of the other randomized attributes/profiles which is known by design. This "marginal component effect" (MCE) of setting attribute l to level t_1 versus t_0 can be defined for each individual as:

Definition 3. *Marginal component effect*

$$MCE_{il}(t_1, t_0, p(\mathbf{t})) = \sum_{(t,\mathbf{t})\in\mathcal{T}} [Y_i(t_1, t, \mathbf{t}) - Y_i(t_0, t, \mathbf{t})] \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}})$$

where $\tilde{\mathcal{T}}$ *is the intersection of the supports of* $p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijkl} = t_1)$ *and* $p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijkl} = t_0)$

One can interpret the MCE as a way of reducing the dimensionality of an individual's full set of preferences in tasks that involve either t_1 or t_0 to a single scalar value that averages over the design distribution of those tasks. As we highlight later in this paper, interpreting the MCE further as a comparison *between* t_1 and t_0 requires some additional assumptions.

If individuals could be subjected to an arbitrarily large number of tasks where a response to each possible set of profile combinations is observed, this quantity can be directly identified for each respondent. However, in practice, conjoint designs will expose respondents to only a handful of choice tasks and any given individual may have only one or zero tasks where a profile takes on a particular attribute-level, making it impossible to identify the MCE of a specific attribute-level pair for each individual.[4] Hainmueller et al. (2014) therefore define a more feasible target of inference, the "Average marginal component effect" or AMCE, which is the expected MCE in the sample.

Definition 4. *Average marginal component effect*

$$\begin{aligned} AMCE_l(t_1, t_0, p(\mathbf{t})) &= \mathbb{E}[MCE_{il}(t_1, t_0, p(\mathbf{t}))] \\ &= \sum_{(t,\mathbf{t})\in\mathcal{T}} \mathbb{E}\left[Y_i(t_1, t, \mathbf{t}) - Y_i(t_0, t, \mathbf{t})\right] \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}}) \end{aligned}$$

Our exposition of the AMCE, and introduction of the intermediate quantity of the individual-

[4]Hainmueller et al. (2014) discuss this briefly in footnote 9, which mentions that individual-level effects are potentially identifiable for a very small number of attribute combinations that are actually observed. However, this will not be the *same* MCE for each individual, which limits the utility of this result.

level MCE, differs somewhat from how it is introduced in Hainmueller et al. (2014). We clarify that there are *two* ways in which the AMCE can be thought of as "averaging" over causal quantities of interest. The first, which we label *marginalizing*, is the averaging of component effects over the distribution of all other attribute levels and profiles. This gives a quantity that can be interpreted as a comparison just between two levels of a given attribute, but critically, it is a coherent and well-defined quantity at the individual level. The second, which we refer to as *averaging* due to its connection to average treatment effects, involves taking an expectation of component effects over units in the sample. Hence, the MCE is defined as an individual-level quantity that *marginalizes* over the distribution of attributes and profiles in the design but does not involve *averaging* over the sample.

What is gained by choosing the AMCE as a quantity of interest is the ability to identify the effect of *any* attribute-level comparison and to estimate them with some degree of precision using relatively easy-to-implement estimators. This requires an additional ignorability assumption on the treatment assignment mechanism which can be guaranteed from the experimental design.

Assumption 3. *(Treatment Ignorability and Positivity).*

$$Y_{ijk}(\mathbf{t}) \perp\!\!\!\perp T_{ijkl}$$

for all i, j, k, l, and $\mathbf{t}$.

$$0 < p(\mathbf{T}_{ik} = \mathbf{t}) < 1$$

for all $\mathbf{t}$ within the support of possible tasks

Hainmueller et al. (2014, p. 11) show that under Assumptions 1, 2, and 3, the AMCE is nonparametrically identified from the observed data.

$$\widehat{\text{AMCE}}_l(t_1, t_0, p(\mathbf{t})) = \sum_{(t,\mathbf{t})\in\mathcal{T}} \Big\{ E[Y_{ijk}|T_{ijkl} = t_1, T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t})] - E[Y_{ijk}|T_{ijkl} = t_0, T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t})] \Big\} \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}|T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}})$$

Additionally, under an assumption of complete randomization the AMCE can be estimated by the simple difference-in-means of Y_{ijk} between profiles with $T_{ijkl} = t_1$ and $T_{ijkl} = t_0$ (Hainmueller et al., 2014, p. 16).

Assumption 4. *(Complete randomization).*

$$T_{ijkl} \perp\!\!\!\perp T_{ijk[-l]}, \mathbf{T}_{i[-j]k}$$

for all i, j, k,

The analysis in the remainder of this paper focuses on the common conjoint design where complete randomization holds and the difference-in-means therefore identifies the AMCE. While this does not encompass all possible conjoint designs, it is one of the most frequently used designs and is the default randomization approach when researchers do not have any attribute combinations that they wish to restrict from appearing in a task. We note that the MCE decomposition results we present in the subsequent section do continue to hold under the weaker conditional randomization assumption (Hainmueller et al., 2014, p. 13) that allows for some sharp within-profile cross-attribute restrictions. However, this design introduces some additional challenges in interpreting the AMCE, since the averages of binary comparisons into which the AMCE can be decomposed are defined over different randomization distributions. We discuss this issue more in Appendix A.

2.2 Binary choice designs and the AMCE

Even before we consider the conceptual problems that arise from *averaging* over units in the sample, it is important to understand how to interpret marginal component effects and how they relate to the actual quantities that researchers are interested in when discussing individual preferences. We focus in this section, and in the remainder of the paper, on a very common conjoint design, which we call the "binary choice design." In this design, each task contains two profiles and respondents are forced to indicate a preference for one profile or the other. Researchers typically want to use these types of conjoint designs to make statements about whether respondents *prefer* profiles with a particular level of some attribute over those with a different level of the same attribute when forced to make a choice between them. While the average marginal component effect is often

discussed in terms of these binary comparisons, it itself is *not* a binary comparison. Rather, it is a difference across *counterfactual tasks*. It captures how the expected respondent choice behavior would change in a task, if the task instead contained a set of profiles where one attribute in one profile were set to a different level. Its power comes from its generality. It is a valid estimand for any conjoint design that satisfies the (conditional) randomization assumptions, including those designs with more than two profiles or non-binary/continuous outcomes. The AMCE is also a very simple estimand. It does not require any additional modeling of the choice task and requires only design-based assumptions for non-parametric identification. But researchers often discuss the AMCE in terms of the forced choice comparisons that comprise the conjoint survey. It is common to see a positive AMCE interpreted as implying that a respondent will, on average, choose a profile with that attribute level over a profile with the baseline attribute level when presented with two profiles that differ on that attribute.

In this section, we highlight that this is not necessarily the case. Whether a respondent will choose one profile over another is a different quantity of interest that is distinct from the marginal component effect. We define this quantity, the "feature choice probability" or FCP, discuss its connection to the marginal component effect, and formalize the assumptions necessary to use the MCE in order to draw inferences about the FCP. We then discuss the additional assumptions necessary to draw inferences about the average feature choice probability (AFCP) using the AMCE, highlighting that heterogeneity in marginal component effects across individuals can lead the two quantities to imply differing conclusions about respondent preferences.

Formally, a "binary choice design" is one with two profiles per task ($J = 2$) where respondents are forced to assign in each task a value of 1 to their preferred profile and a value of 0 to the profile that they do not prefer. While this does not encompass the full scope of all possible conjoint designs—it does not include designs where respondents choose among many profiles or assign ratings to individual profiles rather than choices—it does represent one of the most common conjoint designs, especially in political science applications.

Denote an individual's **feature choice** as the joint potential outcome $Y_{ij}(t_1, t, t_0, t')$ where attribute l of profile j is set to level t_1 and that same attribute l of profile $-j$ is set to level t_0, holding the remainder of profile j constant at t and profile $-j$ at t'. As above, we can sum over

the profiles to define $Y_i(t_1, t, t_0, t')$:

$$Y_i(t_1, t, t_0, t') = \frac{1}{J} \sum_{j=1}^{J} Y_{ij}(t_1, t, t_0, t')$$

Again, under Assumptions 1 and 2, this value is equal to 1 if a respondent would select the profile with t_1 over t_0, 0 if they prefer the profile with t_0 over t_1, and $\frac{1}{2}$ if the profiles are identical ($t_1 = t_0$ and $t = t'$).

The individual **feature choice probability** (FCP) is defined as the feature choice marginalized over the distribution of t and t'. In other words, it is the share of comparisons involving one profile with t_1 and another with t_0 where the profile with t_1 is selected.

Definition 5. *Feature Choice Probability*

$$FCP_{il}(t_1, t_0, p(t, t')) = \sum_{(t,t')} Y_i(t_1, t, t_0, t') \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_1, T_{i[-j]kl} = t_0)$$

We can interpret the FCP as capturing the extent to which one feature is preferred against another when considering all possible tasks that involve a comparison of those two features. If a respondent *only* decides on the basis of one attribute and only prefers feature t_1 to t_0, then the FCP will be equal to 1. Conversely, if that feature is never preferred, the FCP will equal 0. An FCP between 0 or 1 reflects the possibility that a respondent may still nevertheless choose a profile with a disfavored feature if the remaining attributes are sufficiently preferable to those that appear in the other profile. Typically, the question of interest is whether the FCP is greater than or less than .5. An FCP greater than .5 can be directly interpreted as a *preference* for feature t_1 over t_0 since it can be understood as the expected proportion of tasks involving t_1 and t_0 in which t_1 is selected.[5]

Since an individual respondent may answer only a handful of tasks, estimates of individual-level FCPs will be highly variable and potentially unidentifiable if a particular attribute-level pair does not appear in the set of comparisons evaluated by a respondent. A more feasible target of

[5]Of course, because the FCP reduces the entire set of an individual's pairwise preferences to a single scalar quantity, it cannot necessarily capture all of the complexity of the full preference ranking. For example, an FCP of .5 may denote that an attribute is irrelevant in that a respondent only chooses on the basis of t and t' or it may be due to an interaction where a feature is strictly preferred in half of the profile comparisons and strictly not preferred in the other half.

inference, as in the case of the MCE, is the average FCP, defined as the expected FCP among units in the sample.

Definition 6. *Average Feature Choice Probability*

$$\begin{aligned} AFCP_l(t_1, t_0, p(t, t')) &= E[FCP_{il}(t_1, t_0, p(t, t')] \\ &= \sum_{(t,t')} Y_i(t_1, t, t_0, t') \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_1, T_{i[-j]kl} = t_0) \end{aligned}$$

The AFCP can be non-parametrically identified under Assumptions 1, 2, and 3 simply by the observed mean response Y_{ijk} among tasks where $T_{ijkl} = t_1$ and $T_{i-jkl} = t_0$.

For simplicity, we will suppress the profile distribution notation $p(t, t')$ when writing and discussing the FCP and the MCE in the remainder of this paper. The feature choice probability of level t_1 against itself is $FCP_{il}(t_1, t_1) = .5$. Likewise, $FCP_{il}(t_1, t_0) = 1 - FCP_{il}(t_0, t_1)$ since choosing a profile with t_1 implies *not* choosing the profile with t_0.

In contrast to the MCE, which involves a difference in choice outcomes across two different hypothetical task sets, the FCP is simply the choice we observe between profiles in tasks where both t_1 and t_0 are present. However, the set of FCPs across levels of an attribute and the MCEs are directly related. Proposition 1 states that the MCE of level t_1 versus t_0 is a function of all feature choice probabilities involving *either* t_1 or t_0. When there are only two levels in a given attribute, the MCE is simply a rescaling of the FCP such that positive values indicate a feature choice probability greater than .5 and negative values indicate a feature choice probability below .5. However, when there are more than two levels, the marginal component effect incorporates *indirect* comparison tasks that have respondents compare a profile with one of the two levels being considered and a level that is not t_1 or t_0.

Proposition 1. *MCE decomposition*

Under Assumption 4

$$\begin{aligned} MCE_{il}(t_1, t_0) = \left[FCP_{il}(t_1, t_0) - \frac{1}{2} \right] \times \left[p(T_{i[-j]kl} = t_1) + p(T_{i[-j]kl} = t_0) \right] + \\ \sum_{q \neq (0,1)} \left[FCP_{il}(t_1, t_a) - FCP_{il}(t_0, t_a) \right] \times p(T_{i[-j]kl} = t_a) \end{aligned}$$

See Appendix A for the full proof.

If there are only two features in an attribute, the MCE equals the FCP minus one half. Therefore, a positive MCE at the individual level implies an FCP greater than one half. If there are more features, then the MCE also contains a term that sums the difference in the FCP between each other feature and t_1, t_0. Intuitively, this stems from the fact that the MCE also incorporates all indirect comparisons with tasks that contain only t_1 or t_0: $FCP_{il}(t_1, t_a)$ and $FCP_{il}(t_0, t_a)$.

It is straightforward to connect the AMCE to the AFCP by linearity of expectations.

$$\text{AMCE}_{il}(t_1, t_0) = \left[\text{AFCP}_{il}(t_1, t_0) - \frac{1}{2}\right] \times \left[p(T_{i[-j]kl} = t_1) + p(T_{i[-j]kl} = t_0)\right] + \sum_{q \neq (0,1)} \left[\text{AFCP}_{il}(t_1, t_a) - \text{AFCP}_{il}(t_0, t_a)\right] \times p(T_{i[-j]kl} = t_a)$$

By plugging in the simple non-parametric difference-in-means estimator for the AMCE along with the corresponding sample mean estimators of each AFCP, it is straightforward to see that the AMCE estimator has lower variance. While each task contributes only to the estimation of a single AFCP, it can contribute to multiple AMCE estimates.

Intuitively, the AMCE uses *more* observations by incorporating not only the direct pairwise comparisons between t_1 and t_0, but also all indirect comparisons from tasks that have only one profile with t_1 or t_0. The variance benefits grow as we consider designs with attributes that have many possible levels. While there may be zero tasks in-sample where attribute level t_1 is directly compared to attribute level t_0, making it impossible to directly estimate the AFCP, the presence of indirect comparisons permits researchers to estimate the AMCE with some precision. This is the primary advantage of estimating the marginal component effect in a conjoint design: it utilizes information from the indirect comparisons to provide a more precise estimate of the relative preference between two attribute levels in a conjoint. However, this comes at the cost of interpretability. Conceptualizing a positive MCE in terms of a preference for that level in a binary choice is not possible without any additional assumptions unless there exist only two levels of the attribute. Otherwise, the presence of cycles can potentially result in misleading inferences.

2.3 Comparing the FCP and MCE when attributes have more than two levels

This section compares the FCP and the MCE in the case of attributes that can take at least three levels. In particular, it demonstrates how the MCE (AMCE) glosses over individual (aggregate) preference cycles to produce a linear ordering and how this can lead to misleading inferences. This happens because the MCE implicitly assumes that preferences are acyclic. We argue that while this assumption is likely innocuous at the individual level and at the aggregate level when preferences are single-peaked, in many if not most applications of conjoint experiments in political science (e.g. race, occupation, country of origin) it is implausible. The FCP, in contrast, does not assume preferences are acyclic, and can capture such cycles.

When trying to understand choices between two alternatives, the FCP focuses on pairwise comparisons involving these two alternatives. This allows researchers to capture individual or aggregate cycles and make out-of-sample predictions about how a profile with a given set of features would fare in a race against another. In contrast, the MCE borrows information from races with alternatives irrelevant to the comparison at hand to force preferences into scalars. In what follows we present examples of preference cycles at the individual and aggregate levels to highlight how the assumptions built into the AMCE can potentially lead to inaccurate inferences.

In this example, for the purposes of illustration, we will focus on comparing the MCE and FCP from a simple conjoint design with uniform, complete randomization over all of the attributes in a profile and independence across profiles. In this setting, where each level of an attribute has the same probability of appearing as any other level of that same attribute, the MCE decomposition can be written in a more simplified form as:

$$\text{MCE}_{il}(t_1, t_0) = \frac{2}{D_l}\left[\text{FCP}_{il}(t_1, t_0) - \frac{1}{2}\right] + \frac{D_l - 2}{D_l} \sum_{t_a \neq (t_1, t_0)} \left[FCP_{il}(t_1, t_a) - \text{FCP}_{il}(t_0, t_a)\right] \quad (1)$$

where D_l is the number of levels that the attribute can take on.

For an example of an individual preference cycle, consider a voter who has intransitive preferences about politicians' experience in office. This voter generally prefers more experienced politicians because she believes they are more likely to be competent. However, when faced with a contrast between a veteran politician and a complete outsider, she prefers the latter, who is

more likely to push for significant changes or "shake things up." This voter's preferences over candidate experience exhibit a cycle: she prefers newcomers to veterans, veterans to inexperienced politicians, and inexperienced politicians to newcomers.

As another example, consider a voter who wants a high marginal tax rate for the wealthy, but is concerned that if it is too high this will cause capital flight with negative downstream consequences. Suppose he does not know the actual tax rate, but tries to infer it from the question he observes in a survey experiment. When faced with two hypothetical candidates proposing 30% and 50% respectively, he infers the current rate is probably in this range. He does not think a small increase would result in a significant capital flight, and says he prefers the candidate proposing the higher marginal tax rate. He would make a similar inference when faced with a decision between candidates proposing 50% and 70%, and prefer the candidate proposing 70%. But when he is asked to choose between two candidates proposing 30% and 70%, the possible range for the current rate is much wider. He decides that it is probably better to play it safe and choose the candidate who likely offers a small cut than a candidate who might increase the marginal tax rate dramatically, potentially causing flight. Thus, this voter prefers the candidate proposing 30% to 70%, 70% to 50%, and 50% to 30%.

For both of the voters described above, their preferences exhibit cycles of the form:[6]

$$a \succ b \succ c \succ a$$

Now let us consider how the MCE and the FCP treat such preferences. First for simplicity, suppose that there is only one attribute; all our results carry through when there are multiple attributes. When looking at the FCP, we have $FCP_i(a,b) = FCP_i(b,c) = FCP_i(c,a) = 1$. So the feature choice probability captures accurately the cycle this voter exhibits. When we instead calculate the MCE using Equation (1), we find that:

$$MCE_i(a,b) = \frac{1}{3}FCP_i(a,b) + \frac{1}{6} + \frac{1}{3}FCP_i(a,c) - \left(\frac{1}{3}FCP_i(b,a) + \frac{1}{6} + \frac{1}{3}FCP_i(b,c)\right) = 0.$$

Similar calculations reveal $MCE_i(b,c) = MCE_i(c,a) = 0$. A researcher looking at the MCE would then conclude that these voters are indifferent, despite having well-defined, and possibly

[6]Here, we define the preference relation $\succ$ as $x \succ y \iff Y(x,y) = 1$.

intense, preferences. This happens because the MCE borrows information from races against c when making inferences for voters' preferences between a and b. By including indirect comparisons, the MCE dilutes information about individual preferences.

Consider a slightly more involved example with two attributes, party $p \in \{D, R\}$ (Democrat or Republican) and experience $e \in \{N, S, V\}$ (Newcomer, Some experience, Veteran). Consider a voter whose preferences satisfy the following:

$$
\begin{aligned}
&pN \succ pV \succ pS \succ pN \quad \text{for} \quad p \in \{D, R\} \\
&De \succ Re \quad \text{for} \quad e \in \{N, S, V\} \\
&RV \succ DS \succ RN \succ DV \succ RS \succ DN \\
&RV \succ DN
\end{aligned}
$$

Thus, holding party fixed, this voter has the same cyclic preferences over experience as before. Comparing across candidates with equal experience, she prefers Democrats to Republicans. When comparing candidates from different parties, she prefers candidates with more experience, except when comparing a Republican newcomer to a Democrat veteran, in which case she prefers the Republican.

Let us first calculate this voter's FCPs. We have $FCP_i(D, R) = 5/9$, $FCP_i(N, V) = 3/4$, $FCP_i(V, S) = FCP_i(S, N) = 1$. Thus, the FCP accurately captures that this voter always votes for veterans over candidates with some experience, whom she chooses over newcomers; and that most of the time she votes for Democrats over Republicans and newcomers over veterans.

When we calculate the MCEs, we get that $MCE_i(D, R) = 1/9$, $MCE_i(N, V) = -1/6$, and $MCE_i(V, S) = MCE_i(S, N) = 1/12$. A researcher using the MCE may therefore conclude that this voter prefers a veteran Democrat (DV) to a newcomer Republican (RN) because both the marginal component effect of Democrat over Republican and Veteran over Newcomer are positive, despite the fact that this voter chose RN over DV when presented with this pairwise comparison. This happens because the MCE implicitly imposes transitivity on this voter's preferences.

When indeed individual preferences are transitive, we can prove that the sign of the MCE corresponds to whether the FCP is greater than one half. Substantively, this means that researchers can use individual MCEs, which are more precise than FCPs, when they are confident that indi-

vidual preferences satisfy this assumption. Before proving this result, let us first formally define the following.

Definition 7. *Transitivity*

Voter i's preferences are transitive if for all X, Y, and Z, we have that $X \succ Y$ and $Y \succ Z$ implies $X \succ Z$, or equivalently, $Y_i(X, t, Y, t') = 1$ and $Y_i(Y, t', Z, t'') = 1 \implies Y_i(X, t, Z, t'') = 1$ for all t, t', and t''.

Further, in the context of FCPs:

Definition 8. *FCP-Transitivity*

Voter i's preferences are FCP-transitive if for all X, Y, and Z, we have that $FCP_i(X, Y) > FCP_i(Y, X) \implies FCP_i(X, Z) \geq FCP_i(Y, Z)$.

In words, FCP-transitivity says that if an option t_1 is chosen more than half of the time in a pairwise comparison against t_0, t_1 must be chosen at least as often against any other alternative than t_0 is against the same alternative.

With these definitions, we are ready to state our result on the correspondence of the individual MCEs and FCPs.

Lemma 2.1. *When preferences are FCP-transitive, the MCE is positive if and only if the corresponding FCP is greater than one half.*

Next, let us consider aggregate or Condorcet cycles. It is a well-known result in social choice theory that when there are at least three voters and three candidates, majority rule may fail to pick a winner. This happens when aggregate preferences exhibit what is known as a Condorcet cycle, named after the 18th-century French philosopher Marquis of Condorcet. Even if individual preferences are transitive and do not exhibit cycles, when aggregated they may produce a cycle where each candidate loses to another in a head-to-head match-up.[7]

To get around this problem, researchers often assume that preferences are single-peaked: that alternatives have a natural ordering and two alternatives cannot be *both* preferred to a third alternative that lies between them. In other words, single-peakedness asserts that an intermediate

[7]For example, a poll leading up to the 2016 GOP primary found that the majority of responses indicated a preference for Scott Walker over Jeb Bush, Jeb Bush over Ted Cruz, and Ted Cruz over Scott Walker.

option cannot be any voter's least favorite. For preferences over ordinal domains, this assumption is reasonable. For instance, if a voter says they prefer candidates with graduate degrees over high school graduates, we can infer that they also prefer candidates with Bachelor's degrees over high school graduates. However, over domains where there is no clear ordering of alternatives, the assumption of single-peakedness is harder to justify. Suppose researchers want to consider voters' evaluation of candidates who are Black, Latino, and white. There is no clear way to order these features. As such, any voter can place any feature at the top, middle, or bottom of their rankings. Put differently, voter preferences over race are not single-peaked.

Without single-peakedness, aggregate preferences over features may exhibit cycles. Consider the following example. Suppose voter A prefers Black to Latino to white candidates, or $B \succ_A H \succ_A W$. Voter B prefers Latino to white to Black candidates, so that $H \succ_B W \succ_B B$. Finally, suppose voter C prefers white candidates the most and Latino candidates the least: $W \succ_C B \succ_C H$. Here, a Black candidate defeats a Latino candidate in a match-up between the two, a Latino candidate beats a white candidate, who in turn beats a Black candidate. In other words, despite transitivity of individual preferences, the majority exhibits a cycle not dissimilar to a game of rock-paper-scissors.

The FCP captures this cycle: We have $FCP_A(H,W) = FCP_B(H,W) = 1$ and $FCP_C(H,W) = 0$. Aggregated, these reveal $AFCP(H,W) = 2/3$, precisely the vote share of a Latino candidate facing a white candidate. The same holds for $AFCP(B,H)$ and $AFCP(W,B)$.

Let us now calculate the MCEs. Using Equation (1), we have

$$\begin{aligned}
MCE_A(H,W) &= \frac{1}{3}\left(FCP_A(H,W) + FCP_A(H,B) - FCP_A(W,H) - FCP_A(W,B)\right) = \frac{1}{3} \\
MCE_B(H,W) &= \frac{1}{3}\left(FCP_B(H,W) + FCP_B(H,B) - FCP_B(W,H) - FCP_B(W,B)\right) = \frac{1}{3} \\
MCE_C(H,W) &= \frac{1}{3}\left(FCP_C(H,W) + FCP_C(H,B) - FCP_C(W,H) - FCP_C(W,B)\right) = -2/3
\end{aligned}$$

Thus, when we calculate the AMCE, we get that

$$AMCE(H,W) = \frac{1}{3}\sum_{i\in\{A,B,C\}} MCE_i(H,W) = \frac{1}{3}\left(\frac{1}{3} + \frac{1}{3} - 2/3\right) = 0.$$

In words, the marginal component effect of Latino, when white is used as a benchmark, is zero,

despite the fact that two of the three voters have a preference in favor of the Latino candidate. This is because the AMCE takes into account voters' relative rankings of Black candidates, even though they are irrelevant in a race between H and W. This holds regardless of which race we focus on and which we take as the benchmark: the AMCE of any race in this population of voters A, B, and C is zero. Regardless of the intensity of preferences, and despite the clear predictions about electoral outcomes that we can derive from FCPs, the average marginal component effect of each race is zero. This is because the AMCE implicitly assumes preferences over alternatives are single-peaked, including over those domains where this assumption is not justified.[8]

While it is often reasonable to assume that individuals exhibit transitive preferences such that the direction of their individual MCE aligns correctly with whether their FCP is greater or less than 1/2, the above discussion highlights that this alone is insufficient to establish a clear relationship between the average MCE and the average FCP. Even if it is rare for individual preferences to exhibit cycles, it is not impossible to see such cycles in the aggregate, especially in political science applications (e.g. Bochsler, 2010; Kurrild-Klitgaard, 2001). It may therefore be useful for researchers analyzing conjoint experiments to obtain estimates of the more high-variance AFCPs in addition to the estimated AMCEs. While both are estimated with uncertainty, and therefore a positive AMCE and a corresponding AFCP less than .5 may not be indicative of cycling, a clear divergence between the two estimates should raise some concerns.[9]

The divergence between AMCE and AFCP in the above example arises partly from *heterogeneity* in the preferences of individuals in the sample—different individuals have different preference orderings and therefore different marginal component effects and feature choice probabilities. Aggregating these preferences by averaging across them leads to clearly misleading estimates, but even when single-peakedness is satisfied, the presence of heterogeneity across individual effects makes the expectation of MCEs with respect to the distribution of respondents in the sample diverge from the quantity of interest that a researcher may have in mind. In the next section, we discuss the issue identified by AKM in the context of heterogeneous treatment effects. We highlight that while the AMCE is an easily identified quantity in a conjoint experiment, it may not

[8]Reversals are straightforward to construct. For example, consider the same example but with $FCP_C(B, W) = 0.9$, that is, voter X votes for a white candidate over a Black 10% of the time. Then, the AMCE of Latino when white is used as a benchmark is $-1/30$, despite the fact that Latinos always defeat whites.

[9]A formal hypothesis test for cycling in the style of the Hausmann test (Hausman, 1978) is a useful subject for future research.

reflect the relevant individual or group-level preference that a researcher wishes to make inferences on in substantive applications of the design.

3 The AMCE & Treatment Effect Heterogeneity

With a sufficiently large number of tasks and assumptions about the stability of respondents' preferences across repeated tasks, it would be possible to recover estimates of the MCE for each individual. Given known individual preferences, any form of aggregation over all respondents' preferences is straightforward as the entire distribution of individual-level MCEs, and by extension FCPs, is known. This would permit researchers to obtain not just the average MCE, but any quantile of the sample distribution such as the median.

However, in nearly all conjoint experiments the number of tasks assigned to an individual is fixed at a number that is much smaller than would be needed to recover individual MCEs. While studies suggest that respondents can handle experiments with upwards of twelve tasks without much loss of data quality (Bansak et al., 2018), in experiments with many levels per attribute, it will often be the case that any individual will not be exposed to any tasks with a particular level. Moreover, longer surveys are typically more expensive to field per respondent and researchers may instead want to allocate their budget to expanding the number of individuals surveyed.

Since individual effects are typically not directly identifiable due to the fundamental problem of causal inference (Holland, 1986), averages of either the marginal component effect or the feature choice probability with respect to the distribution of individuals in the sample may represent a more feasible quantity of interest. Notably, these effects are identifiable in a standard conjoint experiment without the need for untestable assumptions, e.g. when the treatment assignment probabilities are known and controlled by the researcher and ignorability is assured by design. This is the key advantage provided by the potential outcomes reframing of conjoint experiments in Hainmueller et al. (2014) and an important extension of the traditional conjoint literature in marketing, which typically takes a model-based approach to preference elicitation, making explicit assumptions about the underlying individual utility functions.

A key advantage of the model-based approach to conjoint analysis is its ability to recover preference heterogeneity. Indeed, this is the central goal of conjoint analysis in marketing, which

seeks to make tailored, individual-level predictions about consumer preferences and willingness to pay for product features. To obtain these estimates, however, researchers typically rely on hierarchical modeling techniques that incorporate individual-level characteristics and covariates to estimate a model of utility that captures heterogeneity across respondents.[10]

While it is not necessary for researchers accustomed to design-based inference in experiments to fully embrace the latent utility modeling framework for conjoint analysis, they should also not simply ignore heterogeneity in responses to the treatments. As AKM illustrate, the average of marginal component effects is an average over both the intensity of individual preferences (the magnitude of the individual MCE) and their prevalence in the sample (the number of individuals with the same MCEs). However, researchers interested in making claims about the direction of majority preferences are only interested in averaging over the *sign* of the individual MCEs and not necessarily their magnitudes. Alternatively, they may be interested in the median of the MCE distribution or other relevant quantiles depending on the relevant voting or decision rule. The consequence of this is that the AMCE may not necessarily be directly informative about these other quantities of interest without the extreme assumption of no heterogeneity in MCEs (such that the average MCE is equal to the individual MCE). When there is heterogeneity in MCEs, a small subgroup with a particularly strong positive MCE can lead to the AMCE estimate also being positive even when a majority of respondents have negative MCEs. This problem exists even if the individual MCE is properly reflective of an individual's feature choice probability—it is an effect heterogeneity problem.

Recent work by Bansak et al. (2020) contextualizes the AMCE by relating it to the expected increase in vote share resulting from a change in one attribute that a hypothetical profile (e.g. candidate) would receive averaged over the distribution of challenger profiles. Connecting this to our interpretation of the AMCE, the "vote share" part is the expectation of the binary outcome over the distribution of voters/respondents in the sample for a fixed comparison (the "average" part

[10]In particular, hierarchical Bayes models (Huber & Train, 2001; Lenk et al., 1996) are a common approach to extending the simple conditional logit model to permit variation in estimating individual part-worths (the marketing analogue to component effects) and are considered the recommended method of analysis in the commercial conjoint analysis literature (Orme, 2010). These models leverage the fact that conjoint designs obtain repeated measures on each individual and parameterize the latent individual utility functions with individual-level part-worth parameters that are assumed to be drawn from some higher-order model that is common to all respondents. Estimation and inference is conducted in a fully Bayesian setting, obtaining posterior estimates of the higher-order parameters along with the individual part-worths/utilities of interest. Often, this higher-order model will incorporate respondent-level covariates that are likely to predict variation in part-worths (Orme & Howell, 2009).

of the AMCE) and the "expected increase" in that share is the expectation over the distribution of both possible challengers and possible other attributes of the candidate profile (the "marginal" part of the AMCE). However, this interpretation does not avoid the heterogeneous treatment effect problem. As AKM show, identical AMCEs can be produced from substantively different preference distributions, and the same is true of vote shares: the same average can be produced by one landslide election and a reduction in vote share across all other contests, or a consistent increase in vote share across a large number of contests.

As shown in the previous section, an individual's marginal component effect *can* be directly informative about their preference over different attribute levels when transitivity is assumed. However, in the presence of treatment effect heterogeneity, the average will not necessarily map on to the researcher's quantity of interest. Luckily, the developments in conceptualizing conjoint designs as factorial experiments and conjoint quantities of interest as treatment effects imply a clear path forward in addressing this challenge. Researchers are often interested not only in estimating average treatment effects, but in tailoring their estimates to specific covariate profiles to estimate conditional average treatment effects (CATEs). Inference on CATEs is useful even for researchers working with experimental data to better understand how treatment effects may differ across subjects and to generate individualized recommendations and treatment regimes.

We start with a related quantity for the conjoint setting, the Conditional Average Marginal Component Effect (CAMCE), that denotes the expected marginal component effect for units observed with covariate level $X_i = x$.

Definition 9. *Conditional Average Marginal Component Effect*

$$AMCE_l(t_1, t_0, p(\mathbf{t}, x) = \mathbb{E}[MCE_{il}(t_1, t_0, p(\mathbf{t}))|X_i = x]$$

When X_i is high-dimensional or continuous, non-parametric identification of the CAMCE is typically infeasible as only a handful of observations will share a common value of X_i. Additionally, the specific CAMCE of a particular individual in the sample is rarely the quantity that researchers are interested in; rather, the goal is to characterize the preferences of the majority or a particularly salient fraction of respondents. We outline a set of assumptions that will permit researchers to

make statements about the fraction of respondents that hold a positive or negative preference for a particular attribute level against a baseline. This approach is motivated by the *sorted effects* approach proposed by Chernozhukov, Fernández-Val, et al. (2018) with the goal of performing inference on percentile groups of the CAMCE.

Given a known set of covariates, we suggest a binning strategy that defines k sub-groups $\{G_1, G_2, \ldots, G_k\}$ via a coarsening of X_i. Within each group, we can define a group-level AMCE which is non-parametrically identified by the standard AMCE estimator within the subset of units in group G_k

Definition 10. *Group Average Marginal Component Effect*

$$GAMCE_{kl}(t_1, t_0) = \mathbf{E}[MCE_{il}(t_1, t_0)|G_k]$$

These groups are ordered such that the following monotonicity condition on the treatment effect is satisfied:

Assumption 5. *(Monotonicity).*

$$GAMCE_{1l}(t_1, t_0) \leq GAMCE_{2l}(t_1, t_0) \leq \ldots \ldots \leq GAMCE_{K-1l}(t_1, t_0) \leq GAMCE_{Kl}(t_1, t_0)$$

This definition of grouping connects the ordering of the GAMCEs to the distribution of treatment effects in the sample: the first group contains the units with the most negative AMCE, the second group contains units with an AMCE that is greater than the first, but smaller than the rest, and the Kth contains the units with the largest AMCE. Comparing differences in treatment effects between the bins allows researchers to characterize the degree of heterogeneity in the treatment effects. It also permits statements about the shares of respondents with positive or negative preferences towards a given attribute level based on the signs of the AMCEs within each group combined with the relative sizes of these groups.

We make one final assumption, which is a conditional version of the assumption we would need in order to interpret the AMCE as a sample-wide preference: conditional homogeneity of the MCEs within subgroups.

Assumption 6. *(Conditional homogeneity).*

Let G_i be an indicator denoting a unit's group membership ($i = \{1, 2, \ldots, K\}$)

$$GAMCE_{kl}(t_1, t_0) = MCE_{il}(t_1, t_0))$$

for all $i : G_i = k$.

In other words, for units within effect group G_k, MCEs are constant. This is, of course, a very strong assumption, essentially stating that the grouping variable captures all of the treatment effect heterogeneity in the data. While obviously a less restrictive assumption than assuming constant effects for the entire sample, this approach does require us to choose covariates and coarsenings that are sufficiently narrow as to capture enough of the within-sample effect variation while containing enough observations to enable reasonably precise estimation of the treatment effects. Optimizing with respect to this bias-variance trade-off for the specific task of preference estimation is an interesting topic for future research.

Under conditional homogeneity, the direction of the GAMCE$_{kl}$ allows us to make an inference on the direction of the marginal component effect for the share of respondents with $G_i = K$. For a negative GAMCE$_{kl}$, conditional homogeneity implies that the proportion of respondents with $G_i = k$ have a preference against that particular feature and in favor of the baseline while the converse holds if that value is positive. Researchers can then aggregate the total proportion of units with estimated positive and negative GAMCEs to report the share of the sample that would favor or disfavor a particular feature.

Even if these specific quantities are not of interest, estimating subgroup AMCEs can give a sense of how well the AMCE captures the overall preference direction and intensity of respondents in the sample. Substantial variation among conditional AMCEs between different subgroups should suggest to researchers that a sample average effect will be a poor summary of feature preferences.

3.1 Causal forests to detect heterogeneous effects

While the theoretical discussion above centers on the case where the groups can be specified ex-ante conditional on particular values of X_i, it is rare that researchers will know exactly which covariates drive heterogeneity in treatment effects. With a large variety of possible choices, there is a risk

that researchers will search for the set of subgroup analyses that will yield the most preferred results, which has led many to be cautious about the estimation of conditional treatment effects.

A growing body of literature has explored applying machine learning methods to uncover heterogeneous treatment effects while guarding against the potential for specification searching and selective subgroup analyses. Among the many techniques proposed in this literature are modifications and/or extensions of SVM classifiers (Imai & Ratkovic, 2013), LASSO regression (Ratkovic, Tingley, et al., 2017) and other penalized/low-rank regression methods (Chernozhukov et al., 2019), mixture models (Shiraito, 2016), Bayesian additive regression trees (BART) (Green & Kern, 2012) and random forests (Wager & Athey, 2018). Additionally, researchers have suggested ensembles (Grimmer et al., 2017) or metalearners (Künzel et al., 2019) to pool and aggregate across multiple methods. Similar sorts of methods have already been applied in the conjoint setting to detect higher-order interactions among treatment components while regularizing away from too many false positives (Egami & Imai, 2019), but to our knowledge this is the first such application to the question of marginal component effect heterogeneity across respondent characteristics.

While machine learning techniques have been shown to be very powerful for uncovering within-sample heterogeneity in high-dimensional settings, the properties of ML estimators are often difficult to pin down and the inferential properties of predicted conditional AMCEs may be somewhat poor. In the vein of suggestions developed in Chernozhukov, Demirer, et al. (2018), we recommend machine learning as a first step to obtaining predicted treatment effects for each unit and then post-processing to conduct inference on features of the conditional average treatment effect distribution. Specifically, we implement an extension of the "sorted group average treatment effects" (GATES) approach described in Chernozhukov, Demirer, et al. (2018) to the conjoint setting. We first use a machine learning technique to estimate the conditional AMCE for each unit. We then use these predictions to define the homogeneous subgroups G_k defined in the previous section, grouping the units based on quantiles of their predicted CAMCE. The choice of the size of the bins is driven by the researcher, with smaller bins potentially capturing greater heterogeneity in treatment effects but also yielding higher-variance estimates. We then obtain an estimated GAMCE for each of the subgroups and draw inferences about the shares of the sample that holds a particular preference based on the directions of the estimated group marginal effects.[11] This approach can

[11] Because the bins are defined with respect to the predicted conditional AMCE, the monotonicity assumption will hold asymptotically if the conditional effects estimators are consistent and the true effect distribution is continuous

be understood as a kind of smoothing technique with respect to the noisy and potentially biased ML estimates.

The specific machine learning technique we use here to obtain predicted estimates of individual effects is the causal forest algorithm developed in Athey and Imbens (2016), Athey et al. (2019), and Wager and Athey (2018). This approach extends the popular random forest classification/prediction algorithm (Breiman, 2001) to the setting where the objective function is not predictive accuracy but rather heterogeneity in treatment effects. The algorithm works by recursively generating "trees" based on splits on values of each covariate, with the observations remaining at the end of each of the splits comprising a "leaf" of the tree. Within each leaf, the treatment effect is estimated and the quality of the splits is evaluated on the basis of the difference in treatment effects among leaves. This is in contrast to regular random forest algorithms, which focus on predictive accuracy for determining the splits.

Critically, the set of observations on which the splits are decided and the set used to estimate the leaf-level treatment effects are separate, a property often referred to as "honest" inference. For each tree, the "double-sampling" algorithm in Wager and Athey (2018) randomly subsamples the observations to allocate to each of these two sets (deciding the splits and evaluating the effect at the nodes). With this procedure, an ensemble (forest) of trees is grown, each with a different train-evaluation split of the data, and the predicted conditional treatment effect for each unit can be obtained as the average of predicted CATEs across trees where that unit was "held out" and not used for constructing the tree splits. Because each unit will be used some of the time to construct the tree and in other times for evaluation, this approach allows for sample splitting without any loss of data—all predicted effects on in-sample units can be thought of as "out-of-sample" with respect to the underlying algorithm. Wager and Athey (2018) establish consistency and asymptotic normality for the conditional average treatment effects using the causal forest algorithm.

Using this approach, we obtain an estimate of the conditional average marginal component effect at each value of X_i in the sample. This allows for a simple and easy-to-visualize summary of the overall distribution of treatment effects in-sample via a histogram of these features. We

(Chernozhukov, Demirer, et al., 2018, p. 12). However, because of estimation uncertainty and possible finite-sample biases this ordering may not be true for the estimates obtained in the sample, particularly between groups that have negligibly different treatment effects. Re-arrangement techniques (Chernozhukov et al., 2009) can help improve estimates here.

then obtain estimates of the group CATEs for 10 subgroups in the sample defined by evenly sized splits of the predicted CAMCE for each unit (e.g. units in the 0-10th percentile, 10th-20th, 20th-30th and so on). Using these estimates, we assess the number of subgroups with positive and negative estimated treatment effects to draw conclusions about sample preferences. Finally, as a diagnostic, we examine the covariates that the random forest algorithm found to have a high "variable importance" for each treatment effect to better understand the sources of treatment effect heterogeneity.

4 Simulation: Estimating the Mean-Median Gap

Before applying the method to an actual dataset, we want to evaluate the quality of the random forest predictions in a simulated conjoint setting to better understand how our approach performs under greater and greater violations of the conditional homogeneity assumption. We conduct a series of simulations to assess how robust the machine learning algorithm is in recovering the true individual level MCEs when the covariates are not necessarily perfect predictors of group membership. We want to assess the extent to which the machine learning predictor will detect heterogeneity, and in particular a divergence between the average and the median MCEs under different levels of noise in the covariates. Not surprisingly, we find that more noisy covariates that weakly predict the true underlying heterogeneity groups reduce the quality of the ML predictions. We find no bias in the random forest predictions for features of the CAMCE distribution, particularly the median, when the covariates perfectly discriminate between the heterogeneous subgroups.

The simulation proceeds as follows. First, we construct a population of voters belonging to one of two types, as summarized in Table 1. These voters have complete and transitive preferences over candidates who are defined according to three binary attributes: gender (female or male), denoted by $G \in \{F, M\}$; race (Black or white), denoted by $R \in \{B, W\}$; and age (old or young), denoted by $A \in \{O, Y\}$. Voters of the first type care about gender above all else, and they prefer women over men, so they rank any female candidate above any male candidate. Holding gender fixed, they evaluate a candidate's race—they prefer Black over white candidates—and finally their preferred age, old over young. By contrast, voters of the second type prefer men over women, but

they evaluate candidates first by their preferred race—white over Black—then by their gender, and finally by age.

Rank	Type 1 (40%)	Type 2 (60%)
1.	FBO	MWO
2.	FBY	MWY
3.	FWO	FWO
4.	FWY	FWY
5.	MBO	MBO
6.	MBY	MBY
7.	MWO	FBO
8.	MWY	FBY

Table 1: Preferences Over Candidate Profiles

We assume that our population of voters contains 40% of Type 1 and 60% of Type 2; thus, a minority of voters intensely prefer women, while the majority moderately prefers men. We generate a fully observed dataset of conjoint tasks, in which every voter makes every possible binary comparison and reports a candidate choice derived from the preference orderings in Table 1. We then use this data to compute the true distribution of MCEs, as well as their median and mean. In this population, a correlation between the intensity and direction of preferences drives the true AMCE of female over male, 0.05, to have the opposite sign from the median MCE of -0.25. Thus, even though we can recover an unbiased estimate of the AMCE from an experiment in which some subset of all possible comparisons can be observed, our interest in this exercise is to go one step further: to recover the *distribution* of MCEs, from which we can derive informative statistics for electoral outcomes.

We do so by taking advantage of additional covariates that are informative about voter type. Having generated a population of voters and their preferences over candidates according to Table 1, we construct three additional respondent-level variables centered on respondent type, $T \in \{1, 2\}$, with some random noise. Then, we simulate a survey experiment by randomly sampling a reasonable number of conjoint tasks per respondent from this fully observed dataset. Finally, we apply a nonparametric causal forest approach for estimating heterogeneous treatment effects (Wager & Athey, 2018) to this simulated experimental data, with a binary indicator for female as the treatment and a binary indicator for candidate choice as the outcome, as in a standard conjoint

analysis, and with X_1, X_2, and X_3 as predictors. Repeating this process—both resampling the experimental data and reestimating the causal forest—over many iterations enables us to assess how well this method recovers the underlying MCE distribution, and how its performance is affected by the number of survey respondents, tasks per respondent, and predictiveness of the covariates.

First, we fix the number of survey respondents and tasks at reasonable values that are well within the bounds of standard researcher practice—1000 respondents and 4 choice tasks per person—and vary only the predictiveness of the covariates. We generate our three predictors according to Table 2: as a baseline, covariates that perfectly predict group membership (column b of Table 2 and panel b of Figure 1 below), so that the only source of uncertainty is the sampling of questions; then, noisier covariates that are still highly discriminating (column/panel c); and finally, poorly discriminating covariates, where one standard deviation of the least noisy variable (0.5) is equal to half the distance between the two group means (column/panel d).

	(b)	(c)	(d)
X_1	$\mathcal{N}(T, 0)$	$\mathcal{N}(T, 0.25)$	$\mathcal{N}(T, 0.5)$
X_2	$\mathcal{N}(T, 0)$	$\mathcal{N}(T, 0.5)$	$\mathcal{N}(T, 1)$
X_3	$\mathcal{N}(T, 0)$	$\mathcal{N}(T, 1)$	$\mathcal{N}(T, 2)$

Table 2: Construction of Covariates

Figure 1 gives a sense of the performance of the causal forest for one iteration. With perfectly predictive covariates and only sampling uncertainty, it nearly exactly recovers the MCE distribution, and it still performs quite well when the standard deviation of the most predictive covariate (0.25) is a quarter of the distance between the two group means ($\mu_1 = 1$ and $\mu_2 = 2$). As the noise in the covariates grows, the average remains unbiased but the median is pulled toward the mean.

In Figure 2, we repeat the analysis in Figure 1 over 1000 iterations, varying the predictiveness of the most predictive covariate on the x-axis as well as the number of survey respondents (across panels). Throughout, we fix the total number of binary attributes at 3, and the number of tasks per respondent at 4. On the left, the true difference between subgroup MCEs is maximized at 0.75 (with Type 1's AMCE of female vs. male at 0.5 and Type 2's at -0.25, as in Table 1 above); on the right, we examine the case where there is less differentiation between the two groups, where the 40% minority has an AMCE of 0.25 and the 60% majority has an AMCE of -0.25.

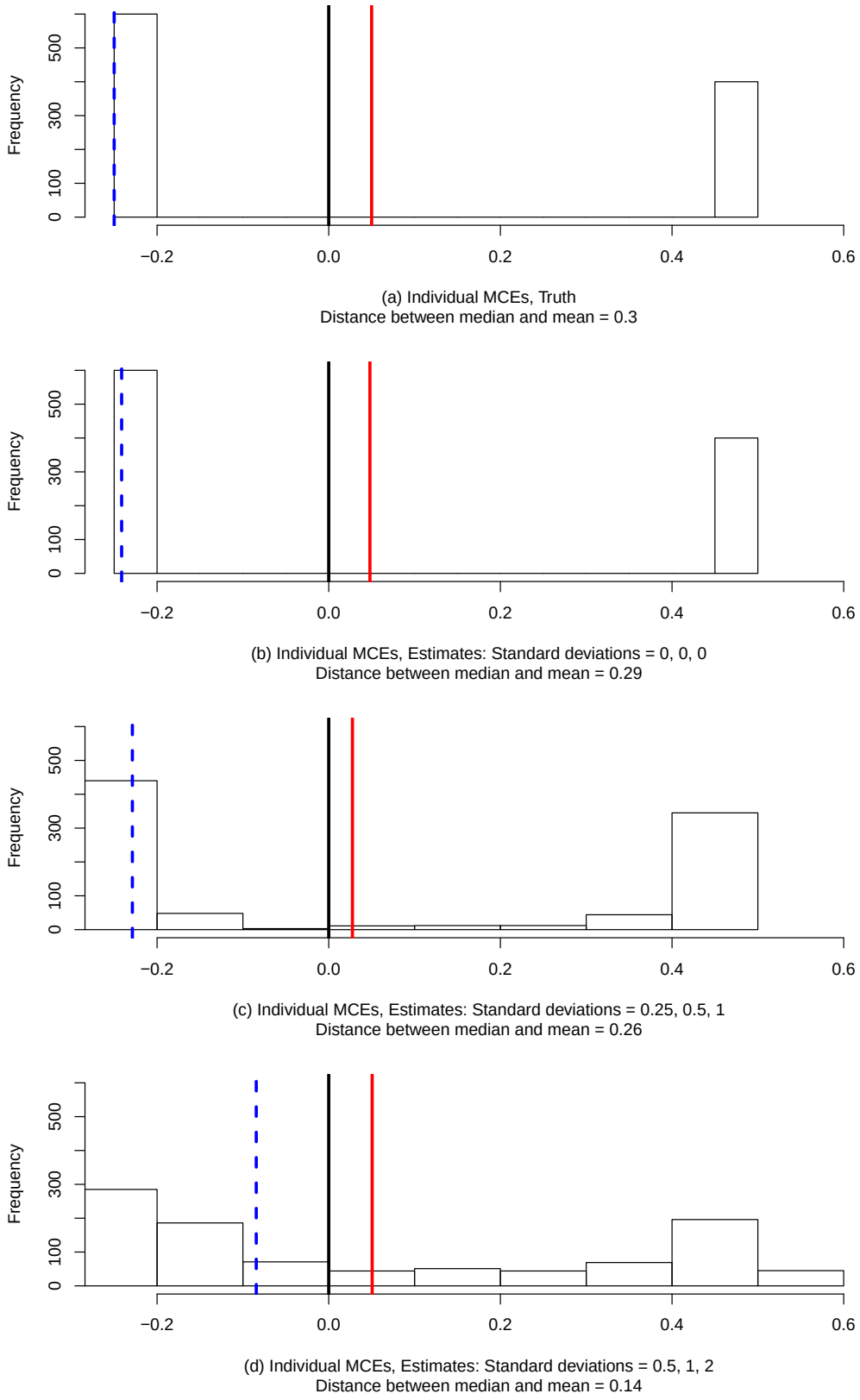


Figure 1: Distribution of MCEs, True vs. Estimated. Blue lines indicate median; red lines mean.

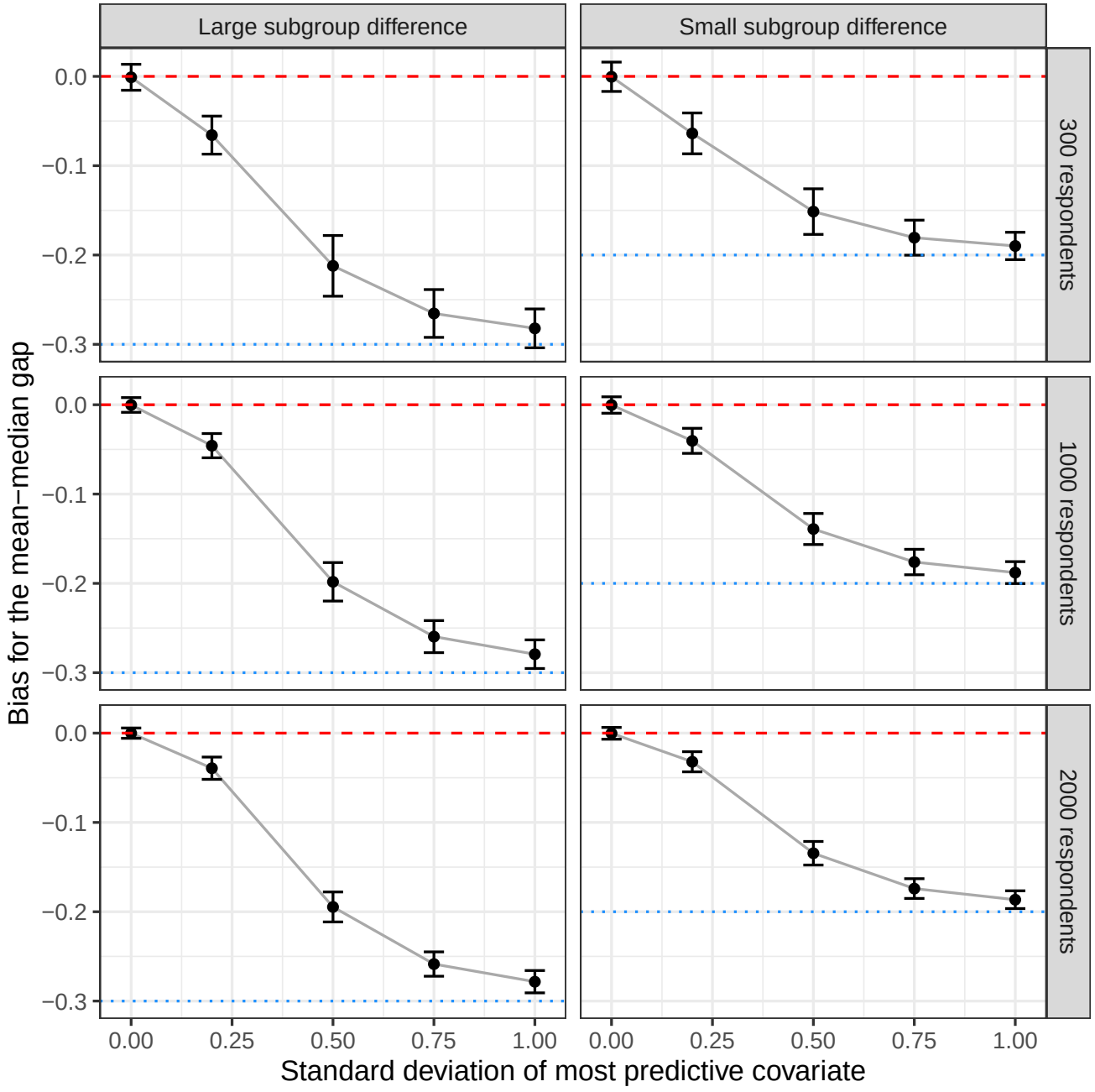


Figure 2: Causal Forest Performance, Detecting the Mean-Median Gap

Across the board, we show that we can recover the mean-median gap without bias when there are perfectly discriminating covariates, and that an attentuation bias emerges, and grows with, noise in the covariates. The analysis shows that, in the worst case, with noisy covariates we will fail to detect heterogeneities that exist; barring random estimation error, the approach will not pick up subgroup treatment effects that are not there.

5 Application: Conjoint Vote Choice Experiments

In this section, we apply our method to the analysis of a recent survey from Data for Progress, which was fielded on a sample of Democratic voters in June-July 2019 (Schaffner & Green, 2019). The conjoint experiment was designed to understand what attributes of candidates Democratic primary voters care about when assessing potential nominees for the 2020 Presidential election. The design presented respondents with two hypothetical candidates comprised of 7 attributes: gender, race, age, strategy for the general election (persuade moderates vs. mobilize the progressive base), outsider vs. establishment background, and positions on health care and climate policy. Each of 2,216 respondents evaluated 5 conjoint tasks, indicating which of the two candidates in a pairwise contest they prefer. In addition to a conjoint experiment, respondents were presented with a large battery of questions about their backgrounds, their political views, and their assessments of the Democratic primary candidates at the time. The data thus provides a rich set of covariates to help us detect heterogeneity in the MCE.

We focus our analysis here on three salient attributes: gender, race and general election strategy. The latter we take as reflective of political ideology, with the more "moderate" approach being focused on persuasion of potential swing voters and the "progressive" approach focusing on base turnout. We also discuss the possible presence of cycling in the age attribute, as it could take on nine possible levels (from 35-75). Our analysis yields important insights about Democratic primary voters that are obscured by a focus on average treatment effects alone. On average, AMCE estimates in this survey are positive for female over male candidates, Black over white candidates, and moderates over progressives. The magnitudes of these estimates are roughly comparable, about 0.02 to 0.04, but the distributions of individual MCEs reveal important differences in how voters assess gender, race, and ideology. Whereas the vast majority of respondents are more likely

to select female over male candidates, with no clear group exhibiting a strong negative preference against a female candidate, race is somewhat more polarizing and ideology/election strategy is *very* polarizing. A small subset of respondents exhibit a negative MCE of a Black candidate compared to a white candidate. With respect to election strategy, there exist two groups of roughly equal size with intense preferences on either side. The pro-persuasion respondents appear to have a slightly more intense effect, resulting in the average indicating a slight positive preference in favor of the persuasion approach.

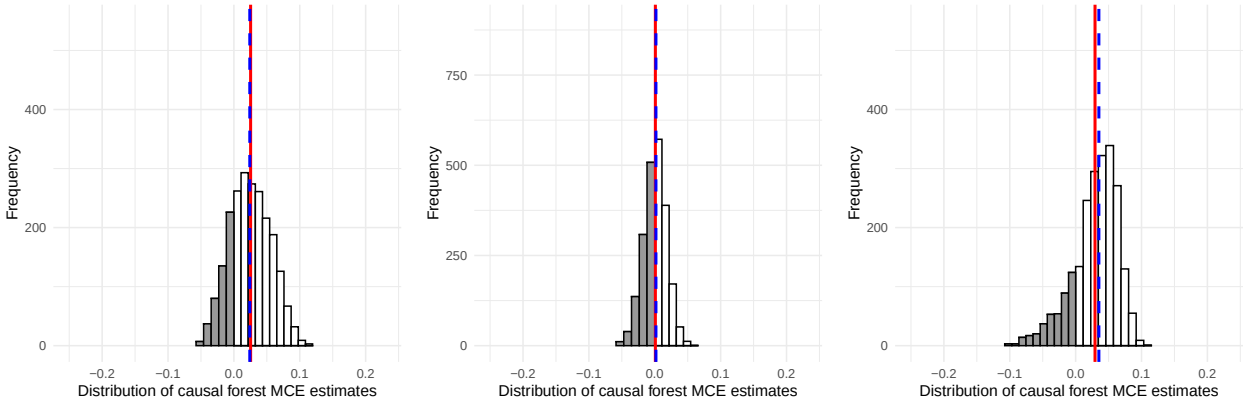
In what follows, we first validate our approach by reporting the most predictive covariates for our treatment effects of interest. Across the board, these top predictors are substantively meaningful and drive theoretically consistent variation in the MCEs. Then, we conduct two further exercises. First, we test for correlations between individual MCE estimates, and find a cluster of voters who strongly prefer Black, female, and progressive candidates and another who prefer white, male centrists. Finally, we show how to use comparisons between the AMCE and the AFCP to evaluate whether preferences over multivalued attributes exhibit any preference cycling.

We first present estimates of the AMCE and AFCP for each attribute-level pairing for these three attributes. We then turn to a heuristic approach suggested in Wager and Athey (2019) for initially assessing heterogeneity in the treatment effects: grouping the observations based on whether their predicted MCE lies above or below the median MCE and estimating the group AMCE for each of these two subgroups. The individual MCEs are estimated for each attribute-level pairing using the causal forest method described above, generating 2000 trees and clustering at the level of the respondent. In addition to yielding cluster-robust variance estimates, this approach ensures that in each "honest" split of the data, all tasks associated with a respondent will be used either for the tree construction or for effect estimation. In total, we have 114 covariates that we use in the causal forest algorithm. Since this is a rather large number of covariates relative to the sample size, we follow the recommendation from Wager and Athey (2019) motivated by Basu et al. (2018) to fit an initial forest with all of the covariates and select those covariates that exhibit above-average "importance" (in that they are frequently used to make splits) and then a second forest using only those selected covariates from the first round.

Table 3 reports the AFCP, AMCE, and the group AMCEs of the above-median and below-median units. In the last column, we compute the proportion of survey respondents who have a

Variable	AFCP	AMCE	Group AMCE, MCE above median	Group AMCE, MCE below median	Proportion w/positive MCE
Latino vs. white	0.524 [0.505, 0.543]	0.025 [0.007, 0.041]	0.058 [0.034, 0.083]	-0.008 [-0.033, 0.016]	0.781
Black vs. Latino	0.505 [0.472, 0.538]	0.005 [-0.017, 0.025]	0.002 [-0.032, 0.035]	0.005 [-0.028, 0.039]	0.544
Black vs. white	0.528 [0.508, 0.549]	0.029 [0.011, 0.046]	0.073 [0.047, 0.098]	-0.013 [-0.037, 0.011]	0.813
Female vs. male	0.540 [0.526, 0.553]	0.039 [0.026, 0.053]	0.080 [0.061, 0.100]	-0.001 [-0.019, 0.018]	0.826
Moderate vs. progressive	0.519 [0.504, 0.534]	0.019 [0.005, 0.033]	0.080 [0.060, 0.099]	-0.042 [-0.062, -0.022]	0.649

Table 3: Summary of average and heterogeneous effects, Data for Progress Conjoint Survey

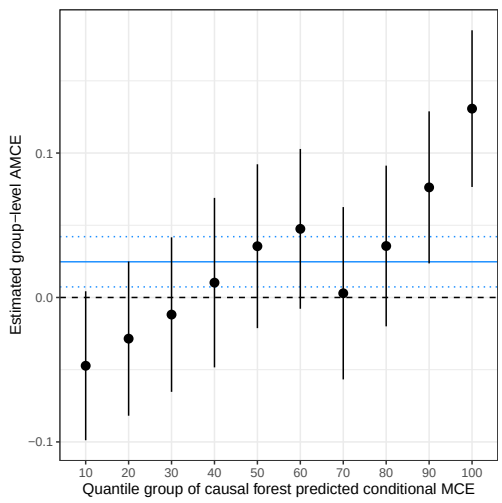


(a) Latino vs. white (b) Black vs. Latino (c) Black vs. white

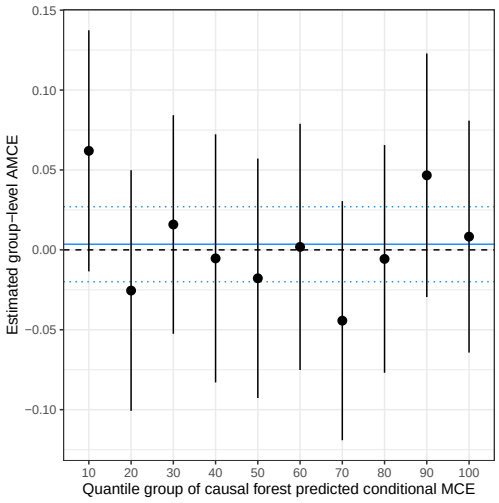
(d) Female vs. male (e) Moderate vs. progressive

Figure 3: Distributions of MCE Estimates

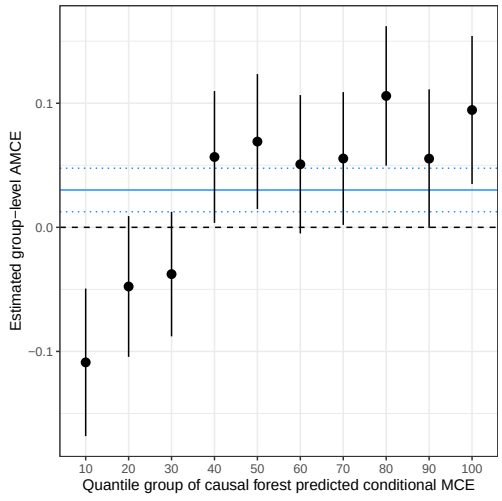
Notes: Blue line indicates median; red line indicates mean. Part of distribution that falls below 0 is shown in gray.



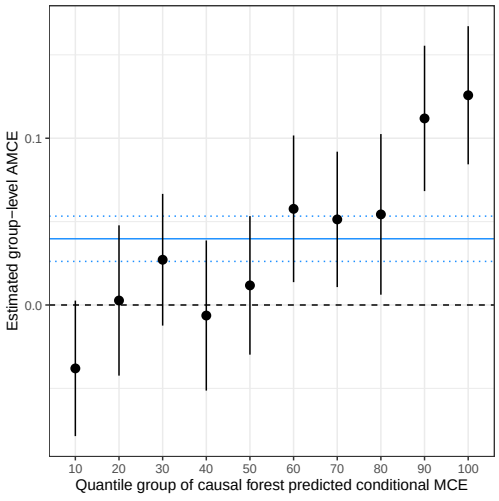
(a) Latino vs. white



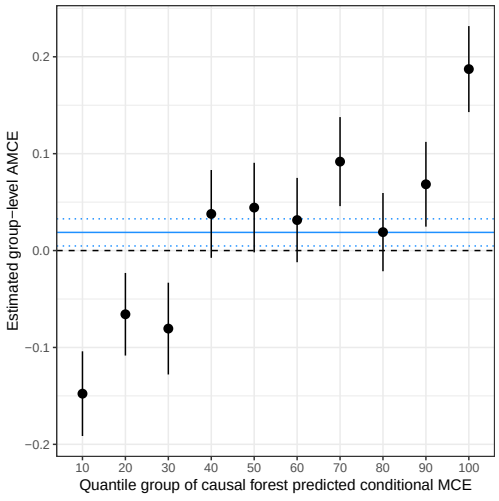
(b) Black vs. Latino



(c) Black vs. white



(d) Female vs. male



(e) Moderate vs. progressive

Figure 4: Estimated Group AMCEs by quantiles of the predicted causal forest MCEs

Notes: Horizontal blue lines indicate estimated AMCE and 95% confidence intervals.

positive out-of-sample MCE estimate, which can give researchers a rough sense of the magnitude of heterogeneity in the preference distribution. All confidence intervals are computed by clustering standard errors on the respondent. An initial glance at the estimates reveals potentially interesting heterogeneity for two attributes with similar AMCE estimates: gender and strategy. The group AMCEs for units with above-median predicted MCEs for these two attributes are nearly identical (0.08). However, in contrast to the gender attribute where there is no clear countervailing *negative* MCE group that prefers men to women, there is clear evidence of a subgroup of respondents who oppose a moderate persuasion strategy and favor a progressive base mobilization strategy.

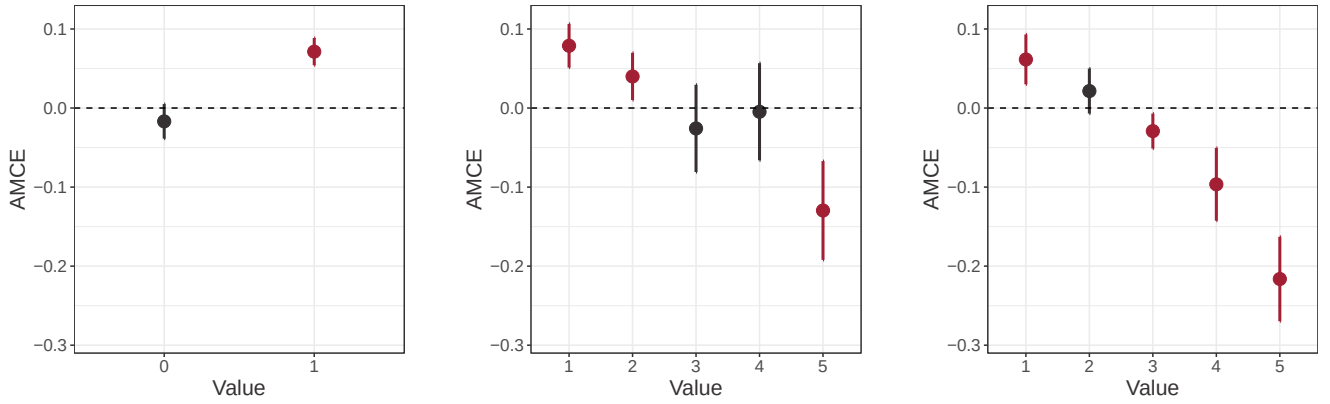
The plotted distributions of the individual MCE estimates in Figure 3 reveal further the variation in how respondents assess race, gender, and ideology. While the AMCE of Black over white is positive, and a majority of respondents prefer Black over white candidates, a long left tail indicates intense preferences in the opposite direction among a minority of the sample. By the same token, a small but statistically significant positive treatment effect for moderate over progressive candidates masks intense preferences on both sides, and the relatively small margin by which moderates outnumber progressives suggests that it is crucial to take differential turnout between the two groups into account when forecasting electoral outcomes. By contrast, the MCE distribution for female over male candidates is symmetric and mostly positive, suggesting that preferences for a female candidate are relatively homogeneous within the sample, or at least that some respondents prefer a female candidate and the remainder are largely indifferent.

Figure 4 implements our proposed application of the GATES sorted group average treatment effects method. We bin respondents into disjoint groups based on their predicted MCEs from the random forest model for each of the treatment comparisons and estimate the group AMCE for each of these quantiles. We find that the within-group estimates are mostly increasing with respect to the quantiles, as expected. However, slight deviations in ordering are expected, especially for effects that are close to zero. In the most extreme case, Black vs. Latino, the estimates all essentially bounce around zero, consistent with no treatment effect heterogeneity being detected by our covariates. Conversely, for the moderate vs. progressive component effect, we find that three out of ten of our bins exhibit strong, statistically significant negative effects and three out of ten exhibit strong, statistically significant positive effects, with the remaining groups all having confidence intervals crossing zero. This provides a clear example where the positive and statistically

significant pooled AMCE estimate is clearly misleading regarding majority preference. Rather, we see two roughly equally sized groups with intense preferences in countervailing directions and a plurality with very weak opinions on strategy. By contrast, for the female vs. male treatment effect, we see groups with strong positive preferences for a female candidate on the top end of the predicted effect distribution but no countervailing negative treatment effect. Here, the direction of the AMCE appears to coincide well with majority preference. Preferences for Black vs. white candidates are also generally positive, but here we do see a small minority of respondents with a predicted negative treatment effect, i.e. voters who would prefer a white candidate.

To validate whether our approach has successfully discovered politically meaningful preference heterogeneities and to better understand the types of individuals falling in each of the predicted MCE bins, we examine the most predictive covariates for the reported treatment effects based on their variable importance (see Appendix Table B1 for a detailed list of the five most discriminating survey questions for each effect). For instance, support for female candidates is best predicted by support for Elizabeth Warren and Alexandria Ocasio-Cortez, as well as reactions to the statement that "women are too easily offended." Support for Black over white candidates is predicted by support for the Black Lives Matter movement and reactions to the claims that slavery and discrimination have created conditions that make it difficult for African Americans to work their way out of the lower class, and that "political correctness" has "gone too far." Not surprisingly, support for progressive candidates is predicted by support for the Democratic Socialists of America and Alexandria Ocasio-Cortez, by political ideology and by self-identification with the "progressive" label. In Figure 5, we plot subgroup AMCE estimates by value of the most predictive covariate for three estimated effects: female vs. male, Black vs. white, and moderate vs. progressive. These estimates move dramatically, and in the expected direction, with values of these covariates.

The MCE estimates also reveal ideological coherence among respondents. Figure 6 presents correlations between individual-level MCEs for eight estimated effects of gender, race, age, and ideology. Strong correlations in respondents' preference structures are observed across the board, with some of the largest correlations for female over male and progressive over moderate candidates (0.74); female over male and Black over white (0.73) as well as Latino over white (0.8) candidates; and progressive over moderate and Black over white (0.61) as well as Latino over white (0.75) candidates.



(a) Female vs. male
1 = Considering Warren,
0 = Not considering Warren.

(b) Black vs. white
Generations of slavery and discrimination have created conditions that make it difficult for African Americans to work their way out of the lower class. 1 = Strongly agree, 5 = Strongly disagree.

(c) Progressive vs. moderate
View of Democratic Socialists of America: 1 = Strongly favorable, 5 = Strongly unfavorable.

Figure 5: AMCE by Subgroup for the Most Predictive Covariate

Notes: Vertical lines represent 95% confidence intervals. See the top question for each effect in Appendix Table B1 for exact question wording.

Finally, in Figure 7, we show AFCP estimates side by side with AMCE estimates for all candidate age comparisons that can be computed from the data—that is, all possible pairs of treatment (x-axis) and baseline (y-axis) values of candidate age that were part of the experiment. The right panel demonstrates how the AMCE implicitly imposes single-peakedness, as indicated by the color gradient always moving in the same direction over any given column or row. However, the AFCP estimates on the left allow for the detection of cycles, such as the one suggested by a positive AFCP of age 35 over 60 (0.53), age 60 over 40 (0.55), and 40 over 35 (0.6). Nevertheless, because these estimates are not all statistically distinguishable from 0.5, and because we do not observe many such cycles, we cannot rule out that age may satisfy single-peakedness.

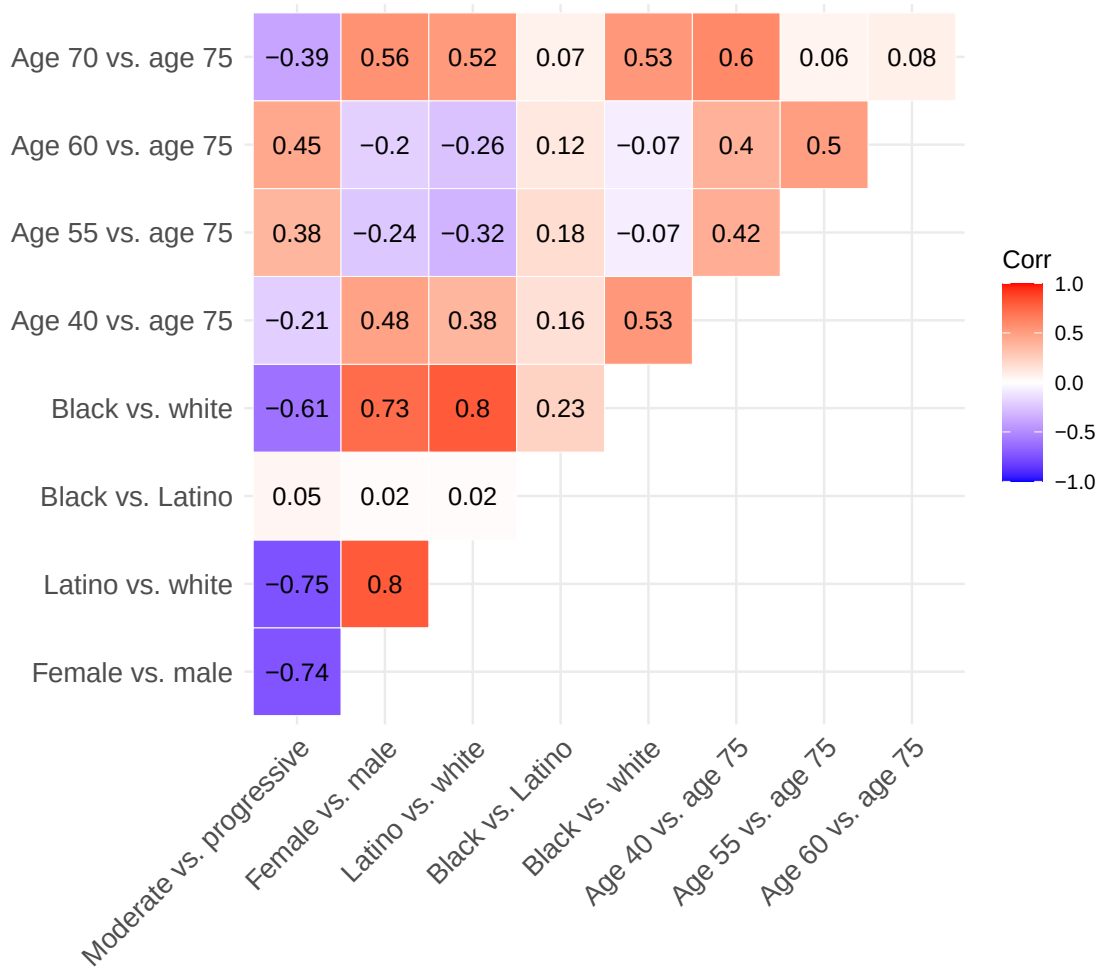


Figure 6: Correlations Between MCE Estimates

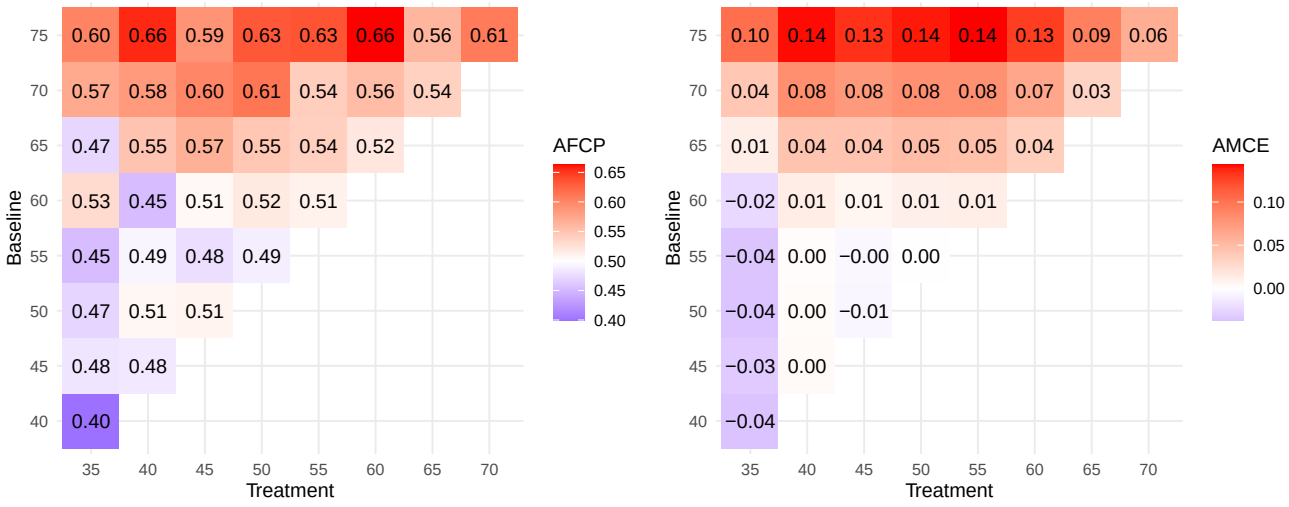


Figure 7: AFCP and AMCE Estimates for All Age Comparisons

6 Conclusion

Preference elicitation via conjoint analysis in political science has, to this point, largely focused on average treatment effects. However, heterogeneity in treatment effects across respondents in a conjoint design is not just an interesting curiosity, but poses a significant problem to interpreting average effects from conjoint designs in a theoretically coherent manner. In this paper we have shown that, in a standard conjoint design, researchers can recover a wide range of meaningful quantities of interest using off-the-shelf machine learning tools. Of course, there is no free lunch and our approach requires the additional assumption of conditional preference homogeneity. However, the theoretical purchase in many instances will outweigh whatever costs researchers may face in terms of this potentially restrictive assumption.

Moreover, we have shown that the AMCE, the estimand focused on by most political scientists, also requires the potentially strong assumption of single-peaked preferences to ensure the non-existence of Condorcet-like preference cycles. For many unordered and multi-valued attributes, this is a questionable assumption. We have characterized an alternative quantity, the AFCP, that ameliorates and uncovers the key theoretical problem we identify. Related problems occur even without averaging over individual effects. We show that the individual MCE implicitly assumes the transitivity of individual preferences. That is, the individual MCE may seemingly evince a transitive preference ordering even when transitivity does not maintain. By contrast, the set of FCPs will reflect non-transitive preferences. However, since the AFCP exploits fewer comparisons, it will be less precisely estimated and therefore comparisons of the AFCP and AMCE may be driven by preference cycling or by statistical noise. We are in the process of constructing a formal Hausmann-like statistical test for the presence of Condorcet-like cycles based upon comparisons of the set of all AMCEs and AFCPs.

The approach we outlined in this paper will allow researchers to obtain a wider range of relevant quantities from already implemented conjoint designs. Future research should develop and adapt algorithmic and adaptive approaches common in the literature on preference measurement in marketing to the specific context of voter preferences elicitation. When coupled with theoretical models of choice, these methods may provide researchers with even more refined tools for understanding political behavior.

References

Abramson, S. F., Kocak, K., & Magazinnik, A. (2020). *What do we learn about voter preferences from conjoint experiments?* (Tech. rep.).

Achen, C. H. (1975). Mass political attitudes and the survey response. *The American Political Science Review*, *69*(4), 1218–1231.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, *89*(1-2), 57–78.

Ansolabehere, S., Rodden, J., & Snyder Jr, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 215–232.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., Wager, S., Et al. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2018). The number of choice tasks and survey satisficing in conjoint experiments. *Political Analysis*, *26*(1), 112–119.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2019). Conjoint survey experiments. In J. N. Druckman & D. P. Green (Eds.), *Cambridge handbook of advances in experimental political science*. Cambridge University Press.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2020). Using conjoint experiments to analyze elections: The essential role of the average marginal component effect (amce). *Available at SSRN*.

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, *115*(8), 1943–1948.

Bochsler, D. (2010). The marquis de condorcet goes to bern. *Public Choice*, *144*(1-2), 119–131.

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

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The american voter*. University of Chicago Press.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogenous treatment effects in randomized experiments* (tech. rep.). National Bureau of Economic Research.

Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, *96*(3), 559–575.

Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.

Chernozhukov, V., Hansen, C. B., Liao, Y., & Zhu, Y. (2019). *Inference for heterogeneous effects using low-rank estimations* (tech. rep.). CEMMAP working paper.

De la Cuesta, B., Egami, N., & Imai, K. (2019). *Improving the external validity of conjoint analysis: The essential role of profile distribution* (tech. rep.).

Egami, N., & Imai, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, *114*(526), 529–540.

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

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, *25*(4), 413–434.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, *22*(1), 1–30.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Heckman, J. J., & Snyder Jr, J. M. (1997). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The Rand Journal of Economics*, *28*, S142.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Huber, J., & Train, K. (2001). On the similarity of classical and bayesian estimates of individual mean partworths. *Marketing Letters*, *12*(3), 259–269.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470.

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

Kurrild-Klitgaard, P. (2001). An empirical example of the condorcet paradox of voting in a large electorate. *Public Choice*, *107*(1-2), 135–145.

Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, *15*(2), 173–191.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J. C., Et al. (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, *19*(3-4), 337.

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

Orme, B., & Howell, J. (2009). *Application of covariates within sawtooth software's cbc/hb program: Theory and practical example* (tech. rep.). Sawtooth Software Research Paper Series.

Ratkovic, M., Tingley, D. Et al. (2017). Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis*, *25*(1), 1–40.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American statistical association*, *81*(396), 961–962.

Schaffner, B., & Green, J. (2019). What attributes do democratic primary voters value? *Data for Progress*. `https://www.dataforprogress.org/blog/2019/7/11/what-attributes-do-democratic-primary-voters-value`

Scholz, S. W., Meissner, M., & Decker, R. (2010). Measuring consumer preferences for complex products: A compositional approach based on paired comparisons. *Journal of Marketing Research*, *47*(4), 685–698.

Shiraito, Y. (2016). Uncovering heterogeneous treatment effects. *Working Paper*.

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

Wager, S., & Athey, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv:1902.07409*.

A Proofs

Proof of the relationship between the MCE and FCP (Proposition 1). Start with the definition of the marginal component effect

$$\text{MCE}_{il}(t_1, t_0) = \sum_{(t,\mathbf{t})\in\mathcal{T}} [Y_i(t_1, t, \mathbf{t}) - Y_i(t_0, t, \mathbf{t})] \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}})$$

For ease of exposition, we suppress the conditioning notation $T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}}$.

Take the first part of the expression and decompose it by splitting $\mathbf{t}$ into the value for level l and the values for all levels t'. Let $\mathcal{D}_l$ denote the set of levels that comprise attribute l, $\mathcal{D}_l = \{0, 1, \ldots, D_l - 1\}$.

$$\sum_{(t,\mathbf{t})\in\mathcal{T}} Y_i(t_1, t, \mathbf{t}) \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}) =$$

$$\sum_{a\in\mathcal{D}_l} \sum_{(t,t')\in\mathcal{T}} Y_i(t_1, t, t_a, t') \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t^* | T_{i[-j]kl} = t_a) p(T_{i[-j]kl} = t_a)$$

By complete randomization (Assumption 4)

$$= \sum_{a\in\mathcal{D}_l} \left[\sum_{(t,t')\in\mathcal{T}} Y_i(t_1, t, t_a, t') \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t^* | T_{ijkl} = t_1, T_{i[-j]kl} = t_a) \right] p(T_{i[-j]kl} = t_a)$$

By the definition of the FCP (Definition 5)

$$= \sum_{q\in\mathcal{D}_l} \text{FCP}(t_1, t_a) \times p(T_{i[-j]kl} = t_a)$$

Splitting the sum into three components yields

$$= \text{FCP}(t_1, t_0) \times p(T_{i[-j]kl} = t_0) + FCP(t_1, t_1) \times p(T_{i[-j]kl} = t_1) + \sum_{q\neq(0,1)} FCP(t_1, t_a) \times p(T_{i[-j]kl} = t_a)$$

where $\sum_{a\neq(0,1)}$ denotes a sum over all levels that are not t_1 or t_0

Applying the same logic to the other half yields

$$\sum_{(t,\mathbf{t})\in\mathcal{T}} Y_i(t_0,t,\mathbf{t}) \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}) =$$

$$\text{FCP}(t_0,t_1) \times p(T_{i[-j]kl} = t_1) + FCP(t_0,t_0) \times p(T_{i[-j]kl} = t_0) + \sum_{a\neq(0,1)} \text{FCP}(t_0,t_a) \times p(T_{i[-j]kl} = t_a)$$

By construction, $\text{FCP}(t_1,t_1) = \text{FCP}(t_0,t_0) = \frac{1}{2}$ and $\text{FCP}(t_0,t_1) = 1 - \text{FCP}(t_1,t_0)$. Therefore

$$\sum_{(t,\mathbf{t})\in\mathcal{T}} Y_i(t_0,t,\mathbf{t}) \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}) =$$

$$p(T_{i[-j]kl} = t_1) - \text{FCP}_{il}(t_1,t_0) \times p(T_{i[-j]kl} = t_1) + \frac{1}{2} p(T_{i[-j]kl} = t_0) + \sum_{a\neq(0,1)} \text{FCP}_{il}(t_0,t_a) \times p(T_{i[-j]kl} = t_q)$$

Substituting into the MCE expression and rearranging terms:

$$\text{MCE}_{il}(t_1,t_0) = \left[\text{FCP}_{il}(t_1,t_0) - \frac{1}{2}\right] \times \left[p(T_{i[-j]kl} = t_1) + p(T_{i[-j]kl} = t_0)\right] + \sum_{q\neq(0,1)} \left[\text{FCP}_{il}(t_1,t_a) - \text{FCP}_{il}(t_0,t_a)\right] \times p(T_{i[-j]kl} = t_a)$$

If we further assume that the distribution of the attribute levels T_{i-jkl} is uniform: $p(T_{i-jkl} = t_a) = p(T_{ijkl} = t_a^*) = \frac{1}{D_l}$ for all levels $t_a, t_a^* \in \mathcal{D}_l$, this simplifies to

$$\text{MCE}_{il}(t_1,t_0) = \frac{2}{D_l}\left[\text{FCP}_{il}(t_1,t_0) - \frac{1}{2}\right] + \frac{D_l - 2}{D_l} \sum_{t_a\neq(t_1,t_0)} \left[FCP_{il}(t_1,t_a) - \text{FCP}_{il}(t_0,t_a)\right]$$

It is worth noting that the key equality obtained by the complete randomization assumption is

$$p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t^* | T_{i[-j]kl} = t_a) = p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t^* | T_{ijkl} = t_1, T_{i[-j]kl} = t_a)$$

which can hold under the less restrictive conditionally independent randomization (Assumption 4, Hainmueller et al. (2014)) since we are also conditioning on the other attributes being in the common support $T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}}$

The relevant randomization restrictions are that all attribute levels in j are independent of those in $-j$ (such that $T_{ijkl} \perp\!\!\!\perp T_{i[-j]kl}, T_{i[-j]k[-l]}$) and that conditional on the common support $\tilde{\mathcal{T}}$, $T_{ijkl} \perp\!\!\!\perp T_{ijk[-l]}$ (since the attribute-levels in the common support are the "unrestricted" attribute levels).

However, one caveat is that conditioning on the "common support" $\tilde{\mathcal{T}}$ may imply different distributions of the other attributes $T_{ijk[-l]}$, $\mathbf{T}_{i[-j]k[-l]}$ for each of the FCPs that comprise a given MCE. For example, if t_a cannot appear with certain other attribute combinations, the $\text{FCP}(t_1, t_0)$ will be defined as an average over all possible other attribute combinations, but $\text{FCP}(t_1, t_a)$ and $\text{FCP}(t_0, t_a)$ are averages only for those tasks where the attribute restrictions hold on the profile with t_a. □

Proof of Lemma 2.1. For sufficiency, without loss of generality take $FCP_i(x, y) \geq 1/2$. From FCP-transitivity, it must be that $FCP_i(x, z) \geq FCP_i(y, z)$. It follows then from the definition of $MCE_i(x, y)$:

$$\begin{aligned} MCE_i(x,y) &= \frac{1}{3}FCP_i(x,y) + \frac{1}{6} + \frac{1}{3}FCP_i(x,z) - \left(\frac{1}{3}FCP_i(y,x) + \frac{1}{6} + \frac{1}{3}FCP_i(y,z)\right) \\ &= \frac{1}{3}\underbrace{(FCP_i(x,y) - FCP_i(y,x))}_{\geq 0} + \underbrace{\left(\frac{1}{6} - \frac{1}{6}\right)}_{0} + \frac{1}{3}\underbrace{(FCP_i(x,z) - FCP_i(y,z))}_{\geq 0} \geq 0. \end{aligned}$$

Necessity follows from the same argument as above by way of contraposition. That is, to show that $MCE_i(y, x) \geq 0 \implies FCP(y, x) \geq 1/2$, we prove its contrapositive, $FCP(y, x) < 1/2 \implies MCE_i(y, x) < 0$. This is equivalent to $FCP(x, y) \geq 1/2 \implies MCE_i(x, y) \geq 0$ which we have shown in the first part. □

B Additional Tables and Figures

Effect	Question
Female vs. male	Thinking about the 2020 Democratic presidential [primary/caucus] in your state, which candidate or candidates are you considering voting for? Select all that apply: 1 = Selected Warren, 0 = Did not select Warren.
	Would you say that, in general, you have a favorable or unfavorable view of Alexandria Ocasio-Cortez? 1 = Strongly favorable, 5 = Strongly unfavorable.
	Please indicate the extent to which you agree with the following statement. Women are too easily offended. 1 = Strongly agree, 5 = Strongly disagree.
	Please indicate the extent to which you agree with the following statement: Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. 1 = Strongly agree, 5 = Strongly disagree.
	For each of the following groups, please say whether most people in the group have more money than they deserve, less money than they deserve, or about the right amount of money: Poor people. 1 = A lot more money than they deserve, 7 = A lot less money than they deserve.
Black vs. white	Please indicate the extent to which you agree with the following statement. Generations of slavery and discrimination have created conditions that make it difficult for African Americans to work their way out of the lower class. 1 = Strongly agree, 5 = Strongly disagree.
	Would you say that, in general, you have a favorable or unfavorable view of Black Lives Matter? 1 = Strongly favorable, 5 = Strongly unfavorable.
	Please indicate the extent to which you agree with the following statement: Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. 1 = Strongly agree, 5 = Strongly disagree.
	There's been a lot of talk lately about 'political correctness'. Some people think that the way people talk needs to change with the times to be more sensitive to people from different backgrounds. Others think that this has already gone too far and many people are just too easily offended. Which is closer to your opinion? People are much too easily offended. 1 = Selected, 0 = Did not select.
	Please indicate the extent to which you agree with the following statement: White people in the U.S. have certain advantages because of the color of their skin. 1 = Strongly agree, 5 = Strongly disagree.

Table B1: Five Most Predictive Covariates, Data for Progress Conjoint Survey

Effect	Question
Progressive vs. moderate	Would you say that, in general, you have a favorable or unfavorable view of the Democratic Socialists of America? 1 = Strongly favorable, 5 = Strongly unfavorable.
	Would you say that, in general, you have a favorable or unfavorable view of Alexandria Ocasio-Cortez? 1 = Strongly favorable, 5 = Strongly unfavorable.
	Please indicate the extent to which you agree with the following statement: Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. 1 = Strongly agree, 5 = Strongly disagree.
	Which of the following words apply to you? Check all that apply: Progressive. 1 = Checked, 0 = Not checked.
	In general, how would you describe your own political viewpoint?. 1 = Very Liberal, 5 = Very Conservative

Table B1 (cont.): Five Most Predictive Covariates, Data for Progress Conjoint Survey