Detecting Preference Cycles in Forced-Choice Conjoint Experiments

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

September 11, 2022

Abstract

In this paper we describe implications of an under-explored theoretical property of the Average Marginal Component Effect (AMCE): its violation of *independence*. We show how this results from the AMCE's incorporation of information about *irrelevant attributes* by averaging over both direct and indirect comparisons of features. In doing so, the AMCE can impose an artificial *transitivity* and produce positive AMCEs even when respondents are *less* likely to choose a profile with one feature over the baseline in direct comparisons. We introduce an alternative estimand, the Average Feature Choice Probability (AFCP), which only considers direct comparisons and corresponds to the frequency an attribute is chosen in paired comparisons. We decompose the AMCE into a weighted average of AFCPs and we describe the necessary conditions under which a positive AMCE implies that an AFCP is greater than one-half. Finally, we develop a statistical test for the presence of preference intransitivities and illustrate our method with the reanalysis of a number of conjoint experiments.

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

[†]Assistant Professor, Department of Political Science, New York University Abu Dhabi, email: kkocak@nyu.edu

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

[§]Assistant Professor, Department of Political Science, University of Chicago, email: astrezhnev@uchicago.edu

1 Introduction

Preference measurement and preference aggregation constitute two of the most well-developed branches of contemporary political science (Ansolabehere et al., 2008; Austen-Smith & Banks, 1998; Berinsky, 2017; Patty & Penn, 2019). They are, nevertheless, too frequently studied in isolation. Practitioners interested in the elicitation of preferences rarely consider the aggregation mechanisms implied by the statistical tools they employ. This can be problematic since statements about social preferences derived from modern elicitation techniques (at least as practiced in political science) are frequently difficult to justify on theoretical grounds.

In this paper we describe a heretofore unexplored theoretical property of typical estimands targeted in conjoint experiments in political science. These experiments focus on the Average Marginal Component Effect (AMCE), which is nearly always interpreted as a measure of social preference (Bansak et al., 2021; Hainmueller et al., 2014; Leeper et al., 2020). We show that the AMCE incorporates information about irrelevant attribute levels when making comparisons across feature-levels and, as such, can artificially create a transitive ranking of attributes. This can occur when, in fact, the underlying distribution of preferences does not support such an ordering in pairwise comparisons. That is, with three attribute levels a, b, & c, the AMCE incorporates information about how a and b both perform against c when constructing an estimate of social preference for a relative to b.

This is an undesirable property for a preference relation if we believe that judgements about any pair of attributes should not depend upon preferences relative to a third attribute (Chernoff, 1954; Sen, 1993). It is potentially even less desirable for empirical scholars who want to use conjoint experiments to make valid statements about voter preferences as observed in the world. The AMCE will incorporate information from indirect comparisons and mechanically produce a transitive ordering of attributes even when such an order is not warranted. That is, we show that the AMCE can produce a transitive ranking even when, in paired comparisons of attributes, there are preference cycles where a is preferred to b, b preferred to c, but c preferred to a.

Of course, preference aggregation problems are among the most well-studied in political science. Indeed, our paper directly relates to canonical results in social choice (Arrow, 1950; Arrow, 1963). As a preference aggregation mechanism, the AMCE trivially satisfies three of Arrow's four conditions: unanimity, universal admissibility, and non-dictatorship. An innovation of ours is to describe the AMCE's violation of the independence assumption and its mechanical production of a transitive

ranking of attributes in data produced by conjoint experiments.

The main contribution of our paper, however, is to present an alternative estimand, the Average Feature Choice Probabilty (AFCP), which satisfies independence. It represents the probability that a given feature is chosen in *direct* comparisons, eschewing information from indirect comparisons, reflecting the probability of choosing a profile with feature a over a profile with feature b only among those tasks where a and b are present. That is, unlike the AMCE, the AFCP of a vs b does not incorporate any information about their performance against c.

Like the AMCE, the AFCP fails to satisfy all Arrow's criteria, though this is true of any aggregation rule. We compare the properties of the AFCP and the AMCE from a social choice perspective — a comparison whose historical meaning we highlight in the conclusion. The key theoretical trade-off researchers face is that in targeting the AFCP, they give up transitivity to obtain independence. In our view, this trade is worth making, since the goal of empirical preference elicitation is to *describe* preferences as they are, rather than construct an aggregation rule that ranks outcomes in a manner consistent with normatively appealing axioms. That is, if there exist preference cycles in the target population, the estimand of a preference elicitation experiment should reflect this. The AFCP reflects cycles (when they exist) while the AMCE imposes an artificial transitivity.

Still, the benefits of the AFCP do not come without some cost. By disregarding information from all indirect comparisons, researchers will inevitably lose statistical power. This loss of precision may very well be considerable in standard designs when attributes are drawn uniformly, thus resulting in few direct comparisons. As a first-best, we therefore recommend that researchers implement randomization schemes that maximize the number of direct comparisons of interest.

However, as a second-best, we show that the AMCE can be characterized as a weighted average of the AFCP generated by the *direct* comparisons between features of interest and all other AFCPs generated by *indirect* comparisons involving features of interest and irrelevant features. We use this result to describe the necessary conditions under which a positive (negative) AMCE implies that an AFCP greater (less) than one-half and then to provide a method for uncovering the presence of preference cycles in data generated in conjoint experiments.

Our reanalysis of several published experiments highlights three ways in which preference cycles commonly emerge in conjoint designs:

i. *Single-peaked preferences cannot be assumed*

It is well known that when preferences are single-peaked, cycles of the sort our method can

uncover are ruled out (Black, 1948). In the first place, when there is no natural ordering of features it is unlikely that single-peakedness is likely to maintain. For example, a common conjoint design attempts to elicit preferences on immigration by asking respondents to select between profiles of hypothetical immigrants (e.g. Arias & Blair, 2022; Denney & Green, 2021; Hainmueller & Hopkins, 2015). In such experiments there are many features of immigrants, such as country of origin or ethnicity, over which we cannot simply assume that respondents will have well-behaved preferences of this sort. Even when features have a natural ordering, the assumption of single-peakedness may not hold. Consider age, an attribute frequently included in candidate choice conjoint designs. Although there is a clear ordering of age, it is plausible that voters do not maintain single-peaked preferences in this domain. For instance, some voters may very well prefer both young and old candidates to middle aged ones.

ii. *Attributes over which respondents have multi-dimensional preferences*
For attributes like candidates' political parties, also commonly included in candidate choice experiments, we may obtain intransitivities because these features are genuinely multidimensional. Again, it is a well known result that when voter preferences exist in a multi-dimensional space it is possible to construct preference cycles (McKelvey, 1979; McKelvey & Schofield, 1987; Plott, 1967). As such, features that themselves are multidimensional in nature are very likely the sort of attributes where we might expect to find intransitivities in data generated by conjoint experiments.

iii. *The existence of similar attribute-features*
When researchers include multiple similar features for one attribute in their experiment, there is the potential that, while respondents are indifferent between the two similar features in direct comparisons, these features perform differently against other options.

Broadly, our paper contributes to the growing literature on the "theoretical implications of empirical models," the aim of which is to link credibly identified effect estimates to parameters relevant to theories of politics (Ashworth et al., 2021; De Mesquita & Tyson, 2020; Izzo et al., 2018). We advance this literature by showing that even in the most credible of research designs — a randomized trial — additional, potentially strong, assumptions are required to make credible statements about the most central feature of politics, preferences. More specifically, we contribute to the literature in political methodology on the interpretation of results from conjoint experiments (Abramson et al.,

2022; Ganter, 2021). Most closely, we address the interpretation of the AMCE of Bansak et al. (2021) as the average effect of manipulating a given feature on a candidate's vote-share. Our findings make clear that this averaging is over all of the direct comparisons between attributes of interest *as well as* their indirect comparisons with irrelevant attributes.

The remainder of the paper is structured as follows: Section 2 reviews the definition of the AMCE in terms of counterfactual responses. It then defines an alternative quantity, the Average Feature Choice Probability (AFCP), for designs involving a binary choice between two profiles. It derives the relationship between the AMCE and the AFCPs. Section 3 discusses the assumptions necessary to interpret the marginal component effect as a pairwise preference and provides a definition of preference cycling in terms of the feature choice probabilities. Section 4 describes the procedure for estimating the AFCPs from the data and develops a statistical test for the presence of a divergence between the direct and indirect AFCP comparisons, a graphical diagnostic tool to assess the sensitivity of AMCEs to the inclusion of different indirect comparisons, and a generalized method of moments estimator for the AFCP that provides for improved precision if researchers are in fact willing to make the assumption that the direct and indirect comparisons are equivalent. Section 5 applies these methods to three recently published conjoint experiments. Finally, 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 t_a and

some alternative t_b, but *also* indirect pairwise comparisons between t_a, t_b and each of the other attribute levels, t_c. 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 necessarily indicative of whether a feature is preferred to an alternative on average.

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. We observe for each profile Y_{ijk}, which denotes the score, rating, or — in the case of a forced choice design — a choice indicator that is assigned to that profile.

The respondent's potential outcomes in each task can be written as the vector $Y_{ik}(\bar{\mathbf{T}}_i)$ with profile-level components $Y_{ijk}(\bar{\mathbf{T}}_i)$, which denote the outcome that respondent i would assign 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 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.[1] What this allows us to do is establish that the potential outcome for an individual profile rating Y_{ijk} depends on the treatment assigned for that profile and the unordered set of treatments for the other comparison profiles.

We focus in this paper on a very common version of the conjoint experiment, 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. Formally, this is a design with $J = 2$ where the preferred profile is assigned an outcome of 1 while the non-preferred profile is assigned an outcome of 0. For example, if in task k, respondent i preferred the first profile over the second profile, we would observe $Y_{i1k} = 1$ and $Y_{i2k} = 0$. Since $J = 2$, assumption 2 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. 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, following Hainmueller et al. (2014), we split the treatment associated with profile j, T_{ijk} into 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 $Y_{ijk} = Y_{ijk}(t_{ijkl}, T_{ijk[-l]}, \mathbf{T}_{i[-j]k})$.

We define an individual-level average $\bar{Y}_i()$ of respondent i's profile and task-level potential outcomes

[1] 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).

$Y_{ijk}()$. Under Assumption 2 respondents' potential outcomes are the same across tasks if those tasks have identical profile sets. Therefore, we can simply suppress the k task index. While Assumption 2 also implies that profile order is irrelevant in most tasks, this is not the case in tasks where the two profiles assigned are identical.

Definition 1. *Individual potential outcomes*

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

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, a scenario that we do not examine in this paper since $J = 2$, 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*

$$CE_{il}(t_a, t_b, t, \mathbf{t}) = \bar{Y}_i(t_a, t, \mathbf{t}) - \bar{Y}_i(t_b, t, \mathbf{t}).$$

The component effect represents the change in an individual's response if the lth attribute of the jth profile were set to t_a versus t_b, 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 and profiles which is known by design. This "marginal component effect" (MCE) of setting attribute l to level t_a versus t_b can be

defined for each individual as:

Definition 3. *Marginal component effect*

$$MCE_{il}(t_a, t_b, p(\mathbf{t})) = \sum_{(t,\mathbf{t})\in\mathcal{T}} \left[\bar{Y}_i(t_a, t, \mathbf{t}) - \bar{Y}_i(t_b, t, \mathbf{t})\right] \times p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t} | T_{ijk[-l]}, \mathbf{T}_{i[-j]k} \in \tilde{\mathcal{T}})$$

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

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_a or t_b 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_a and t_b 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 can 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.[2] 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 with the expectation taken over the sample.

Definition 4. *Average marginal component effect*

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

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

[2]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.

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. The implications of this dual averaging for interpreting the AMCE in terms of majority preference in the sample are discussed more extensively in Abramson et al. (2022). For the purposes of this paper, our analysis does not focus on heterogeneity in MCEs across respondents and as such we discuss the AMCE and MCE interchangeably.

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 those effects with some degree of precision using 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.

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

Additionally, under an assumption of complete randomization the AMCE can be estimated by the

simple difference-in-means of Y_{ijk} between profiles with $T_{ijkl} = t_a$ and $T_{ijkl} = t_b$ (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 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

Although the AMCE is defined in terms of cross-task counterfactuals, researchers often use conjoints to make claims about cross-profile comparisons — whether a respondent will select a profile with one feature over another profile with a different feature. In the particular conjoint design we focus on here, the binary forced choice design, we show that the AMCE can be decomposed into an average over cross-profile comparisons. However, this average is not restricted to *only* those tasks comparing the two features of interest, but to all indirect comparisons involving at least one of those features. We show that when attributes contain many levels, the presence of preference cycles may result in AMCEs that misrepresent the underlying pairwise preference between two attributes.

We start this section by operationalizing the concept of a pairwise preference in terms of potential outcomes. First, write the potential outcomes for a binary choice task with $J = 2$ as $Y_{ij}(t_a, t, t_b, t')$. which denotes the choice assigned to profile j when attribute l of profile j is set to level t_a and that

same attribute l of the contrasting profile $-j$ is set to level t_b, holding the remainder of profile j constant at t and profile $-j$ at t'. As above, under Assumptions 1 and 2 we can sum over the profile order to define a **feature choice** $Y_i(t_a, t, t_b, t')$

$$\bar{Y}_i(t_a, t, t_b, t') = \frac{1}{J} \sum_{j=1}^{J} Y_{ij}(t_a, t, t_b, t').$$

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

As with the marginal component effect, we define an individual's **feature choice probability** (FCP) as the feature choice marginalized over the distribution of t and t'.[3] In other words, it is the share of comparisons involving one profile with t_a and another with t_b where the profile with t_a is selected.

Definition 5. *Feature Choice Probability*

$$FCP_{il}(t_a, t_b, p(t, t')) = \sum_{(t,t')} \bar{Y}_i(t_a, t, t_b, t') \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b).$$

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_a to t_b, 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_a over t_b since it can be understood as the proportion of tasks involving t_a and t_b in which t_a is selected.[4]

[3] As we are considering the case with no cross-attribute restrictions, we suppress the common support term in the conditioning set $\tilde{\mathcal{T}}$ that appears in the definition of the MCE.

[4] 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.

As with the MCE, we define an average FCP with the expectation taken over the respondents in the sample that serves as a more feasible target of inference.

Definition 6. *Average Feature Choice Probability*

$$\begin{aligned} AFCP_l(t_a, t_b, p(t,t')) &= E[FCP_{il}(t_a, t_b, p(t,t')] \\ &= \sum_{(t,t')} E[\bar{Y}_i(t_a, t, t_b, t')] \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b) \Bigg]. \end{aligned}$$

The AFCP can be non-parametrically identified under Assumptions 1, 2, and 3 simply by the mean of the observed response Y_{ijk} among tasks where $T_{ijkl} = t_a$ and $T_{i[-j]kl} = t_b$.

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_a against itself is $\text{FCP}_{il}(t_a, t_a) = .5$. Likewise, $\text{FCP}_{il}(t_a, t_b) = 1 - \text{FCP}_{il}(t_b, t_a)$ since choosing a profile with t_a implies *not* choosing the profile with t_b.

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_a and t_b 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_a versus t_b is a function of all feature choice probabilities involving *either* t_a or t_b. 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_a or t_b.

Proposition 1. *MCE decomposition*

Under Assumption 4

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

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 than two features, then the MCE also contains a term that sums the difference in the FCP between each other feature and t_a, t_b. Intuitively, this stems from the fact that the MCE also incorporates all indirect comparisons with tasks that contain only t_a or t_b: $\text{FCP}_{il}(t_a, t_c)$ and $\text{FCP}_{il}(t_b, t_c)$.

Equivalently, the AMCE is a function of AFCPs.

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

Intuitively, the AMCE consists of both the direct pairwise comparisons between t_a and t_b *and* all indirect comparisons from tasks that have only one profile with t_a or t_b. Conversely, the AFCP is comprised of only those tasks involving *both* feature t_a and the other feature t_b. But as the decomposition illustrates, these are not entirely disconnected quantities. The AMCE has a clear descriptive interpretation as an average over AFCPs, most of which are not direct comparisons of t_a and t_b but rather indirect comparisons involving *all of the other* levels in the attribute. As a consequence, the AMCE of feature t_a versus feature t_b may change depending on the number levels a researcher chooses to use for a particular attribute as well as the content of those levels. This is essentially 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. The efficiency gains are sizeable when the number of levels is large. However, this benefit 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.

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, for many if not most attributes in political science conjoint experiments (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 exclusively uses tasks that involve pairwise comparisons between those alternatives. In contrast, the MCE borrows information from tasks with alternatives irrelevant to the comparison at hand. In what follows we present examples of preference cycles at the individual and aggregate levels to highlight how the assumptions built into the marginal component effect 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:

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

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

For an example of an individual preference cycle, consider a respondent who has intransitive preferences about politicians' experience in office. This respondent 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 respondent's preferences over candidate experience exhibit a cycle: she prefers outsiders to veterans, veterans to inexperienced politicians,

and inexperienced politicians to outsiders.

As another example, consider a respondent 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% rates 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% rates, 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 respondent prefers the candidate proposing 30% to 70%, 70% to 50%, and 50% to 30%.

For both of the respondents described above, their preferences exhibit cycles of the form:[5]

$$t_a \succ t_b \succ t_c \succ t_a.$$

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 $\text{FCP}_{il}(t_a, t_b) = \text{FCP}_{il}(t_b, t_c) = \text{FCP}_{il}(t_c, t_a) = 1$. The full set of feature choice probabilities between t_a, t_b, and t_c capture the this respondent's cyclical preferences over the features. When we instead calculate the MCE using Equation (1), we find:

$$\text{MCE}_{il}(t_a, t_b) = \frac{1}{3}\left[2\text{FCP}_{il}(t_a, t_b) - 1 + \text{FCP}_{il}(t_a, t_c) - \text{FCP}_{il}(t_b, t_c)\right] = 0$$

Similar calculations reveal $MCE_i(t_b, t_c) = MCE_i(t_c, t_a) = 0$. A researcher looking at the MCE would then conclude that these voters are indifferent, despite having well-defined, and possibly intense, preferences. This happens because the MCE borrows information from tasks incorporating t_c when making inferences for respondents' preferences between t_a and t_b. By including indirect comparisons, the MCE dilutes information about an individual's direct pairwise preference between features.

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

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 respondent 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 respondent has the same cyclical 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 $\text{FCP}_{il}(D, R) = 5/9$, $\text{FCP}_{il'}(N, V) = 3/4$, $\text{FCP}_{il'}(V, S) = \text{FCP}_{il'}(S, N) = 1$. Thus, the FCP accurately captures that this respondent always selects veterans over candidates with some experience, whom she chooses over newcomers; and that most of the time she chooses Democrats over Republicans and newcomers over veterans.

When we calculate the MCEs, we get $MCE_{il}(D, R) = 1/9$, $MCE_{il'}(N, V) = -1/6$, and $MCE_{il'}(V, S) = MCE_{il'}(S, N) = 1/12$. 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 individual preferences are separable and 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 individual preferences satisfy these assumptions. Before proving this result, let us first formally define the following.

Definition 7. *Full support*

A set of profiles has full support if whenever profiles (t_a, t) and (t_b, t') are included in the randomization distribution, so do profiles (t_a, t') and (t_b, t), for all t_a, t_b, t, and t'.

This assumption states that there are no combinations of levels from the attribute of interest and other attributes that are removed from a task, for example for being unrealistic. This assumption is satisfied by default in a completely randomized design.

Definition 8. *Separability*

Respondent i's preferences are separable when for all t_a and t_b, we have $Y_i(t_a, t, t_b, t) = Y_i(t_a, t', t_b, t')$ for all t and t'.

In other words, when respondents face two profiles that are otherwise identical except for differing in the attribute of interest, they will always either choose the profile with feature t_a or t_b. This restricts the presence of interactions between attributes: the preference for feature t_a over feature t_b does not depend on the values of the other attributes.

Definition 9. *Connexity*

Respondent i's preferences are connected if for any two distinct profiles (t_a, t) and (t_b, t') with $t_a \neq t_b$ or $t \neq t'$, we have either $(t_a, t) \succ (t_b, t')$ or $(t_b, t') \succ (t_a, t)$; or equivalently, $Y_i(t_a, t, t_b, t') = 1$ or $Y_i(t_b, t', t_a, t) = 1$ for all t_a, t_b, t, and t'.

Definition 10. *Asymmetry*

Respondent i's preferences are asymmetric if for any two profiles (t_a, t) and (t_b, t'), $(t_a, t) \succ (t_b, t') \implies (t_b, t') \not\succ (t_a, t)$; or equivalently, $Y_i(t_a, t, t_b, t') = 1 \iff Y_i(t_b, t', t_a, t) = 0$ for all t_a, t_b, t, and t'.

Definition 11. *Transitivity*

Voter i's preferences are transitive if for any profiles (t_a, t), (t_b, t'), and (t_c, t''), we have that $(t_a, t) \succ (t_b, t')$ and $(t_b, t') \succ (t_c, t'')$ implies $(t_a, t) \succ (t_c, t'')$; or equivalently, $Y_i(t_a, t, t_b, t') = 1$ and $Y_i(t_b, t', t_c, t'') = 1 \implies Y_i(t_a, t, t_c, t'') = 1$ for all t_a, t_b, t_c, t, t', and t''.

These definitions are helpful in defining a version of transitivity in the context of FCP's, which we then use to connect FCP and MCE:

Definition 12. *FCP-Transitivity*

Voter i's preferences are FCP-transitive if for all t_a, t_b, and t_c, $FCP_i(t_a, t_b) > FCP_i(t_b, t_a) \implies FCP_i(t_a, t_c) \geq FCP_i(t_b, t_c)$.

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

Proposition 2. *When the set of profiles has full support, and preferences over profiles are connected, asymmetric, transitive, and separable, FCP-transitivity is satisfied.*

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

Proposition 3. *When preferences are FCP-transitive, the MCE is positive if and only if the corresponding FCP is greater than one half.*

Even when we assume homogeneity across respondents, we still need additional assumptions on their behavior in order to interpret the MCEs as direct preferences for one feature over another. While the above assumptions may be reasonable at the individual level, similar assumptions at the aggregate level may be more difficult to sustain when preferences over features are heterogeneous. Here we consider the possibility of 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.

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 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 1 prefers Black to Latino to white candidates, or $B \succ_1 H \succ_1 W$. Voter 2 prefers Latino to white to Black candidates, so that $H \succ_2 W \succ_2 B$. Finally, suppose voter 3 prefers white candidates the most and Latino candidates the least: $W \succ_3 B \succ_3 H$. 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 AFCP captures this cycle: We have $\text{FCP}_1(H, W) = \text{FCP}_2(H, W) = 1$ and $\text{FCP}_3(H, W) = 0$.[6] Aggregated, these reveal $\text{AFCP}(H, W) = 2/3$, precisely the vote share of a Latino candidate facing a white candidate. The same holds for $\text{AFCP}(B, H)$ and $\text{AFCP}(W, B)$.

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

$$
\begin{aligned}
MCE_1(H, W) &= \frac{1}{3}\left[\text{FCP}_1(H, W) + \text{FCP}_1(H, B) - \text{FCP}_1(W, H) - \text{FCP}_1(W, B)\right] = \frac{1}{3} \\
MCE_2(H, W) &= \frac{1}{3}\left[\text{FCP}_2(H, W) + \text{FCP}_2(H, B) - \text{FCP}_2(W, H) - \text{FCP}_2(W, B)\right] = \frac{1}{3} \\
MCE_3(H, W) &= \frac{1}{3}\left[\text{FCP}_3(H, W) + \text{FCP}_3(H, B) - \text{FCP}_3(W, H) - \text{FCP}_3(W, B)\right] = -\frac{2}{3}
\end{aligned}
$$

Thus, when we calculate the AMCE, we get that

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

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 1, 2, and 3 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.[7]

[6] Here with a single attribute, we suppress the l notation denoting the attribute of interest.

[7] Reversals are straightforward to construct. For example, consider the same example but with $\text{FCP}_1(B, W) = 0.9$, that is, voter 1 votes for a white candidate over a Black candidate 10% of the time. Then, the AMCE of Latino when

While it is often reasonable to assume that individuals exhibit transitive and separable 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. In the next section we outline a procedure for researchers to estimate both sets of effects and conduct hypothesis tests on violations of transitivity.

4 Estimation

As with the AMCE, there exists a straightforward, unbiased estimator of $\text{AFCP}_l(t_a, t_b)$ which — in the case of a completely randomized design — is equal to the sample average of Y_{ijkl} among those tasks with $t_{ijkl} = t_a$ and $t_{i-jkl} = t_b$. In other words, we can estimate the AFCP of feature t_a versus feature t_b using the share of pairwise tasks involving a profile with feature t_a and a profile with feature t_b in which the profile with feature t_a is selected.

Proposition 4. *Estimation of the AFCP under complete randomization*

Under Assumptions 1, 2, 3, and 4, and when $J = 2$, the sample mean estimator $\hat{\pi}(t_a, t_b, p(\boldsymbol{t}))$ is unbiased for the AFCP

$$\widehat{AFCP}_l(t_a, t_b, p(\boldsymbol{t})) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} Y_{ijk} \mathbf{1}(T_{ijkl} = t_a, T_{i-jkl} = t_b)}{n_{ab}}$$

where n_{ab} is the number of tasks where $T_{ijkl} = t_a, T_{i-jkl} = t_b$)

The proof is straightforward. See Appendix A.

Comparing this estimator with the difference-in-means estimator from Proposition 3 in Hainmueller et al. (2014), it is straightforward to see that the AMCE estimator generally has lower variance due to its smaller sample size. While the AMCE estimator uses any task involving a profile

white is used as a benchmark is $-1/30$, despite the fact that Latinos always defeat whites.

with *either* feature t_a or feature t_b, the unbiased AFCP estimator uses only those tasks involving *both* features t_a and t_b. The variance benefits grow as we consider designs with attributes that have many levels. As the number of levels grows extremely large, there may be zero tasks in-sample where attribute level t_a is directly compared to attribute level t_b, resulting in an undefined sample mean estimator. In such a situation, the AMCE could still be estimated as long as there are some tasks involving *either* t_a or t_b.

While the AMCE estimator may be more precise than the corresponding AFCP estimator, the AMCE *estimand* may be less appropriate than the AFCP for characterizing the preferences for one feature against another. How should researchers resolve this trade-off, especially after a conjoint experiment has been run? In particular, how should researchers reconcile a divergence between an estimated AMCE and a higher-variance AFCP?

First, we propose a statistical test for the divergence of the direct and indirect preference comparisons. Our testing strategy permits researchers to test for any divergence between the direct AFCP between features a and b and the indirect comparison via the difference in AFCPs involving c. This involves a joint test for a combined restriction on the three parameters: $\text{AFCP}_l(t_a, t_b)$, $\text{AFCP}_l(t_b, t_c)$, and $\text{AFCP}_l(t_a, t_c)$. Under the null hypothesis where the indirect and direct preferences are equivalent, the use of the AMCE as a stand-in for the preference for one feature over another is unproblematic. However, when these quantities diverge, the interpretation of the AMCE becomes slightly more complicated. In the extreme case where we observe a different sign for an indirect comparison relative to the direct comparison, a researcher should be very concerned about the possibility of preference cycling. But even in a situation where the direct and indirect comparisons are of the same sign, a difference in magnitude has substantial consequences for interpretation.

The hypothesis test that we propose for assessing violations of transitivity tests the null that the direct and indirect AFCP comparisons are equivalent against the unrestricted alternative. Specifically, our null hypothesis assumes

Assumption 5. *Equivalence of direct and indirect AFCP comparisons*

For any three features t_a, t_b, and t_c

$$AFCP_l(t_a, t_b) - \frac{1}{2} = AFCP_l(t_a, t_c) - AFCP_l(t_b, t_c).$$

As with the AMCE, a standard nonparametric estimator for the three AFCPs can be implemented via a regression model. First subset the data to only those tasks involving a direct comparison of t_a versus t_b or a comparison involving t_c and either t_a or t_b. Then, define a treatment indicator T_{ikl} as an unordered pair two levels of l assigned for the task (t_q, t'_q). The task-level outcome variable, $\tilde{Y}_{ik}$, takes on a value of 1 when the profile with the first of the two assigned features in T_{ikl} is selected and a 0 if the profile with the second feature is selected in that task. Finally, estimate a fully saturated regression model:

$$E[\tilde{Y}_{ik}] = \beta_0 + \beta_1 \mathcal{I}[T_{ikl} = (t_a, t_c)] + \beta_2 \mathcal{I}[T_{ikl} = (t_b, t_c)]$$

where $\mathcal{I}(T_{ikl} = (t_q, t_{q'}))$ denotes an indicator for whether attribute l of task k takes on level q in one profile and level q' in the other.

In this setup, $\hat{\beta}_0$ corresponds to our estimate of $\text{AFCP}_l(t_a, t_b)$, while the other coefficients denote the differences between an AFCP and the baseline $\text{AFCP}_l(t_a, t_b)$. Obtaining the variance-covariance matrix via standard cluster-robust methods allows us to generate a procedure for carrying out hypothesis tests for the presence of FCP-transitivity violations.

In order to test for a violation of transitivity, the AFCP constraint can be re-written in terms of the regression parameters.

$$\beta_0 - \beta_1 + \beta_2 = \frac{1}{2}. \tag{2}$$

This allows us to conduct a standard Wald or likelihood ratio hypothesis test under a single constraint on the regression parameters.[8] It also generalizes easily to test for multiple constraints on all indirect FCP comparisons involving features t_a and t_b. This would require a total of $D_l - 2$ constraints where D_l is the total number of levels in the attribute of interest l — one constraint for each of the levels not included in the main AFCP of interest. The fully saturated regression model estimated on all observations containing either feature t_a or feature t_b contains $2(D_l - 2) + 1$

[8] An alternative approach would be to instead assume a set null that the signs of $\beta_0 - \frac{1}{2}$ and $\beta_1 - \beta_2$ are equivalent and test the implied nonlinear inequality constraints. Although there exists an extensive literature on hypothesis tests of multiple inequalities in regression models (Gourieroux et al., 1982; Wolak, 1987, 1989), as noted in Wolak (1991), the problem of testing non-linear inequalities is particularly challenging due to the difficulty of finding the least-favorable set of parameters under the null to compute critical values. Additionally, we consider the case where the direct $\beta_0 - \frac{1}{2}$ and indirect $\beta_1 - \beta_2$ significantly diverge while still having the same sign to still be of interest to researchers diagnosing their AMCE estimates as it provides additional information on the precise sets of comparisons that contribute to the AMCE.

parameters.

$$
\begin{aligned}
E[\tilde{Y}_{ik}] = \beta_0 &+ \beta_1 \mathcal{I}[T_{ikl} = (t_a, t_c)] + \beta_2 \mathcal{I}[T_{ikl} = (t_a, t_d)] + \dots \\
&+ \beta_{D_l-2} \mathcal{I}[T_{ikl} = (t_a, t_L)] + \beta_{D_l-1} \mathcal{I}[T_{ikl} = (t_b, t_c)] + \beta_{D_l} \mathcal{I}[T_{ikl} = (t_b, t_d)] + \dots \\
&+ \beta_{2(D_l-2)} \mathcal{I}[T_{ikl} = (t_b, t_L)].
\end{aligned}
$$

Under this formulation of the regression, we have $L - 2$ constraints on each indirect comparison of features a and b. For $k = 1, 2, \dots, L - 2$

$$\beta_0 - \beta_k + \beta_{k-L+2} = \frac{1}{2}.$$

We collect these constraints into the form of a $D_l - 2$ x $2(D_l - 2) + 1$ matrix R and write the null hypothesis that all constraints hold as $R\beta = \frac{1}{2}$. The standard Wald test statistic (Wald, 1943) is

$$\left(R\hat{\beta} - \frac{1}{2}\right)' \left(R\hat{\Sigma}R'\right)^{-1} \left(R\hat{\beta} - \frac{1}{2}\right)'$$

where R is the $D_l - 2$ x $2(D_l - 2) + 1$ matrix of constraints on the regression parameters and $\hat{\Sigma}$ is a consistent estimate of the variance-covariance matrix of $\hat{\beta}$, estimated using standard cluster-robust approaches. Asymptotically, the test statistic has a χ^2 distribution with degrees of freedom equal to the number of constraints (1 for the test of a single indirect comparison and $D_l - 2$ for testing all indirect comparisons).

Second, researchers may wish to generally know how sensitive their AMCE estimates are to the choice of which levels are included as part of the indirect comparisons. If there was only a single feature for which the indirect preferences were believed to diverge significantly from the direct preferences, then a straightforward fix would be to remove those tasks containing this feature and re-estimate the AMCE.[9] This may make Assumption 5 more well-justified for the remaining features while avoiding some of the loss of power resulting from estimating only $\text{AFCP}_l(t_a, t_b)$.

For an attribute with D_l levels, there are 2^{D_l-2} possible AMCEs arising from different combinations of held-out levels. This can be seen as a continuum from the lowest-variance AMCE which uses all

[9]We note that this otherwise should not change the randomization distribution of the other attributes under designs with complete randomization. However, if there are additional cross-attribute randomization restrictions involving the removed feature, this approach will also affect the randomization distribution of any linked features. In our applications in section 5, all analyzed experiments use complete randomization.

levels and indirect comparisons to the highest-variance AMCE which leaves out all but the two features of interest t_a and t_b and relies only on direct comparisons. If all of these estimates remain roughly the same, then that is at least suggestive evidence that the AMCE is not sensitive to the problem of indirect comparisons. Conversely, if researchers observe substantial swings in the magnitude and direction of the estimated AMCE depending on which indirect comparisons are included, that suggests a need to reexamine the underlying quantity of interest. We illustrate this approach using a graphical diagnostic, which we apply to multiple conjoint experiments in Section 5, that arranges all possible "held-out" AMCE estimates in their order of magnitude and plots each estimate and corresponding confidence interval.

Finally, if a researcher is willing to make the assumptions necessary to interpret the AMCE as a direct preference — namely Assumption 5 — then it is possible to obtain a more precise estimator of the AFCP by incorporating the additional moment constraints implied by the indirect comparisons using a generalized method of moments (GMM) estimator (L. P. Hansen, 1982). GMM estimators specify an estimator of a parameter using terms of a series of moment conditions. The simple sample mean estimator in Proposition 4 can be interpreted as a "just-identified" method of moments estimator with the following moment constraint:

$$\text{AFCP}_l(t_a, t_b) - \frac{1}{N_{ab}} \sum_{i=1}^{N} \sum_{j,k:T_{ijkl}=t_a, T_{i-jkl}=t_b} Y_{ijkl} = 0$$

where N_{ab} is the number of tasks involving a comparison between feature t_a and feature t_b.

Our proposed generalized method of moments estimator augments the just-identified estimator with a set of $D_l - 2$ additional moment conditions for each indirect comparison

$$\text{AFCP}_l(t_a, t_b) - \frac{1}{2} - \frac{1}{N_{ac}} \sum_{i=1}^{N} \sum_{j,k:T_{ijkl}=t_a, T_{i[-j]kl}=t_c} Y_{ijkl} + \frac{1}{N_{bc}} \sum_{i=1}^{N} \sum_{j,k:T_{ijkl}=t_b, T_{i[-j]kl}=t_c} Y_{ijkl} = 0$$

We follow the standard framework for efficient estimation of overidentified GMM from L. P. Hansen (1982) with the caveat that our observations are not independent and identically distributed but rather clustered by respondents. To address this, we use the cluster-robust variance estimator for two-step GMM proposed by B. E. Hansen and Lee (2019).

5 Application

In this section, we reanalyze data from three recently published conjoint experiments in order to demonstrate when preference cycling, and thus AMCE-AFCP divergence, is likely to occur. Our analysis highlights three conditions under which researchers ought to be especially wary of transitivity violations: when preferences are not single-peaked, either because the attribute lacks a natural ordering or because single-peakedness is violated *despite* there being a natural ordering; when preferences over attributes are multidimensional; and when there are multiple attribute-features that are similar to one another. In what follows, we walk through some applications of our diagnostic tools, illustrating each type of FCP-transitivity violation and how it can be detected in data.

5.1 Single-peaked preferences cannot be assumed

As discussed in Section 3, aggregate preferences may exhibit cycles when individual preferences fail to satisfy single-peakedness. By far the most common failure of single-peakedness that arises in conjoint experiments occurs when attributes have no natural ordering, as is the case for race, profession, or country of origin. We begin with an illustration of this type of violation using country of origin in Hainmueller and Hopkins (2015). In this experiment, American survey respondents were asked to compare profiles of hypothetical immigrants and select the ones they would prefer to admit to the United States. We highlight a number of FCP-transitivity violations around the country of origin values of France and Poland, which drive the AMCE of France over Poland to zero despite an average preference for Poland over France in direct comparisons.

In the first column of Table 1, we compute the AFCP of France against each other alternative (other than Poland) as given by Proposition 4, and in the second column we do the same for Poland. Our reported AFCPs are recentered around 0 — that is, we subtract 0.5 from all estimates — to aid in interpretability and comparability to the AMCE. The third column presents the difference between the first two columns. We can then compare this quantity to the direct comparison, AFCP(France, Poland), given in the fourth column. Any statistically significant divergence between the third and fourth columns constitutes evidence of a transitivity violation. The final column presents the p-value from our statistical test of such a violation — the equivalence test of the direct and indirect comparisons given in Equation 2. We find three statistically significant FCP-transitivity violations: France and Poland with, respectively, Germany ($p = 0.02$), Sudan ($p = 0.01$), and Somalia ($p = 0.00$).

Table 1: Centered AFCPs for Direct and Indirect Comparisons for Country of Origin in Hainmueller and Hopkins (2015)
a = France
b = Poland

	AMCE(a, b)	Centered AFCP(a, b) sample mean	Centered AFCP(a, b) GMM
Estimate	-0.015	-0.115	-0.020
CRSE	(0.020)	(0.043)	(0.028)
p-value	(0.46)	(0.008)	(0.49)

	Indirect Comparisons			**Direct Comparison**	
c	$AFCP(a, c)$	$AFCP(b, c)$	$AFCP(a, c) - AFCP(b, c)$	$AFCP(a, b)$	p-value
Philippines	-0.02	0.04	-0.06	-0.11	0.45
India	-0.00	0.05	-0.05	-0.11	0.40
China	0.01	0.02	-0.01	-0.11	0.19
Mexico	-0.03	-0.03	0.00	-0.11	0.12
Iraq	0.18	0.17	0.01	-0.11	0.15
Sudan	0.09	0.03	0.06	-0.11	0.04
Germany	0.00	-0.06	0.06	-0.11	0.02
Somalia	0.13	-0.00	0.13	-0.11	0.00
				p-value from joint test of all restrictions:	0.08

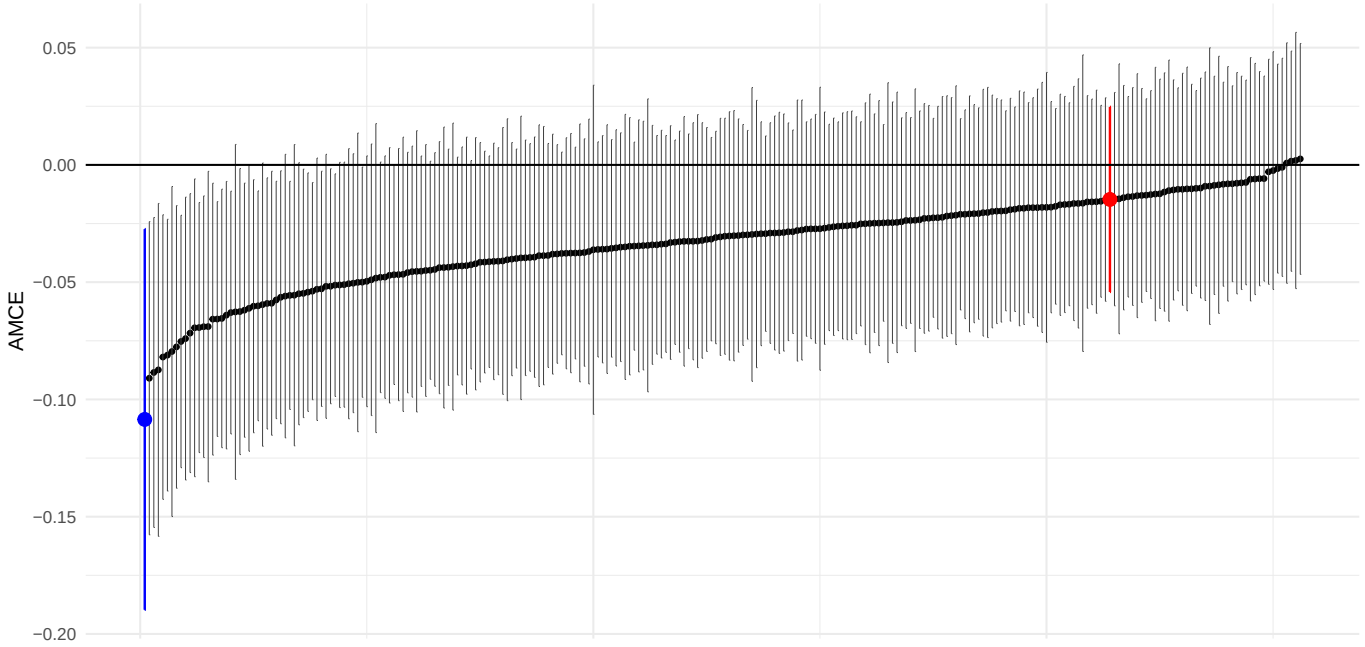
Notes: This table presents AFCP estimates centered around zero, computed by subtracting 0.5 from the AFCP estimate. p-value column presents the p-value associated with the Wald statistic from the three-level likelihood ratio hypothesis test given in Equation 2, for levels a=France, b=Poland, and c as given in each row of the table. First table presents estimates and standard errors for the AMCE and our two suggested AFCP estimators.

In each case, France performs better against the third alternative than does Poland, even though France loses to Poland in a direct comparison. Accordingly, the AMCE of France vs. Poland is close to 0, whereas the direct comparison clearly favors Poland.

As a test of the AMCE's sensitivity to which levels are included in the indirect comparisons, Figure 1 presents our graphical diagnostic that plots all possible "held-out" AMCE estimates as described in Section 4 above. First, we generate all possible combinations of left-out features, from zero (giving us the AMCE reported in the original study) through all eight other countries besides France and Poland (giving us the AFCP). For a given set of left-out countries, we eliminate from the data any *questions* (not just profiles) that include any of those countries and reestimate the AMCE.[10] We order

[10]Throughout, we also drop any comparisons involving a profile assigned to "facing persecution" for the reason for immigration. This is because the randomization scheme has a cross-attribute restriction where immigrants facing persecution are only allowed to be from Sudan, Somalia, Iraq, and China. Failing to account for this dependence across attributes results in a partial aliasing of the country of origin treatments with the reason for immigration, as a

Figure 1: AMCE Estimates for France vs. Poland for All Possible Combinations of Left-Out Attribute Levels, Country of Origin in Hainmueller and Hopkins (2015)



Notes: Blue point estimate indicates AFCP (equivalent to AMCE with all other attribute levels left out). Red point estimate indicates AMCE in the original study (with no other attribute levels left out). 95% confidence intervals are shown in gray.

and plot these AMCE estimates with their associated confidence intervals, highlighting the AMCE (no left-out features) in red and the AFCP (all other features left out) in blue. Due to the amount of data used, the AFCP has one of the largest confidence intervals of the estimates, while the AMCE has one of the smallest.

Our analysis yields some cause for concern: it is not only the AFCP that diverges from the AMCE, but a small (but not insubstantial) range of left-out estimates. Figure 1 also shows that while the AMCE is lower-variance than the AFCP, it is nonetheless possible for the AFCP to detect a significant effect where the AMCE does not, due to the AMCE's averaging over preference cycles.

Even attributes that do have a natural ordering, such as age, may still fail to satisfy single-peakedness. We have already proposed one hypothetical but entirely plausible example of such a violation at the individual level: a voter who prefers both young and old candidates to middle-aged ones. Reanalyzing an experiment targeting the political preferences of Tunisian voters by Blackman and Jackson (2021), we highlight an aggregate FCP-transitivity violation for candidate age:

profile with an immigrant from Sudan, Somalia, Iraq or China has a different distribution of the reason for immigration attribute compared to an immigrant from any other country. However, conditional on dropping these comparisons, the experiment is essentially completely randomized.

respondents are nearly indifferent between 30-year-old and 60-year-old candidates, but 30-year-olds are significantly favored to 60-year-olds in indirect comparisons.

In Table 2, we can see an FCP-transitivity violation that is not an aggregate cycle of the form $a \succ b \succ c \succ a$, but that is nonetheless problematic, and that can nonetheless drive a wedge between direct and indirect preferences. Here, respondents are nearly indifferent between candidates aged 30 and 60: the AFCP of 30 vs. 60 is 0.04 ($p = 0.43$). But respondents *significantly* favor both 50-year-olds and 40-year-olds over 60-year-olds, and also favor 30-year-olds over both categories, such that FCP-transitive preferences ought to exhibit a much bigger positive preference for 30 over 60 in the direct comparison.

Table 2: Centered AFCPs for Direct and Indirect Comparisons for Age in Blackman and Jackson (2021)
$a = 30$
$b = 60$

	AMCE(a, b)	Centered AFCP(a, b) sample mean	Centered AFCP(a, b) GMM
Estimate	0.11	0.037	0.11
CRSE	(0.031)	(0.047)	(0.039)
p-value	(0.00052)	(0.43)	(0.0040)

	Indirect Comparisons			**Direct Comparison**	
c	$AFCP(a, c)$	$AFCP(b, c)$	$AFCP(a, c) - AFCP(b, c)$	$AFCP(a, b)$	p-value
70	0.09	0.05	0.04	0.04	0.93
50	0.07	-0.14	0.20	0.04	0.04
40	0.05	-0.17	0.22	0.04	0.02
			p-value from joint test of all restrictions: 0.03		

Notes: This table presents AFCP estimates centered around zero, computed by subtracting 0.5 from the AFCP estimate. p-value column presents the p-value associated with the Wald statistic from the three-level likelihood ratio hypothesis test given in Equation 2, for levels a=30 years old, b=60 years old, and c as given in each row of the table. AMCE(a, b) is 0.11 ($p = 0.00$).

The diagnostic plot shown in Appendix Figure B1 again reveals some sensitivity of the AMCE to left-out attribute values — this time, with the full AMCE reflecting a preference that is not present in a head-to-head comparison of 30-year-old to 60-year-old candidates. We note, further, that it is possible for some feature pairs to exhibit this problematic behavior while others do not: Figure B2 presents an example of the same diagnostic plot for age 30 vs. 70, in which the AMCE estimate is remarkably stable (and close to the AFCP) across all left-out feature combinations. We

therefore encourage researchers to check *every* pair of features for such sensitivities before concluding respondents have transitive preferences over an attribute.

5.2 Respondents have multidimensional preferences

Table 3: Centered AFCPs for Direct and Indirect Comparisons, Party in Blackman and Jackson (2021)
a = Current of Love
b = Popular Front

	AMCE(a,b)	Centered AFCP(a,b) sample mean	Centered AFCP(a,b) GMM
Estimate	0.049	-0.086	0.044
CRSE	(0.043)	(0.086)	(0.057)
p-value	(0.25)	(0.312)	(0.436)

	Indirect Comparisons			**Direct Comparison**	
c	$AFCP(a,c)$	$AFCP(b,c)$	$AFCP(a,c) - AFCP(b,c)$	$AFCP(a,b)$	p-value
Democratic Current	-0.02	0.17	-0.18	-0.09	0.55
Initiative Party	-0.19	-0.12	-0.07	-0.09	0.92
Afek Tounes	-0.10	-0.14	0.04	-0.09	0.43
Nidaa Tounes	0.08	0.00	0.08	-0.09	0.29
UPL	0.05	-0.06	0.11	-0.09	0.25
Machrouu Tounes	0.00	-0.25	0.25	-0.09	0.05
Ennahda	0.24	-0.06	0.30	-0.09	0.01
al-Irada Movement	0.20	-0.11	0.30	-0.09	0.02
				p-value from joint test of all restrictions:	0.03

Notes: This table presents AFCP estimates centered around zero, computed by subtracting 0.5 from the AFCP estimate. p-value column presents the p-value associated with the Wald statistic from the three-level likelihood ratio hypothesis test given in Equation 2, for levels a=Current of Love, b=Popular Front, and c as given in each row of the table. First table presents estimates and standard errors for the AMCE and our two suggested AFCP estimators.

Conjoint designs that seek to elicit political preferences must often contend with parties or positions that are arranged along a multidimensional spectrum — fertile ground for transitivity violations. Looking once more at Blackman and Jackson's (2021) experiment, we highlight a number of aggregate cycles in voters' preferences over Tunisian parties in Table 3. Whereas the Popular Front is preferred to the Current of Love party in direct comparisons, Current of Love tends to perform better than Popular Front against most third alternatives, with three of these (Machrouu Tounes, Ennahda, and al-Irada Movement) being statistically significant at $p < .05$. The diagnostic plot in Appendix Figure

B3 shows that the AMCE for Current of Love vs. Popular Front is somewhat sensitive to the other included features, though most estimates are not statistically distinguishable from zero.

5.3 The existence of similar attribute-features

Researchers often design experiments that include many attribute-features, as presenting respondents with a variety of candidates is seen to increase the realism and external validity of the experiment. However, with many possible attribute values on the table, there will often be categories that are difficult for respondents to distinguish from one another. This situation also holds potential for FCP-transitivity violations. In particular, while respondents may be indifferent or near-indifferent between two features when presented with them in a direct comparison, the same features may perform differently against other features, driving AMCE-AFCP divergence.

Table 4: Centered AFCPs for Direct and Indirect Comparisons for Profession in Clayton et al. (2021)
a = Financial analyst
b = Professor

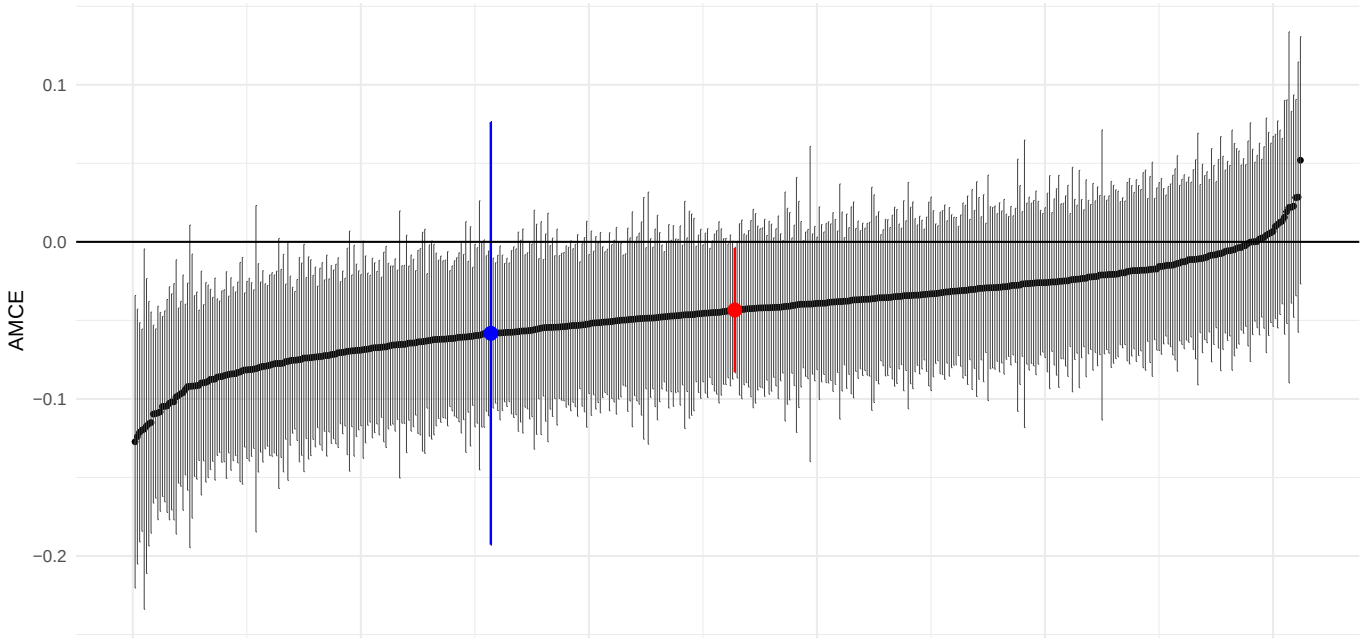
	AMCE(a,b)	Centered AFCP(a,b) sample mean	Centered AFCP(a,b) GMM
Estimate	-0.043	-0.06	-0.052
CRSE	(0.020)	(0.071)	(0.029)
p-value	(0.032)	(0.40)	(0.075)

	Indirect Comparisons			**Direct Comparison**	
c	$AFCP(a,c)$	$AFCP(b,c)$	$AFCP(a,c) - AFCP(b,c)$	$AFCP(a,b)$	p-value
Computer programmer	-0.17	0.10	-0.27	-0.06	0.10
Doctor	-0.12	0.02	-0.14	-0.06	0.47
Nurse	-0.07	0.04	-0.11	-0.06	0.62
Child care provider	-0.04	0.07	-0.11	-0.06	0.56
Janitor	0.05	0.15	-0.10	-0.06	0.67
Construction worker	0.02	0.05	-0.03	-0.06	0.73
Gardener	0.06	0.09	-0.02	-0.06	0.67
Waiter	0.10	0.02	0.08	-0.06	0.11
Research scientist	0.07	-0.12	0.19	-0.06	0.04
			p-value from joint test of all restrictions:		0.02

Notes: This table presents AFCP estimates centered around zero, computed by subtracting 0.5 from the AFCP estimate. p-value column presents the p-value associated with the Wald statistic from the three-level likelihood ratio hypothesis test given in Equation 2, for levels a=Financial analyst, b=Professor, and c as given in each row of the table. First table presents estimates and standard errors for the AMCE and our two suggested AFCP estimators.

In our final application, we analyze such a scenario for profession in Clayton et al. (2021), an experiment eliciting preferences for immigrants to France. In Table 4, we focus on the comparison between two professions, financial analyst and professor — both high-skilled, somewhat technical white-collar jobs. Although the AFCP of financial analyst vs. professor is an imprecisely estimated -0.06 ($p = 0.40$), there is a strong, statistically significant *positive* difference between the financial analyst's and the professor's performance against a research scientist ($p = 0.04$). As shown in Figure 2, this intransitivity generates a wide range of possible AMCE estimates depending on which features are left out (though the AFCP and AMCE themselves do not dramatically diverge).

Figure 2: AMCE Estimates for Financial Analyst vs. Professor for All Possible Combinations of Left-Out Attribute Levels, Profession in Clayton et al. (2021)



Notes: Blue point estimate indicates AFCP (equivalent to AMCE with all other attribute levels left out). Red point estimate indicates AMCE in the original study (with no other attribute levels left out). 95% confidence intervals are shown in gray.

6 Conclusion

Preference elicitation via conjoint analysis in political science has, to this point, largely focused on the recovery of average treatment effects. However, the mapping from these effects to statements about aggregate preferences is not straightforward. We have shown that the AMCE, the estimand focused on by most political scientists, requires transitive and single-peaked preferences when there are more

than two attribute levels. For many unordered and multi-valued attributes, these are potentially very strong assumptions. If they fail, Condorcet-like preference cycles may exist, rendering AMCE estimates largely uninterpretable.

The problem of preference cycles in conjoint experiments occurs because the AMCE violates the independence axiom while forcing estimates into a linear order. That is, it incorporates information about irrelevant attributes, mechanically producing a transitive ordering of attributes even when preferences may be cyclic. We have characterized an alternative quantity, the AFCP, that remedies this problem by only considering pairwise comparisons of features. However, since the AFCP exploits fewer comparisons, it will be less precisely estimated. As a consequence, comparisons of the AFCP and AMCE may be driven by preference cycling or by statistical noise.

We have proposed adapting a conventional statistical test of equality constraints on linear regression coefficients to formally evaluate the potential for preference cycles in conjoint data. We have shown in several conjoint experiments the presence of cycles. Moreover, we highlight how the AMCE masks them by incorporating indirect, irrelevant, comparisons.

As a point of intellectual history, the relationship we highlight between the AMCE and the AFCP parallels a debate between Jean-Charles de Borda and the Marquis de Condorcet. As established by Abramson et al. (2022), the AMCE corresponds to a perturbation of the Borda Rule (or ranked-choice voting). Similar to the Borda Rule, the AMCE assigns scores to features by counting the number of alternatives they are preferred to. Like the AMCE, the AFCP can also be viewed a preference aggregation rule. And like the AMCE and Borda Rule, there are also strong similarities between the AFCP and another well-known social choice rule: the Condorcet Method — or majority rule. The Condorcet Method focuses on pairwise head-to-head comparisons between each pair of alternatives. Because of this, it yields a straightforward interpretation as a majority preference.

The Marquis de Condorcet and Jean-Charles de Borda — contemporaries and arch-rivals — proposed their methods as electoral rules that would choose among alternatives. This is straightforward for the Borda Rule: the option with the highest Borda score wins.[11] Moreover, because each option's Borda score is a natural number — and because natural numbers are ordered — the Borda rule produces a transitive social preference ranking of the set of alternatives. For the Condorcet Method, matters are, potentially, more complicated because an alternative that is majority preferred to every

[11] Of course, ties are possible for almost any choice rule including Borda, but they become increasingly unlikely as the number of voters grow. Like most of the literature, we ignore the possibility of ties.

other alternative — a "Condorcet winner" — simply may not exist. The pairwise comparisons may produce cycles, meaning that the social preferences it produces are not transitive. The Condorcet Method is thus said to not be *decisive*.

Subsequently, others have modified the Condorcet Method to ensure decisiveness. Duncan Black, for example, proposed that the Condorcet winner should be picked if one exists, otherwise, the Borda winner can be picked instead — forming a bridge between the two choice rules. The approach we have put forward in this paper echoes Black's rule but in the converse. Using our tools researchers can estimate the relevant AFCPs and test for cycles. If none exist — and transitivity is likely to maintain in their data — they should then feel more comfortable presenting results based upon the AMCE, taking advantage of greater statistical power.

References

Abramson, S. F., Kocak, K., & Magazinnik, A. (2022). What do we learn about voter preferences from conjoint experiments? *American Journal of Political Science*. https://doi.org/https://doi.org/10.1111/ajps.12714

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.

Arias, S. B., & Blair, C. W. (2022). Changing tides: Public attitudes on climate migration. *The Journal of Politics*, *84*(1), 560–567.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of political economy*, *58*(4), 328–346.

Arrow, K. J. (1963). *Social choice and individual values*. Yale University Press.

Ashworth, S., Berry, C. R., & de Mesquita, E. B. (2021). *Theory and credibility: Integrating theoretical and empirical social science*. Princeton University Press.

Austen-Smith, D., & Banks, J. S. (1998). Social choice theory, game theory, and positive political theory. *Annual Review of Political Science*, *1*(1), 259–287.

Bansak, K., Hainmueller, J., Hopkins, D. J., Yamamoto, T., Druckman, J. N., & Green, D. P. (2021). Conjoint survey experiments. *Advances in Experimental Political Science*, *19*.

Berinsky, A. J. (2017). Measuring public opinion with surveys. *Annual Review of Political Science*, *20*, 309–329.

Black, D. (1948). On the rationale of group decision-making. *Journal of Political Economy*, *56*(1), 23–34. Retrieved September 6, 2022, from http://www.jstor.org/stable/1825026

Blackman, A. D., & Jackson, M. (2021). Gender stereotypes, political leadership, and voting behavior in tunisia. *Political Behavior*, *43*, 1037–1066.

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

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica: journal of the Econometric Society*, 422–443.

Clayton, K., Ferwerda, J., & Horiuchi, Y. (2021). Exposure to immigration and admission preferences: Evidence from france. *Political Behavior*, *43*, 175–200.

De Mesquita, E. B., & Tyson, S. A. (2020). The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior. *American Political Science Review*, *114*(2), 375–391.

Denney, S., & Green, C. (2021). Who should be admitted? conjoint analysis of south korean attitudes toward immigrants. *Ethnicities*, *21*(1), 120–145.

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

Gourieroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica: journal of the Econometric Society*, 63–80.

Hainmueller, J., & Hopkins, D. J. (2015). The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. *American Journal of Political Science*, *59*(3), 529–548.

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.

Hansen, B. E., & Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of econometrics*, *210*(2), 268–290.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 1029–1054.

Izzo, F., Dewan, T., & Wolton, S. (2018). Cumulative knowledge in the social sciences: The case of improving voters' information. *Available at SSRN 3239047*.

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

Leeper, T. J., Hobolt, S. B., & Tilley, J. (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis*, *28*(2), 207–221.

McKelvey, R. D. (1979). General conditions for global intransitivities in formal voting models. *Econometrica: journal of the Econometric Society*, 1085–1112.

McKelvey, R. D., & Schofield, N. (1987). Generalized symmetry conditions at a core point. *Econometrica: journal of the Econometric Society*, 923–933.

Patty, J. W., & Penn, E. M. (2019). Measuring fairness, inequality, and big data: Social choice since arrow. *Annual Review of Political Science*, *22*, 435–460.

Plott, C. R. (1967). A notion of equilibrium and its possibility under majority rule. *The American Economic Review*, *57*(4), 787–806.

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

Sen, A. (1993). Internal consistency of choice. *Econometrica: Journal of the Econometric Society*, 495–521.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*(3), 426–482.

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, *82*(399), 782–793.

Wolak, F. A. (1989). Testing inequality constraints in linear econometric models. *Journal of econometrics*, *41*(2), 205–235.

Wolak, F. A. (1991). The local nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrica: Journal of the Econometric Society*, 981–995.

A Proofs

Proof of Proposition 1. Start with the definition of the marginal component effect

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

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, $\mathcal{D}_l = \{0, 1, \ldots, D_l - 1\}$.

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

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

By complete randomization (Assumption 4)

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

By the definition of the FCP (Definition 5)

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

Splitting the sum into three components yields

$$= \text{FCP}(t_a, t_b) \times p(T_{i[-j]kl} = t_b) + FCP(t_a, t_a) \times p(T_{i[-j]kl} = t_a) + \sum_{q\notin\{a,b\}} \text{FCP}(t_a, t_q) \times p(T_{i[-j]kl} = t_q)$$

where $\sum_{q\notin\{a,b\}}$ denotes a sum over all levels that are not t_a or t_b

Applying the same logic to the other half yields

$$\sum_{(t,\mathbf{t})\in\mathcal{T}} Y_i(t_b,t,\mathbf{t}) \times p(T_{ijk[-l]}=t,\mathbf{T}_{i[-j]k}=\mathbf{t}) =$$
$$\text{FCP}(t_b,t_a)\times p(T_{i[-j]kl}=t_a)+\text{FCP}(t_b,t_b)\times p(T_{i[-j]kl}=t_b)+\sum_{q\notin\{a,b\}}\text{FCP}(t_b,t_q)\times p(T_{i[-j]kl}=t_q)$$

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

$$\sum_{(t,\mathbf{t})\in\mathcal{T}} Y_i(t_b,t,\mathbf{t}) \times p(T_{ijk[-l]}=t,\mathbf{T}_{i[-j]k}=\mathbf{t}) =$$
$$p(T_{i[-j]kl}=t_a)-\text{FCP}_{il}(t_a,t_b)\times p(T_{i[-j]kl}=t_a)+\frac{1}{2}p(T_{i[-j]kl}=t_b)+\sum_{q\notin\{a,b\}}\text{FCP}_{il}(t_b,t_q)\times p(T_{i[-j]kl}=t_q)$$

Substituting into the MCE expression and rearranging terms:

$$\text{MCE}_{il}(t_a,t_b)=\left[\text{FCP}_{il}(t_a,t_b)-\frac{1}{2}\right]\times\left[p(T_{i[-j]kl}=t_a)+p(T_{i[-j]kl}=t_b)\right]+$$
$$\sum_{q\notin\{a,b\}}\left[\text{FCP}_{il}(t_a,t_q)-\text{FCP}_{il}(t_b,t_q)\right]\times p(T_{i[-j]kl}=t_q)$$

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

$$\text{MCE}_{il}(t_a,t_b)=\frac{2}{D_l}\left[\text{FCP}_{il}(t_a,t_b)-\frac{1}{2}\right]+\frac{1}{D_l}\sum_{q\notin\{a,b\}}\left[FCP_{il}(t_a,t_q)-\text{FCP}_{il}(t_b,t_q)\right]$$

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

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

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 FCP(t_a, t_b) will be defined as an average over all possible other attribute combinations, but FCP(t_a, t_q) and FCP(t_b, t_q) are averages only for those tasks where the attribute restrictions hold on the profile with t_q. □

Proof of Proposition 2. We will assume complete, transitive, and separable preferences on a set of profiles with full support, and show the contra-positive of FCP-transitivity, i.e. FCP$(t_b, t_c) >$ FCP$(t_a, t_c) \implies$ FCP$(t_a, t_b) <$ FCP(t_b, t_a).

First, note that FCP$(t_a, t_b) >$ FCP$(t_b, t_a) \iff$ FCP$(t_a, t_b) > 0.5$. We start by proving the following lemma:

Lemma A.1. *FCP*$(t_a, t_b) > 0.5 \iff$ t_a *defeats* t_b *in all all-else-equal comparisons.*

Proof. Suppose there exists some combination of other attributes t such that $(t_b, t) \succ (t_a, t)$. Then, by separability this must be true for all combinations of other attributes. Next, take some t' and t'' such that $(t_a, t') \succ (t_b, t'')$, we know some must exist because FCP$(t_a, t_b) > 0$ and preferences are connected. We also know that $(t_b, t') \succ (t_a, t')$ and $(t_b, t''c) \succ (t_a, t'')$ by separability. By transitivity, it must also be that $(t_b, t') \succ (t_a, t'')$. Thus, for every pair j, k such that $(t_a, j) \succ (t_b, k)$, we must have $(t_b, j) \succ (t_a, k)$. It follows that there are at least as many comparisons that are not all-else-equal in which t_b defeats t_a as vice versa. Because t_b also wins all all-else-equal comparisons, FCP$(t_a, t_b) < 0.5$. Swapping t_a and t_b proves necessity. □

Next, take three attribute values t_a, t_b, and t_c, and define the following:

$$X_1 \equiv \{t, b : (t_a, t) \succ (t_c, t')\} \qquad X_0 \equiv \{t, t' : (t_a, t) \prec (t_c, t')\}$$
$$Y_1 \equiv \{t, t' : (t_b, t) \succ (t_c, t')\} \qquad Y_0 \equiv \{t, t' : (t_b, t) \prec (t_c, t')\}$$

By completeness, we have that every pair of other attribute combinations must belong to exactly one of X_1 and X_0, and exactly one of Y_1 and Y_0; and $|X_1| + |X_0| = |Y_1| + |Y_0|$.

Suppose that FCP$(t_b, t_c) >$ FCP(t_a, t_c). Then, $Y_1 > X_1$, which in turn implies $X_0 > Y_0$. It follows that there must exist a pair of attribute combinations j, k such that $(j, k) \in Y_1, X_0$, that is,

$(t_b, j) \succ (t_c, k)$ and $(t_a, j) \prec (t_c, k)$. By transitivity, it follows that $(t_b, j) \succ (t_a, j)$, and by Lemma A.1 we know that $\text{FCP}(t_b, t_a) > 0.5$. □

Proof of Proposition 3. For sufficiency, without loss of generality take $\text{FCP}_i(t_a, t_b) \geq 1/2$. From separability, it follows that for any vector of $-l$ other attributes t, we have $Y(t_a, t, t_b, t) = 1$. By transitivity, for any t_j,

$$Y_i(t_b, t, t_j, t) = 1 \implies Y_i(t_a, t, t_j, t) = 1.$$

When preferences are separable, for any vector of $-l$ other attributes t and t', we have that

$$Y_i(t_b, t, t_j, t') = 1 \implies Y_i(t_a, t, t_j, t') = 1$$

which implies that $\text{FCP}_i(t_a, t_j) \geq \text{FCP}_i(t_b, t_j)$ for all t_j. It follows then from the definition of $MCE_i(t_a, t_b)$:

$$\text{MCE}_i(t_a, t_b) = \frac{2}{D_l}\Bigg(\underbrace{\text{FCP}_i(t_a, t_b) - \frac{1}{2}}_{\geq 0}\Bigg) + \frac{D_l - 2}{D_l} \sum_{t_j \neq (t_a, t_b)} \Bigg(\underbrace{\text{FCP}_i(t_a, t_j) - \text{FCP}_i(t_b, t_j)}_{\geq 0}\Bigg).$$

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

Proof of Proposition 4. We focus on the case where $t_a \neq t_b$. By construction, $AFCP_l(t_a, t_a) = AFCP_l(t_b, t_b) = .5$. Conditioning on the randomization distribution:

$$\begin{aligned}
E[\widehat{\text{AFCP}}_l(t_a, t_b, p(\mathbf{t}))] &= E\left[\frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} Y_{ijk}\mathbf{1}(T_{ijkl} = t_a, T_{i-jkl} = t_b)}{n_{ab}}\right] \\
&= \frac{1}{n_{ab}}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} E[Y_{ijk}\mathbf{1}(T_{ijkl} = t_a, T_{i-jkl} = t_b)] \\
&= \frac{1}{n_{ab}}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} E[Y_{ijk}|T_{ijkl} = t_a, T_{i[-j]kl} = t_b]p(T_{ijkl} = t_a, T_{i[-j]kl} = t_b)
\end{aligned}$$

Where the last step follows from law of total expectation. Again, applying law of total expectation

$$= \frac{1}{n_{ab}} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{(t,t')} E[Y_{ijk} | T_{ijkl} = t_a, T_{ijk[-l]} = t, T_{i[-j]kl} = t_b, T_{i[-j]k[-l]} = t'] \times$$
$$p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b) p(T_{ijkl} = t_a, T_{i[-j]kl} = t_b)$$

Under assumptions 1 and 2

$$= \frac{1}{n_{ab}} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{(t,t')} E[Y_{ij}(t_a, t, t_b, t')] \times$$
$$p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b) p(T_{ijkl} = t_a, T_{i[-j]kl} = t_b)$$

For $t_a \neq t_b$, under Assumption 2 $\bar{Y}_i(t_a, t, t_b, t') = Y_{ij}(t_a, t, t_b, t')$

$$= \frac{1}{n_{ab}} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{(t,t')} E[\bar{Y}_i(t_a, t, t_b, t')] \times$$
$$p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b) p(T_{ijkl} = t_a, T_{i[-j]kl} = t_b)$$

Cancelling yields:

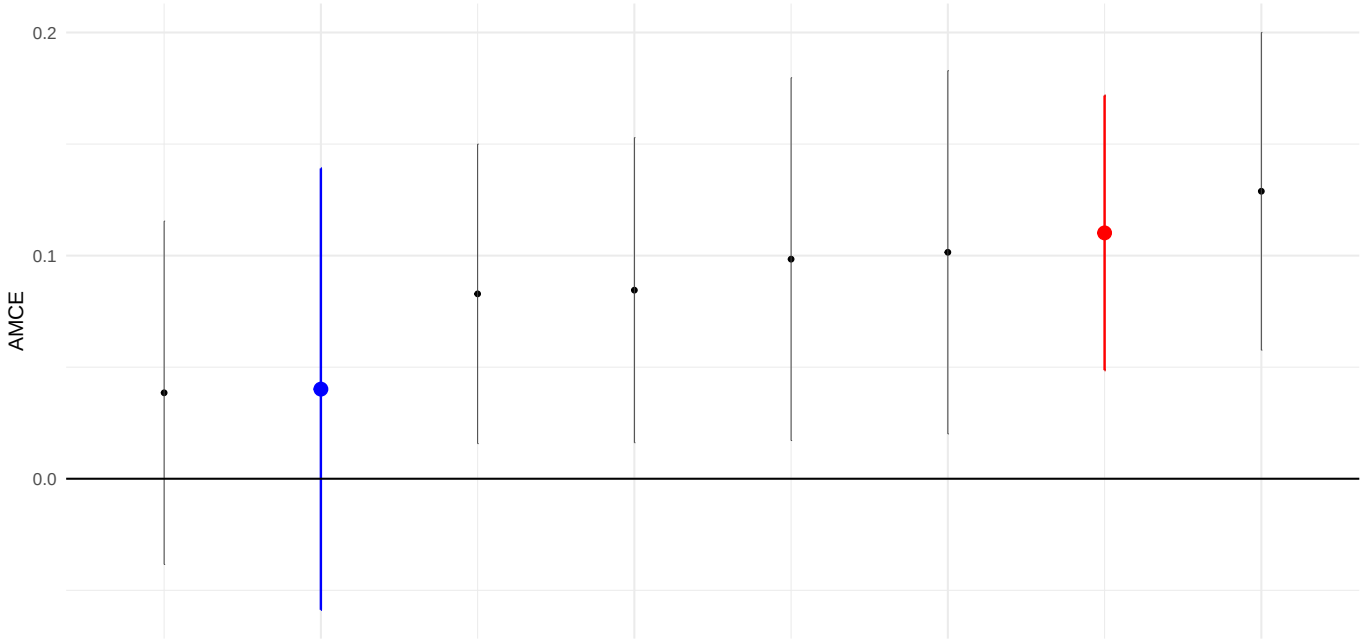
$$= \sum_{(t,t')} E[\bar{Y}_i(t_a, t, t_b, t')] \times p(T_{ijk[-l]} = t, T_{i[-j]k[-l]} = t' | T_{ijkl} = t_a, T_{i[-j]kl} = t_b)$$

which is the definition of the AFCP.

□

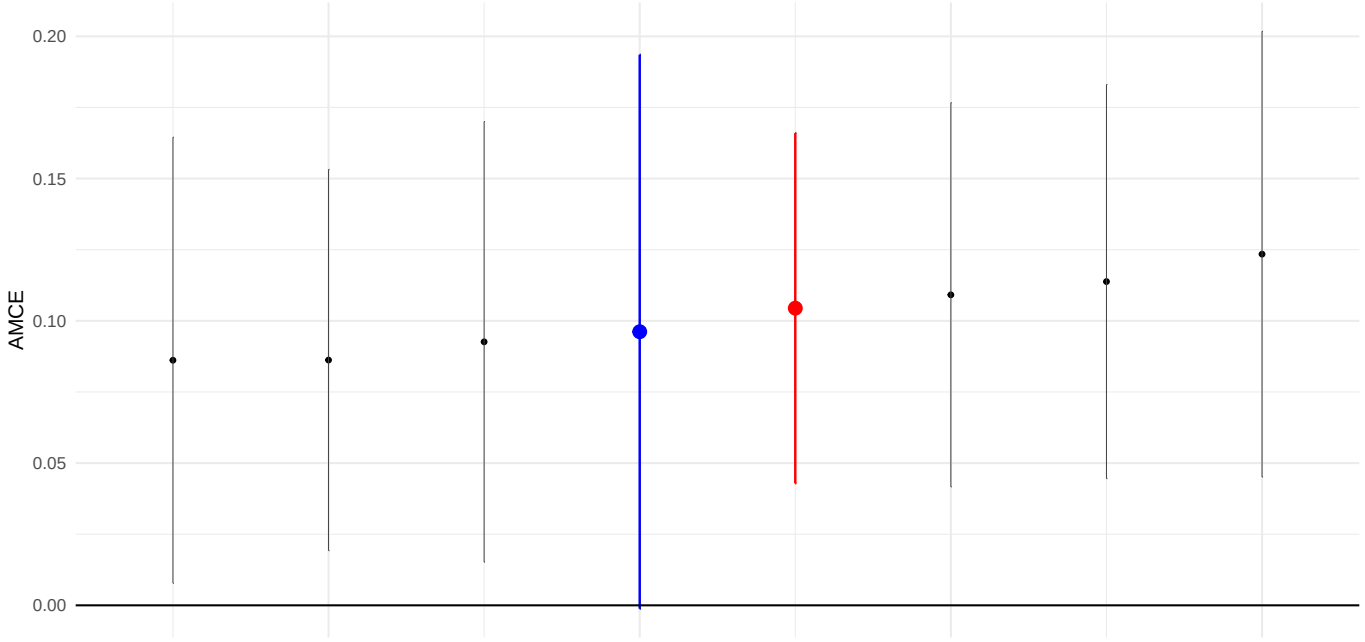
B Additional Tables and Figures

Figure B1: AMCE Estimates for 30 vs. 60 Years Old for All Possible Combinations of Left-Out Attribute Levels, Age in Blackman and Jackson (2021)



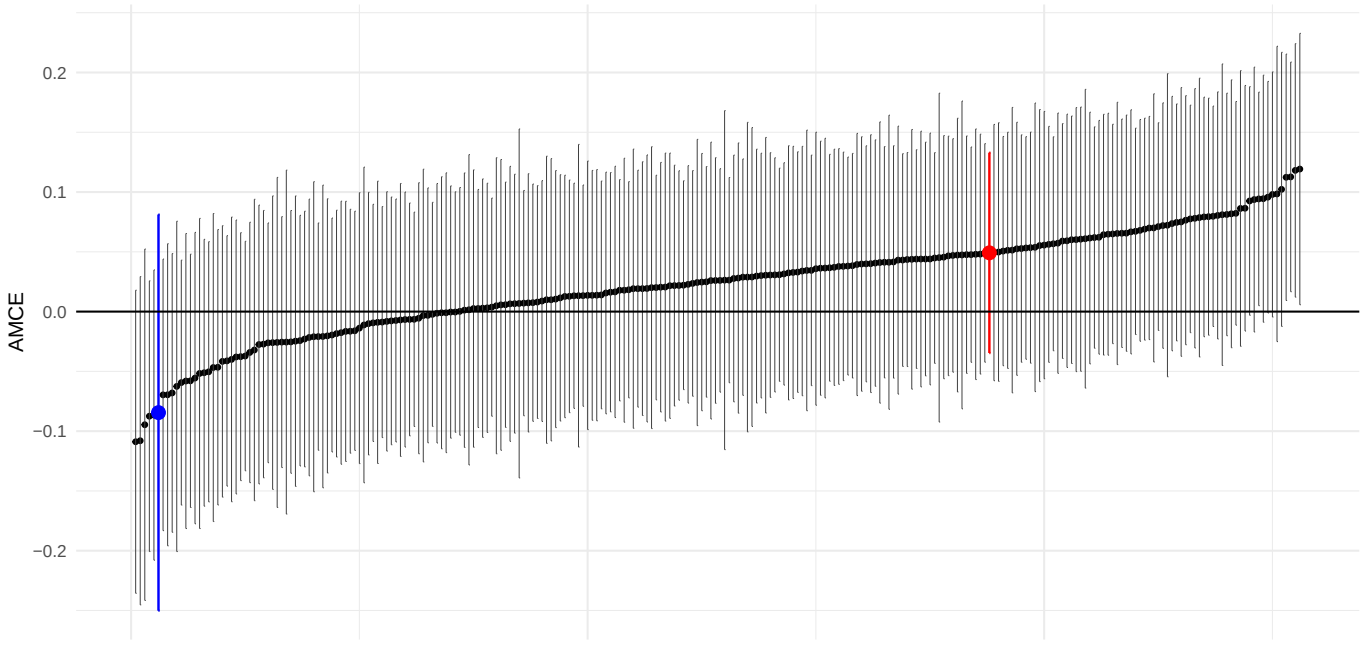
Notes: Blue point estimate indicates AFCP (equivalent to AMCE with all other attribute levels left out). Red point estimate indicates AMCE in the original study (with no other attribute levels left out). 95% confidence intervals are shown in gray.

Figure B2: AMCE Estimates for 30 vs. 70 Years Old for All Possible Combinations of Left-Out Attribute Levels, Age in Blackman and Jackson (2021)



Notes: Blue point estimate indicates AFCP (equivalent to AMCE with all other attribute levels left out). Red point estimate indicates AMCE in the original study (with no other attribute levels left out). 95% confidence intervals are shown in gray.

Figure B3: AMCE Estimates for Current of Love vs. Popular Front for All Possible Combinations of Left-Out Attribute Levels, Party in Blackman and Jackson (2021)



Notes: Blue point estimate indicates AFCP (equivalent to AMCE with all other attribute levels left out). Red point estimate indicates AMCE in the original study (with no other attribute levels left out). 95% confidence intervals are shown in gray.