What Do We Learn About Voter Preferences From Conjoint Experiments?

By Scott F Abramson, Korhan Koçak, & Asya Magazinnik*

Political scientists frequently interpret the results of conjoint experiments as reflective of voter preferences. In this paper we show that the target estimand of conjoint experiments, the AMCE, is not well-defined in these terms. Even with individually rational experimental subjects, unbiased estimates of the AMCE can indicate the opposite of the true preference of the majority. To show this, we characterize the preference aggregation rule implied by AMCE and demonstrate its several undesirable properties. With this result we provide a method for placing sharp bounds on the proportion of experimental subjects with a strict preference for a given candidate-feature. We provide a testable assumption to show when the AMCE corresponds in sign with the majority preference. Finally, we offer a structural interpretation of the AMCE and highlight that the problem we describe persists even when a model of voting is imposed.

Word Count: 8,053

FIRST DRAFT: 6/11/2019

THIS DRAFT: 8/14/2019

Conjoint experiments have become a standard part of the political scientist's toolkit. Across the top scholarly journals political scientists regularly interpret the results of these experiments to make empirical claims about both voter preferences and electoral outcomes. In this paper, we show that the target estimand of conjoint experiments, the average marginal component effect (AMCE), does not support such claims. We do so by characterizing the preference aggregation rule implied by the AMCE and, in line with well known results (Arrow, 1950; Gibbard, 1973), demonstrate its undesirable properties for making inferences about voter preferences and electoral outcomes.

The goal of factorial designs like those in forced-choice conjoint experiments is to mimic the comparisons individual voters make at the ballot box. By randomizing a large number of candidate and platform features, researchers seek to construct realistic approximates of the choices voters face. With a simple difference-in-means or least-squares regression researchers compare the attributes of

* Abramson: Assistant Professor, Department of Political Science, University of Rochester, email: sabramso@ur.rochester.edu; Kocak: PhD Candidate, Department of Politics, Princeton University, email: kkocak@princeton.edu; Magazinnik: Instructor, Department of Political Science, Massachusetts Institute of Technology, email: asyam@princeton.edu

candidates most frequently chosen to the attributes of the candidates least frequently chosen to make empirical claims about the preferences of voters.

For example, experimental results from conjoint experiments are used to make claims about voters' preferences for particular policies like: "Americans express a pronounced preference for immigrants who are well educated, are in high-skilled professions, and plan to work upon arrival (Hainmueller and Hopkins, 2015); and "[there is] strong evidence for progressive preferences over taxation among the American public" (Ballard-Rosa, Martin and Scheve, 2017). Even more frequently, conjoint results are used to make statements about candidates for elected office like: "voters prefer experienced or locally born politicians, but do not prefer politicians affiliated with a major political party... and are indifferent with regard to dynastic family ties and gender (Horiuchi, Smith and Yamamoto, 2018);" and "voters and legislators do not seem to hold female candidates in disregard; all else equal, they prefer female to male candidates (Teele, Kalla and Rosenbluth, 2018)."

Put simply, political scientists use conjoint results to make statements about a *binary preference relation* for a representative voter in the context of elections. Moreover, researchers interpret findings from conjoints as evidence that candidates with particular features are most preferred and thereby more likely to win elections (Carnes and Lupu, 2016; Teele, Kalla and Rosenbluth, 2018). This common interpretation has even migrated to the public discourse. CBS News and POLITICO, for example, have both highlighted results from conjoint experiments, asserting that the "[Democratic] party's primary voters prefer female candidates of color in 2020 (Magni and Reynolds, 2019)" and that [Democratic] "voters showed a clear preference for females, all else equal (Khanna, 2019)." By way of example and formal proof, we show that the AMCE produces a representative voter that does not support empirical claims about electoral contests.

The AMCE is defined as the average effect of varying one attribute of a candidate profile, e.g. the race or gender of the candidate, from A to A', on the probability that the candidate will be chosen by a respondent, where the expectation is defined over the distribution of the other attributes. To be clear, we do not dispute that the estimators proposed by Hainmueller, Hopkins and Yamamoto (2014) for this quantity are unbiased under their assumptions. Rather, we show that even when these assumptions hold, a positive AMCE of candidate-feature A over A' does not indicate: 1.) A majority of voters prefer candidates with feature A to those with A'; 2.) all else equal the median

voter prefers candidates with A to those with A'; nor 3.) candidates with feature A beat candidates with feature A' in most elections.

This occurs because the AMCE averages over two aspects of individual preferences: their direction (whether or not an individual prefers A to A') and their intensity (how much they prefer A to A'). Because the AMCE produces a literally average voter, it assigns greater weight to voters who intensely prefer a particular outcome, the consequence of which can be inaccurate out-of-sample predictions. For example, a large majority of people may have a strict preference for male candidates over female candidates, but the AMCE can, nevertheless, be positive for female candidates if there is a small minority of voters who have an intense preference for women. Far from being a statistical accident, this structure undergirds numerous political questions where the direction and intensity of preferences are potentially correlated.

Our point is not merely semantic. In the field of market research, where the tools of conjoint experiments were first developed, scholars are typically interested in the demand for a given product, which is determined by both the intensive and extensive margins of consumer choice. By contrast, political scientists typically care about elections, which are won on the extensive margin. Indeed, outside of fantastical institutional designs (e.g, Lalley and Weyl (2018)) electoral contests are not swayed by *how much* a subset of voters prefer a given candidate but, rather, *how many voters* have a strict preference for each candidate. By averaging over both margins of choice, the AMCE can prove largely uninformative with respect to the questions of interest to political scientists.

Since the objective of conjoint experiments is to construct a mapping from individual to aggregate preferences, we build on the literature in positive political theory that formally evaluates mechanisms that do just that. That is, we characterize the AMCE as a preference aggregation rule — a mapping from individual to aggregate preferences (Austen-Smith and Banks, 2000, p. 26). In doing so, we show that the AMCE is a perturbation of the Borda rule and, as such, inherits some of its undesirable properties. Namely, we demonstrate that the AMCE does not satisfy the majority or independence of irrelevant alternatives (IIA) criteria.

Having characterized the preference aggregation rule of the AMCE, we then use results from this exercise to provide a method that, for a given AMCE estimate, allows researchers to place sharp bounds on the proportion of experimental subjects that maintain a strict preference for a candidate-

feature. Using this method, we re-evaluate the findings of every conjoint experiment published in the *American Political Science Review, American Journal of Political Science*, and *the Journal of Politics* between 2016 and 2019 and show that, with two exceptions, their results are consistent with either a majority or minority of respondents holding a strict preference for the candidate-feature that yielded each study's largest estimated effect.

Finally, we explore the relationship between the AMCE and a simple model of voting. In providing a structural interpretation of the AMCE we show that it reflects an average of individual ideal points over candidate-features. This highlights how conjoints combine information about both the intensity and direction of preferences and demonstrates the need to impose additional structure in order to obtain estimates of theoretically relevant quantities of interest. We conclude with some directions for future research on how to make conjoints more informative about voter preferences.

I. An Example

Part I

To start, we work through a toy example of how the AMCE aggregates preferences. We aim to make as few assumptions about the underlying preferences of individual voters as possible. While we view our assumptions as benign, we note that if the AMCE exhibits undesirable properties under these assumptions, placing even less structure will not rectify whatever problems we identify and only obscure what drives these results. Furthermore, we emphasize that we are agnostic with respect to the content of voters preferences. Individuals may be self-interested, other-regarding, or some mixture thereof. *We impose only that individual preferences are complete and transitive.*[1] Without completeness and transitivity we can learn about neither individual nor aggregate preferences. As such, these are the minimal assumptions about individual preferences we can make and still hope to recover meaningful insight into the AMCE.

Since, fundamentally, the object researchers seek to describe concerns a preference relation over candidate-features, the primitives we begin with are over these features. For simplicity, consider an electorate of five voters (V1, V2, V3, V4, V5), whose preferences over candidates we would like to

[1]Formally, completeness is defined as preferences satisfy $x \succsim y$, $y \succsim x$ or both for all $x, y \in X$. Transitivity is defined as $x \succsim y$ & $y \succsim z$, then $x \succsim z$ for all $x, y, z \in X$ where $\succsim$ denotes the weak preference relation.

study with a conjoint experiment. To eliminate concerns about estimation, suppose we can fully observe every potential choice between candidates made by every member of this population. In this world, there are two attributes of candidates that are important to voters: their gender (female or male) denoted by $G \in \{F, M\}$, and their party (Democrat or Republican) denoted $P \in \{D, R\}$. Each candidate is an ordered pair of gender and age, so that there are four different candidate profiles: FD, FR, MD, and MR. The voters' preferences over attributes are a strict partial order $\succ$, and are given in the following table:

V1	V2	V3	V4	V5
$M \succ F$	$M \succ F$	$M \succ F$	$F \succ M$	$F \succ M$
$R \succ D$	$R \succ D$	$R \succ D$	$D \succ R$	$D \succ R$

Table 1—: Preferences over attributes

It can easily be seen that a majority of voters prefer male candidates to female candidates, and a majority of voters prefer Republican candidates to Democratic candidates.

We construct preferences over candidates from preferences over attributes in the following way: Voters prefer candidates that have both of the attributes they like to those that have one attribute they like, which in turn they prefer to candidates who have neither of the attributes they like. Notice that there are two types of candidates that have only one attribute that matches a voter's preference. For these candidates, whether a voter prefers one or the other depends on which attribute the voter places a greater weight on. For example, if a voter places more weight on gender, we would expect them to choose a candidate who has their preferred gender but not their preferred party over a candidate who has the voter's preferred party but not the gender.

Formally, such preferences over candidate profiles can be written as the lexicographic preference relation $\succsim$, where for each voter one attribute is given a greater weight in determining the preference ordering. Accordingly, we assume that voters 1, 2, and 3 place more weight on the candidate's party, $P \succsim G$, whereas voters 4 and 5 place more on the candidate's gender, $G \succsim D$. Combining weights with preferences over attributes, we can produce voters' preferences over candidate profiles. These are presented in Table 2.

Given these preferences, in Table 3 we present the votes candidates would obtain in each head-to-

Rank	V1	V2	V3	V4	V5
1.	MR	MR	MR	FD	FD
2.	FR	FR	FR	FR	FR
3.	MD	MD	MD	MD	MD
4.	FD	FD	FD	MR	MR

Table 2—: Preferences over candidate profiles

head election for every possible pairwise comparison. Note that in this example men win three of the four elections when they face off against a woman and four of the six total contests (the winner is bolded in the first column).

Comparison	V1	V2	V3	V4	V5	Tally
MR,FR	MR	MR	MR	FR	FR	3, 2
MR,FD	MR	MR	MR	FD	FD	3, 2
MR,MD	MR	MR	MR	MD	MD	3, 2
MD,**FR**	FR	FR	FR	FR	FR	0, 5
MD,FD	MD	MD	MD	FD	FD	3, 2
FR,FD	FR	FR	FR	FD	FD	3, 2

Table 3—: Aggregate preferences over candidate profiles

In this simple setting, the AMCE is derived as in Hainmueller, Hopkins and Yamamoto (2014), Proposition 3. The intuition behind the comparisons being made when estimating the AMCE is given in Table 4. Here, $\bar{Y}(C_1, C_2)$ denotes the fraction of votes that candidate C_1 obtains when run against candidate C_2. For each contest we can obtain $\bar{Y}$ from the last column of Table 3. To obtain the AMCE for males we compare how male candidates (column 1) fare relative to female candidates (column 2) when they run against the same opponent, then sum this difference over all possible opponents. This sum is finally normalized by the number of possible profiles minus one (3) times the number of possible values for gender (2). The procedure yields an AMCE for male equal to $-1/15$, meaning that the average probability of being chosen is higher for female candidates than it is for male candidates.

Our toy example illustrates the intuition driving our main result. Notice that the AMCE for men is *negative*, and yet we know that by construction a majority of the voters prefer male to female candidates and that men will beat women in a majority of head-to-head elections. Holding all else

1.		2.		
$\bar{Y}(MR, MD)$	$-$	$\bar{Y}(FR, MD)$	$=$	$-2/5$
$\bar{Y}(MR, FD)$	$-$	$\bar{Y}(FR, FD)$	$=$	0
$\bar{Y}(MR, MR)$	$-$	$\bar{Y}(FR, MR)$	$=$	$1/10$
$\bar{Y}(MR, FR)$	$-$	$\bar{Y}(FR, FR)$	$=$	$1/10$
$\bar{Y}(MD, MD)$	$-$	$\bar{Y}(FD, MD)$	$=$	$1/10$
$\bar{Y}(MD, FD)$	$-$	$\bar{Y}(FD, FD)$	$=$	$1/10$
$\bar{Y}(MD, MR)$	$-$	$\bar{Y}(FD, MR)$	$=$	0
$\bar{Y}(MD, FR)$	$-$	$\bar{Y}(FD, FR)$	$=$	$-2/5$
				$-2/5$
(# of profiles $-$ 1) $\times$ (# of features -1) $\times$# of values for gender			$=$	6
AMCE			$=$	$-1/15$

Table 4—: Obtaining the AMCE

constant (in the case of this example, party), a male candidate would always win.[2] Furthermore, men win more electoral contests.[3] The AMCE produces an estimate that indicates the opposite of the true majority preference because the minority, who place the greatest weight on the gender dimension, also have a preference for female candidates, while the majority, who prefer men, do not place much weight on gender when making their decisions. When aggregating preferences over gender, the AMCE mechanically assigns greater weight to the minority that strongly prefer women.

Crucially, this result is a feature of the target estimand and is not a problem of estimation. Our example is analogous to a survey in which each respondent is asked to evaluate all possible head-to-head comparisons. To highlight this, we conduct a simulation exercise where we run a three question conjoint experiment on a population characterized by the distribution of voter preferences in our toy example. That is, we take a population of five voters with the preferences detailed in Table 3. Then, we randomly construct pairs of candidates, perturbing their gender and age. Knowing voter preferences for candidate profiles we then obtain a winner in each contest and estimate the AMCE for male candidates. In Figure 1 we present results from conducting this exercise 1,000 times. Of course, because the AMCE is unbiased, the effect is centered on -1/15, despite being generated from a population of voters where 3/5 prefer men.

[2]Note that in rows 1 and 5 of Table 3 the male candidate gets three votes against the female candidate's two.
[3]From Table 3 men win rows 1, 2, 3, and 5 and women win rows 4, and 6.

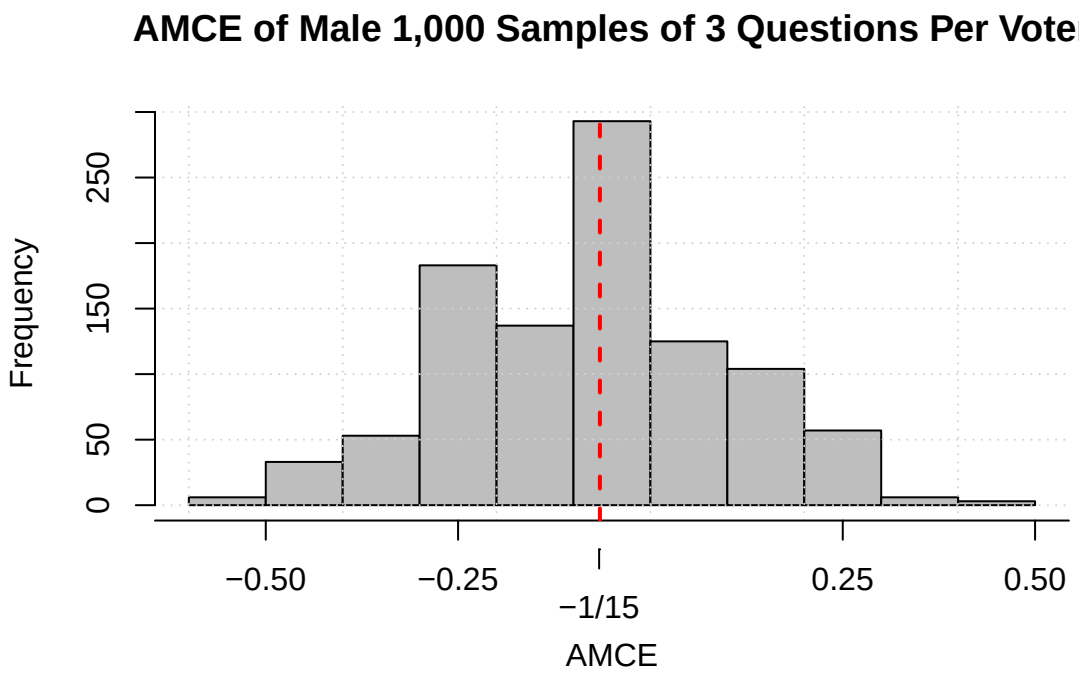

Figure 1. : This figure presents an exercise where we conduct 1,000 conjoint experiments on a population of 5 voters with preferences as detailed in Table 3 and where candidates are randomly generated from a combination of gender and age.

Part II

Consider the same population of five voters. However, instead of conducting an experiment where we randomize only party and gender, we now include an additional third feature. Call this third feature race, which for simplicity takes on only two values, black or white. Denote this $R \in \{B, W\}$. Let voters 1, 2, and 3 have the preference $W \succ B$ and voters 4 and 5 have the preference $B \succ W$. Furthermore, let voters 1, 2, and 3 place the greatest weight on party, the second most on gender, and least weight on race, i.e. $P \succsim G \succsim R$, and let voters 4 and 5 place the greatest weight on race, the next most on gender, and the least on party, i.e. $R \succsim G \succsim P$.

As in the previous section, with this combination of weights and preferences for features we can produce a full ranking of preferences for candidates. Voters most prefer candidates with all three of their preferred features and least prefer those with none of their preferred features. Among

candidates that have two of the three features voters prefer, they rank candidates with their first and second most preferred feature first, first and third most preferred features second, and second and third most preferred features third. Finally, we assume that voters prefer all candidates with two preferred features to all candidates with just one feature. Preferences over candidates are given in Table 5.

Rank	V1	V2	V3	V4	V5
1.	MRW	MRW	MRW	FDB	FDB
2.	MRB	MRB	FRW	FRB	FRB
3.	FRW	FRW	MDW	MDB	MDB
4.	MDW	MDW	FRB	FDW	FDW
5.	FRB	FRB	MDB	MRW	MRW
6.	MDB	MDB	FDW	FRW	FRW
7.	FDW	FDW	FDB	MDW	MDW
8.	FDB	FRB	FRW	MRW	MRW

Table 5—: Preferences over candidate profiles - Example Part II

Again, with these preferences and weights we can fully characterize vote shares in each possible electoral contest. These are presented in Table A1 in the supplemental appendix.

Since these are the same exact voters from the previous example, their preferences with respect to gender have not changed: 3/5 of them prefer men to women. As before, in this example men win a large majority of elections.[4] However, by contrast to our previous example, instead of always ranking female candidates above male candidates, voters 4 and 5 will now sometimes vote for a man. Because they place more weight on race than gender, in some contests, voters 4 and 5 will be willing to accept a man even though, all else equal, they prefer women. Since including race changes the relative ranking of male and female candidates, it changes the AMCE researchers would derive from this experiment.

As before, we follow Hainmueller, Hopkins and Yamamoto (2014) Proposition 3 to calculate the AMCE. In Table A2 we show that it is equal to 1/14, yielding the exact opposite substantive result compared to the previous experiment where we considered only gender and party. That is, with the same set of experimental subjects, just by adding an additional feature we can flip the sign of

[4]Thirteen of the sixteen elections where men run head to head against women are won by the male candidate as are nineteen of the twenty-eight overall contests.

the AMCE. This highlights our second main result: the sign and magnitude of the AMCE depend upon the features included in the experimental design even though individual preferences over these features remain constant across experiments. This occurs because the inclusion of the additional feature changes the relative rankings of candidates with respect to other unrelated, but potentially theoretically important, attributes.

In other words, even with identical subjects, the results researchers obtain from conjoint experiments depend upon the specific set of features included in their experimental design. In the supplemental appendix we describe an example that highlights a variant of this problem that is relevant for applied researchers. There, we show that the exclusion of "unrealistic" feature-combinations alters the AMCE for the same reason as above. For instance, the exclusion of uneducated doctors in a candidate-choice experiment where education, occupation, and gender are randomized changes the AMCE a researcher would obtain for male candidates. This occurs because the exclusion of uneducated doctors can alter the relative ranking of female and male candidates, in turn changing the AMCE we would obtain from this restricted randomization relative to a uniform randomization of features.

II. The AMCE as a Preference Aggregation Rule

In this section, we show that the above example characterizes general features of the AMCE. To accomplish this we start by showing that the AMCE has a direct correspondence to the Borda rule, a voting system that assigns points to candidates according to their order of preference. Borda rule voting is implemented as follows. With n candidates, the Borda rule assigns zero points to each voter's least preferred candidate, one point to the candidate preferred to that but no other, and so on until the most preferred candidate receives $n-1$ points. Thus for each voter, the Borda score contributed to a candidate corresponds to the number of other candidates to whom he or she is preferred. This in turn is equal to the number of times that candidate would be chosen if the voter was presented with every possible binary comparison. A candidate's Borda score is the sum of the individual Borda scores assigned to that candidate by each voter, and is equal to the total number of times that candidate would be chosen if each voter was subjected to each binary comparison. This is summarized in Lemma 1:

LEMMA 1: *The Borda score of each profile is equal to the total number of times that profile is chosen in all pairwise comparisons.*

PROOF:

All proofs are in the appendix.

In the context of conjoint experiments, we further define the Borda score of a feature as the sum of the Borda scores of each profile that has that feature. For example, the Borda score of "female" is the sum of the Borda scores of all female candidates. This definition is useful, because we can prove the following result that relates Borda scores of features to Borda scores of profiles:

LEMMA 2: *When there are no interactions and only binary attributes, a profile has the highest Borda score if and only if all its features have the highest Borda scores for their respective attributes.*

We can now state our result pertaining to the equivalence of Borda and AMCE:

PROPOSITION 1: *The difference of the Borda scores of a feature and the benchmark is proportional to the AMCE of that attribute.*

The proof of Proposition 1 follows from Lemma 1 and the observation that Borda and AMCE measure aggregate preferences in analogous ways. They both tally the number of alternatives that are defeated by candidates with a given feature, then use that tally to compare across features. The AMCE estimates are constructed by taking the difference of these tallies and normalizing them to be between -1 and 1. In the proof we formally walk through the steps of how to get to AMCE from Borda counts, and produce the same expression as the AMCE in Equation 5 of Hainmueller, Hopkins and Yamamoto (2014).

This equivalence is important, because it is well known in the social choice literature that the Borda rule has several undesirable properties. We have shown that these properties extend to the AMCE. For example, the Borda rule violates the irrelevance of independent alternatives (IIA) criterion, which states that the relative ranking of two candidates should not depend on the presence of another candidate. In the second part of our example we showed that the AMCE violates IIA for similar reasons — that is, that the AMCE of a given feature depends on the other features included in the experiment. In our example, the estimated AMCE on male versus female depended

upon whether or not we included race. Of course, this is deeply problematic since the AMCE is only internally valid with respect to the features included, or feature-combinations excluded, in the particular experimental design that produced it.[5]

A second social choice property of the AMCE, which it also inherits from the Borda rule, tells us that we should be wary of standard interpretations of conjoint results in political science. Specifically, like the Borda rule, the AMCE violates the majority criterion. This states that if a majority of voters prefer one candidate, then that candidate must win. Our example shows that this feature of the Borda rule extends to attributes, where a majority of voters prefer male candidates to female candidates, but the Borda score of F is greater than that of M. Here, we establish this result more generally. Specifically, we show that when a majority of respondents prefer a feature, the AMCE may still indicate that feature has a negative effect on the probability of being chosen. This discrepancy is driven by respondents assigning different weights, or importance, to attributes. For example, if respondents who like a feature also put more weight on it than those who dislike it, the AMCE estimate will be higher than the margin of respondents who strictly prefer that attribute. More importantly, a small minority that cares intensely about an attribute can overtake a much larger majority that has the opposite preference but cares less intensely about it. This may result in an AMCE in favor of the feature the minority prefers, even if that feature would in fact lead to a large electoral disadvantage between otherwise similar candidates.

We leverage the correspondence between the AMCE and the Borda rule to derive sharp bounds on the fraction of the population that prefers a feature over the benchmark and show that the potential divergence with AMCE grows in the number of unique candidate profiles, K. More precisely, for any given value of the AMCE for a feature, total number of candidate profiles, and the number of values the attribute can take, we define the maximum and minimum fractions of voters who prefer that feature over the benchmark attribute. These bounds are given in our next result.

PROPOSITION 2: *Let y denote the fraction of voters who prefer t_1 over t_0. Given an AMCE*

[5]This finding, furthermore, provides a theoretical foundation for the paper of de la Cuesta, Egami and Imai (2019) who demonstrate the sensitivity of the AMCE to changes in the distribution of features randomized.

estimate of $\pi(t_1, t_0)$*, it must be that*

$$y \in \left[\max\left\{\frac{\pi(t_1,t_0)\tau(K-1)+\tau}{K(\tau-1)+\tau}, 0\right\}, \min\left\{\frac{\pi(t_1,t_0)\tau(K-1)+K(\tau-1)}{K(\tau-1)+\tau}, 1\right\}\right]$$

where τ *is the number of distinct values the attribute of interest can take.*

To find these bounds, we calculate the highest and lowest possible Borda scores a respondent can contribute to a feature as a function of the total number of possible profiles, and the number of distinct values the attribute of interest can take. We first assume that for all proponents of a feature, the attribute involved is the top priority. This means that all profiles with that feature are preferred to all profiles without that feature. This results in the highest possible Borda score to the feature, and minimum possible Borda score to the benchmark. Thus we obtain the maximum net Borda score a proponent can contribute to a feature. In contrast, we assume for all opponents of that feature, the attribute has the lowest priority. This means that the attribute in question only factors in what an opponent chooses if the profiles are otherwise identical. This results in the highest possible Borda score for the feature, subject to the constraint that opponents prefer the benchmark to it. This yields the minimum net Borda score an opponent can subtract from a feature. Having calculated the maximum Borda score for a feature per proponent and opponent, we can invoke Proposition 1 to calculate the maximum possible AMCE estimate for a given fraction of opponents and proponents. Inverting this function yields the lowest possible fraction of proponents for a given AMCE estimate. The upper bound is calculated analogously. Interested readers can find the details in the proof, where we formally state and carefully trace the arguments summarized here.

In Figure 2, we apply this proposition to compute the bounds for AMCEs of 0.05, 0.10, 0.15, and 0.25 for a binary feature, plotting the upper and lower bounds of the proportion of experimental subjects who prefer a binary feature on the y-axis against the number of potential candidate profiles that respondents can choose from on the x-axis. As the figure shows, even for AMCE estimates of a fairly large magnitude, it takes fewer than five possible profiles for these bounds to grow to a completely uninformative range. Of course, nearly all conjoint experiments exceed five possible candidate profiles. For instance, with six attributes taking two possible values each — still a conservative design by recent standards — there are already $2^6 = 64$ possible profiles. Only when the

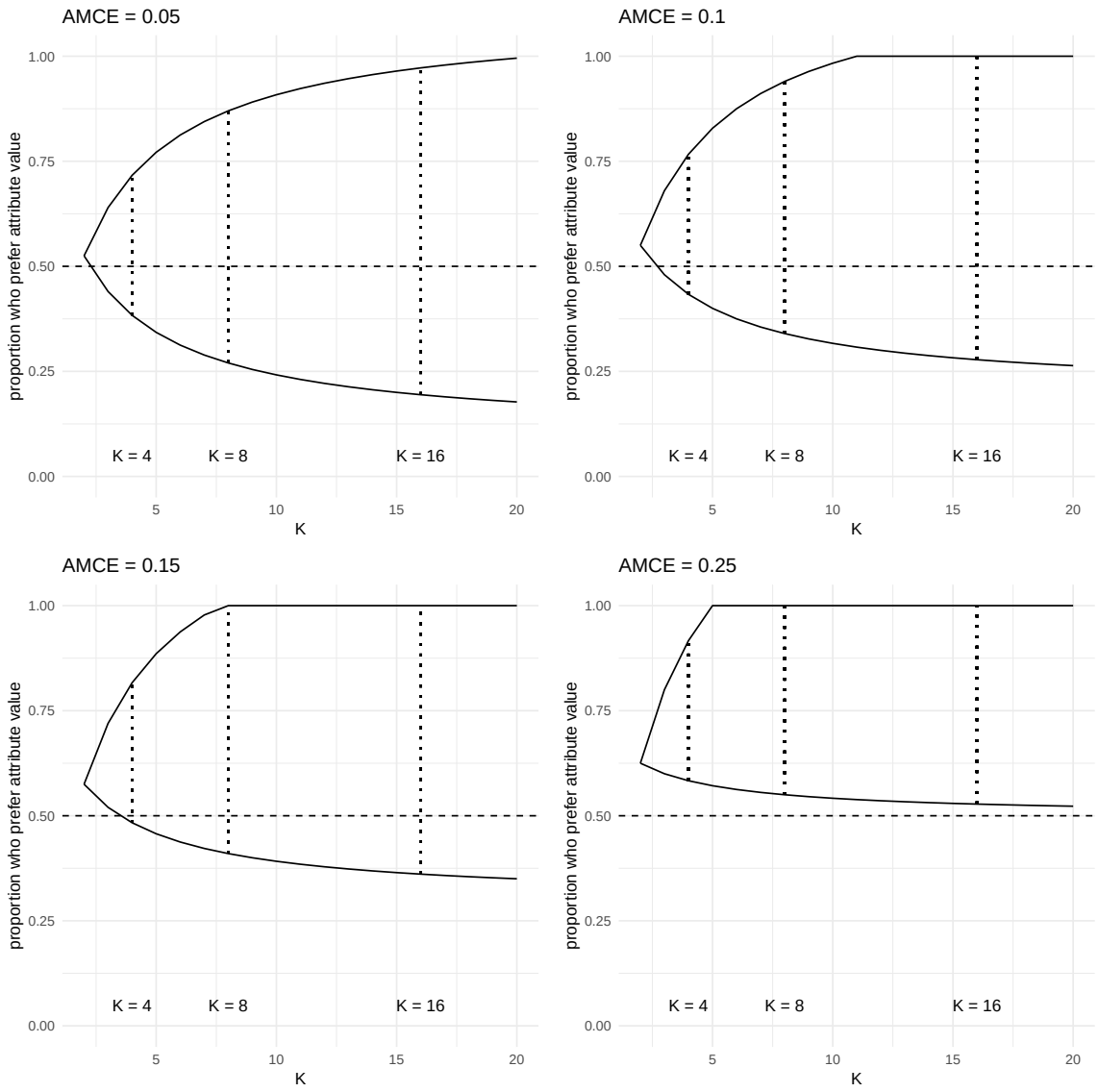

Figure 2. : Upper and lower bounds on fraction of people who prefer a binary feature, consistent with an AMCE of .05, .10, .15, and .25, respectively, as a function of number of possible candidate profiles.

AMCE is extremely large — an effect size of 0.25, which is rarely achieved by anything other than controls such as a candidate's partisanship or experience — does the bounding exercise assure a majority preference. Even then, if the feature of interest were ternary instead of binary, this would no longer be the case even at an effect size of 0.25.

In Table 6, we conduct this exercise for every forced-choice conjoint experiment in the *APSR,*

AJPS, and *JOP* published between 2016 and the first quarter of 2019. We construct our bounds for the largest estimated effect presented in each of these papers. From the eleven papers we analyze, only two — those of Mummolo (2016) and Hemker and Rink (2017) — produce bounds that guarantee a majority preference. In both of these papers, the effect sizes are quite large — 0.30 and 0.33, respectively — and the number of possible candidate profiles is comparatively small (6 and 32, respectively). Furthermore, in both cases, the attribute of interest is binary; note that, by contrast, the very largest effect size of 0.35, found in Newman and Malhotra (2018), produces bounds that do not guarantee a majority preference due to the large number of possible profiles (over 120,000) and relevant features (9).[6]

The bounding exercise we propose contains the entire range of preferences that are consistent with a given AMCE. In other words, the upper and lower bounds reflect a worst-case scenario for researchers, which is realized when preferences over features and weights over attributes are highly correlated. Thus, Proposition 2 underscores the dangers of making statements about aggregate preferences with so little structure on individual choices.

Of course, we may be unlikely to achieve this worst-case scenario, and the correlation between preferences over features and weights may not be so large. Researchers may therefore want to know how the AMCE performs in the best-case scenario. We can use the logic underlying Proposition 2 to show that when voters have homogeneous weights — that is, when every respondent has the same priorities over attributes — the AMCE and the majority preference must point in the same direction. That is, when all subjects assign the same priority ranking to a binary attribute, we show that the sign of the AMCE must correspond to the sign of the margin of victory for the relevant feature over the baseline. Usefully for researchers, under these conditions the AMCE will be smaller in magnitude than the size of the margin, thus providing a downwardly biased — and therefore conservative — estimate for that quantity. Furthermore, we find that as the weight assigned to an attribute relative to other attributes grows, the distance between the AMCE and the size of the margin shrinks.

COROLLARY 1 (Homogeneous weights): *When voters assign homogeneous weights to attributes, the AMCE of a binary attribute has the same sign as the majority preference, but underestimates*

[6]Here, Newman and Malhotra (2018) re-analyze data from a conjoint experiment conducted by Hainmueller and Hopkins (2015).

Table 6—: Bounds on Proportion of Sample Having Preferences Consistent with AMCE, Computed for Recent Papers in the Top Political Science Journals

Paper	Estimated effect	AMCE (π)	Number of profiles (K)	Number of relevant features (τ)	Bounds on proportion with consistent preference
APSR					
Ward (2019)	Proportion of group comprised of young men on support for immigration, 100% vs. 50% baseline	-0.18	6,840	5	[0.00, 0.77]
Auerbach and Thachil (2018)	Broker education on support, high (BA) vs. none	0.13	1,296	3	[0.20, 1.00]
Hankinson (2018)	Proximity of new housing on homeowners' support for new construction, 2-minute walk vs. 40 minutes	-0.09	6,144	4	[0.00, 0.88]
Teele, Kalla, and Rosenbluth (2018)	Experience on candidate support, 8 years vs. 0 years	0.15	864	4	[0.20, 1.00]
Carnes and Lupu (2016)	Experience on candidate support, some vs. none	0.09	32	2	[0.22, 1.00]
JOP					
Newman and Malhotra (2018)*	Skill on support for immigrants, high vs. low	0.35	120,960	9	[0.39, 1.00]
Ballard-Rosa, Martin, and Scheve (2016)	Tax rate on those earning <10k on support for plan, 25% vs. 0%	-0.24	38,400	4	[0.00, 0.68]
Mummolo and Nall (2016)	White proportion of community on Republicans' choice to live there, 96% vs. 50%	0.11	3,456	4	[0.15, 1.00]
Mummolo (2016)	Relevant information on choice to consume, vs. irrelevant	0.30	6	2	[0.63, 1.00]
AJPS					
Hemker and Rink (2017)	Nationality on quality of response, foreign vs. German	0.33	32	2	[0.66, 1.00]
Huff and Kertzer (2017)	Perpetrator's organization on labeling attack as terrorism, foreign ties vs. no info	0.19	108,000	6	[0.23, 1.00]

* Newman and Malhotra (2019) re-analyze data from a conjoint experiment conducted by Hainmueller and Hopkins (2015)

the size of the margin. The size of the underestimation grows as the relative weight assigned to the attribute of interest falls. In the limit as the relative weight of the attribute of interest goes to zero, so does AMCE; even when the margin is arbitrarily close to one.

Proof of Corollary 1 follows closely the logic of Proposition 2: when weights are identical across proponents and opponents, each proponent contributes as many net Borda points to a feature as an opponent takes away from it. As such, when the points contributed by proponents and opponents cancel out, the remainder corresponds to the margin of victory for the feature preferred by the majority. Because Borda scores are increasing in the weight assigned to an attribute, the remainder

also increases. Thus the AMCE is sensitive to the weight assigned to that attribute and therefore captures the size of this margin, even when it has the correct sign.

Allowing for Interactions Between Features

For clarity of exposition, we have focused on the scenario where voters have unconditional preferences over candidate features. We now relax this assumption, allowing for interactions between features such that voters may prefer, for instance, men over women when the candidate is a Republican, and the reverse when the candidate is a Democrat.[7] In this section, we derive a summary statistic for aggregate feature preferences that captures this more complex and potentially more realistic preference structure, and we show that the bounds derived in Proposition 2 are *also* the bounds on this quantity.

We define an **individual feature preference** for feature t_1 over feature t_0 as the proportion of the time respondent i selects a profile with feature t_1 over an otherwise identical profile with feature t_0, over all all-else-equal head-to-head contests that can be constructed from all values of the other features. Formally:

$$\Pi_i(t_1, t_0) = \frac{1}{K/2} \sum_{j=1}^{K/2} Y_i(x_{j1}, x_{j0})$$

where K is the total number of possible profiles, as before. Just as we do in our proofs, we denote by $Y_i(x_{j1}, x_{j0}) = 1$ if voter i chooses profile x_{j1} with feature t_1 over an otherwise identical profile x_{j0} with feature t_0 in a pairwise comparison, and $Y_i(x_{j1}, x_{j0}) = 0$ otherwise. To aggregate over voters, we simply compute the average of this proportion:

$$\Pi(t_1, t_0) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{K/2} \sum_{j=1}^{K/2} Y_i(x_{j1}, x_{j0}) \right\}$$

In Table 7 below, we reproduce Table 2 and compute Π_i for the two kinds of voters already discussed. We also include a new type of voter, $V6$, whose preferences are $M \succ F$ when the candidate is a Republican, and $F \succ M$ when the candidate is a Democrat.[8]

[7] That is, the *feature* they prefer is a function of the other features — not their preferred candidate profile, which is, of course, also a function of the other features in our main example.

[8] We thank Naoki Egami for encouraging us to consider this case.

Rank	V1	V2	V3	V4	V5	V6
1.	MR	MR	MR	FD	FD	MR
2.	FR	FR	FR	FR	FR	FR
3.	MD	MD	MD	MD	MD	FD
4.	FD	FD	FD	MR	MR	MD
$\Pi_i(M,F)$:	1	1	1	0	0	$\frac{1}{2}$

Table 7—: Illustrations of Π_i

Here, $\Pi_i(M,F)$ is computed as follows:

$$\Pi_{\{1,2,3\}}(M,F) = \frac{Y_i(MR,FR) + Y_i(MD,FD)}{2} = \frac{1+1}{2} = 1$$
$$\Pi_{\{4,5\}}(M,F) = \frac{Y_i(MR,FR) + Y_i(MD,FD)}{2} = \frac{0+0}{2} = 0$$
$$\Pi_6(M,F) = \frac{Y_i(MR,FR) + Y_i(MD,FD)}{2} = \frac{1+0}{2} = \frac{1}{2}$$

The aggregate preference in this population is $\Pi(M,F) = \frac{1}{6}\left(3+0+\frac{1}{2}\right) = \frac{7}{12}$.

As shown after the proof of Proposition 2 in the supplementary appendix, the bounds we derived on the proportion of the population that prefers feature t_1 over t_0 also apply to the proportion $\Pi(t_1,t_0)$ in a framework that allows for interactions, since the additional preference orderings introduced by the possibility of interactions are not the extreme cases that establish the bounds. We also show, however, that this quantity does not necessarily imply an electoral advantage for t_1 over t_0 in the presence of interactions. In other words, under a more complex interactive preference structure, our bounding exercise speaks only to the specific quantity $\Pi(t_1,t_0)$, and analysts should still take care about making claims about winners of electoral contests.

III. Structural Interpretation of the AMCE

Although the proposed estimator of the AMCE of Hainmueller, Hopkins and Yamamoto (2014) is "model free," in this section we demonstrate how it relates to an underlying model of choice. Our purpose in providing this simple structural interpretation of the AMCE is to illustrate from another angle the same aggregation problem that we have already identified in the preceding sections, wherein we cannot disentangle the intensity and direction of individual preferences. To start, consider two

candidates $c \in \{1, 2\}$ running in contest j who offer platforms $\mathbf{x}_{ijc}$ to voter i. A platform $\mathbf{x}_{ijc}$ is a vector of policies of length M that fully characterizes a candidate in contest j, which we will eventually recast as a vector capturing all the features (e.g. female, white, Republican) of that candidate. Let b_i represent an M length vector of voter i's preferred policy locations (e.g., their issue-specific ideal-points), and assume that voters have quadratic utility functions. Thus, voter i's utility is maximized when candidate c offers a platform that exactly matches her preferred policy positions, and the loss she obtains is a function of the distance between the candidate's policies and her ideal platform. Her utilities from the Candidate 1 and 2's respective platforms is given by:

$$
\begin{aligned}
U_i(\mathbf{x}_{ij1}) &= -\left(b_i - \mathbf{x}_{ij1}\right)^2 + \eta_{ij1} \\
U_i(\mathbf{x}_{ij2}) &= -\left(b_i - \mathbf{x}_{ij2}\right)^2 + \eta_{ij2}
\end{aligned} \tag{1}
$$

While the imposition of quadratic loss utilities may seem restrictive, in the appendix we show that our results are numerically identical if we assume an absolute linear loss utility function. Regardless, it follows that:

$$
\begin{aligned}
\Pr(y_{ij1} = 1) &= \Pr(U_i(\mathbf{x}_{ij1}) > U_i(\mathbf{x}_{ij2})) \\
&= \Pr(-(b_i - \mathbf{x}_{ij1})^2 + \eta_{ij1} > -(b_i - \mathbf{x}_{ij2})^2 + \eta_{ij2}) \\
&= \Pr(\eta_{ij2} - \eta_{ij1} < 2(b_i'(\mathbf{x}_{ij1} - \mathbf{x}_{ij2}) + \mathbf{x}_{ij2}'\mathbf{x}_{ij2} - \mathbf{x}_{ij1}'\mathbf{x}_{ij1})
\end{aligned} \tag{2}
$$

where y_{ij1} is a binary indicator that equals 1 when respondent i chooses Candidate 1 in contest j and 0 otherwise. Now consider data generated from a conjoint experiment, where $\mathbf{x}_{ij1}$ and $\mathbf{x}_{ij2}$ are vectors of randomized candidate attributes that have been discretized into binary indicators with an omitted category.

Typically, we would estimate Equation 2 with a probit or logit-like regression. Instead consider a linear model of the form:

$$
\begin{aligned}
y_{ij1} &= 2(b_i'(\mathbf{x}_{ij1} - \mathbf{x}_{ij2}) + \mathbf{x}_{ij2}'\mathbf{x}_{ij2} - \mathbf{x}_{ij1}'\mathbf{x}_{ij1}) + \eta_{ij1} - \eta_{ij2} \\
&= \sum_k \left(2b_{im}(x_{ijm1} - x_{ijm2}) + x_{ijm2}^2 - x_{ijm1}^2\right) + \eta_{ij1} - \eta_{ij2} \\
&= \sum_k (2b_{im} - 1)(x_{ijm1} - x_{ijm2}) + \eta_{ij1} - \eta_{ij2} \\
&= \sum_k \beta_{im}\Delta x_{ijm} + \epsilon_{ij}
\end{aligned} \tag{3}
$$

where $\mathbb{E}(\epsilon_{ij}) = \mathbb{E}(\eta_{ij1} - \eta_{ij2}) = 0$ follows from the randomization of $\mathbf{x}_{ij1}$ and $\mathbf{x}_{ij2}$, and the third line follows from the fact that $x_{ijmc}^2 = x_{ijmc}$, as this is a dummy. The slope, $\beta_{im} = 2b_{im} - 1$, gives the change in probability for individual i of choosing Candidate 1 when Candidate 1 has feature m and Candidate 2 does not, holding all their other features constant. Implicitly, it also constrains each element of b_i to the $[0, 1]$ line. When $b_{im} = 0$ (and $\beta_{im} = -1$) the manipulation $\Delta x_{ijm} = 1$ holding all other features constant gives a predicted reduction in the probability of choosing Candidate 1 of one-hundred percent. When $b_{im} = 1$ (and $\beta_{im} = 1$), the same manipulation gives a predicted increase in the probability of choosing Candidate 1 of one-hundred percent. When $b_{im} = \frac{1}{2}$ (and $\beta_{im} = 0$), this indicates that voter i is perfectly indifferent.

Finally, averaging over all individuals, we obtain $\mathbb{E}(\beta_{im})$ as the coefficient from the regression:

$$
y_{ij1} = \sum_m \Delta x_{ijm}\beta_m + \epsilon_{ij} \tag{4}
$$

where the estimated coefficient $\hat{\beta}_m$ recovers the AMCE for feature m.[9] Thus we see that, under this simple model of choice, the AMCE can be interpreted as an average of respondents' ideal points. This insight illuminates why the AMCE is such an inappropriate summary statistic for making claims about winners of elections or representative voters' preferences. Under majority rule, elections are won by the median voter, and the magnitudes of the ideal points of the most extreme voters should do nothing to change the probability of a given candidate winning the election. Measuring the probability of winning as a function of preference intensity essentially gives citizens voting power

[9]For a simple proof, see the supplemental appendix.

commensurate with the strength of their opinions — a feature almost never observed in real-life institutional designs.[10]

IV. Conclusion

In this paper we have shown that the AMCE, the target estimand of most conjoint experiments, does not support most interpretations made by political scientists. A positive AMCE for a particular candidate-feature does not imply that the majority of respondents prefer that feature over the baseline. It does not indicate that they prefer a candidate with that feature to a candidate without it, all else equal. It does not mean that voters are more likely to elect a candidate with that feature than candidates without it. None of this is the consequence of uncertainty introduced by sampling or measurement; rather, it is inherent to the AMCE's properties as an aggregation mechanism. Even when the universe of respondents is fully observed and every conceivable contest between candidates is assessed carefully and honestly, claims about voter preferences and electoral outcomes are not generally supported by the results from conjoint experiments.

Instead, what we have demonstrated is that the AMCE can be thought of as an average of the direction and intensity of voters' preferences, or essentially an average of ideal points. As a consequence, it can point in the opposite direction as the majority preference when there is a minority that intensely prefers a feature and a majority that feels the opposite, but less strongly. The larger the correlation between direction and intensity, the more misleading the AMCE. Far from a statistical accident, this preference structure pervades the sorts of issues that interest political scientists (and for which conjoints are often deployed), such as gender parity in elected office (Teele, Kalla and Rosenbluth, 2018) or the sorts of people who should be favored by the nation's immigration policy (Hainmueller and Hopkins, 2015).

Building on well-known results from the literature on social choice, we have characterized the AMCE as a preference aggregation mechanism and shown its relationship to the Borda rule. Using this correspondence, we then derived sharp bounds on the proportion of a sample that prefers a feature based on a given AMCE. Unfortunately, the vast majority of findings published in the top

[10]It is true that voters with intense preferences may be more likely to turn out to vote or to be politically active, and may therefore exercise outsized influence on electoral outcomes. However, these are not the mechanisms that researchers are currently thinking about when interpreting the results of conjoint experiments, and if we think they are at work, we should study and model them more explicitly.

political science journals in the past few years fail to support claims about majority preferences. That said, we have also shown that if one is willing to assume homogeneity in preference intensity, then at the very least the sign of the AMCE from these experiments corresponds to the will of the majority. Since this assumption is unlikely to hold, the discipline must reevaluate what we have learned from conjoint experiments with this clearer understanding of the AMCE in mind. We do not know whether most voters prefer male or female candidates; we have only learned that the "average preference" for women is positive (Teele, Kalla and Rosenbluth, 2018). We have no idea what features of immigrants are popular with American voters; we can only say what characteristics evoke positive and negative "average reactions" (Hainmueller and Hopkins, 2015).

A simple corrective, which we strongly encourage, is for applied researchers to use precise language when interpreting the results of conjoint experiments, placing the "representative voter" implied by the AMCE in the appropriate context. While common interpretations such as "voters prefer A to A'" are not well-defined,[11] typically researchers interpret conjoint results to evoke some notion of a majority — one that is not supported by the target estimand of these experiments. By the same token, political scientists should, on the whole, stop making inferences about electoral contests from the AMCE unless these claims are supported by further evidence about the homogeneity of voters' priorities.

Nevertheless, if researchers must rely upon forced-choice conjoint experiments, our paper suggests they may find themselves in a bind. On the one hand, our results indicate that if they want to interpret their findings with respect to a majority or plurality preference, then researchers should restrict themselves to conservative randomization schemes that limit the number of attributes and potential candidate-profiles. Only with a conservative design and a small number of binary attributes is there hope of producing sufficiently small bounds on an estimated AMCE to conclusively reflect a majority preference. On the other hand, because the AMCE violates the independence of irrelevant alternatives axiom, for an experimental result to be externally valid researchers must include the full set of theoretically relevant attributes in their randomization scheme. That is, for a conjoint experiment to provide substantively relevant results and, moreover, for it to recover point estimates that are stable with respect to the inclusion of additional features, researchers must get the distri-

[11] Does "voters prefer" mean all voters? A subset of voters? If so, what subset?

bution of randomized attributes exactly right. Unfortunately, it may prove difficult to construct a "Goldilocks" experimental design that both randomizes a conservative number of features — to enable researchers' claims about a majority preference — and includes a sufficiently large number of attributes — to be assured that results are not sensitive to IIA violations.

If researchers want to make claims about majority preferences from conjoint experiments, one potential way forward may be to combine them with experiments designed to recover voters' priorities. As we have shown in Corollary 1, if respondents have homogeneous weights on the dimensions of choice, claims about a majority preference can be sustained with existing research designs. However, this may not be a fruitful avenue since the likelihood of homogeneous priorities in realistic political contexts is limited. Instead, we suggest that researchers be willing to trade off stronger assumptions with an ability to make claims about electoral outcomes. A fully structural approach to conjoint analysis may prove most capable of combining the realistic approximations of candidates that randomizing a large number of candidate-features provides with an ability to make claims about electoral contests. By imposing and estimating a model of voter choice, researchers may be able to have their cake and eat it too.

REFERENCES

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

Austen-Smith, David, and Jeffrey S Banks. 2000. *Positive Political Theory I: Collective Preference.* Vol. I, University of Michigan Press.

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

Carnes, Nicholas, and Noam Lupu. 2016. "Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class." *American Political Science Review*, 110(4): 832–844.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2019. "Experimental Design and Statistical Inference for Conjoint Analysis: The Essential Role of Population Distribution." Working Paper.

Gibbard, Allan. 1973. "Manipulation of voting schemes: a general result." *Econometrica*, 41(4): 587–601.

Hainmueller, Jens, and Daniel J Hopkins. 2015. "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science*, 59(3): 529–548.

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

Hemker, Johannes, and Anselm Rink. 2017. "Multiple dimensions of bureaucratic discrimination: Evidence from German welfare offices." *American Journal of Political Science*, 61(4): 786–803.

Horiuchi, Yusaku, Daniel M Smith, and Teppei Yamamoto. 2018. "Identifying voter preferences for politicians' personal attributes: A conjoint experiment in Japan." *Political Science Research and Methods*, 1–17.

Khanna, Kabir. 2019. "What traits are Democrats prioritizing in 2020 candidates?" *CBS News*. https://www.cbsnews.com/news/democratic-voters-hungry-for-women-and-people-of-color-in-2020-nomination/.

Lalley, Steven P, and E Glen Weyl. 2018. "Quadratic voting: How mechanism design can radicalize democracy." *AEA Papers and Proceedings*, 108: 33–37.

Magni, Gabriele, and Andrew Reynolds. 2019. "Democrats Don't Want to Nominate a Candidate Who Looks Like Bernie or Joe." *POLITICO*. https://www.politico.com/magazine/story/2019/05/24/democrats-dont-want-to-nominate-another-white-man-for-president-226977.

Mummolo, Jonathan. 2016. "News from the other side: How topic relevance limits the prevalence of partisan selective exposure." *The Journal of Politics*, 78(3): 763–773.

Newman, Benjamin J., and Neil Malhotra. 2018. "Economic Reasoning with a Racial Hue: Is the Immigration Consensus Purely Race Neutral?" *The Journal of Politics*, 81(1): 153–166.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review*, 112(3): 525–541.

Appendix

PROOF OF LEMMA 1:

Suppose there are N voters and K profiles. Consider voter i's preference ranking over profiles. For any x_j, x_k, denote by $Y_i(x_j, x_k) = 1$ if i chooses profile x_j over x_k in a pairwise comparison, and $Y_i(x_j, x_k) = 0$ otherwise. Without loss of generality, reorder the profiles such that the profile most preferred by i is x_1, the second most preferred is x_2, and so on such that the least preferred is x_K. Assign i's most preferred profile a Borda score of $b_i(x_1) = K - 1$, their second most preferred profile a score of $b_i(x_2) = K - 2$, and so on such that their least preferred profile has a score of zero. Notice that when i is presented with each pairwise comparison, their most preferred profile x_1 will be chosen every time it is on the ballot, so

$$\sum_{j \neq 1} Y_i(x_1, x_j) = \underbrace{1 + 1 + 1 + ... + 1}_{K-1 \text{ times}} = K - 1$$

times. The second most preferred will be chosen each pairing except with the most preferred profile, so

$$\sum_{j \neq 2} Y_i(x_2, x_j) = 0 + \underbrace{1 + 1 + 1 + ... + 1}_{K-2 \text{ times}} = K - 2$$

times. Going this way, we see that individual Borda scores over profiles match perfectly with the number of times each profile is chosen in pairwise comparisons. Finally, the least preferred profile will never be chosen in every pairwise comparison made by voter i, $\sum_{j \neq K} Y_i(x_K, x_j) = 0+0+0+...+0 = 0$. Thus, for each individual voter, the Borda score of a profile is equal to the number of times it is chosen when that voter makes all pairwise comparisons, $b_i(x_m) = \sum_{j \neq m} Y_i(x_m, x_j)$.

The aggregate Borda score of a profile is the sum of individual voters' Borda scores of that profile. When we sum across voters the times each profile x_m is chosen in all pairwise comparisons, their sums must be equal to the sum of individual Borda scores. Formally,

$$b(x_m) \equiv \sum_{i \in N} b_i(x_m) = \sum_{i \in N} \sum_{j \neq m} Y_i(x_m, x_j)$$

■

PROOF OF LEMMA 2:

Before we prove this lemma, let us first state the formal definition of no interactions. Voter i's preferences have no interactions when for all $t_1, t_0, \alpha, \beta, \gamma, \alpha', \beta'$, and γ', we have

$$t_1\alpha\beta\gamma \succ t_0\alpha\beta\gamma \Leftrightarrow t_1\alpha'\beta'\gamma' \succ t_0\alpha'\beta'\gamma'.$$

were α, β, and γ represent some other relevant features of a candidate.

Recall next that we defined the Borda score of a feature as the total number of times all the profiles with that feature are chosen in all pairwise comparisons. Formally, let Borda score of a feature t_1, $B(t_1)$ be

$$B(t_1) \equiv \sum_{i\in N} \sum_{x_1\in\kappa(t_1)} \sum_{x_j\neq x_1} Y_i(x_1, x_j)$$

where $\kappa(t_1)$ denotes the set of all profiles that have the feature t_1.

No interactions implies

$$b_i(t_1\alpha\beta\gamma) - b_i(t_1\alpha'\beta'\gamma') = b_i(t_0\alpha\beta\gamma) - b_i(t_0\alpha'\beta'\gamma')$$

for all $t_1, t_0, \alpha, \beta, \gamma, \alpha', \beta'$, and γ' by a straightforward application of Lemma 1. Then, summing these up, we observe

$$\sum_{i\in N} b_i(t_1\alpha\beta\gamma) - \sum_{i\in N} b_i(t_0\alpha\beta\gamma) = \sum_{i\in N} b_i(t_1\alpha'\beta'\gamma') - \sum_{i\in N} b_i(t_0\alpha'\beta'\gamma')$$

Suppose now $t_1\alpha\beta\gamma$ is the profile with the highest Borda score. It follows that for all $\alpha, \beta, \gamma, \alpha', \beta'$, and γ' we have

$$\sum_{i\in N} b_i(t_1\alpha\beta\gamma) - \sum_{i\in N} b_i(t_0\alpha\beta\gamma) \geq 0 \qquad \text{and} \qquad \sum_{i\in N} b_i(t_1\alpha'\beta'\gamma') - \sum_{i\in N} b_i(t_0\alpha'\beta'\gamma') \geq 0$$

This implies

$$B_i(t_1) = \sum_{\alpha\beta\gamma} \sum_{i\in N} b_i(t_1\alpha\beta\gamma) \geq \sum_{\alpha\beta\gamma} \sum_{i\in N} b_i(t_0\alpha\beta\gamma) = B_i(t_0).$$

PROOF OF PROPOSITION 1:

Recall the definition of the Borda score of a feature $B(t_1)$ from Lemma 2. Dividing $B(t_1)$ by the total number of pairwise comparisons t_1 appears in, $|\kappa(t_1)|N(K-1)$, and taking the difference with the Borda score of the benchmark attribute t_0, divided by $|\kappa(t_0)|N(K-1)$ yields exactly AMCE as defined in Hainmueller et al (2014):

$$\pi(t_1,t_0) \equiv \frac{B(t_1)}{|\kappa(t_1)|N(K-1)} - \frac{B(t_0)}{|\kappa(t_0)|N(K-1)} = \frac{\sum_{i\in N}\sum_{x_1\in\kappa(t_1)}\sum_{x_j\neq x_1} Y_i(x_1,x_j)}{|\kappa(t_1)|N(K-1)} - \frac{\sum_{i\in N}\sum_{x_0\in\kappa(t_0)}\sum_{x_j\neq x_0} Y_i(x_0,x_j)}{|\kappa(t_0)|N(K-1)}$$

■

PROOF OF PROPOSITION 2:

Since we have already established the equivalence of Borda and AMCE in Proposition 1, we prove this proposition by finding the range of Borda scores of t_1 and t_0 that can be rationalized for some proportion of voters who prefer t_1 over t_0; and then inverting to find the minimum and maximum proportions for a given AMCE.

Let us find the minimum fraction of voters who prefer t_1 over t_0 that is consistent with an AMCE estimate. Notice first that for a fixed fraction of voters, AMCE is maximized when voters in favor of t_1 assign the highest priority to the attribute, they rank t_1 the best, and t_0 the worst; whereas those who prefer t_0 like t_1 second, and assign the lowest priority to it. In other words, when those who prefer t_1 rank all profiles with t_1 at the top, and all profiles with t_0 at the bottom, this drives the AMCE estimate up. To help with the intuition, the preferences of such a voter might look like:

$$\underbrace{t_1\alpha\beta\gamma}_{K-1} \succ \underbrace{t_1\alpha'\beta\gamma}_{K-2} \succ \ldots \succ \underbrace{t_1\alpha'\beta'\gamma'}_{K-\frac{K}{\tau}} \succ t_2\alpha\beta\gamma \succ \ldots \succ t_2\alpha'\beta'\gamma' \succ \ldots \succ \underbrace{t_0\alpha\beta\gamma}_{\frac{K}{\tau}-1} \succ \underbrace{t_0\alpha'\beta\gamma}_{\frac{K}{\tau}-2} \succ \ldots \succ \underbrace{t_0\alpha'\beta'\gamma'}_{0}$$

were α, β, and γ represent some other relevant features of a candidate. Holding constant the other features, the difference in Borda scores of a profile with t_1 and with t_0 is thus $K-\frac{K}{\tau}$. Formally, the maximum difference $b_i(t_1,x)-b_i(t_0,x) = K-\frac{K}{\tau}$, for any arbitrary combination of other attributes, x. Since each voter makes $\frac{K}{\tau}$ such comparisons between t_1 and t_0, each voter who prefers t_1 maximally generates $\frac{K^2(\tau-1)}{\tau^2}$ scores in favor of t_1.

Similarly, when those who prefer t_0 assign the lowest priority to this attribute, their preferences

might look like:

$$\underbrace{t_0\alpha\beta\gamma}_{K-1} \succ \underbrace{t_1\alpha\beta\gamma}_{K-2} \succ t_2\alpha\beta\gamma \succ \ldots \succ \underbrace{t_0\alpha'\beta\gamma}_{K-\tau-1} \succ \underbrace{t_1\alpha'\beta\gamma}_{K-\tau-2} \succ t_2\alpha'\beta\gamma \succ \ldots \succ \underbrace{t_0\alpha'\beta'\gamma'}_{\tau-1} \succ \underbrace{t_1\alpha'\beta'\gamma'}_{\tau-2} \succ t_2\alpha'\beta'\gamma' \succ \ldots$$

By holding constant the other features, the difference in Borda scores of a profile with t_1 and with t_0 is -1. For these voters, the maximum difference is $b_j(t_1, x) - b_j(t_0, x) = -1$, for any arbitrary combination of other attributes, x. Therefore, each voter who prefers t_0 maximally generates $-\frac{K}{\tau}$ scores in favor of t_1.

Thus, for a given AMCE $\pi(t_1, t_0)$, we can derive the minimum fraction y of voters who prefer t_1 by summing these scores and normalizing.

$$\pi(t_1, t_0) = \frac{(y^{\min})\frac{K^2(\tau-1)}{\tau^2} - (1 - y^{\min})\frac{K}{\tau}}{\binom{K}{2}\frac{2}{\tau}}$$

Simple algebra reveals

$$y^{\min} = \max\left\{\frac{\pi(t_1, t_0)\tau(K-1) + \tau}{K(\tau-1) + \tau}, 0\right\}$$

A very similar argument establishes the upper bound of y.

■

PROPOSITION 2 (INTERACTIVE CASE):

The bounds established above also correspond to the bounds on the quantity:

$$\Pi(t_1, t_0) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{K/2}\sum_{j=1}^{K/2} Y_i(x_{j1}, x_{j0})\right\} \tag{A.11}$$

in the case allowing for interactive preferences, e.g. a voter who prefers men to women when they are Republicans but women to men when they are Democrats. This is because the additional possibilities created by allowing for interactions are all interior to the bounds; in other words, the extreme cases that establish the bounds are the same whether or not we allow for interactions.

To see this, first note that if there are no interactions, an individual $\Pi_i(t_1, t_0)$ collapses to 1 or

0, corresponding to whether she prefers t_1 or t_0, respectively. In that case, $\Pi(t_1, t_0)$ is simply the proportion of the population that prefers t_1 to t_0, our original bounded quantity. So we must only make sure that introducing the possibility of interactions does not expand the bounds derived above. Note, again, that AMCE is maximized when voters in favor of t_1 assign the highest priority to this attribute, whereas those who prefer t_0 assign it the lowest priority, so that they rank all profiles with t_1 at the top and all profiles with t_0 at the bottom. This can only happen when there are no interactions with the attribute of interest, i.e. when no profile having t_0 can be ranked above any profile having t_1.

ELECTORAL ADVANTAGE IN INTERACTIVE AND NON-INTERACTIVE SETTINGS:

While the quantity $\Pi(t_1, t_0)$ conforms to the derived bounds when we allow for interactions, it is no longer necessarily indicative of an electoral advantage for t_1 over t_0 under this more complex preference structure. We define **electoral advantage** of t_1 over t_0 as the difference between the proportion of the time t_1 beats t_0 in an all-else-equal contest, over all possible all-else-equal contests, and one-half:

$$A(t_1, t_0) = \frac{1}{K/2} \sum_{j=1}^{K/2} \mathbb{1} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_i(x_{j1}, x_{j0}) > \frac{1}{2} \right\} - \frac{1}{2} \tag{A.12}$$

First consider the baseline case with no interactions. Here, note that whenever a majority of voters prefers t_1 to t_0, x_{j1} will beat x_{j0} in any j, and $A(t_1, t_0)$ will be $\frac{1}{2}$. Thus, without interactions, whenever $\Pi(t_1, t_0) > \frac{1}{2}$, we can be confident that t_1 carries an electoral advantage.

This is no longer the case when we allow for interactions. To see this, recall that:

$$\begin{aligned} \Pi(t_1, t_0) &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{K/2} \sum_{j=1}^{K/2} Y_i(x_{j1}, x_{j0}) \right\} \\ &= \frac{1}{K/2} \sum_{j=1}^{K/2} \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(x_{j1}, x_{j0}) \right) \end{aligned} \tag{A.13}$$

and define the **vote share** won by t_1 over t_0 in contest j as $v_{j10} = \frac{1}{N} \sum_{i=1}^{N} Y_i(x_{j1}, x_{j0})$. So, $\Pi(t_1, t_0) > \frac{1}{2}$ when $\bar{v}_{10} > \frac{1}{2}$.

We can easily construct a case where $\bar{v}_{10} > \frac{1}{2}$, but $\frac{1}{K/2}\sum_{j=1}^{K/2} \mathbb{1}\left\{v_{j10} > \frac{1}{2}\right\} < \frac{1}{2}$, yielding an electoral disadvantage. Consider, for example, the following set of outcomes with three all-else-equal races:

$$\begin{aligned} v_{110} &= 1 \\ v_{210} &= \frac{1}{4} \\ v_{310} &= \frac{2}{5} \end{aligned} \tag{A.14}$$

In this case, $\bar{v}_{10} = 11/20$, which exceeds $\frac{1}{2}$. However, t_1 only wins $\frac{1}{3}$ of all-else-equal races against t_0, so does not have an electoral advantage.

PROOF OF COROLLARY 1:

When weights are homogeneous, all voters who prefer a feature contribute the same amount of net points to it; whereas others with the opposite preference take away as many net points each. Formally, suppose there is at least one voter i who prefers t_1 to t_0, and swap labels if there is not. Then, for any j who prefers t_0 to t_1, we have that for all combinations of other attributes x; $b_i(t_1, x) - b_i(t_0, x) = b_j(t_0, x) - b_j(t_1, x)$. Each voter makes $\frac{K}{2}$ comparisons involving t_1 and t_0. Therefore, if there are y voters who prefer t_1 to t_0 and $1 - y$ voters who prefer t_0 to t_1, we can write that for any x,

$$B(t_1) - B(t_0) = (b_i(t_1, x) - b_i(t_0, x))(2y - 1)N\frac{K}{2} \tag{A.15}$$

We know from the proof of Proposition 2 that if i prefers t_1 to t_0, the maximum value $b_i(t_1, x) - b_i(t_0, x)$ can take for a binary attribute is $\frac{K}{2}$, which obtains when the weight assigned to t is so high that i prefers any profile with t_1 to any profile without. The minimum value it can take is 1, which obtains when the weight assigned to t is so low that there is no profile that is ranked lower than (t_1, x) but higher than (t_0, x), for any x. Notice further that $b_i(t_1, x) - b_i(t_0, x)$ is monotone increasing in the relative weight of t. This is because profiles with t_1 and some less preferred values on other attributes become preferred to some profiles that have t_0 and better preferred values on

other attributes with lower weights. This drives the ranking of all profiles with t_1 higher, and those with t_0 lower.

Next, recall that $\pi(t_1, t_0) = \frac{B(t_1)}{|\kappa(t_1)|N(K-1)} - \frac{B(t_0)}{|\kappa(t_0)|N(K-1)}$. Combining this with Equation A.15 gives:

$$\pi(t_1, t_0) = (b_i(t_1, x) - b_i(t_0, x))\frac{2y-1}{K-1}$$

Since $b_i(t_1, x) - b_i(t_0, x) > 0$, the sign of $\pi(t_1, t_0)$ is be positive if and only if $y > 1/2$. That is, under homogeneous weights, the AMCE returns a positive estimate for t_1 if and only if there are more people who prefer t_1 to t_0. When the weight assigned to t is highest, we have $\pi(t_1, t_0) = \frac{K(2y-1)}{2(K-1)}$, and so the AMCE corresponds to roughly half the size of the margin. As attributes with higher weights are added, AMCE falls. In particular, when the weight assigned to t is lowest, we have $\pi(t_1, t_0) = \frac{2y-1}{K-1}$. It is clear that for any y, as the number of profiles grows, AMCE goes to zero in the limit.

■

PROOF THAT EQUATION 4 IS EQUIVALENT TO THE AMCE:

To show that the estimation of Equation 4 would yield the AMCE note first that Hainmueller, Hopkins and Yamamoto (2014) show that the following regression recovers an unbiased estimate of the AMCE:

$$y_{ijc} = \delta + x_{jmc}\rho_k + \upsilon_{ijmc}$$

where $\hat{\rho}_m$ gives the AMCE for feature m. From the randomization of x, it follows from standard results that the vector of coefficients β from Equation 4 can be obtained from the separate regression of the outcome y_{ij1} on each column k of the matrix ΔX_{ij}, e.g. $y_{ij1} = \Delta x_{ijm}\beta_m + \epsilon_{ijm}$. It is sufficient to show that $\hat{\rho}_m = \hat{\beta}_m$. The above equation implies $\hat{\rho}_m = \frac{Cov(x_{ijmc}, y_{ijc})}{Var(x_{ijmc})}$. Similarly, estimating Equation 4 via least squares without an intercept implies $\hat{\beta}_m = \frac{\mathbb{E}(\Delta x_{ijm} y_{ij1})}{\mathbb{E}(\Delta x_{ijm}^2)}$. Since $\mathbb{E}(\Delta x_{ijm}) = 0$, it follows that $\hat{\beta}_m = \frac{Cov(x_{ijm1} - x_{ijm2}, y_{ij1})}{Var(x_{ijm1} - x_{ijm2})}$.

Consider the numerator.

$$\begin{aligned} Cov(x_{ijm1} - x_{ijm2}, y_{ij1}) &= Cov(x_{ijm1}, y_{ij1}) - Cov(x_{ijm2}, y_{ij1}) \\ &= Cov(x_{ijm1}, y_{ij1}) - Cov(x_{ijm2}, 1 - y_{ij2}) \\ &= 2Cov(x_{ijmc}, y_{ijmc}) \end{aligned}$$

The last line follows from the fact that $Cov(x_{ijm1}, y_{ij1}) = Cov(x_{ijm2}, y_{ij2})$

Next consider the denominator.

$$\begin{aligned} Var(x_{ijm1} - x_{ijm2}) &= Var(x_{ijm1}) + Var(-x_{ijm2}) - 2Cov(x_{ijm1}, x_{ijm2}) \\ &= 2Var(x_{ijmc}) \end{aligned}$$

Which again follows from the randomization of features. It directly follows that $\hat{\beta}_m = \hat{\rho}_m = AMCE$.

■

PROOF OF THE EQUIVALENCE OF THE QUADRATIC LOSS AND ABSOLUTE LOSS:

$$\begin{aligned} U_i(x_{j1}) &= -|x_{j1} - b_i| + \eta_{ij} \\ U_i(x_{j2}) &= -|x_{j2} - b_i| + \nu_{ij} \end{aligned} \tag{A.16}$$

Assume $0 \leq b_i \leq 1$

$$\begin{aligned} \Pr(y_{ij1} = 1) &= \Pr(U_i(\mathbf{x}_{ij1}) > U_i(\mathbf{x}_{ij2})) \\ &= \Pr\left(\eta ij - \nu_{i2} < |x_{j2} - b_i| - |x_{j1} - b_i|\right) \end{aligned} \tag{A.17}$$

Since x_{j1} & x_{j2} can take on only two values $\{0, 1\}$, it follows $x_{j1} \leq b_i \leq x_{j2}$ or $x_{j2} \leq b_i \leq x_{j1}$ This yields:

$$\Pr(y_{ij1} = 1) = Pr\left(\eta_{j1} - \nu_{j2} < \Delta x_j(2b_i - 1)\right) \tag{A.18}$$

If we were to estimate this via a linear probability model we obtain

$$\begin{aligned} y_{ij1} &= \Delta x_j(2b_i - 1) + \eta_{ij} - \nu_{ij} \\ &= \Delta x_j \beta_i + \epsilon_{ij} \end{aligned} \tag{A.19}$$

PROOF:

Tables For Example Part II

Comparison	V1	V2	V3	V4	V5	Tally
FDB,**FDW**	FDW	FDW	FDW	FDB	FDB	2,3
FDB,**FRB**	FRB	FRB	FRB	FDB	FDB	2,3
FDB,**FRW**	FRW	FRW	FRW	FDB	FDB	2,3
FDB,**MDB**	MDB	MDB	MDB	FDB	FDB	2,3
FDB,**MDW**	MDW	MDW	MDW	FDB	FDB	2,3
FDB,**MRB**	MRB	MRB	MRB	FDB	FDB	2,3
FDB,**MRW**	MRW	MRW	MRW	FDB	FDB	2,3
FDW,**FRB**	FRB	FRB	FRB	FRB	FRB	0,5
FDW,**FRW**	FRW	FRW	FRW	FDW	FDW	2,3
FDW,**MDB**	MDB	MDB	MDB	MDB	MDB	0,5
FDW,**MDW**	MDW	MDW	MDW	FDW	FDW	2,3
FDW,**MRB**	MRB	MRB	MRB	FDW	FDW	2,3
FDW,**MRW**	MRW	MRW	MRW	FDW	FDW	2,3
FRB,**FRW**	FRW	FRW	FRW	FRB	FRB	2,3
FRB,MDB	FRB	FRB	FRB	FRB	FRB	5,0
FRB,**MDW**	MDW	MDW	MDW	FRB	FRB	2,3
FRB,**MRB**	MRB	MRB	MRB	FRB	FRB	2,3
FRB,**MRW**	MRW	MRW	MRW	FRB	FRB	2,3
FRW,MDB	FRW	FRW	FRW	MDB	MDB	3,2
FRW,MDW	FRW	FRW	FRW	FRW	FRW	5,0
FRW,**MRB**	MRB	MRB	MRB	MRB	MRB	0,5
FRW,**MRW**	MRW	MRW	MRW	FRW	FRW	2,3
MDB,**MDW**	MDW	MDW	MDW	MDB	MDB	2,3
MDB,**MRB**	MRB	MRB	MRB	MDB	MDB	2,3
MDB,**MRW**	MRW	MRW	MRW	MDB	MDB	2,3
MDW,**MRB**	MRB	MRB	MRB	MRB	MRB	0,5
MDW,**MRW**	MRW	MRW	MRW	MDW	MDW	2,3
MRB,**MRW**	MRW	MRW	MRW	MRB	MRB	2,3

Table A1—: Aggregate preferences over candidate profiles - Example Part II

1.		2.		
$\bar{Y}(MDB, MDB)$	$-$	$\bar{Y}(FDB, MDB)$	$=$	$1/10$
$\bar{Y}(MDB, FDB)$	$-$	$\bar{Y}(FDB, FDB)$	$=$	$1/10$
$\bar{Y}(MDB, MRB)$	$-$	$\bar{Y}(FDB, MRB)$	$=$	0
$\bar{Y}(MDB, FRB)$	$-$	$\bar{Y}(FDB, FRB)$	$=$	$-2/5$
$\bar{Y}(MDB, MDW)$	$-$	$\bar{Y}(FDB, MDW)$	$=$	0
$\bar{Y}(MDB, FDW)$	$-$	$\bar{Y}(FDB, FDW)$	$=$	$3/5$
$\bar{Y}(MDB, MRW)$	$-$	$\bar{Y}(FDB, MRW)$	$=$	0
$\bar{Y}(MDB, FRW)$	$-$	$\bar{Y}(FDB, FRW)$	$=$	0
$\bar{Y}(MRB, MDB)$	$-$	$\bar{Y}(FRB, MDB)$	$=$	$-2/5$
$\bar{Y}(MRB, FDB)$	$-$	$\bar{Y}(FRB, FDB)$	$=$	0
$\bar{Y}(MRB, MRB)$	$-$	$\bar{Y}(FRB, MRB)$	$=$	$1/10$
$\bar{Y}(MRB, FRB)$	$-$	$\bar{Y}(FRB, FRB)$	$=$	$1/10$
$\bar{Y}(MRB, MDW)$	$-$	$\bar{Y}(FRB, MDW)$	$=$	$3/10$
$\bar{Y}(MRB, FDW)$	$-$	$\bar{Y}(FRB, FDW)$	$=$	$-2/5$
$\bar{Y}(MRB, MRW)$	$-$	$\bar{Y}(FRB, MRW)$	$=$	0
$\bar{Y}(MRB, FRW)$	$-$	$\bar{Y}(FRB, FRW)$	$=$	$3/5$
$\bar{Y}(MDW, MDB)$	$-$	$\bar{Y}(FDW, MDB)$	$=$	$3/5$
$\bar{Y}(MDW, FDB)$	$-$	$\bar{Y}(FDW, FDB)$	$=$	0
$\bar{Y}(MDW, MRB)$	$-$	$\bar{Y}(FDW, MRB)$	$=$	$-2/10$
$\bar{Y}(MDW, FRB)$	$-$	$\bar{Y}(FDW, FRB)$	$=$	$3/5$
$\bar{Y}(MDW, MDW)$	$-$	$\bar{Y}(FDW, MDW)$	$=$	$1/10$
$\bar{Y}(MDW, FDW)$	$-$	$\bar{Y}(FDW, FDW)$	$=$	$1/10$
$\bar{Y}(MDW, MRW)$	$-$	$\bar{Y}(FDW, MRW)$	$=$	0
$\bar{Y}(MDW, FRW)$	$-$	$\bar{Y}(FDW, FRW)$	$=$	$-2/5$
$\bar{Y}(MRW, MDB)$	$-$	$\bar{Y}(FRW, MDB)$	$=$	0
$\bar{Y}(MRW, FDB)$	$-$	$\bar{Y}(FRW, FDB)$	$=$	0
$\bar{Y}(MRW, MRB)$	$-$	$\bar{Y}(FRW, MRB)$	$=$	$3/5$
$\bar{Y}(MRW, FRB)$	$-$	$\bar{Y}(FRW, FRB)$	$=$	0
$\bar{Y}(MRW, MDW)$	$-$	$\bar{Y}(FRW, MDW)$	$=$	$-2/5$
$\bar{Y}(MRW, FDW)$	$-$	$\bar{Y}(FRW, FDW)$	$=$	0
$\bar{Y}(MRW, MRW)$	$-$	$\bar{Y}(FRW, MRW)$	$=$	$1/10$
$\bar{Y}(MRW, FRW)$	$-$	$\bar{Y}(FRW, FRW)$	$=$	$1/10$
				2
(# of profiles $-$ 1) $\times$ (# of features -1) $\times$# of values for gender			$=$	28
AMCE			$=$	$1/14$

Table A2—: Obtaining the AMCE - Example II

AN EXAMPLE OF THE AMCE VIOLATING IIA:

Consider three types of voters with preferences over three candidate-features, Gender (M or F), Age (O or Y), and Race (B or W). Preferences over features are given in Table A3.

V1	V2	V3
$M \succ F$	$F \succ M$	$F \succ M$
$O \succ Y$	$Y \succ O$	$Y \succ O$
$B \succ W$	$B \succ W$	$W \succ B$

Table A3—: Preferences over attributes

Assume priorities over features as follows. V1: $R \succsim A \succsim G$; V2: $A \succsim R \succsim G$; V3: $A \succsim G \succsim R$. With this information we can construct preferences over candidates for each type as presented in Table A4.

Rank	V1	V2	V3
1.	MOB	FYB	FYW
2.	FOB	MYB	FYB
3.	MYB	FYW	MYW
4.	MOW	FOB	FOW
5.	FYB	MYW	MYB
6.	FOW	MOB	FOB
7.	MYW	FOW	MOW
8.	FYW	MOW	MOB

Table A4—: Preferences over attributes

Consider a population of five V1s, two V2s, and two V3s. Table A5 gives the AMCE estimate with the full set of candidate-features and then restricting each combination of Age and Race. We see that the sign flips when we omit either OB and YW, indicating that the AMCE for Male is dependent upon the other feature-combinations, violating IIA.

Omitted Features	O	Y
B	1/168	-4/189
W	-4/189	1/168

No Omitted Features: -1/126

Table A5—: AMCE Estimates of Male, restricting Age-Race feature combinations