What Do We Learn About Voter Preferences From Conjoint Experiments?

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

Political scientists frequently interpret the results of conjoint experiments as reflective of majority 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, the AMCE can indicate the opposite of the true preference of the majority. To show this, we characterize the preference aggregation rule implied by the AMCE and demonstrate its several undesirable properties. With this result we provide a method for placing bounds on the proportion of experimental subjects who prefer a given candidate-feature. We describe conditions under which 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: 8,319

Verification Materials:

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/DR0YF2.

* Abramson: Associate Professor, Department of Political Science, University of Rochester, email: sabramso@ur.rochester.edu; Kocak: Assistant Professor, Division of Social Science, New York University Abu Dhabi, email: kkocak@nyu.edu; Magazinnik: Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, email: asyam@mit.edu. Kocak gratefully acknowledges the support of the Research Program in Political Economy at Princeton University. The authors thank Naoki Egami, Matias Iaryczower, Kosuke Imai, John Londregan, Nolan McCarty, Teppei Yamamoto, and seminar audiences at Harvard, NYUAD, Princeton, Rochester, UCL, UdeM, the APSA Annual Meeting, the 2019 Conference of the Society for Political Methodology, and the 2019 Toronto Political Behavior Workshop for useful comments and encouragement.

I. Introduction

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 majority preferences and electoral outcomes. In this paper we show that the target estimand of conjoint experiments, the average marginal component effect (AMCE), does not typically support such claims. 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'). In so doing, it assigns greater weight to voters who intensely prefer a particular outcome.

This paper clarifies the connection between the AMCE and the substantive quantities of interest that political scientists frequently seek to recover in preference-elicitation experiments. First, we illustrate by way of a simple example how the AMCE aggregates individual choices. In so doing, we show how it can prove misleading for identifying proportions of voters who favor particular features—an inference researchers implicitly make when they summarize population preferences ("Americans prefer highly-educated immigrants"), characterize electoral majorities ("voters want a lower tax rate"), or project winners of elections ("the Democratic party would be well-served to nominate a female candidate").

Having established the substantive claims that it cannot support, we turn to an analysis of three valid interpretations of the AMCE. The first is as a change in expected vote share. We advise caution for researchers who wish to proceed with this interpretation. First, the average vote share does not imply other, more intuitive measures of electoral advantage. We show through our running example that there can exist underlying preference distributions that produce an average vote share that favors A over A' where, nevertheless, A' beats A in the vast preponderance of elections. Far from being a statistical artifact, preference distributions of this form are pervasive: they reflect populations

in which a minority intensely prefers an alternative, while a majority has a mild preference for its opposite. Second, the average vote share is only valid with respect to the particular randomization scheme defined by the experimenter. This is to say, the averaging implied by this interpretation is over the set of electoral contests between candidates defined internally to the experimental design. So, unless researchers have theoretical reasons to care about the mean election in the particular set of contests their experiment implies, this interpretation is unlikely to prove informative.

Next, we highlight a second valid interpretation of the AMCE characterizing the mapping between it and the Borda rule, a preference aggregation mechanism that picks a winner based on voters' rankings of alternatives. We prove that the AMCE can be used to make statements about winners of Borda-rule elections. Of course, not many real-world contests are decided by this procedure.[1] This leads us to ask what further inferences about the underlying distribution of voter preferences we can draw from the AMCE. In doing so we provide a method that, for an estimated AMCE, allows researchers to place bounds on the proportion of experimental subjects that maintain a strict preference for a candidate-feature. We close this discussion with a sufficient condition under which the AMCE indicates a majority preference: when the direction and intensity of voters' preferences are uncorrelated.

Finally, we explore the relationship between the AMCE and a simple model of choice. In providing this structural foundation for the AMCE, we show that it supports a third interpretation: an average of individual ideal points over candidate-features. Typically, elections are decided by the median voter's ideal point. Although a class of probabilistic voting models does rely on the mean preference to characterize equilibria, the relevance of the mean ideal point in these models depends on a set of strong assumptions. Amongst these is, crucially, that candidates know voters' preferences up to a

[1]Examples include some elections in Slovenia and Kiribati and voting for the Heisman Trophy and Eurovision Song Contest.

random shock. The purpose of conjoint experiments, however, is to uncover exactly these preferences.

Our analysis highlights the importance of placing theoretical structure on the estimands used in applied empirical work. While methods in this literature are often lauded for being "model-free," we emphasize that any estimand that aggregates preferences is by its very nature a social choice rule. Without theories describing the mapping of individual preferences to observed choices, as well as their aggregation, nonparametric estimates of population preferences are difficult to interpret substantively.

II. Invalid Interpretations of the AMCE

The goal of factorial designs like those in forced-choice conjoint experiments is to mimic the complex comparisons faced by real-world decision-makers.[2] By randomizing a large number of candidate and platform features, political scientists aim to construct realistic approximations of the choices voters face. With repeated observations of these randomized features and respondents' choices, the AMCE can be computed via a simple difference-in-means or least-squares regression, and 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, where the expectation is taken over the distribution of the other attributes as well as over respondents.

This quantity is commonly used to make claims about voters' preferences for particular policies,

[2]Throughout, we focus on forced-choice conjoint experiments as the most common implementation in political science. Another popular implementation involves using scales (or thermometers) as the response variable. We are unaware of a microfoundation of choice behavior when responses take a range of values such that it would allow a theoretical exploration similar this paper. This does not imply that our critique only applies to forced-choice conjoints. If, for example, respondents partition the scale such that there is a one-to-one mapping between disjoint ranges of scores and unique candidates, then the results we present for the forced-choice setup carry through exactly.

such as: "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). Conjoint results are also used to make statements about candidates for elected office, such as: "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).[3]

While statements of the form "voters prefer A to A'" have many possible meanings,[4] a reasonable interpretation is that there are more voters who prefer A to A' than vice versa. To make such a statement, it would suffice to say that the median voter prefers A to A'. But the representative voter whose preferences are captured by the AMCE is not the median; it is the average over both the *intensive* and *extensive* margins of choice. Outside of fantastical institutional designs (e.g, Lalley and Weyl (2018)), electoral contests are not typically swayed by *how much* a subset of voters prefer a given candidate; rather, elections are won—and voting populations are most straightforwardly described—by *how many* voters prefer each candidate.

Here, we work through an example that begins with voter preferences, translates those preferences into observed choices, and aggregates those choices to the AMCE. The example is designed to build

[3]We conducted a review of conjoint analyses published in top political science journals and found that 83% of all papers using conjoint experiments make direct reference to voter preferences and 51% interpret their findings in the context of elections. This is described in Table G1 in the supplemental material.

[4]Do researchers mean to say that there exist some voters who prefer A to A'? That most voters prefer A to A'? That all voters prefer A to A'?

intuition around the AMCE's underlying preference aggregation mechanism, and to illustrate how a positive AMCE can be inconsistent with a number of majoritarian claims. Throughout, we aim to make as few assumptions about the underlying preferences of individual voters as possible. While we view the assumptions we make as benign, we note that if the AMCE exhibits undesirable properties under these assumptions, placing even less structure on the problem will not rectify whatever issues we identify and only obscure what drives them. Furthermore, we emphasize that we are agnostic about 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.*[5]

Since researchers who use conjoint experiments seek to characterize preference relations over candidate-features, we define our primitives over this space. For simplicity, consider an electorate of five voters (V1, V2, V3, V4, V5). Candidates possess two attributes that are relevant 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 party, so that there are four different candidate profiles: FD, FR, MD, and MR. The voters' preferences over attributes are given in Table 1. 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.

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

We construct preferences over candidates from preferences over attributes in the following way:

[5]Formally, completeness is defined as $x \succeq y$, $y \succeq x$ or both, and transitivity is defined as if $x \succeq y$ & $y \succeq z$, then $x \succeq z$. We define a *strict preference relation* as $x \succ y$ if and only if $x \succeq y$ and *not* $y \succeq x$ and henceforth refer to this definition when we write "preference." To vastly simplify the presentation, we rule out indifference as is standard in the social choice literature.

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 their preferred party but not gender.

In this simple setting, we can use the weight relation $>>$ to indicate that an attribute is given greater weight in determining a voter's preference ordering. Accordingly, we assume that voters 1, 2, and 3 place more weight on the candidate's party ($P >> G$), whereas voters 4 and 5 place more on the candidate's gender ($G >> P$).[6] 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-head election for every possible pairwise comparison; the winner is bolded in the first column.

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

Next, we derive the AMCE for male over female candidates following Hainmueller, Hopkins and

[6]Note that these *relative* weights are meaningful within individuals but cannot be compared across respondents. That the minority cares more intensely about gender than party does not imply that it cares more intensely about gender than does the majority. The weights therefore cannot speak to which group's turnout or candidate choice will be more influenced by a change along the relevant dimension.

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

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 number of votes candidate C_1 obtains when running 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 (4) times the number of possible profiles with a given gender (2) times the number of voters (5). The procedure yields an AMCE for male equal to $-1/20$, meaning that the average probability of being chosen is higher for female candidates than it is for male candidates.

$\bar{Y}(MR, MD)$	$-$	$\bar{Y}(FR, MD)$	$=$	-2
$\bar{Y}(MR, FD)$	$-$	$\bar{Y}(FR, FD)$	$=$	0
$\bar{Y}(MR, MR)$	$-$	$\bar{Y}(FR, MR)$	$=$	$1/2$
$\bar{Y}(MR, FR)$	$-$	$\bar{Y}(FR, FR)$	$=$	$1/2$
$\bar{Y}(MD, MD)$	$-$	$\bar{Y}(FD, MD)$	$=$	$1/2$
$\bar{Y}(MD, FD)$	$-$	$\bar{Y}(FD, FD)$	$=$	$1/2$
$\bar{Y}(MD, MR)$	$-$	$\bar{Y}(FD, MR)$	$=$	0
$\bar{Y}(MD, FR)$	$-$	$\bar{Y}(FD, FR)$	$=$	-2
				-2
(# of profiles) $\times$ (# of voters) $\times$(# of profiles with a given gender)			$=$	40
AMCE			$=$	$-1/20$

Table 4—Obtaining the AMCE

Our toy example illustrates the intuition driving our main result. Notice that the AMCE for male candidates is *negative* (thus the AMCE for female candidates is positive), and yet the following statements do not hold:

1) *A majority of voters prefer female to male candidates*
   As Table 1 indicates, a majority of voters (3 out of 5) prefer males to females.
2) *A majority of voters prefer female to male candidates, all else equal*
   As Table 2 indicates, fixing party at R, 3 out of 5 voters prefer the male candidate (MR) to the female candidate (FR). The same goes for MD over FD.
3) *Female candidates beat male candidates in the majority of possible head-to-head electoral contests*
   As Table 3 indicates, men win 4 of the 6 possible elections.
4) *Female candidates beat male candidates in the majority of possible all-else-equal head-to-head electoral contests*
   Table 3 also shows that in all-else-equal races (MR vs. FR and MD vs. FD), the male candidate always wins.

The AMCE produces an estimate that indicates the opposite of these majoritarian statements 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, place less weight on gender than party 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.

III. Three Valid Interpretations of the AMCE

A. Expected Vote Share

One might think to interpret the AMCE as the expected change in vote share associated with a given candidate-feature.[7] To see this, note that each row of Table 4 is simply the difference in votes

[7]On this point, see Bansak et al. (2021).

otherwise identical men and women receive in pairs of elections with a fixed opponent. Averaging over the total number of voters and the set of elections defined by the experiment yields the average change in vote share associated with a candidate being male. It is also exactly equivalent to the AMCE of male over female.

While it is correct to interpret the AMCE as a change in expected vote share, doing so runs into the same aggregation problem that we highlighted in our example. The negative change in expected vote share in our example is driven by one landslide election, MD vs. FR, where the female candidate wins 5–0. In all other contests, the female candidate loses—just by a smaller margin. Thus, out-of-sample predictions and claims about the relative electability of specific candidates are no more warranted under this interpretation of the AMCE than any of the others we have discussed.

What is more, the change in expected vote share is defined over the specific set of elections determined by the randomization scheme. The sign and magnitude of the AMCE vary with the attributes included in the experimental design, *holding fixed the experimental subjects and their preferences*. This occurs because the inclusion of a new attribute may change the relative rankings of candidates with respect to the other, previously included attributes.

To see this, consider the same population of five voters as in our previous example. However, instead of conducting an experiment where we randomize only party and gender, we now include a third attribute, 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, then gender, then race ($P >> G >> R$), and let voters 4 and 5 place the greatest weight on race, then gender, then party ($R >> G >> P$). As in the previous section, we can produce a full ranking of candidate profiles using this combination of weights and preferences. 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 they 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 preferred feature. Preferences over candidates are given in Table 5.

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

Table 5—Preferences Over Candidate Profiles—Example Part II

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, men win a large majority of elections.[8] However, in contrast with our previous example, instead of always ranking female candidates above male candidates, voters 4 and 5 will now be willing to accept a man in some contests because they place more weight on race than gender. Since including race changes the relative ranking of male and female candidates, it changes the AMCE researchers would derive from this experiment. Again we calculate the AMCE, yielding 1/16—the exact opposite of the substantive result from the previous experiment where we considered only gender and party.[9]

We have therefore shown that, even with identical subjects, the results researchers obtain from

[8]Male candidates win thirteen of the sixteen elections in which they face off against female candidates, and nineteen of the twenty-eight overall contests.

[9]Trivially, we could add a fourth attribute, and again flip the sign of the AMCE. In Section B of the supplemental materials we provide simple R code to perform this and similar calculations.

conjoint experiments depend upon the specific set of attributes included in their experimental design and thereby the particular set of elections implied by this design. In the next section, we provide further insight into this sensitivity by showing a direct mapping between the AMCE and the Borda rule, which fails to satisfy the independence of irrelevant alternatives axiom (IIA). That is, the Borda winner—and thereby the AMCE—a researcher obtains from a given experiment changes when she removes particular candidates from the contest, for instance when she restricts the randomization to exclude particular feature combinations.[10] In this way, we provide a microfoundation for the results of de la Cuesta, Egami and Imai (2021), who highlight the sensitivity of the AMCE to the randomization scheme imposed by researchers. We show that this is not just a *statistical* property of the AMCE, but a core *theoretical* feature of the aggregation mechanism that generates this quantity.

B. Borda Rule Elections

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). This exercise reveals that the AMCE is closely related to the Borda rule, a voting system that assigns points to candidates according to their order of preference. We build on this result to provide a method that, for a given AMCE estimate, allows researchers to place bounds on the proportion of experimental subjects that maintain a strict preference for a candidate-feature.

Borda rule voting is implemented as follows. With K candidates, the Borda rule assigns zero points to each voter's least preferred candidate, one point to the candidate preferred to that but

[10]In Section B of the supplementary materials we also provide an example of this IIA violation wherein the sign of the AMCE changes depending on which feature-combinations are excluded.

no other, and so on until the most preferred candidate receives $K-1$ points. Thus for each voter, the Borda score contributed to a candidate corresponds to the number of other candidates to whom they are 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 thus 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 allows us to state our first main result that connects the AMCE to the Borda rule:

Proposition 1. *The difference of the Borda scores of two features is proportional to the AMCE.*

The intuition for the proof of Proposition 1 follows from Lemma 1 and the observation that Borda and AMCE 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 is constructed by taking the difference of these tallies and normalizing them. In the appendix we walk through the steps of how to get to AMCE from Borda scores, and produce the same expression as the AMCE in Equation 5 of Hainmueller, Hopkins and Yamamoto (2014).

This connection between the Borda rule and the AMCE is important, because the Borda rule has several undesirable properties that the AMCE inherits—properties that were already revealed in our

initial example. The Borda rule violates the independence of irrelevant alternatives (IIA) criterion, which states that the relative ranking of two candidates should not depend on the inclusion of another candidate. In our example, we showed that the sign of the AMCE of male versus female depends upon whether or not we include race, because the inclusion of race changes the relative ranking of male and female candidates. A second social choice property of the Borda rule is that it violates the majority criterion, which states that if a majority of voters prefer one candidate then that candidate must win. This property also extends to attributes. In our example, we showed that a majority of voters prefer male to female candidates, but the AMCE of male over female is negative. In linking the AMCE to the Borda rule, we have now shown that this violation of the majority criterion is a more general property of the AMCE's underlying preference aggregation mechanism.

The relationship between the AMCE and the Borda rule can usefully be leveraged to derive bounds on the fraction of the population that prefers a feature. That is, for a given AMCE, total number of possible candidate profiles in the experiment, and number of values the attribute of interest can take, we can characterize the maximum and minimum fractions of voters who might prefer that feature over the baseline. Our next result presents these bounds. For simplicity, we assume that preferences are separable, that is, voters have unconditional preferences over candidate features; we discuss what happens when we relax this assumption at the end of this section.[11]

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

---

[11]Formally, voter i's choices are separable when for all t_1 and t_0, we have

$$Y_i\left((t_1, T_{[-l]}), (t_0, T_{[-l]})\right) = Y_i\left((t_1, T'_{[-l]}), (t_0, T'_{[-l]})\right)$$

where $T_{[-l]}$ and $T'_{[-l]}$ denote two arbitrary vectors of other treatment components.

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

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

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

To find these bounds, all we need to calculate are the range of 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. First, we assume that the attribute of interest has the highest possible importance for all supporters of the feature of interest, i.e. the respondents who prefer it over the baseline. For this group, all profiles with the feature of interest are preferred to all profiles without that feature, yielding the highest possible Borda score for the feature of interest and the minimum possible Borda score for the baseline. Thus we obtain the maximum net Borda score a supporter can contribute to a feature.

Second, we assume that the attribute of interest is least important for all opponents of that feature, i.e. the respondents who prefer the baseline. When this is the case the feature of interest will factor into the respondent's choice only if the profiles are otherwise identical. Subject to the constraint that opponents prefer the baseline, this results in the highest possible Borda score for the feature and the lowest for the baseline, yielding the minimum net Borda score an opponent can subtract from a feature. Having calculated the maximum Borda score for a feature per supporter and opponent, we can invoke Proposition 1 to calculate the maximum possible AMCE for a given fraction of opponents and supporters. Inverting this function yields the lowest possible fraction of supporters for a given AMCE. 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. We also provide simple R code to compute these bounds for given values of π, τ, and K in Appendix C.

In Figure 1, 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 AMCEs of a fairly large magnitude, it takes fewer than five possible profiles for these bounds to grow to a range that is inconclusive about the preference of the majority. 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 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 attribute of interest were ternary instead of binary, this would no longer be the case even at an effect size of 0.25.

In Appendix Table C1, we conduct this exercise for every forced-choice conjoint experiment in the top three political science journals 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 eight papers we analyze, only one, Mummolo (2016), produces bounds that guarantee a majority preference. In this paper, the estimated effect is quite large (0.30), the attribute of interest is binary, and the number of possible profiles is the smallest by far of all the included experiments. In Supplementary Appendix C, we demonstrate how researchers can exploit the separability assumption further and use the structure of conjoint data to compute bounds that are guaranteed to be weakly narrower than those given in Proposition 2. However, when we incorporate uncertainty estimates, this approach does not produce sufficiently narrow bounds to change any of the substantive conclusions in Table C1.

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

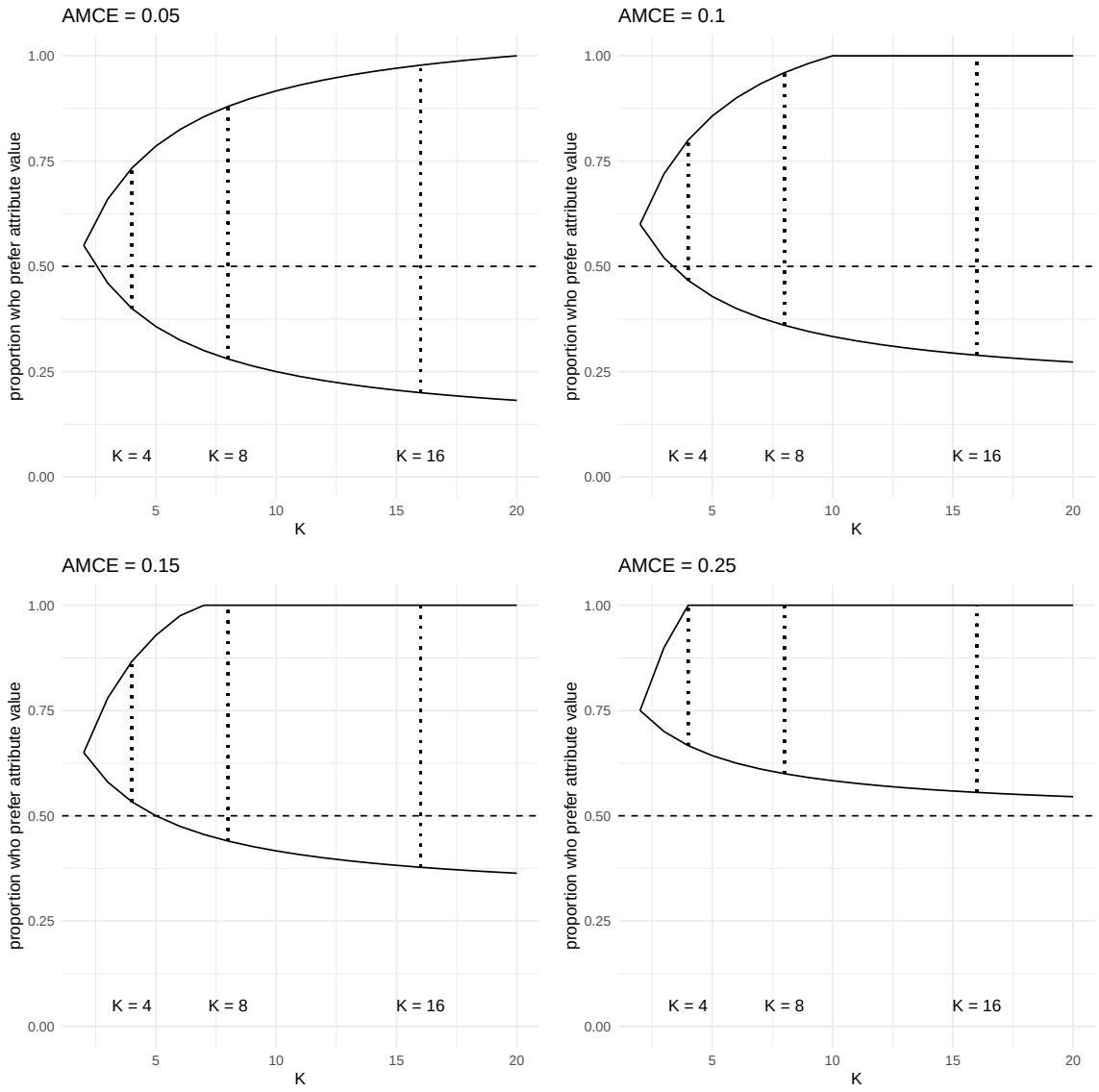

Figure 1. Upper and lower bounds on fraction of people who prefer a feature, consistent with an AMCE of .05, .10, .15, and .25, respectively, as a function of number of possible candidate profiles.

researchers, which is realized when preference direction and intensity are highly correlated. Thus, Proposition 2 underscores the dangers of making statements about aggregate preferences with so little structure on individual choices.

Of course, this worst-case scenario may be unlikely. To show how the correlation between the intensity and direction of preferences relates to the proportion of voters who prefer a given candidate-feature, we work through a toy example with two binary attributes and fifteen voters, where we are interested in the proportion of voters with a preference for men over women. Define the *intensity* of preferences for a feature t_1 over t_0 as the absolute value of the difference between the Borda scores of the two features. The *direction* of preferences is simply a binary indicator for whether voters prefer t_1 over t_0, i.e. $\mathbb{1}\{Y_i(t_1, t_0) = 1\}$. In Figure 2 we plot the Pearson correlation coefficient between direction and intensity of preferences for gender on the x-axis and every possible proportion of the voters that prefer men over women that is consistent a given AMCE on the y-axis, for AMCEs of 0.05 (the left panel) and 0.10 (the right panel).[12] In the left panel, we see that for an AMCE of men over women of 0.05, a correlation of less than 0.4 is required to infer a majority preference for men; for an AMCE of 0.10, all but a correlation of 1 assures that the sign of the AMCE indicates the majority preference. Note, however, that Figure 2 corresponds to the most charitable case, as pictured for $K = 4$ in the bounds in Figure 1. As the number of possible profiles grows to $K = 16$ (only four binary attributes), even small positive correlations can be sufficient to make the AMCE indicate the opposite of the majority preference. Thus, Figure 2 illustrates a general rule of thumb for

[12]Specifically, we generate all the combinations (with replacement) of 15 voters that can be constructed from the eight possible non-interactive preference orderings for the four candidates given in Table 2. We use 15 voters because that number is both informative and computationally feasible, yielding $C^R(8, 15) = 170,544$ combinations to evaluate. For each possible voter set, we compute an AMCE of male over female, a proportion of the sample that prefers males over females, and a correlation of direction and intensity. Figure 2 displays all these possibilities for a given AMCE.

researchers: for a positive (negative) AMCE, a positive (negative) correlation between respondents' direction and intensity vectors may lead to the failure of the AMCE to correspond in sign to the majority preference. Just how strong that correlation must be is a function of where the relevant upper/lower bound is located relative to the 0.5 threshold.

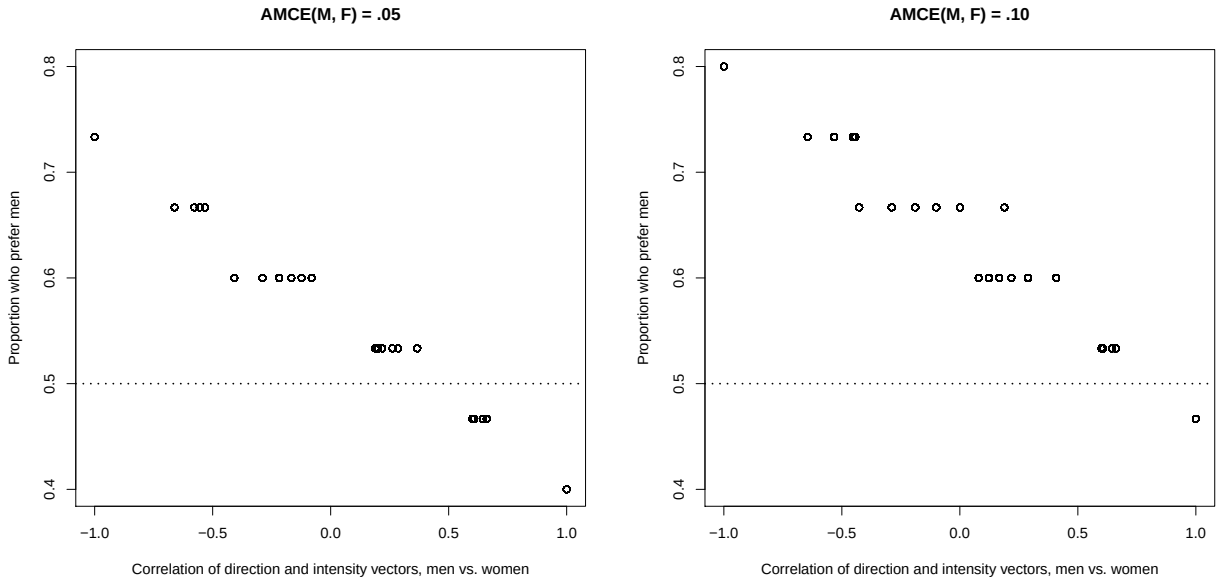

Figure 2. Combinations of proportions of respondents who prefer males to females and correlations between direction and intensity of male-female preference consistent with a given AMCE, computed for fifteen respondents and two binary attributes.

Furthermore, it can be seen in Figure 2 that when the correlation of direction and intensity is zero, the AMCE corresponds in sign with the majority preference. Using the logic underlying Proposition 2, we show that this holds in general, allowing researchers to assess how the AMCE performs in the best-case scenario, when there is no systematic relationship between preference intensity and direction for the feature of interest. When this is the case—that is, when our expectation about the importance of an attribute to a respondent does not change when we learn about the direction of their preference—the sign of the AMCE must correspond to the feature preferred by the majority. However, under these conditions the AMCE will be smaller in magnitude than the size of the margin,

thus providing a conservative estimate of that quantity.

Proposition 3. *When the direction and intensity of preferences across respondents are uncorrelated, the AMCE of a binary attribute has the same sign as the majority preference, but underestimates the size of the margin.*

Proof of Proposition 3 follows closely the logic of Proposition 2: when intensity and direction are uncorrelated, on the net, each supporter contributes as much to a feature as an opponent contributes to the baseline. As such, the points contributed by supporters and opponents cancel out, and the remainder corresponds in sign to the margin of victory for the feature preferred by the majority.

How realistic is the assumption of no correlation between the direction and intensity of attribute preferences? To answer this we turn to survey data from the 2016 American National Election Studies (ANES) and assess the degree to which there is a correlation in the expressed direction and intensity on a wide range of survey items. Specifically, the ANES asks about both direction and intensity of preferences for twenty-two issue areas; across these issues, respondents assess both whether they support or oppose a position, and how much importance they attach to the question. On seventeen of these questions—that is, for the vast majority of the issues in the ANES for which we have a measure of both direction and intensity of preferences—we find evidence that the supporters of a given policy or issue area have a meaningfully different assessment of its importance than its opponents. Indeed, the ANES provides strong evidence of the very dynamic that drives our stylized example: self-described "feminists" attach much more importance to this identity than do self-described "anti-feminists." See Supplementary Appendix D for a full discussion and results of this analysis.

We conclude this discussion with one final consideration: what happens when we relax the separability assumption and allow for arbitrary interactions between feature preferences? For instance, rather than assuming that voters unconditionally prefer men or women, we now allow for the possi-

bility that a voter prefers Republican men to Republican women, but Democrat women to Democrat men. In Supplementary Appendix E, we derive a summary statistic for aggregate feature preferences that captures this more complex, and potentially more realistic, preference structure, providing the necessary scaffolding for our final result:

Proposition 4. *When separability is relaxed, the bounds on the fraction of voters who prefer t_1 over t_0 are wider for any given AMCE.*

We also show that when separability is relaxed, the proportion of experimental subjects who prefer t_1 to t_0 is no longer indicative of an electoral advantage. In other words, without separability, even with tight bounds indicating a majority of respondents preferring t_1 to t_0, we cannot conclude that candidates with feature t_1 will beat candidates with feature t_0 in most all-else-equal contests. Furthermore, without separability, individual feature preferences do not necessarily satisfy transitivity.[13] Put simply, relaxing separability makes the very notion of a preference over features difficult to pin down from a theoretical perspective.

C. Average of Ideal Points

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 highlighted 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. 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

[13] A simple example is provided in Supplementary Appendix E.

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 utility from Candidate c's platform is given by:

$$U_i(\mathbf{x}_{ijc}) = -(b_i - \mathbf{x}_{ijc})^2 + \eta_{ijc} \tag{1}$$

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. In Supplementary Appendix F we walk through the steps that relate Equation (2) to a linear regression model estimated on 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. Letting $\Delta\mathbf{x}_{ij}$ represent the difference between the vectors $\mathbf{x}_{ij1}$ and $\mathbf{x}_{ij2}$, and m represent a given feature or element of this vector, one can estimate the regression:

$$y_{ij1} = \sum_m \beta_{im}\Delta x_{ijm} + \epsilon_{ij} \tag{3}$$

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, and we obtain the AMCE for feature m by averaging β_{im} over individuals.

Under this simple model of choice, the AMCE can be interpreted as an average of respondents' ideal points. The usefulness of the mean voter's preference, however, depends upon the particular model of elections that applied researchers have in mind. As is well known, the median voter's preference characterizes the unique equilibrium in a large number of probabilistic and deterministic voting models, and under a broad set of conditions (Calvert, 1985; Duggan, 2006; Bernhardt, Duggan and Squintani, 2007). By contrast, mean voter results maintain in a limited class of probabilistic voting models (Hinich, 1977; Lin, Enelow and Dorussen, 1999; Schofield, 2007) that require stronger assumptions about the motivations of candidates, the shape of voters' utility functions, and symmetry in the distribution of voter preferences, the latter of which is akin to our uncorrelated weights assumption.[14] Most importantly, these models require that parties know each voter's ideal point and only face uncertainty about voters' preference "shocks" or "biases"—additively separable error terms distributed independently of ideal points. Unfortunately, political scientists employ conjoint experiments precisely because we do not know voters' preferences.[15]

IV. Conclusion

We have shown that the AMCE, the target estimand of many conjoint experiments, does not support many interpretations ascribed to it 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

[14]For an extensive discussion of necessary and sufficient conditions for the existence of mean-voter equilibria, see Banks and Duggan (2005).

[15]In another class of probabilistic voting models candidates face uncertainty about voters' ideal points. In these models, convergent equilibria have candidates placing themselves at the expected position of the median voter.

than candidates without it. How, then, should researchers interpret the AMCE?

First, as shown by Bansak et al. (2021), the AMCE reflects the effect of changing an attribute on the expected vote share, where the average is taken with respect to the distribution of other attributes. As demonstrated in our main example, identical expected vote shares can be generated from a preference distribution that results in a single landslide in favor of women and most other contests resolving in favor of men, as well as from a preference distribution where female candidates win nearly all elections. Because it averages over the intensive and extensive margins of voter preferences, this expected vote share cannot speak to theoretically important questions such as which feature most voters prefer or which feature would dominate in most elections.

Second, we have characterized the AMCE as a preference aggregation mechanism and shown its relationship to the Borda count. Few real-world electoral contests are decided by Borda rule voting, but a more practical application of this insight is that it allows us to derive bounds on the proportion of the experimental sample that prefers a feature over the alternative, given a particular AMCE. Our analysis shows that as the number of possible candidate profiles increases, these bounds quickly expand to a range that is inconclusive about majority preferences for magnitudes of the AMCE that most applied researchers would reasonably encounter.

Third, we have demonstrated that the AMCE can be thought of as an average of the direction and intensity of voters' preferences, or an average of ideal points. Where might this interpretation be of interest? One area is in evaluating hypotheses generated by models of probabilistic voting. Notably, these models require strong additional assumptions for the mean voter's preference to be relevant in characterizing equilibria. Perhaps because of this, we are unaware of a single study that has used a conjoint experiment towards this end.

In general, the problems of interpretation we describe arise when there exists 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 with respect to quantities of interest in a one-person, one-vote setting. Thus, if the researcher has good reasons to believe that her experimental sample has uncorrelated directions and intensities of preferences, then she can proceed with a majoritarian interpretation of her results; that said, correlations of the sort we describe pervade areas of interest to political scientists, from gender parity in elected office (Teele, Kalla and Rosenbluth, 2018) to who should be favored by the nation's immigration policy (Hainmueller and Hopkins, 2015). Moreover, note that while our running example concerns voting in a majoritarian context, our critique applies more broadly to *any* attempt to summarize a population's preferences. Moving from ill-defined claims such as "Population X prefers A to B" to concrete statements concerning *any* proportion of a population requires buttressing the AMCE with very strong assumptions about the distribution of preferences—or developing alternative estimators altogether.

How should applied researchers proceed? Conjoint analysis remains most useful for questions where the *average* preference is of interest. However, scholars seeking answers to *majoritarian* questions may find themselves in a bind. On the one hand, we have shown through our bounding exercise that if they want to interpret their findings with respect to a majority preference, then they 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 is dependent upon the particular features included in an experiment, for a 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, researchers must get the distribution of randomized attributes exactly right. Unfortunately, it may prove difficult to construct a "Goldilocks"

experimental design that serves both goals.

Recently, researchers have begun developing tools for recovering relevant quantities of interest from conjoint and similar designs. Abramson et al. (2020) show that under the assumption of conditional preference homogeneity, researchers can use machine learning tools to recover quantities like the proportion of voters with a strict preference for candidate-features and to generate individual-level predictions for out-of-sample electoral contests. Future avenues for research on preference elicitation in political science should develop experimental designs that can directly recover relevant quantities of interest. For example, there exist experimental and survey designs that can obtain the individual-level estimates of preference intensities (Chen, Cavaillé and Van Der Straeten, 2019; Wiswall and Zafar, 2018). Further developing these tools will allow researchers to make more precise—and theoretically grounded—statements about voters' preferences.

REFERENCES

Abramson, Scott F, Korhan Kocak, Asya Magazinnik, and Anton Strezhnev. 2020. "Improving Preference Elicitation in Conjoint Designs using Machine Learning for Heterogeneous Effects." *Working Paper*, The Society for Political Methodology. Available at: `https://www.korhankocak.com/publication/akms/AKMS.pdf`.

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.

Banks, Jeffrey S, and John Duggan. 2005. "Probabilistic Voting in the Spatial Model of Elections: The Theory of Office-Motivated Candidates." In *Social Choice and Strategic Decisions.* 15–56. Springer.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. 2021. "Using Conjoint Experiments to Analyze Election Outcomes: The Central Role of the Average Marginal Component Effect (AMCE)." *Working Paper*, SSRN. Available at: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588941`.

Bernhardt, Dan, John Duggan, and Francesco Squintani. 2007. "Electoral Competition with Privately-Informed Candidates." *Games and Economic Behavior*, 58(1): 1–29.

Calvert, Randall L. 1985. "Robustness of the Multidimensional Voting Model: Candidate Motivations, Uncertainty, and Convergence." *American Journal of Political Science*, 69–95.

Chen, Daniel L, Charlotte Cavaillé, and Karine Van Der Straeten. 2019. "A Decision-Theoretic Approach to Understanding Survey Response: Likert vs. Quadratic Voting for Attitudinal Research." *University of Chicago Law Review*.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2021. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis*, 29(4).

Duggan, John. 2006. "A note on uniqueness of electoral equilibrium when the median voter is unobserved." *Economics Letters*, 92(2): 240–244.

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.

Hinich, Melvin J. 1977. "Equilibrium in Spatial Voting: The Median Voter Result Is an Artifact." *Journal of Economic Theory*, 16(2): 208–219.

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.

Lalley, Steven P, and E Glen Weyl. 2018. "Quadratic Voting: How Mechanism Design Can Radicalize Democracy." *AEA Papers and Proceedings*, 108: 33–37.

Lin, Tse-Min, James M Enelow, and Han Dorussen. 1999. "Equilibrium in Multicandidate Probabilistic Spatial Voting." *Public Choice*, 98(1): 59–82.

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.

Schofield, Norman. 2007. "The Mean Voter Theorem: Necessary and Sufficient Conditions for Convergent Equilibrium." *The Review of Economic Studies*, 74(3): 965–980.

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.

Wiswall, Matthew, and Basit Zafar. 2018. "Preference for the Workplace, Investment in Human Capital, and Gender." *The Quarterly Journal of Economics*, 133(1): 457–507.

Supplementary Appendix to: What Do We Learn About Voter Preferences From Conjoint Experiments?

A. Proofs

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 of Lemma 1. Suppose there are N voters and K profiles. Consider voter i's preference ranking over profiles. For any pair of profiles 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. Suppose now i is presented with each pairwise comparison. Then, i chooses their most preferred profile x_1 every time it is on the ballot, against every other profile, 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 every time except when compared 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 exactly with the number of times each profile is chosen when every pairwise comparison is made. Finally, the least preferred profile will never be chosen in a 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,

$$b(x_m) \equiv \sum_{i=1}^{N} b_i(x_m) = \sum_{i=1}^{N} \sum_{j\neq m} Y_i(x_m, x_j).$$

□

Lemma 2. *With separable preferences and 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.*

Proof of Lemma 2. Let us first restate the formal definition of separability. Voter i's choices are separable when for all t_1 and t_0, we have

$$Y_i\left((t_1, T_{[-l]}), (t_0, T_{[-l]})\right) = Y_i\left((t_1, T'_{[-l]}), (t_0, T'_{[-l]})\right)$$

where $T_{[-l]}$ and $T'_{[-l]}$ denote two arbitrary vectors of other treatment components.

Formally, Borda score of a feature t_1, $B(t_1)$ is

$$B(t_1) \equiv \sum_{i=1}^{N} \sum_{x_1 \in \kappa(t_1)} \sum_{x_j \neq x_1} Y_i(x_1, x_j)$$

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

$$b_i(t_1, T_{[-l]}) - b_i(t_1, T'_{[-l]}) = b_i(t_0, T_{[-l]}) - b_i(t_0, T'_{[-l]})$$

for all $t_1, t_0, T_{[-l]}$, and $T'_{[-l]}$ by a straightforward application of Lemma 1. Summing these up

$$\sum_{i=1}^{N} b_i(t_1, T_{[-l]}) - \sum_{i=1}^{N} b_i(t_0, T_{[-l]}) = \sum_{i=1}^{N} b_i(t_1, T'_{[-l]}) - \sum_{i=1}^{N} b_i(t_0, T'_{[-l]}).$$

Suppose now $(t_1, T^*_{[-l]})$ is the profile with the highest Borda score. This means:

$$\sum_{i=1}^{N} b_i(t_1, T^*_{[-l]}) - \sum_{i=1}^{N} b_i(t_0, T^*_{[-l]}) \geq 0.$$

By the separability assumption, it follows that for any arbitrary vector of treatments $T_{[-l]}$:

$$\sum_{i=1}^{N} b_i(t_1, T_{[-l]}) - \sum_{i=1}^{N} b_i(t_0, T_{[-l]}) \geq 0$$

Because this is true for each vector of treatments $T_{[-l]}$, it is also true when we sum over them and get the Borda score of t_1. Therefore, the Borda score of t_1 must be greater than that of t_0 because

$$B(t_1) = \sum_{T_{[-l]}} \sum_{i=1}^{N} b_i(t_1, T_{[-l]}) \geq \sum_{T_{[-l]}} \sum_{i=1}^{N} b_i(t_0, T_{[-l]}) = B(t_0).$$

□

PROPOSITION 1: The difference of the Borda scores of two features is proportional to the AMCE.

Proof of Proposition 1. The number of profiles that have t_1 is equal to the number of profiles that have t_0, which is in turn equal to the total number of profiles divided by the number of unique values the attribute of interest can take: $|\kappa(t_1)| = |\kappa(t_0)| = \frac{K}{\tau}$. Then, by dividing the Borda score of a feature, $B(t_1)$ by the total number of pairwise comparisons t_1 appears in, $\frac{K}{\tau}NK$, and taking the difference with the Borda score $B(t_0)$ of the baseline feature t_0, divided by $\frac{K}{\tau}NK$ yields exactly the AMCE of t_1 as defined in Hainmueller, Hopkins and Yamamoto (2014):

$$\pi(t_1, t_0) = \frac{\sum_{i=1}^{N} \sum_{x \in \kappa(t_1)} \sum_{x_j \neq x} Y_i(x, x_j)}{|\kappa(t_1)|NK} - \frac{\sum_{i=1}^{N} \sum_{x \in \kappa(t_0)} \sum_{x_j \neq x} Y_i(x, x_j)}{|\kappa(t_0)|NK} = \frac{\tau}{NK^2} \left(B(t_1) - B(t_0) \right).$$

□

PROPOSITION 2: Let y denote the fraction of voters who prefer t_1 over t_0. Given an AMCE of $\pi(t_1, t_0)$, it must be that

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

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

Proof of Proposition 2. We prove this proposition by finding the range of Borda scores of t_1 and t_0 that can be rationalized for a given proportion of respondents who prefer t_1 over t_0; and then inverting this range to find the minimum and maximum proportions of respondents who prefer t_1 over t_0 for a given AMCE.

Let us find the minimum fraction of respondents who prefer t_1 over t_0 that is consistent with an AMCE. Notice that for a fixed fraction of respondents, the AMCE is maximized when respondents 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 next, 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 up. To help with the intuition, the preferences of such a voter might look like:

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

where α, β, and γ represent a collection of other features of candidates included in the experiment. 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, for any vector of other attributes $T_{[-l]}$, the profile $(t_1, T_{[-l]})$ is maximally chosen $K-\frac{K}{\tau}$ more times than $(t_0, T_{[-l]})$ when every pairwise comparison is made. From Proposition 1 we know that this implies the maximum difference in Borda scores, $b_i(t_1, T_{[-l]}) - b_i(t_0, T_{[-l]}) = K-\frac{K}{\tau}$, for any arbitrary combination of other attributes, $T_{[-l]}$. Because there are $\frac{K}{\tau}$ possible unique combinations of other attributes, each respondent makes $\frac{K}{\tau}$ such comparisons between t_1 and t_0. Thus, each respondent who prefers t_1 maximally generates a $\frac{K^2(\tau-1)}{\tau^2}$ higher Borda score for t_1 than t_0.

Similarly, the maximum AMCE is only obtained when those who prefer t_0 assign the lowest priority to this attribute, and rank profiles with t_1 just below otherwise identical profiles with t_0.

Such preferences might look like:

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

When other features are held constant, the difference in Borda scores of a profile with t_1 and t_0 is -1. In other words, for respondents who prefer t_0 to t_1, the maximum difference is $b_j(t_1, T_{[-l]}) - b_j(t_0, T_{[-l]}) = -1$, for any arbitrary combination of other attributes, $T_{[-l]}$. Again, because there are $\frac{K}{\tau}$ possible combinations of other features and thus as many comparisons between profiles with t_1 and t_0, each respondent who prefers t_0 minimally generates $\frac{K}{\tau}$ more points for t_0 than 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, $y^{\min}$, by summing these scores and normalizing:

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

Simple algebra reveals

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

A very similar argument establishes the upper bound of y. □

PROPOSITION 3: When the direction and intensity of preferences across respondents are uncorrelated, the AMCE of a binary attribute has the same sign as the majority preference, but underestimates the size of the margin.

Proof of Proposition 3. Denote by n_1 the number of respondents who prefer t_1 to t_0. Similarly, let $n_0 = N - n_1$ refer to the number of respondents who prefer t_0 to t_1. Without loss of generality, reorder respondents so those who prefer t_1 to t_0 have the lowest rank, that is $i \in \{1, \ldots, n_1\}$. Suppose direction and intensity of preferences are uncorrelated across respondents. Then, the average net contribution to t_1 from a supporter of t_1 is the same as the average net contribution to t_0 from an opponent of t_1. Formally, we can write this as

$$\frac{1}{n_1}\sum_{i=1}^{n_1} B_i(t_1) - B_i(t_0) = \frac{1}{n_0}\sum_{i=n_1+1}^{N} B_i(t_0) - B_i(t_1). \tag{A1}$$

for any t_1, t_0, and i.

We know from the proof of Proposition 1 that we can write the AMCE as:

$$\pi(t_1, t_0) = \frac{\tau}{NK^2} \sum_{i=1}^{N} B_i(t_1) - B_i(t_0). \tag{A2}$$

Then, we can rewrite expression A2 as

$$\pi(t_1, t_0) = \frac{\tau}{NK^2} \left(\sum_{i=1}^{n_1} B_i(t_1) - B_i(t_0) - \sum_{i=n_1+1}^{N} B_i(t_0) - B_i(t_1) \right)$$

From Equation A1, when preference direction and intensity are uncorrelated:

$$\pi(t_1, t_0) = \frac{\tau \, \mathbb{E}_{i \leq n_1} \left[B(t_1) - B(t_0) \right]}{NK^2} (n_1 - n_0).$$

Thus, $\pi(t_1, t_0)$ is positive if and only if a majority of respondents prefer t_1 to t_0, or $n_1 > 1/2$. □

PROPOSITION 4: When separability is relaxed, the bounds on the fraction of voters who prefer t_1 over t_0 are wider for any given AMCE.

Proof of Proposition 4. When the separability assumption does not hold, the bounds on the fraction of voters who prefer t_1 to t_0 for an AMCE of $\pi(t_1, t_0)$, in an experiment with K possible profiles, and when the attribute of interest can take τ distinct values, are given by

$$y \in \left[\max \left\{ 1 - \frac{\tau(1 - \pi(t_1, t_0)) - 1}{\tau - 1 - \frac{\tau^2}{K^2} \left(\left(\left\lfloor \frac{K}{2\tau} - \frac{1}{2} \right\rfloor \right) \left(K - \left\lfloor \frac{K}{2\tau} - \frac{1}{2} \right\rfloor \right) - \left\lceil \frac{K}{2\tau} + \frac{1}{2} \right\rceil \right)}, 0 \right\} \right.$$
$$\left. \min \left\{ \frac{1 + \tau(1 - \pi(t_1, t_0))}{K^2(\tau - 1) - \frac{\tau^2}{K^2} \left(\left(\left\lfloor \frac{K}{2\tau} - \frac{1}{2} \right\rfloor \right) \left(K - \left\lfloor \frac{K}{2\tau} - \frac{1}{2} \right\rfloor \right) - \left\lceil \frac{K}{2\tau} + \frac{1}{2} \right\rceil \right)}, 1 \right\} \right],$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling functions respectively.[1]

Similarly to the proof of Proposition 2, these bounds obtain when both the voters who prefer t_1 and those who prefer t_0 give the maximum and minimum net Borda scores to t_1 versus t_0. The bounds

[1]The floor and ceiling functions are necessary because of how we define a preference; strictly more than half of all all-else-equal comparisons. If there is an odd (even) number of all-else-equal comparisons, then minimally the profiles with the preferred feature are chosen once (twice) more than those without. The floor and ceiling functions account for this difference.

in this case are wider because interactions allow for more freedom when constructing preferences. Below we lay out the arguments for the lower bound. The upper bound is constructed analogously.

For respondents who prefer t_1, the maximum possible net Borda score given to t_1 versus t_0 without separability is the same as the case with: $\frac{K^2(\tau-1)}{\tau^2}$. Now consider a respondent who prefers t_0. Without separability, such a respondent prefers profiles with t_0 to otherwise identical profiles with t_1 in majority of the cases, but in others they may have a preference for profiles with t_1. Specifically, a respondent who prefers t_0 gives the maximum possible net Borda score to t_1 versus t_0 when her preferences look like the following:

$$\underbrace{t_1\alpha\beta\gamma \succ t_1\alpha'\beta\gamma \succ \dots}_{\lfloor \frac{K}{2\tau}-\frac{1}{2} \rfloor \text{ profiles}} \succ \underbrace{t_0\alpha'\beta'\gamma \succ t_1\alpha'\beta'\gamma \succ \dots \succ t_0\alpha'\beta'\gamma' \succ t_1\alpha'\beta'\gamma'}_{2\lceil \frac{K}{2\tau}+\frac{1}{2} \rceil \text{ profiles}} \succ \underbrace{t_0\alpha\beta\gamma \succ t_0\alpha'\beta\gamma \succ \dots}_{\lfloor \frac{K}{2\tau}-\frac{1}{2} \rfloor \text{ profiles}}$$

where again α, β, and γ represent a collection of other features of candidates included in the experiment. In words, this respondent has the minimal distance of one between the profiles with t_0 she prefers to otherwise identical profiles with t_1, and the maximal distance of $K - \lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor$ between the profiles with t_1 she prefers to otherwise identical profiles with t_0. To check that for this respondent we have $\Psi_i(t_1, t_0) < \frac{1}{2}$, notice there are $\lceil \frac{K}{2\tau} + \frac{1}{2} \rceil$ comparisons where she prefers t_0 over t_1 and $\lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor$ comparisons where t_1 is preferred to t_0. Thus, the maximum net contribution to t_1 of a respondent who prefers t_0 to t_1 is $\left(\lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right)\left(K - \lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right) - \lceil \frac{K}{2\tau} + \frac{1}{2} \rceil$. Notice that for $\frac{K}{\tau} > 2$, we have $\left(\lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right)\left(K - \lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right) > \lceil \frac{K}{2\tau} + \frac{1}{2} \rceil$. This means that without separability, a respondent who prefers t_0 to t_1 may still contribute more Borda points to t_1 than t_0.

When we calculate the bounds as in the proof of Proposition 2, we find that

$$\pi(t_1, t_0) = \frac{(y^{\min})\frac{K^2(\tau-1)}{\tau^2} + (1 - y^{\min})\left(\left(\lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right)\left(K - \lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right) - \lceil \frac{K}{2\tau} + \frac{1}{2} \rceil\right)}{\frac{K^2}{\tau}}.$$

Algebra reveals

$$y^{\min} = \max\left\{1 - \frac{\tau(1-\pi) - 1}{\tau - 1 - \frac{\tau^2}{K^2}\left(\left(\lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right)\left(K - \lfloor \frac{K}{2\tau} - \frac{1}{2} \rfloor\right) - \lceil \frac{K}{2\tau} + \frac{1}{2} \rceil\right)}, 0\right\}$$

It can be confirmed that this is equal to the lower bound in Proposition 2 when $\frac{K}{\tau} = 2$, and strictly lower when $\frac{K}{\tau} > 2$. □

Lemma 3. *The AMCE is equivalent to $y_{ij1} = \sum_m \Delta x_{ijm}\beta_m + \epsilon_{ij}$, or an average ideal point.*

Proof of Lemma 3. To show that the estimation of Equation F4 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 F4 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 $\hat{\rho}_m = \frac{\text{Cov}(x_{ijmc}, y_{ijc})}{\text{Var}(x_{ijmc})}$. Similarly, estimating Equation F4 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 $\hat{\beta}_m = \frac{\text{Cov}(x_{ijm1} - x_{ijm2}, y_{ij1})}{\text{Var}(x_{ijm1} - x_{ijm2})}$. Consider the numerator.

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

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

Next consider the denominator.

$$\text{Var}(x_{ijm1} - x_{ijm2}) = \text{Var}(x_{ijm1}) + \text{Var}(-x_{ijm2}) - 2\text{Cov}(x_{ijm1}, x_{ijm2}) = 2\text{Var}(x_{ijmc})$$

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

Lemma 4. *The result in Lemma 3 holds whether we impose a linear or quadratic loss function.*

Proof of Lemma 4.

$$\begin{aligned}
U_i(x_{j1}) &= -|x_{j1} - b_i| + \eta_{ij} \\
U_i(x_{j2}) &= -|x_{j2} - b_i| + \nu_{ij}
\end{aligned} \tag{A3}$$

Assume $0 \leq b_i \leq 1$

$$\begin{aligned} \Pr(y_{ij1} = 1) &= \Pr(U_i(\mathbf{x}_{ij1}) > U_i(\mathbf{x}_{ij2})) \\ &= \Pr\left(\eta ij - \nu_{i2} < |x_{j2} - b_i| - |x_{j1} - b_i|\right) \end{aligned} \tag{A4}$$

Since x_{j1} and 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:

$$\Pr(y_{ij1} = 1) = Pr\left(\eta_{j1} - \nu_{j2} < \Delta x_j(2b_i - 1)\right) \tag{A5}$$

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

$$\begin{aligned} y_{ij1} = &\ \Delta x_j(2b_i - 1) + \eta_{ij} - \nu_{ij} \\ = &\ \Delta x_j \beta_i + \epsilon_{ij} \end{aligned} \tag{A6}$$

□

B. Robustness of the AMCE to the Inclusion/Exclusion of Additional Treatments

We provide simple R code to generate a fully observed conjoint experiment based on a set of preference orderings for a set of voters, and to use this data to estimate AMCEs—both at the respondent level and over the sample—as described in Section II of the paper. We use this first to demonstrate that the inclusion of an additional attribute while holding constant all respondents' preference orderings over the attribute of interest can change the sign of the estimated AMCE. Then, we show that eliminating certain feature combinations can have the same effect.

```
library(gtools)

# Function to construct matrix of all possible vote choices
construct.vote <- function(ranks) {
  cands <- names(ranks[[1]])
  vote <- data.frame(t(combn(cands, 2)))
  names(vote) <- c("C1", "C2")
  vote <- rbind(vote, data.frame(C1 = cands, C2 = cands))
  vote$C1 <- as.character(vote$C1)
  vote$C2 <- as.character(vote$C2)
  out <- NULL
  for (i in c(1:length(ranks))) {
    choice <- rep(NA, nrow(vote))
```

X

```
    for (j in c(1:nrow(vote))) {
      choice[j] <- ifelse(as.numeric(ranks[[i]][vote$C1[j]]) < as.numeric(ranks[[i]][vote$C2[j]]),
                          vote$C1[j], vote$C2[j])
    }
    tobind <- cbind(vote, choice)
    tobind$type <- i
    out <- rbind(out, tobind)
  }
  return(out)
}

# Function to obtain the AMCE as in Table 4
amce.compute <- function(vote.mat, pos, value.baseline, value.amce, weights = NULL, idvar) {
  df <- vote.mat
  n.atts <- nchar(df$C1[1])

  df$name1 <- paste0(df$C1, "-", df$C2)
  df$name2 <- paste0(df$C2, "-", df$C1)

  # generate all possible comparisons
  combs <- data.frame(C1 = unique(c(df$C1, df$C2)))
  combs$C1 <- as.character(combs$C1)
  both.combs <- data.frame(permutations(n = length(combs$C1), r = 2, v = combs$C1, repeats.allowed = TRUE))

  # restrict to value of interest
  comp1 <- both.combs[substr(both.combs$X1, pos, pos)==value.amce,]
  names(comp1) <- c("C1", "C2")
  comp1$name1 <- paste0(comp1$C1, "-", comp1$C2)
  comp1$name2 <- paste0(comp1$C2, "-", comp1$C1)

  # flip to baseline
  comp2 <- comp1
  comp2$C1 <- paste0(substr(comp1$C1, 0, pos-1), value.baseline, substr(comp1$C1, pos + 1,
                nchar(as.character(comp1$C1))))
  comp2$name1 <- paste0(comp2$C1, "-", comp2$C2)
  comp2$name2 <- paste0(comp2$C2, "-", comp2$C1)

  # compute individual AMCEs
  df1 <- df[,!names(df) %in% "name1"]
  df2 <- df[,!names(df) %in% "name2"]
  names(df1)[ncol(df1)] <- names(df2)[ncol(df2)] <- "name"
  df_all <- rbind(df1, df2)
  amce.ind <- data.frame(voter = unique(df[,idvar]), amce = NA)
  for (i in 1:nrow(amce.ind)) {
    # compute whether C1 wins for every combination
    tomerge <- df_all[df_all[,idvar]==i, c("name", "choice")]
    winstats <- merge(comp1, tomerge, by.x = "name1", by.y = "name", all.x = TRUE)
    win.c1 <- ifelse(winstats$C1==winstats$C2, .5, ifelse(winstats$choice==winstats$C1, 1, 0))
    names(win.c1) <- winstats$name1
    # flip and compute
    winstats <- merge(comp2, tomerge, by.x = "name1", by.y = "name", all.x = TRUE)
    win.cf <- ifelse(winstats$C1==winstats$C2, .5, ifelse(winstats$choice==winstats$C1, 1, 0))
    names(win.cf) <- winstats$name1
    # compute individual amce
    amce.ind$amce[i] <- sum(win.c1 - win.cf)
  }

  # normalize all
  norm <- ((2^n.atts)) * (2^(n.atts - 1))
  amce.ind$amce <- amce.ind$amce/norm

  # compute mean of the difference between the two
```

```
76    if (is.null(weights)) {
77      amce <- mean(amce.ind$amce)
78    } else {
79      amce <- weighted.mean(amce.ind$amce, weights)
80    }
81
82    return(list(amce = amce, amce.ind = amce.ind))
83  }
84
85  # Example in Table 2
86  ranks2 <- list("1" = c("MR" = 1, "FR" = 2, "MD" = 3, "FD" = 4),
87                 "2" = c("MR" = 4, "FR" = 2, "MD" = 3, "FD" = 1))
88  vote.mat2 <- construct.vote(ranks2)
89  amce.compute(vote.mat = vote.mat2,
90                  pos = 1,
91                  value.baseline = "F",
92                  value.amce = "M",
93                  weights = c(3/5, 2/5),
94                  idvar = "type")
95
96  # Example in Table 5
97  ranks3 <- list("1" = c("MRW" = 1, "MRB" = 2, "FRW" = 3, "MDW" = 4, "FRB" = 5, "MDB" = 6, "FDW" = 7, "FDB" = 8),
98                 "2" = c("MRW" = 8, "MRB" = 5, "FRW" = 6, "MDW" = 7, "FRB" = 2, "MDB" = 3, "FDW" = 4, "FDB" = 1))
99  vote.mat3 <- construct.vote(ranks3)
100 amce.compute(vote.mat = vote.mat3,
101                 pos = 1,
102                 value.baseline = "F",
103                 value.amce = "M",
104                 weights = c(3/5, 2/5),
105                 idvar = "type")
```

Running the example in lines 85-94 returns the AMCE of -1/20 computed in Table 4:

```
> amce.compute(vote.mat = vote.mat2,
                 pos = 1,
                 value.baseline = "F",
                 value.amce = "M",
                 weights = c(3/5, 2/5),
                 idvar = "type")

$amce
[1] -0.05

$amce.ind
  voter  amce
1     1  0.25
2     2 -0.50
```

However, when we add a third attribute, $R \in \{B, W\}$, as described in Table 5, without changing the preference orderings of the other two attributes or the distribution of voters, the AMCE changes sign:

```
> amce.compute(vote.mat = vote.mat3,
                 pos = 1,
                 value.baseline = "F",
```

```
                value.amce = "M",
                weights = c(3/5, 2/5),
                idvar = "type")

$amce
[1] 0.0625

$amce.ind
  voter    amce
1     1  0.3125
2     2 -0.3125
```

Similarly, it is straightforward to construct an example where *eliminating* feature combinations, as is standard practice in applied work, changes the sign of the AMCE. Consider three types of voters with preferences as given in Table B1:

V1	V2	V3
$M \succ F$	$F \succ M$	$F \succ M$
$R \succ D$	$D \succ R$	$D \succ R$
$B \succ W$	$B \succ W$	$W \succ B$

Table B1—Preferences over attributes

Assume priorities over attributes as follows. V1: $R >> P >> G$; V2: $P >> R >> G$; V3: $P >> G >> R$ and that each voter prefers candidates with two attributes they like to candidates with only one attribute they like. With this information we can construct preferences over candidates for each type as presented in Table B2.

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

Table B2—Preferences over attributes

Consider a population of five V1s, two V2s, and two V3s. Table B3 gives the AMCE estimate when

we include the full set of candidate features, and when we exclude each combination of party and race. Code to replicate this example is included below.

Omitted Features	R	D
B	0.02	-0.02
W	-0.02	0.02

No Omitted Features: -0.01

Table B3—AMCE Estimates of Male, restricting Party-Race feature combinations

```
#### No Omitted Combinations ####
ranks4 <- list("1" = c("MRB" = 1, "FRB" = 2, "MDB" = 3, "MRW" = 4, "FDB"= 5, "FRW" = 6, "MDW" = 7, "FDW" = 8),
               "2" = c("MRB" = 6, "FRB" = 4, "MDB" = 2, "MRW" = 8, "FDB"= 1, "FRW" = 7, "MDW" = 5, "FDW" = 3),
               "3" = c("MRB" = 8, "FRB" = 6, "MDB" = 5, "MRW" = 7, "FDB"= 2, "FRW" = 4, "MDW" = 3, "FDW" = 1))
vote.mat4 <- construct.vote(ranks4)

#### No RBs  ####
ranks4a <- list("1" = c("MDB" = 1, "MRW" = 2, "FDB" = 3, "FRW" = 4, "MDW" = 5, "FDW" = 6),
                "2" = c("MDB" = 2, "MRW" = 6, "FDB" = 1, "FRW" = 5, "MDW" = 4, "FDW" = 3),
                "3" = c("MDB" = 5, "MRW" = 6, "FDB" = 2, "FRW" = 4, "MDW" = 3, "FDW" = 1))
vote.mat4a <- construct.vote(ranks4a)

#### No RWs  ####
ranks4b <- list("1" = c("MRB" = 1, "FRB" = 2, "MDB" = 3, "FDB" = 4, "MDW" = 5, "FDW" = 6),
                "2" = c("MRB" = 6, "FRB" = 4, "MDB" = 2, "FDB" = 1, "MDW" = 5, "FDW" = 3),
                "3" = c("MRB" = 6, "FRB" = 5, "MDB" = 4, "FDB" = 2, "MDW" = 3, "FDW" = 1))
vote.mat4b <- construct.vote(ranks4b)

#### No DBs  ####
ranks4c <- list("1" = c("MRB" = 1, "FRB" = 2, "MRW" = 3, "FRW" = 4, "MDW" = 5, "FDW" = 6),
                "2" = c("MRB" = 4, "FRB" = 2, "MRW" = 6, "FRW" = 5, "MDW" = 3, "FDW" = 1),
                "3" = c("MRB" = 6, "FRB" = 4, "MRW" = 5, "FRW" = 3, "MDW" = 2, "FDW" = 1))
vote.mat4c <- construct.vote(ranks4c)

#### No DWs ####
ranks4d <- list("1" = c("MRB" = 1, "FRB" = 2, "MDB" = 3, "MRW" = 4, "FDB" = 5, "FRW" = 6),
                "2" = c("MRB" = 4, "FRB" = 3, "MDB" = 2, "MRW" = 6, "FDB" = 1, "FRW" = 5),
                "3" = c("MRB" = 6, "FRB" = 4, "MDB" = 3, "MRW" = 5, "FDB" = 1, "FRW" = 2))
vote.mat4d <- construct.vote(ranks4d)
```

Computing the AMCEs:

```
> amce.compute(vote.mat = vote.mat4,
               pos = 1,
               value.baseline = "F",
               value.amce = "M",
               weights = c(5/9, 2/9, 2/9),
               idvar = "type")

$amce
[1] -0.006944444

$amce.ind
  voter    amce
1     1  0.1875
```

```
2     2 -0.1875
3     3 -0.3125

> amce.compute(vote.mat = vote.mat4a,
                pos = 1,
                value.baseline = "F",
                value.amce = "M",
                weights = c(5/9, 2/9, 2/9),
                idvar = "type")

$amce
[1] 0.01736111

$amce.ind
  voter     amce
1     1  0.15625
2     2 -0.09375
3     3 -0.21875

> amce.compute(vote.mat = vote.mat4b,
                pos = 1,
                value.baseline = "F",
                value.amce = "M",
                weights = c(5/9, 2/9, 2/9),
                idvar = "type")

$amce
[1] -0.01736111

$amce.ind
  voter     amce
1     1  0.09375
2     2 -0.15625
3     3 -0.15625

> amce.compute(vote.mat = vote.mat4c,
                pos = 1,
                value.baseline = "F",
                value.amce = "M",
                weights = c(5/9, 2/9, 2/9),
                idvar = "type")

$amce
[1] -0.01736111

$amce.ind
  voter     amce
1     1  0.09375
2     2 -0.15625
3     3 -0.15625

> amce.compute(vote.mat = vote.mat4d,
                pos = 1,
                value.baseline = "F",
                value.amce = "M",
                weights = c(5/9, 2/9, 2/9),
                idvar = "type")

$amce
[1] 0.01736111

$amce.ind
```

```
  voter     amce
1     1  0.15625
2     2 -0.09375
3     3 -0.21875
```

C. Bounds on Proportion of Experimental Sample Who Prefer a Feature

We can take advantage of the structure of conjoint data to compute tighter bounds on the proportion of survey respondents who prefer a feature over the baseline than the general bounds derived in Proposition 2. To do so, we use the insight that when one or more attributes are held fixed at the same value in a given head-to-head comparison, the respondent makes her decision based only on the values of the remaining attributes (those that differ from one another), assuming that preferences are separable—that is, that the choice between any two features is not contingent on the value of another attribute. Under this key assumption, we can compute tighter bounds as a weighted average of our standard bounds computed within all subsets of the data, where the subsets are defined according to which attributes differ and which are the same in the randomly generated candidate pairings. We recompute π, K, and τ within each subgroup, where K—the number of possible candidate profiles—is computed ignoring the attributes that are the same; thus, it is guaranteed to be smaller than the aggregate K when there is at least one common attribute. Formally, these tighter bounds are given by:

$$\left[\sum_{s=1}^{S} \frac{n_s}{N}\, l_s(\pi_s, K_s, \tau), \sum_{s=1}^{S} \frac{n_s}{N} u_s(\pi_s, K_s, \tau)\right]$$

where l_s and u_s are the lower and upper bounds for a subset s, respectively. To illustrate how we create these subsets of the data, we walk through an example of a conjoint experiment with four attributes: gender (male, female), party (Democrat, Republican), race (white, Black, Hispanic, other), and age (young, middle, and old). Supposing we are interested in the effect of gender (female vs. male), we divide the data into groups based on the three remaining attributes: a group where the candidate pairs have different values of party, race, and age; three groups in which they have the same party, race, and age, respectively; three groups with two matched attributes and a third unmatched (party and race, party and age, and race and age); and a final group with all matched attributes. Generically, this will yield $S = 2^{A-1}$ groups, where A is the number of attributes in the experiment—in other words, the power set of all attributes other than the attribute of interest for the AMCE. Within each of these subsets, we compute an AMCE and a K that ignores the matched attributes: for instance, holding fixed party and race, there are six possible candidate profiles (2 values of gender $\times$ 3 values of age). Finally, we compute a weighted average of these subset-specific

TABLE C1—BOUNDS ON PROPORTION OF SAMPLE HAVING PREFERENCES CONSISTENT WITH AMCE, COMPUTED FOR RECENT PAPERS IN THE TOP THREE POLITICAL SCIENCE JOURNALS.

Paper	Estimated effect	AMCE (π)	Number of profiles (K)	Number of relevant features (τ)	Bounds on proportion with consistent preference	Tighter bounds under separability
APSR						
Ward (2019)	Proportion of group comprised of university graduates on support for immigration, 30% vs. 0%	0.22	20	4	[0.34, 1.00] (0.31, 1.00)	[0.36, 1.00] (0.33, 1.00)
Auerbach and Thachil (2018)	Broker education on support, high (BA) vs. none	0.13	1,296	3	[0.20, 1.00] (0.15, 1.00)	[0.25, 0.94] (0.19, 0.98)
Hankinson (2018)	Height of building on homeowners' support for new construction, 12 vs. 2 stories	-0.16	6,144	4	[0.00, 0.78] (0.00, 0.81)	[0.00, 0.77] (0.00, 0.80)
Teele, Kalla, and Rosenbluth (2018)	Experience on candidate support among legislators, 8 years vs. 0 years	0.18	864	4	[0.24, 1.00] (0.21, 1.00)	[0.25, 1.00] (0.22, 1.00)
Carnes and Lupu (2016)	Liberal party label on candidate support (Argentina)	-0.10	32	2	[0.00, 0.75] (0.00, 0.83)	[0.10, 0.60] (0.04, 0.67)
JOP						
Ballard-Rosa, Martin, and Scheve (2016)	Tax rate on those earning <10k on support for plan, 25% vs. 0%	-0.23	38,400	4	[0.00, 0.70] (0.00, 0.73)	[0.00, 0.70] (0.00, 0.73)
Mummolo and Nall (2016)	Driving time to work on Democrats' choice of community to live, 75 vs. 10 minutes	-0.23	3,456	4	[0.00, 0.69] (0.00, 0.71)	[0.00, 0.45] (0.00, 0.71)
Mummolo (2016)	Relevant information on choice to consume, vs. irrelevant (among seniors)	0.30	6	2	[0.71, 1.00] (0.66, 1.00)	[0.77, 0.96] (0.68, 0.96)

Notes: AMCEs may differ slightly from those reported in paper because we reestimate them without survey weights and only on sample having two candidate profiles per respondent (unmatched profiles appear in some replication datasets). 95% confidence sets computed using a block bootstrap are reported in parentheses below the bounds.

bounds, where the weight is determined by the number of observations in that subset.[2]

Table C1 reports the bounds in Proposition 2 as well as these tighter bounds for all of the forced-choice conjoint experiments published in the *APSR* and the *JOP* between 2016 and the first quarter of 2019.[3] We construct our bounds for the largest estimated effect presented in each paper (thus not necessarily the paper's central finding). To compute uncertainty estimates, we randomly sample individuals (and thus their complete survey responses) and recompute each bound over 1,000 bootstrap replicates, taking the normal approximation 95% confidence interval for each bound. Table C1 reports the lower confidence interval on the lower bound and the upper interval on the upper bound in parentheses below the bounds themselves. In one case (Mummolo and Nall 2016), our tighter bounding exercise produces upper and lower bounds on the same side of the 0.5 threshold (whereas the original bounding approach had not), but these gains in precision are lost once we incorporate the uncertainty of the estimate.

The code below is a simple implementation of the bounds in Proposition 2 in R. Our replication file contains all code needed to construct Table C1, including code for implementing the tighter bounds and for bootstrapping all confidence intervals.

```
bounds <- function(pi, K, tau, se_pi = NULL) {
  # compute lower and upper bound according to proposition 2
  l <- max(((pi * tau * K) + tau) / ((K * (tau - 1)) + tau), 0)
  u <- min(((pi * tau * K) + (K * (tau - 1))) / ((K * (tau - 1)) + tau), 1)
  bounds <- c(l, u)
  names(bounds) <- c("lower", "upper")
  # compute 95% confidence set for the bounds
  if (is.null(se_pi)) {
    # just return analytic bounds if no standard error is provided
    output <- bounds
  } else if (class(se_pi)=="numeric" & length(se_pi)==1) {
    # delta method-computed standard error (same for upper and lower bound)
    se <- sqrt(((tau * (K - 1)) / ((K * (tau - 1)) + tau))^2 * se_pi^2)
    # confidence interval
```

[2]Together, the subsets form a partition of the full dataset. In some cases, a subset may be too small to compute an AMCE, but this will not affect the bounds dramatically precisely because it only has a small number of observations.

[3]We also searched the *AJPS* but there are no forced-choice conjoint experiments appropriate for our analysis published there during this period. Hemker and Rink (2017) have statistically significant findings only when they use non-binary scales as outcomes and Huff and Kertzer (2017) have a binary outcome (labeling an attack as an act of terrorism) that is not a forced choice between two alternatives.

```
    ci.lower = max(0, l + (qnorm(0.025) * se))
    ci.upper = min(1, u + (qnorm(0.975) * se))
    ci <- c(ci.lower, ci.upper)
    names(ci) <- c("lower", "upper")
    output <- list(bounds, ci)
    names(output) <- c("analytic_bounds", "ci_95")
  } else {
    # return an error if standard error is entered incorrectly
    cat("Please provide a numeric value for se_pi \n")
    stop()
  }
  return(output)
}

# example: Ward (2019)
bounds(pi = .22, K = 20, tau = 4)
```

D. Correlations between Direction and Intensity of Preferences in the 2016 ANES

For every question in the 2016 ANES that accommodates such an analysis, we code a direction variable that has a value of 1 if the respondent takes a clear stance in favor of a position and 0 if they are opposed.[4] We also code a measure of intensity that takes on evenly distributed values over the interval $[0, 1]$ depending on how many importance categories were included in the question, where 0 is the lowest level of importance and 1 is the highest.[5] We then compute two summary statistics. The first, shown in the first column of Table D1, is the Pearson correlation between the direction and intensity measures, treating both as continuous variables. The second, shown in the second column, is the test statistic from a χ^2 test of independence of categorical variables. While the χ^2 test is most appropriate when treating both measures as categorical, the Pearson correlation has the advantage of being informative about the direction of the association: a positive correlation means that supporters assign more importance to the policy than opponents, while a negative correlation indicates the opposite. We report both tests and the two agree, rejecting the null hypothesis that directions and intensities are uncorrelated at $p < .001$ for 17 out of 22 questions.

---

[4]We omit respondents who say that they neither favor nor oppose the position, or that they are unsure, because there is no data on the intensity of these respondents' preferences.

[5]For instance, for three importance categories, we code 0 for not important at all, 0.5 for somewhat important, and 1 for very important. Although this is not the same as the intensity measure that we defined for Proposition 3 as the absolute difference in Borda scores between the feature of interest and the baseline, it is another valid way to capture preference intensity and a reasonable proxy for that quantity.

Returning to our running example of the preference for women, we see that the divergence between the preference intensities of supporters and opponents turns out to be more pronounced for espoused support for feminism than for any other question in the ANES. As Figure D1 shows, self-described feminists tend to attach much more importance to this identity than self-described "anti-feminists." On the left side of Figure D1, we take the sample of ANES respondents who answered the question "How well does the term 'feminist' describe you?" with "Very well" or "Extremely well,"[6] and we plot the proportions of this sample who answered the follow-up question "How important is it to you to be a feminist?" with "Not at all important," "A little important," "Somewhat important," "Very important," and "Extremely important," respectively. Nearly half of these feminist identifiers report that this issue is very important to them, with approximately another third calling it extremely important. By contrast, the right side of the figure shows the same distribution for the sample of respondents who answered the question "How well does the term 'anti-feminist' describe you?" with "Very well" or "Extremely well." The distribution of this intensity measure for "anti-feminists" is much flatter than the one for feminists: roughly half of the sample lands between "Not at all important" and "Somewhat important," with the other half reporting "Very important" or "Extremely important." Crucially, the sample on the right is those who identify strongly as *anti-feminists*, not merely those who fail to identify strongly as feminists, who would naturally be expected not to care deeply about the issue. Figure D1 thus presents strong empirical evidence in favor of the very dynamic that drove our stylized running example: there are a majority of voters who prefer men but care little about the issue, with a minority that prefers women but cares a great deal.

[6]The other choices were "Somewhat well," "Not very well," and "Not at all."

FIGURE D1. RESPONDENTS' IDENTIFICATION WITH FEMINIST/ANTI-FEMINIST LABELS, BY ISSUE IMPORTANCE

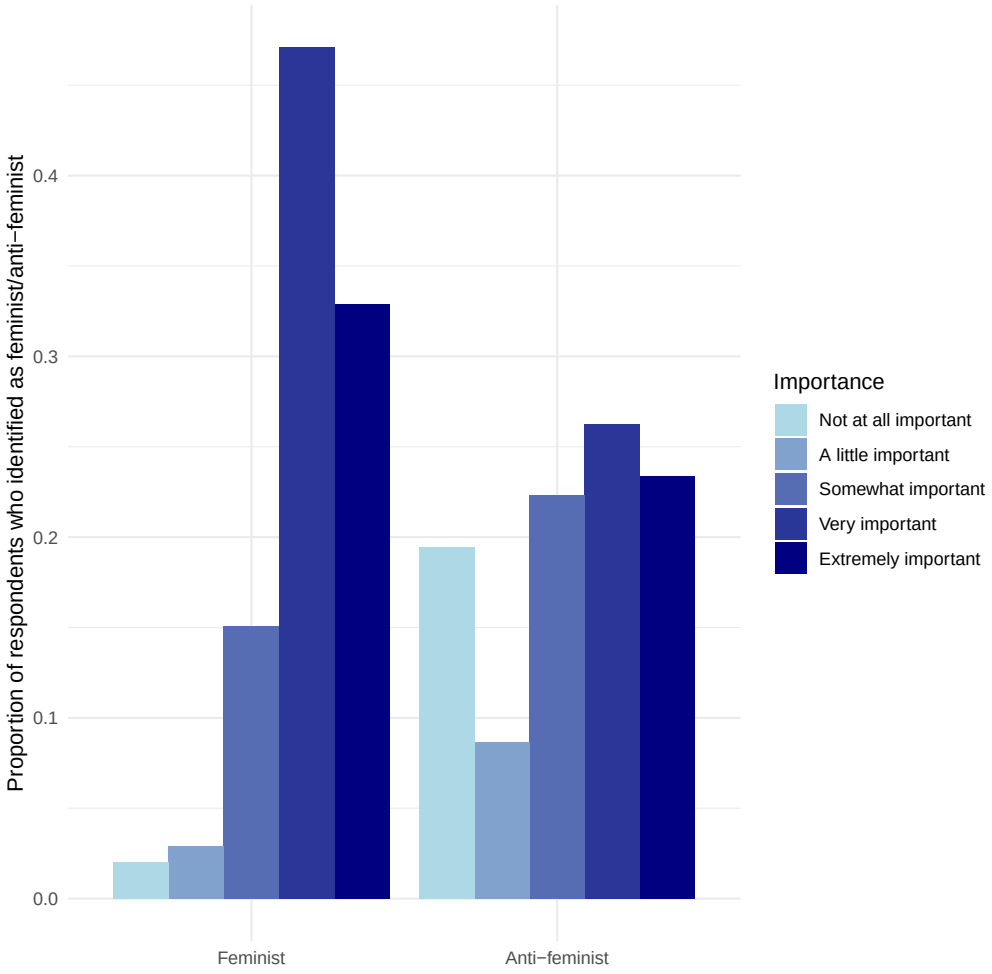

TABLE D1—CORRELATIONS BETWEEN WEIGHTS AND ATTRIBUTE PREFERENCES IN THE 2016 ANES

Question	**Pearson Correlation (p-value)**	χ^2 **statistic (p-value)**	**Number of intensity categories**
Favor allowing use of bathrooms of identified gender	-0.258 (0.000)	309.2 (0.000)	3
Favor torture for suspected terrorists	-0.246 (0.000)	147.8 (0.000)	3
Favor allowing Syrian refugees into US	-0.246 (0.000)	203.6 (0.000)	3
Favor 2010 health care law	-0.182 (0.000)	125.4 (0.000)	3
Support preferential hiring/promotion of blacks	-0.173 (0.000)	104.9 (0.000)	2
Favor building a wall with Mexico	-0.129 (0.000)	73.6 (0.000)	3
Favor affirmative action in universities	-0.098 (0.000)	15.7 (0.000)	2
Favor sending troops to fight ISIS	-0.080 (0.000)	21.7 (0.000)	3
Think economy has gotten better since 2008	-0.065 (0.000)	13.1 (0.000)	2
Agree that children brought illegally should be sent back	-0.025 (0.103)	2.7 (0.260)	3
Think government should make it harder to own a gun	-0.024 (0.289)	7.1 (0.069)	4
Approve of House incumbent	-0.013 (0.472)	0.5 (0.497)	2
Favor ending birthright citizenship	-0.011 (0.550)	1.6 (0.459)	3
Favor requiring provision of services to same-sex couples	0.020 (0.199)	8.8 (0.012)	3
Favor laws protecting gays against job discrimination	0.110 (0.000)	49.9 (0.000)	2
Think government should take more action on climate change	0.132 (0.000)	66.2 (0.000)	3
Favor requiring employers to give paid leave to new parents	0.149 (0.000)	29.1 (0.000)	2
Favor vaccines in schools	0.174 (0.000)	97.7 (0.000)	3
Support requiring equal pay for men and women	0.201 (0.000)	145.2 (0.000)	3
Favor the death penalty	0.211 (0.000)	184.2 (0.000)	2
Believe benefits of vaccination outweigh risks	0.275 (0.000)	251.4 (0.000)	3
The term 'feminist' describes you extremely/very well	0.330 (0.000)	115.1 (0.000)	5

E. Relaxing Separability

We have thus far focused on the scenario where voters had unconditional preferences over candidate features. In this section we explore the implications of altering this definition of preferences for features to allow for arbitrary interactions. For instance, we allow for the possibility that men are preferred to women only when the candidate is a Republican and the reverse when the candidate is a Democrat.[7] We derive a summary statistic for aggregate feature preferences that captures this more complex, and potentially more realistic, preference structure, and show that the bounds derived in Proposition 2 under separability are always smaller than the bounds we can construct when we relax separability. Thus, the AMCE is *less* informative about the fraction of voters who prefer a feature when preferences over features can interact. Furthermore, we discuss some interpretive limitations that applied researchers face when they allow respondents to have interactive preferences over features.

To start, 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 attributes. Formally:

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

where K and τ are defined as before, and thus K/τ represents the number of possible all-else-equal comparisons for the feature of interest. As in our example, 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.

Note that under separability $\Psi_i(t_1, t_0)$ can take only two values, 0 or 1, since voters make the same choice regardless of the other candidate features. Moreover, with separability, averaging the individual feature preference over respondents yields the proportion of individuals who prefer t_1 to t_0. When we relax separability, $\Psi_i(t_1, t_0)$ can take values in the interior of $[0, 1]$. We now define a preference for t_1 over t_0 as having $\Psi_i(t_1, t_0) > 1/2$ in this setting, and we derive the bounds on the

[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.

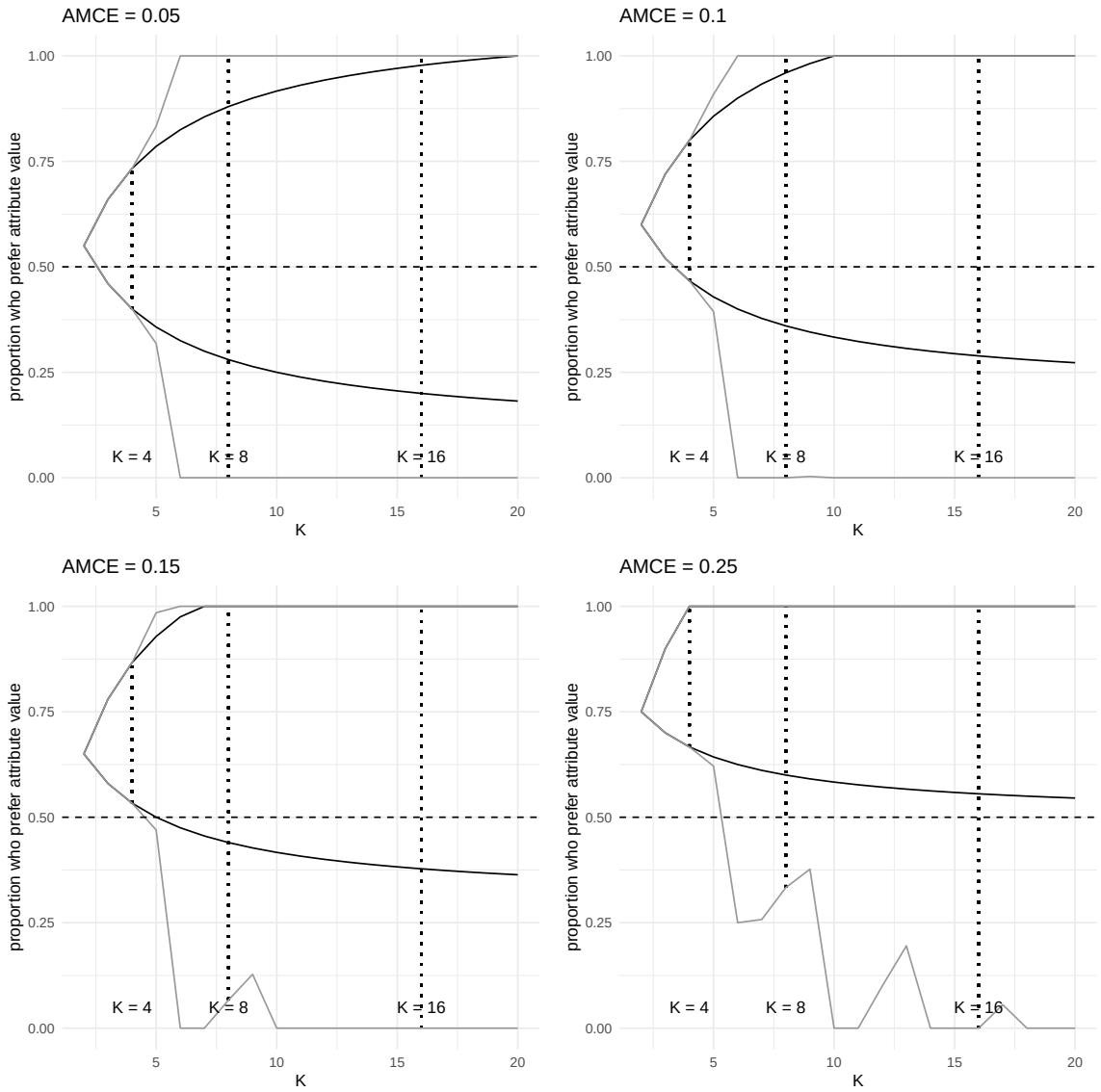

FIGURE E1. UPPER AND LOWER BOUNDS ON FRACTION OF PEOPLE WHO PREFER A BINARY FEATURE, CONSISTENT WITH AN AMCE OF .05, .10, .15, AND .25, RESPECTIVELY, AS A FUNCTION OF NUMBER OF POSSIBLE CANDIDATE PROFILES.

proportion of respondents who prefer t_1 to t_0 according to this definition. (See proof of Proposition 4 in Appendix A.)

In Figure E1, we recreate Figure 1, overlaying these bounds (in gray) over the bounds under the assumption of separability. The jaggedness of the bounds without separability is caused by the ceiling and floor functions in the equation, but regardless, Figure E1 reveals that our bounding exercise can no longer offer a practical remedy to researchers if separability is violated; in that case, they quickly grow to the full [0, 1] interval before $K = 16$ is reached, even with an AMCE as large as 0.25.

Next, we demonstrate that the proportion of respondents who prefer t_1 to t_0, or have an individual feature preference $\Psi_i > 1/2$, is indicative of electoral advantage only when separability holds. That is, without separability, even tight bounds indicating a majority of respondents having $\Psi_i > 1/2$ are not sufficient evidence to conclude that candidates with t_1 will beat candidates with t_0 in most all-else-equal contests.

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, out of all possible all-else-equal contests, and one-half:

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

In other words, $A(t_1, t_0)$ is the difference between the electorate-level analogue of $\Psi_i(t_1, t_0)$—the proportion of the time an electorate selects t_1 over t_0 in a simple-majority vote between all-else-equal alternatives, out of all possible all-else-equal contests—and one-half, and thus it captures the electoral (dis)advantage enjoyed by a candidate with feature t_1 compared to t_0.

First, consider the baseline case under separability. Here, whenever a majority of voters prefers t_1 to t_0, x_{j1} will beat x_{j0} in every all-else-equal contest j, and $A(t_1, t_0)$ will achieve its maximum value of $\frac{1}{2}$, so we can be confident that t_1 carries an electoral advantage over t_0. But this is no longer true when separability fails. We can illustrate this by way of a simple example. Consider a population of three voters with preferences over gender $\in \{M, F\}$ and party $\in \{D, R, I\}$ as in Table E1. Here, $\Psi_i(F, M) > 1/2$ for two out of three respondents, but $A(F, M) = -1/6$, indicating an electoral *disadvantage* for females despite the fact that the majority prefers this feature.

Finally, we show that without separability, the individual feature preference is potentially undesirable because it does not satisfy transitivity. To see this, suppose there are two ternary variables of interest, $P \in \{L, C, R\}$ and $E \in \{H, U, G\}$, and consider a voter whose ranking over candidate

Rank	V1	V2	V3
1.	MD	MR	MI
2.	FR	FI	MD
3.	MR	MI	MR
4.	FI	FD	FI
5.	MI	MD	FD
6.	FD	FR	FR
$\Psi_i(F, M)$	2/3	2/3	0

TABLE E1—PREFERENCES OVER CANDIDATE PROFILES - BOUNDS DO NOT INDICATE ELECTORAL ADVANTAGE WITHOUT SEPARABILITY

profiles is as follows:

$$RG \succ LG \succ CG \succ LU \succ CU \succ RU \succ CH \succ RH \succ LH$$

Looking at all-else-equal comparisons, this voter chooses R over L, L over C, and C over R in two of three comparisons, or $\Psi_i(R, L) = \Psi_i(L, C) = \Psi_i(C, R) = 2/3$. Thus, voter i prefers R to L, L to C, and C to R.

F. Structural Interpretation of the AMCE

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. 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 Candidate 1 and 2's respective platforms are given by:

$$\begin{aligned} U_i(\mathbf{x}_{ij1}) &= -\left(b_i - \mathbf{x}_{ij1}\right)^2 + \eta_{ij1} \\ U_i(\mathbf{x}_{ij2}) &= -\left(b_i - \mathbf{x}_{ij2}\right)^2 + \eta_{ij2} \end{aligned} \tag{F1}$$

While the imposition of quadratic loss may seem restrictive, in Lemma 4 in Appendix A we prove that our results are identical if we assume an absolute linear loss utility function. Regardless, it

follows that:

$$\begin{aligned}\Pr(y_{ij1}=1) &= \Pr(U_i(\mathbf{x}_{ij1}) > U_i(\mathbf{x}_{ij2})) \\ &= \Pr(-(b_i-\mathbf{x}_{ij1})^2+\eta_{ij1} > -(b_i-\mathbf{x}_{ij2})^2+\eta_{ij2}) \\ &= \Pr(\eta_{ij2}-\eta_{ij1} < 2(b_i'(\mathbf{x}_{ij1}-\mathbf{x}_{ij2})+\mathbf{x}_{ij2}'\mathbf{x}_{ij2}-\mathbf{x}_{ij1}'\mathbf{x}_{ij1})\end{aligned} \tag{F2}$$

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 F2 with a probit or logit-like regression. Instead consider a linear model of the form:

$$\begin{aligned}y_{ij1} &= 2(b_i'(\mathbf{x}_{ij1}-\mathbf{x}_{ij2})+\mathbf{x}_{ij2}'\mathbf{x}_{ij2}-\mathbf{x}_{ij1}'\mathbf{x}_{ij1})+\eta_{ij1}-\eta_{ij2} \\ &= \sum_m \left(2b_{im}(x_{ijm1}-x_{ijm2})+x_{ijm2}^2-x_{ijm1}^2\right)+\eta_{ij1}-\eta_{ij2} \\ &= \sum_m (2b_{im}-1)(x_{ijm1}-x_{ijm2})+\eta_{ij1}-\eta_{ij2} \\ &= \sum_m \beta_{im}\Delta x_{ijm}+\epsilon_{ij}\end{aligned} \tag{F3}$$

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:

$$y_{ij1} = \sum_m \Delta x_{ijm}\beta_m + \epsilon_{ij} \tag{F4}$$

where the estimated coefficient $\hat{\beta}_m$ recovers the AMCE for feature m.[8]

G. Additional Tables and Figures

[8]For a simple proof, see Lemma 3 in Appendix A.

Paper	*Journal*	*Voter Preference*	*Election*
Adida et al 2019	PLOS ONE	X	
Arnesen et al 2019	ES	X	X
Atkeson and Hamel 2020	PB	X	X
Auerbach and Thachil 2019	APSR	X	
Badas and Stauffer 2019	ES		X
Ballard-Rosa,Martin, and Scheve 2016	JOP	X	
Bansak et al 2016	Science	X	
Bechtel and Scheve 2013	PNAS	X	
Bechtel et al 2019	BJPS	X	
Berinsky et al 2018	PB	X	
Blackman and Jackson 2019	PB	X	X
Carnes and Lupu 2016	APSR	X	X
Clayton et al 2019	PB	X	X
Crowder-Meyer et al 2020	PB	X	X
de Geus et al 2020	PRQ		X
Dynes and Martin 2019	PB		X
Goggin et al 2019	PB		X
Hainmueller and Hopkins 2015	AJPS	X	
Hainmueller et al 2014	PA	X	X
Hainmueller et al 2015	PNAS	X	
Hankinson 2016	APSR	X	X
Hansen et al 2015	PB	X	X
Hemker and Rink 2017	AJPS		
Horiuchi et al 2018	PA	X	X
Horiuchi et al 2018	PSRM		
Huff and Kertzer 2018	AJPS		
Kirkland and Coppock 2018	PB	X	X
Leeper and Robison 2020	PB	X	X
Liebe et al 2018	PLOS ONE	X	
Martin and Blinder 2020	PB	X	X
Matsuo and Lee 2018	ES	X	X
Mummolo 2016	JOP	X	
Mummolo and Nall 2017	JOP	X	
Mummolo et al 2019	PB	X	
Oliveros and Schuster 2018	CPS	X	
Ono and Burden 2019	PB	X	
Sances 2018	PB	X	X
Sen 2017	PRQ	X	
Shafranek 2019	PB	X	
Smith 2020	PSRM	X	X
Smith et al 2018	PA	X	X
Teele et al 2018	APSR	X	X
Vivyan et al 2020	ES	X	X
Ward 2019	APSR	X	
Wright et al 2015	PB	X	

Table G1—This table describes our literature review describing 45 conjoint experiments by political scientists published between 2015 and 2020. The third column indicates if the authors describe their results with respect to voter preferences. The fourth column indicates if the authors relate their results to outcomes of elections.