AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

What Do We Learn about Voter Preferences from Conjoint Experiments?

Scott F. Abramson University of Rochester
Korhan Kocak New York University Abu Dhabi
Asya Magazinnik Massachusetts Institute of Technology

Abstract: *Political scientists frequently interpret the results of conjoint experiments as reflective of majority preferences. In this article, we show that the target estimand of conjoint experiments, the average marginal component effect (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.*

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.

Conjoint experiments have become a standard part of the political scientist's tool kit. 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 article, 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 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 article 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. Thus, 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 (e.g., "Americans prefer highly educated immigrants"), characterize electoral majorities (e.g., "voters want a lower tax rate"), or project winners of elections (e.g., "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

Scott F. Abramson is Associate Professor, Department of Political Science, University of Rochester, Harkness Hall, 333 Hutchinson Road, Rochester, NY 14627 (sabramso@ur.rochester.edu). Korhan Kocak is Assistant Professor, Division of Social Science, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE (kkocak@nyu.edu). Asya Magazinnik is Assistant Professor, Department of Political Science, MIT, 30 Wadsworth Street, Cambridge, MA 02142 (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.

American Journal of Political Science, Vol. 66, No. 4, October 2022, Pp. 1008–1020

DOI: 10.1111/ajps.12714

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, whereas 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. In other words, 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 who 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. Among these is, crucially, that candidates know voters' preferences up to a 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. Although 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.

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

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, for example, 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, 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, p. 245), and "[there is] strong evidence for progressive preferences over taxation among the American public" (Ballard-Rosa, Martin, and Scheve 2017, p. 14). Conjoint results are also used to make statements about candidates for elected office, such as "voters do not appear to prefer older politicians or celebrities, and are indifferent with regard to dynastic and family ties and gender" (Horiuchi, Smith, and Yamamoto 2020, p. 86), and "voters and legislators do

[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 to this article. This does not imply that our critique only applies to forced-choice conjoint experiments. 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.

not seem to hold female candidates in disregard; all else equal, they prefer female to male candidates" (Teele, Kalla, and Rosenbluth 2018, p. 537).[3]

Although 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; that voter 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 prefers 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 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.

TABLE 1 Preferences over Attributes

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

Note: This table gives voters' preferences over attributes. Candidates are described by two attributes: gender $\in$ {F, M} and party $\in$ {D, R}. Voters 1, 2, and 3 prefer male to female candidates and Republicans to Democrats; voters 4 and 5 have the opposite preferences.

We construct preferences over candidates from preferences over attributes in the following way: Voters prefer candidates who have both of the attributes they like to those who 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 has a greater weight for the voter. 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 $\gg$ 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 ($P \gg G$), whereas voters 4 and 5 place more on the candidate's gender ($G \gg P$).[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

[3]We conducted a review of conjoint analyses published in top political science journals and found that 83% of all articles 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 supporting information (SI).

[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'?

[5]Formally, completeness is defined as $x \succcurlyeq y$, $y \succcurlyeq x$ or both, and transitivity is defined as if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $x \succcurlyeq z$. We define a *strict preference relation* as $x \succ y$ if and only if $x \succcurlyeq y$ and *not* $y \succcurlyeq x$ 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.

[6]Note that these *relative* weights are meaningful within individuals but cannot be compared across respondents. That the minority care more intensely about gender than party does not imply that they care more intensely about gender than do 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.

TABLE 2 Preferences over Candidate Profiles

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

Note: This table presents preferences over profiles constructed from preferences over attributes.

in each head-to-head election for every possible pairwise comparison; the winner is boldfaced in the first column.

Next, we derive the AMCE for male over female candidates, following Proposition 3 from Hainmueller, Hopkins, and Yamamoto (2014). The intuition behind the comparisons being made when estimating the AMCE is given in Figure 1. 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. Finally, this sum is normalized by the number of possible profiles (four) times the number of possible profiles with a given gender (two) times the number of voters (five). 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.

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:

TABLE 3 Aggregate Preferences over Candidate Profiles

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

Notes: First column presents all possible head-to-head comparisons (with the winner of each contest indicated in bold). Columns V1-V5 show each voter's preferred candidate. Tally column summarizes the number of votes received by each candidate in each head-to-head comparison.

FIGURE 1 Obtaining the AMCE

$$
\begin{array}{lcll}
\bar{Y}(MR, MD) & - & \bar{Y}(FR, MD) & = & -2 \\
\bar{Y}(MR, FD) & - & \bar{Y}(FR, FD) & = & 0 \\
\bar{Y}(MR, MR) & - & \bar{Y}(FR, MR) & = & 1/2 \\
\bar{Y}(MR, FR) & - & \bar{Y}(FR, FR) & = & 1/2 \\
\bar{Y}(MD, MD) & - & \bar{Y}(FD, MD) & = & 1/2 \\
\bar{Y}(MD, FD) & - & \bar{Y}(FD, FD) & = & 1/2 \\
\bar{Y}(MD, MR) & - & \bar{Y}(FD, MR) & = & 0 \\
\bar{Y}(MD, FR) & - & \bar{Y}(FD, FR) & = & -2 \\
\hline
 & & & & -2 \\
 & & & & \\
\multicolumn{3}{c}{(\#\text{ of profiles}) \times (\#\text{ of voters})} & = & 40 \\
\multicolumn{3}{c}{\times(\#\text{ of profiles with a given gender })} & & \\
 & & & & \\
\multicolumn{3}{c}{\text{AMCE}} & = & -1/20
\end{array}
$$

Notes: Computing the AMCE for male over female candidates based on the aggregate preferences summarized in Table 3. $\bar{Y}(C_1, C_2)$ denotes the number of votes candidate C_1 obtains when running against candidate C_2.

(1) A majority of voters prefer female to male candidates.

As Table 1 indicates, a majority of voters (three out of five) 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, three out of five 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 four of the six 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, whereas 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, who 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.

Three Valid Interpretations of the AMCE

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 Figure 1 is simply the difference in votes that 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's 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 versus 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 ($P \gg G \gg R$) , and let voters 4 and 5 place the greatest weight on race, then gender, then party ($R \gg G \gg P$) . 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 who 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 4.

Since these are the same exact voters from the previous example, their preferences with respect to gender have not changed: Three out of five 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 conjoint experiments depend upon the specific set of profiles 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, such as 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 (2022), who highlight the sensitivity of the AMCE to the randomization scheme imposed by researchers. We show

[7] On this point, see Bansak et al. (2022).

[8] Male candidates win 13 of the 16 elections in which they face off against female candidates, and 19 of the 28 overall contests.

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

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

TABLE 4 Preferences over Candidate Profiles: Example Part II

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

Note: This table presents preferences over profiles constructed from preferences over attributes, including race.

that this is not just a *statistical* property of the AMCE, but a core *theoretical* feature of the aggregation mechanism that generates this quantity.

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, 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 who 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 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 he or she is preferred. This in turn is equal to the number of times that candidate would be chosen if the voter were 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 it is thus equal to the total number of times that candidate would be chosen if each voter were 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 supporting information.

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 in 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 and 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, producing the same expression as the AMCE in Equation 5 of Hainmueller, Hopkins, and Yamamoto (2014, IV).

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. We demonstrate how

different sets of candidates can lead to different AMCE estimates. A second social choice property of the Borda rule is that it violates the majority criterion, which states that if a majority of voters prefers one candidate, then that candidate must win. This property also extends to attributes. In our example, we showed that a majority of voters prefers 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 $\pi(t_1,\ t_0)$, it must be that*

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

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, that is, 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, that is, 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 SI Appendix C (p. XIX).

In Figure 2, we apply this proposition to compute the bounds for AMCEs of 0.05, 0.10, 0.15, and 0.25 for a binary feature, plotting the upper and lower bounds of the proportion of experimental subjects who prefer a binary feature on the y-axis against the number of potential candidate profiles that respondents can choose from on the x-axis. As the figure shows, even for 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 ensure 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.

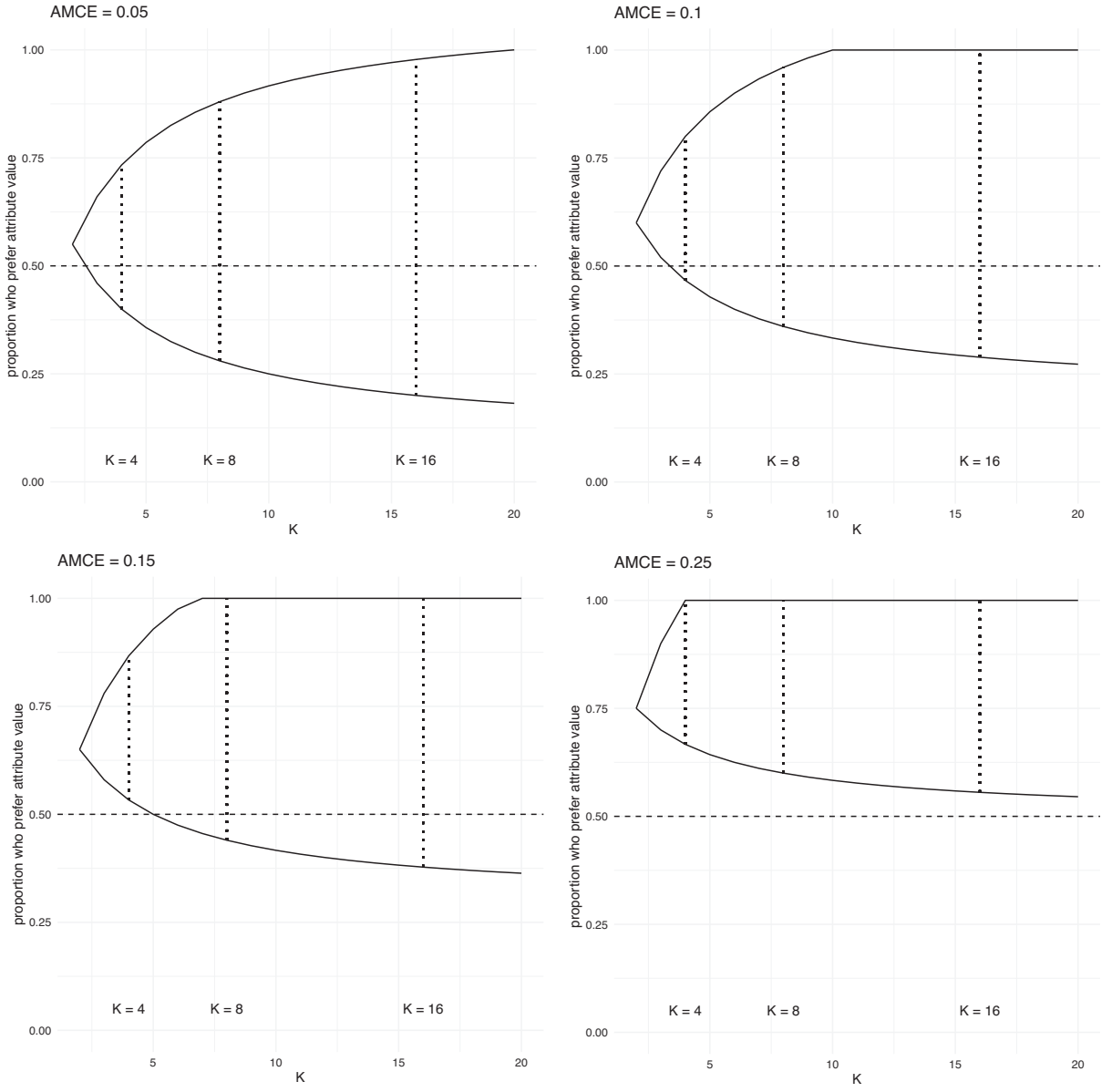
In Appendix Table C1 (p. XVIII), we conduct this exercise for every forced-choice conjoint experiment in the top three political science journals published between 2016 and the first quarter of 2019. We construct our

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

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

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

FIGURE 2 Upper and Lower Bounds on Fraction of People Who Prefer a Candidate Feature

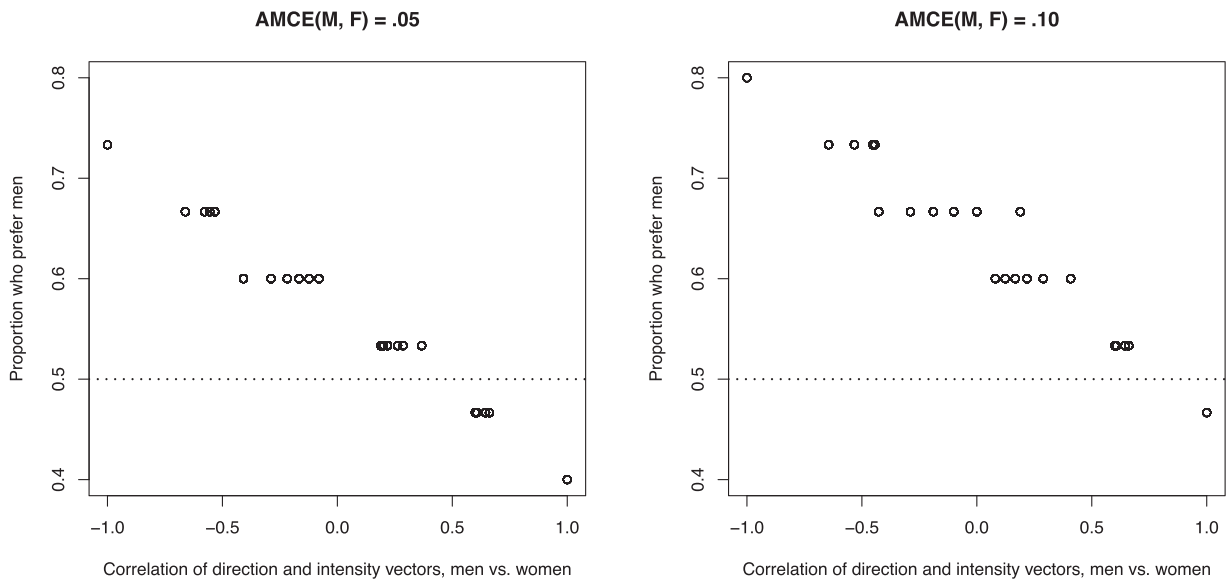


Notes: This figure shows the proportion of respondents who prefer an attribute value over the alternative consistent with an AMCE of 0.05, 0.10, 0.15, and 0.25, respectively, as a function of the number of possible candidate profiles.

bounds for the largest estimated effect presented in each of these articles. From the eight articles we analyze, only one, Mummolo (2016), produces bounds that guarantee a majority preference. In this article, 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 SI Appendix C (p. XVI), 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 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.

FIGURE 3 Distributions of Preferences that Can Generate the Same AMCE



Notes: This figure plots combinations of proportions of respondents who prefer male to female candidates and correlations between direction and intensity of male-female preference for AMCEs of 0.05 (left) and 0.10 (right), computed for 15 respondents and two binary attributes.

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 15 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, that is, $\mathbb{1}\{Y_i\ (t_1,\ t_0) =\ 1\}$. In Figure 3, 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 who prefer men over women that is consistent with 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 ensures that the sign of the AMCE indicates the majority preference. Note, however, that Figure 3 corresponds to the most charitable case, as pictured for $K = 4$ in the bounds in Figure 2. 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 3 illustrates a general rule of thumb for 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.

Furthermore, it can be seen in Figure 3 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 his or her preference—the sign of the AMCE must correspond to

[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 candidates, a proportion of the sample that prefers male over female candidates, and a correlation of direction and intensity. Figure 3 displays all of these possibilities for a given AMCE.

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 it underestimates the size of the margin.*

Proof of Proposition 3 closely follows 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 22 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 17 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 SI Appendix D (p. XX) 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 possibility that a voter prefers Republican men to Republican women, but Democratic women to Democratic men. In SI Appendix E (p. XXIV), 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.

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 $\boldsymbol{x}_{ijc}$ to voter i. A platform $\boldsymbol{x}_{ijc}$ 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., her 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 utility from Candidate c's platform is given by

$$U_i\left(\boldsymbol{x}_{ijc}\right) = -\left(b_i - \boldsymbol{x}_{ijc}\right)^2 + \eta_{ijc}. \tag{1}$$

It follows that

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

[13] A simple example is provided in SI Appendix E (p. XXIV).

where y_{ij1} is a binary indicator that equals 1 when respondent i chooses Candidate 1 in contest j and 0 otherwise. In SI Appendix F (p. XXVII), we walk through the steps that relate Equation (2) to a linear regression model estimated on data generated from a conjoint experiment, where $\boldsymbol{x}_{ij1}$ and $\boldsymbol{x}_{ij2}$ are vectors of randomized candidate attributes that have been discretized into binary indicators with an omitted category. Letting $\Delta\boldsymbol{x}_{ij}$ represent the difference between the vectors $\boldsymbol{x}_{ij1}$ and $\boldsymbol{x}_{ij2}$, and m represent a given feature or element of this vector, one can estimate the following regression:

$$y_{ij1} = \sum_m \beta_{im}\Delta\boldsymbol{x}_{ijm} + \varepsilon_{ij}. \quad (3)$$

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 (Bernhardt, Duggan, and Squintani 2007; Calvert 1985; Duggan 2006). 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]

[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.

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

First, as shown by Bansak et al. (2022), 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 toward 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 such as 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 (Cavaillé, Chen, 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. Society for Political Methodology. https://www.korhankocak.com/publication/akms/AKMS.pdf.

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

Ballard-Rosa, Cameron, Lucy Martin, and Kenneth Scheve. 2017. "The Structure of American Income Tax Policy Preferences." *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*, edited by David Austen-Smith and John Duggan. New York: Springer, 15–56.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2022. "Using Conjoint Experiments to Analyze Election Outcomes: The Essential Role of the Average Marginal Component Effect." *Political Analysis First View*, pp. 1–19.

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* 29(1): 69–95.

Cavaillé, Charlotte, Daniel L. Chen, and Karine van Der Straeten. 2019. "A Decision-Theoretic Approach to Understanding Survey Response: Likert vs. Quadratic Voting for Attitudinal Research." University of Chicago Press: *University of Chicago Law Review*, 87: 22–43.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2022. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis* 30(1): 19–45.

Duggan, John. 2006. "A Note on Uniqueness of Electoral Equilibrium When the Median Voter Is Unobserved." *Economics Letters* 92(2): 240–44.

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): 529–48.

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): 208–19.

Horiuchi, Yusaku, Daniel M. Smith, and Teppei Yamamoto. 2020. "Identifying Voter Preferences for Politicians' Personal Attributes: A Conjoint Experiment in Japan." *Political Science Research and Methods* 8: 75–91.

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." *Journal of Politics* 78(3): 763–73.

Schofield, Norman. 2007. "The Mean Voter Theorem: Necessary and Sufficient Conditions for Convergent Equilibrium." *Review of Economic Studies* 74(3): 965–80.

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): 525–41.

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

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Proofs
Appendix B: Robustness of the AMCE to the Inclusion/Exclusion of Additional Treatments
Appendix C: Bounds on Proportion of Experimental Sample Who Prefer a Feature
Appendix D: Correlations between Direction and Intensity of Preferences in the 2016 ANES
Appendix E: Relaxing Separability
Appendix F: Structural Interpretation of the AMCE
Appendix G: Additional Tables and Figures