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

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

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

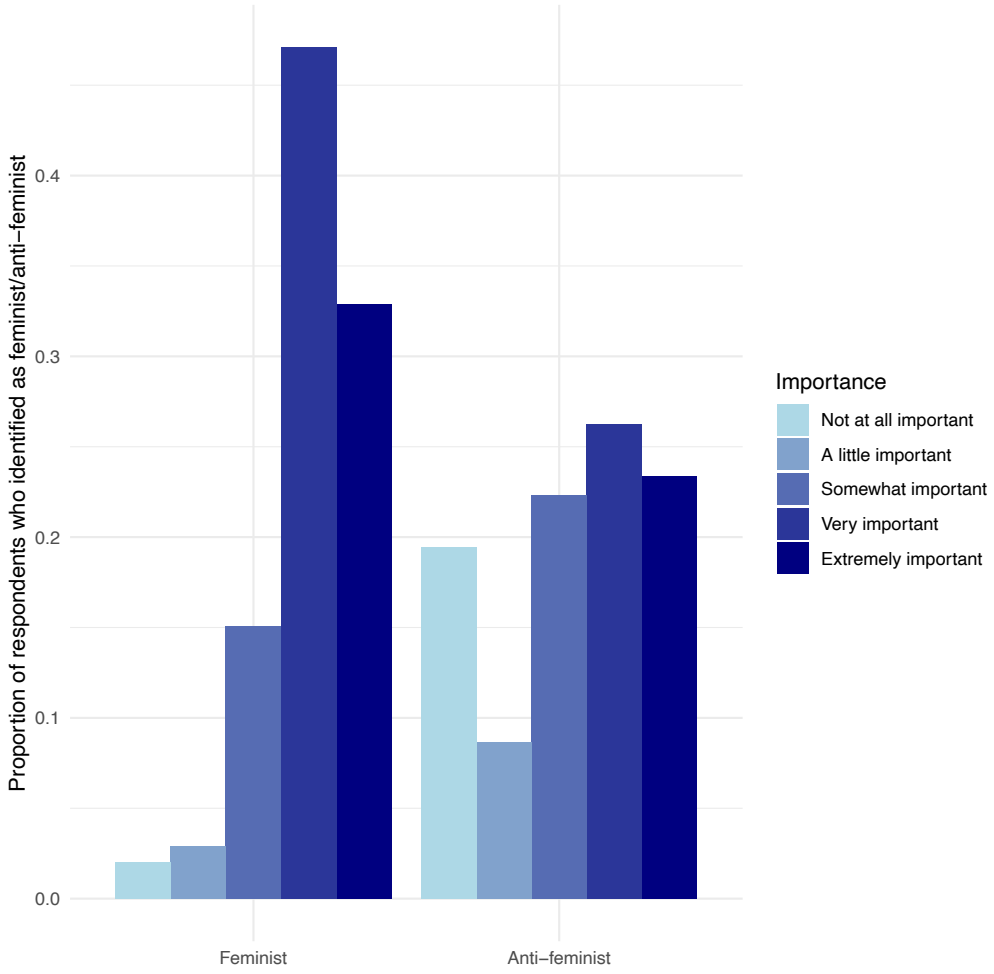

Table D1—Correlations Between Weights and Attribute Preferences in the 2016 ANES

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.

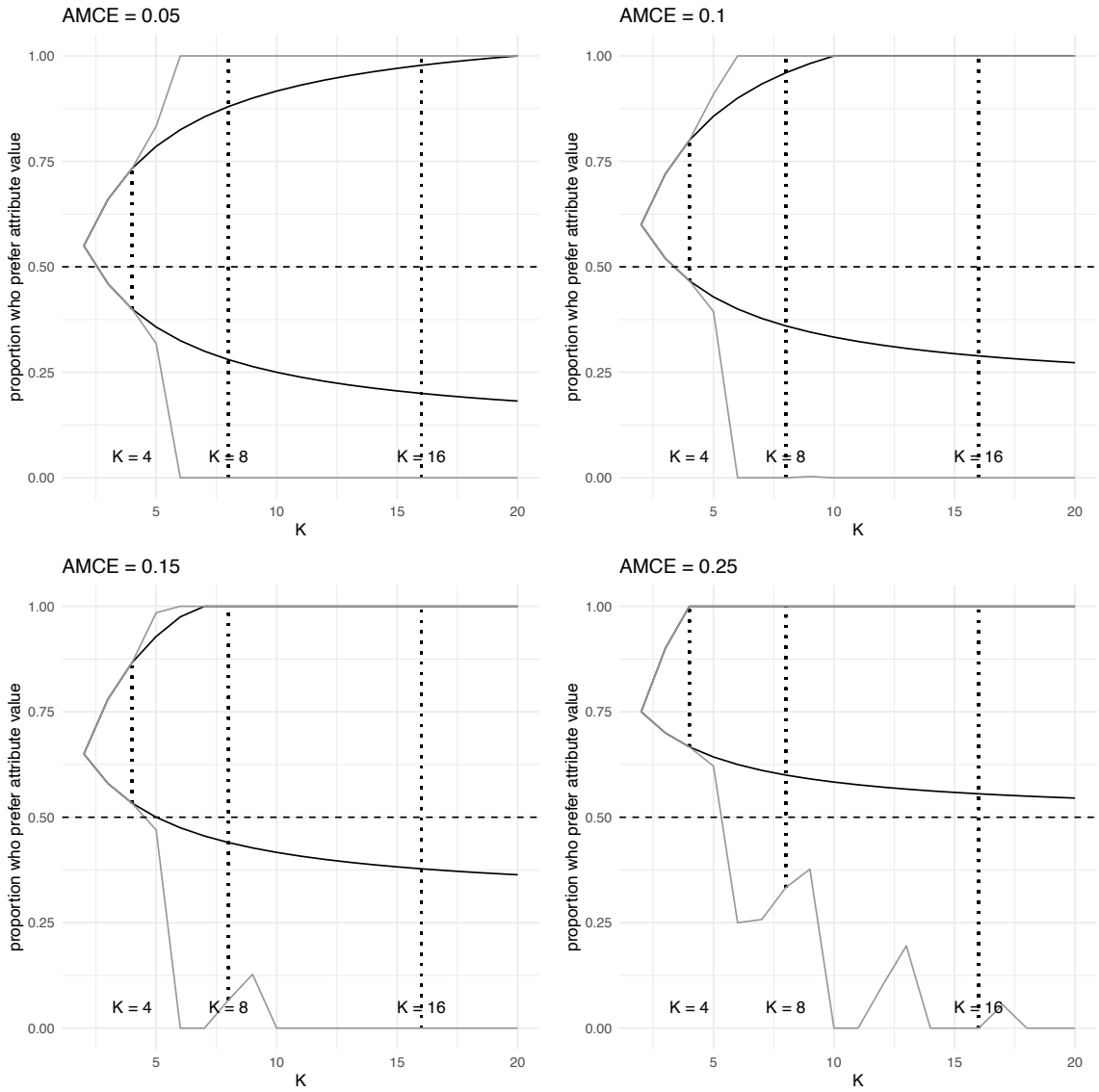

FIGURE E1. UPPER AND LOWER BOUNDS ON FRACTION OF PEOPLE WHO PREFER A BINARY FEATURE, CONSISTENT WITH AN AMCE OF .05, .10, .15, AND .25, RESPECTIVELY, AS A FUNCTION OF NUMBER OF POSSIBLE CANDIDATE PROFILES.

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.