Measuring Subgroup Preferences in Conjoint Experiments*

Thomas J. Leeper, Sara B. Hobolt, and James Tilley

May 24, 2019

Abstract

Conjoint analysis is a common tool for studying political preferences. The method disentangles patterns in respondents' favorability toward complex, multidimensional objects, such as candidates or policies. Most conjoints rely upon a fully randomized design to generate average marginal component effects (AMCEs). These measure the degree to which a given value of a conjoint profile feature increases, or decreases, respondents' support for the overall profile relative to a baseline, averaging across all respondents and other features. While the AMCE has a clear causal interpretation (about the *effect* of features), most published conjoint analyses also use AMCEs to describe *levels* of favorability. This often means comparing AMCEs among respondent subgroups. We show that using conditional AMCEs to describe the degree of subgroup agreement can be misleading as regression interactions are sensitive to the reference category used in the analysis. This leads to inferences about subgroup differences in preferences that have arbitrary sign, size, and significance. We demonstrate the problem using examples drawn from published articles and provide suggestions for improved reporting and interpretation using marginal means and an omnibus F-test. Given the accelerating use of these designs in political science, we offer advice for best practice in analysis and presentation of results.

*We thank Benjamin Lauderdale, Jamie Druckman, Yusaku Horiuchi, the editor, and anonymous reviewers for feedback on this manuscript. Replication data and code for this article are available from the *Political Analysis* Dataverse: `https://doi.org/10.7910/DVN/ARHZU4`. This work was funded, in part, by the United Kingdom Economic and Social Research Council (Grant ES/R000573/1).

One aspect of the dramatic increase in the use of experiments within political science (Druckman et al., 2006; Mutz, 2011) is the establishment of conjoint experimental designs as a prominent methodological tool. While survey experiments have traditionally examined just one or two factors that might shape outcomes (see, for reviews, Gaines, Kuklinski, and Quirk, 2007; Sniderman, 2011), conjoint designs allow researchers to study the independent effects on preferences of many features of complex, multidimensional objects. These include many different types of phenomena, such as political candidates (Campbell et al., 2016; Teele, Kalla, and Rosenbluth, 2018), immigrant admissions (Hainmueller and Hopkins, 2015; Bansak, Hainmueller, and Hangartner, 2016; Wright, Levy, and Citrin, 2016), and public policies (Gallego and Marx, 2017; Hankinson, 2018). Factorial designs of this sort have a long history, but the driving force behind this use of conjoint analysis has been the introduction by Hainmueller, Hopkins, and Yamamoto (2014) of a small-sample, fully randomized conjoint design. The associated analytic approach emphasizes a single quantity of interest: the average marginal component effect (AMCE). By capturing the multidimensionality of target objects, the randomized conjoint design breaks any explicit, or implicit, confounding between features of these objects. This gives the AMCE a clear causal interpretation: the degree to which a given value of a feature increases, or decreases, respondents' favorability towards a packaged conjoint profile relative to a baseline.

While randomization of profile features gives the AMCE a causal interpretation, most published conjoint analyses in political science use AMCEs not only for *causal* purposes (interpreting AMCEs as effect sizes), but also for *descriptive* purposes. The aim is to map levels of favorability toward a multidimensional object across its various features.[1] In this sense, conjoints are often applied like list experiments, using randomization to measure a sample's preferences over something difficult to measure with direct questioning. A positive AMCE for a given feature can be read as a descriptive measure of high favorability towards profiles with that feature. The quantity is causal, but it is often read descriptively.

[1]See Shmueli (2010) for an elaboration on the distinctions between explanatory (causal) modelling, descriptive modelling, and predictive modelling.

This is particularly the case for subgroup analyses of conjoint experiments. Such exercises are an increasingly common feature of experimental analysis (Green and Kern, 2012; Ratkovic and Tingley, 2017; Grimmer, Messing, and Westwood, 2017; Egami and Imai, 2018). For example, the Hainmueller, Hopkins, and Yamamoto (2014) study of immigration attitudes splits the sample in two using a measure of ethnocentrism and then compares AMCEs for the two subgroups. Similarly, Bansak, Hainmueller, and Hangartner (2016) compare preferences toward immigrants across number of binary respondent characteristics: age, education, left-right ideology, and income. Other examples abound. Ballard-Rosa, Martin, and Scheve (2016) compare preferences over tax policies across a number of subgroups defined by demographics and political orientations; Bechtel and Scheve (2013) compare AMCEs on climate agreements across four different countries, and across subgroups of respondents; and Teele, Kalla, and Rosenbluth (2018) compare AMCEs for features of male and female political candidates among male and female respondents. Most of these comparisons are visual or informal. But some involve explicit estimation of the subgroup difference, such as when Kirkland and Coppock (2017) compare conditional AMCEs across hypothetical partisan and nonpartisan elections. Interpretation of subgroup AMCEs thus involves an implied quantity of interest: the *difference* between two conditional AMCEs.

What is not necessarily obvious in such analyses is that differences-in-preferences (that is to say, the difference in degree of favorability toward profiles containing a given feature) are not directly reflected in subgroup differences-in-AMCEs. A difference in effect sizes is distinct from a difference in preferences. We show that a difference in two (or more) subgroups' favorability toward a conjoint feature — like a difference in willingness to support a particular type of immigrant between high and low ethnocentrism respondents — is only rarely reflected in the difference-in-AMCEs. In fact, no information about the similarity of the subgroups' preferences is provided by comparisons of subgroup AMCEs, yet such comparisons are commonly made in practice.

As we will show, where preferences in subgroups toward the experimental ref-

erence category are similar, the difference-in-AMCEs conveys preferences reasonably well. The problem occurs when preferences between subgroups diverge in the reference category. Here, the difference-in-AMCEs is a misleading representation of underlying patterns of favorability. Given most published conjoint studies report results based upon reference categories chosen for *substantive* reasons about the nature or meaning of the levels rather than the configuration of preferences revealed in the experiment, difference-in-AMCEs should not be assumed to be interpretable as differences in subgroup preferences. The root of this error is likely familiar to many researchers: it is simply a matter of regression specification for models involving interactions between categorical regressors. Egami and Imai (2018), for example, provide an extensive discussion of the implications of this property for interpreting causal interactions between randomized features of conjoint profiles. The state of the published literature would suggest the problem remains non-obvious when applied to descriptive analysis of subgroups in conjoint designs.[2]

In what follows, we demonstrate the challenges of conjoint analysis and remind readers of how reference category choice for profile features creates problems for comparing conditional AMCEs across respondent subgroups. We show how the use of an arbitrary reference category means the size, direction, and statistical significance of differences-in-AMCEs have little relationship to the underlying degree of favorability of the subgroups toward profiles with particular features. Reference category choices can make similar preferences look dissimilar and dissimilar preferences look similar. We demonstrate this with examples drawn from the published political science literature (namely experiments by Hainmueller, Hopkins, and Yamamoto 2014; Bechtel and Scheve 2013; Teele, Kalla, and Rosenbluth 2018). The paper then provides suggestions for improved conjoint reporting and interpretation based around two quantities of interest drawn from the factorial experimentation literature: (a) unadjusted marginal means, a quantity measuring favorability toward a given feature, and (b) an omnibus

[2]Since this manuscript has been under review, we have been made aware of one working paper by Clayton, Ferwerda, and Horiuchi (2018), on the topic of immigration preferences, that correctly notes the need to address the arbitrary reference category in order to compare subgroup preferences.

F-test, measuring differences therein. Software for the R programming language to support our findings — and that can be used to examine sensitivity of conjoint analysis to reference category selection, calculate AMCEs and marginal means, perform subgroup analyses, and test for subgroup differences in any conjoint experiment (Leeper, 2018) — is demonstrated throughout using example data (Leeper, Hobolt, and Tilley, 2019). We conclude with advice for best practices in the analysis and presentation of conjoint results.

Quantities of Interest in Conjoint Experiments

Conjoint analysis serves two purposes. One is to assess causal effects. Another is preference description.[3] In causal inference, fully randomized conjoints provide a design and analytic approach that allows researchers to understand the causal effect of a given feature on overall support for a multidimensional object, averaging across other features of the object included in the design. Such inferences can be thought of as statements of the form: "shifting an immigrant's country of origin from India to Poland increases favorability by X percentage points." In descriptive inference, conjoints provide information about both (a) the *absolute* favorability of respondents toward objects with particular features or combinations of features, and (b) the *relative* favorability of respondents toward an object with alternative combinations of features. Such inferences can be thought of as statements of the form "Polish immigrants are preferred by X% of respondents" or "Polish immigrants are more supported than Mexican immigrants, by X percentage points." Thus both causal and descriptive interpretations of conjoints are based upon the distribution of preferences across profile features and differences in preferences across alternative feature combinations.

Analytically, a fully randomized conjoint design without constraints between profile features is simply a full-factorial experiment (with some cells possibly, albeit ran-

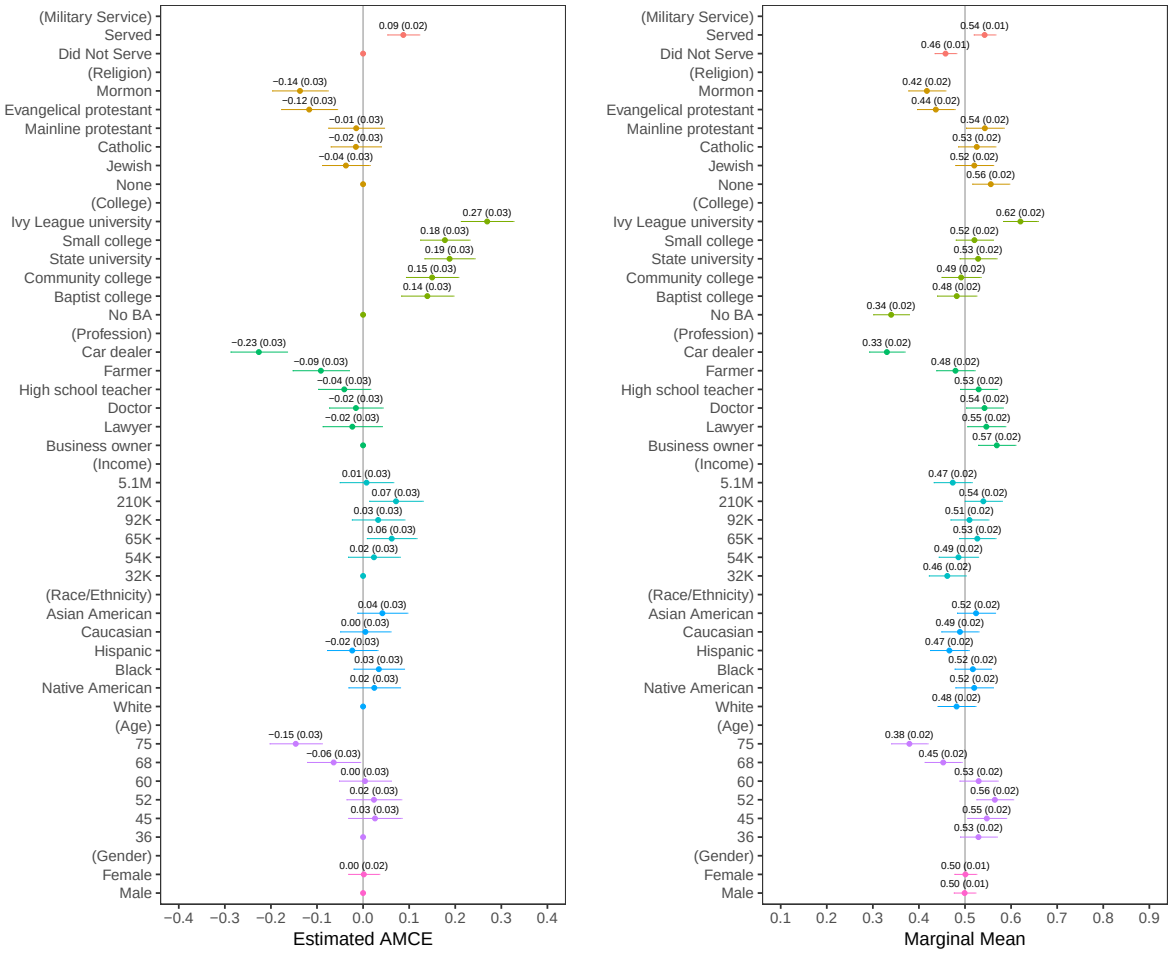
[3]Here we use "preference" as Hainmueller, Hopkins, and Yamamoto (2014) do: that is, as a statement of *favorability* or *support* for a profile, not the more narrow economic definition of a strict rank ordering of objects by favorability.

domly, left unobserved). All quantities of interest relevant to the analysis of conjoint designs therefore derive from combinations of cell means, marginal means, and the grand mean, as in the traditional analysis of factorial experiments. In a forced choice conjoint design, the *grand mean* is by definition 0.5 (i.e., 50% of all profiles shown are chosen and 50% are not chosen). *Cell means* are the mean outcome for each particular combination of feature levels. In the full-factorial design discussed by Hainmueller, Hopkins, and Yamamoto (2014) and now widely used in political science, many or perhaps most cell means are unobserved. For example, in their candidate choice experiment, there are $2 * 6 * 6 * 6 * 2 * 6 * 6 * 6 = 186,624$ cell means, but only 3,466 observations. About 98% of cell means are unobserved. While this would be problematic for attempting to infer pairwise comparisons between cells, conjoint analysts mostly focus on the marginal effects of each feature rather than more complex interactions. Appendix A provides detailed notation and elaborations of these definitions of quantities of interest.

In fully randomized designs, the average marginal component effects (AMCEs) are simply marginal effects of changing one feature level to another, all else constant. AMCEs therefore depend only upon *marginal means*: that is the column and row mean outcomes for each feature level averaging across all other features. A marginal mean describes the level of favorability toward profiles that have a particular feature level, ignoring all other features. For example, in the common forced-choice design with two alternatives, marginal means have a direct interpretation as probabilities. A marginal mean of 0 indicates respondents select profiles with that feature level with probability $P(Y = 1|X = x) = 0$. While a marginal mean of 1 indicates respondents select profiles with that feature level with probability $P(Y = 1|X = x) = 1$, where Y is a binary outcome and X is a vector of profile features.[4] With rating scale outcomes, marginal

[4]It is not possible for the marginal mean to equal zero or one if pairs of profiles shown together are allowed to have the same level of a given feature (for example, both immigrants are from Germany). Instead, the marginal mean can range from the probability of co-occurrence to 1 minus that probability. If there are five levels of a feature, each shown with equal probability, then the probability of co-occurrence is $\frac{1}{5} * \frac{1}{5} = 0.04$ such that the marginal mean can take values in the range $(0.04, 0.96)$. If the design is constrained so that features cannot be the same for both immigrants, then the marginal means fully range from zero to one. This constraint on the range of the marginal means also constrains the range of AMCEs. Notably, many conjoints provide features with only two levels, such as the male-versus-female

Figure 1: Replication of Hainmueller et al. (2014) Candidate Experiment using AMCEs and MMs



means can vary arbitrarily along the outcome scale used.

Because levels of features are randomly assigned, pairwise differences between two marginal means for a given feature (e.g., between candidates who are male versus female) have a direct causal interpretation. For fully randomized designs, the AMCE proposed by Hainmueller, Hopkins, and Yamamoto (2014) is equivalent to the average marginal effect of each feature level for a model where each feature is converted into a matrix of indicator variables with one level left out as a reference category. This is no different from any other regression context wherein one level of any categorical variable must be omitted from the design matrix in order to avoid perfect multi-

candidate feature examined by Teele, Kalla, and Rosenbluth (2018) or Hainmueller, Hopkins, and Yamamoto (2014) in their conjoints on candidate choice. In such cases, the probability of co-occurrence is $\frac{1}{2} * \frac{1}{2} = 0.25$ bounding the AMCE for female (as opposed to male) candidates to the range $(-0.5, 0.5)$ if both candidates can have the same sex. Caution is therefore needed in comparing the relative size of features with few levels to features with many levels given that effects have different bounds.

collinearity.[5] This close relationship between AMCEs and marginal means is visible in Figure 1 which presents a replication of the AMCE-based analysis of the Hainmueller et al. candidate experiment (left panel) and an analogous examination of the results using marginal means (right panel). Note, in particular, how marginal means convey information about the preferences of respondents for all feature levels while AMCEs definitionally restrict the AMCE for the reference category to zero (or undefined). For example, the AMCE for a candidate serving in the military is 0.09 (or a 9-percentage point) increase in favorability, reflecting marginal means for serving and non-serving candidates of 0.46 and 0.54, respectively. Similarly, the zero effect size for candidate gender reflects identical marginal means for male and female candidates (0.50 in each case). AMCEs in fully randomized designs are simply differences between marginal means at each feature level and the marginal mean in the reference category, ignoring other features.

The AMCE is often described as an estimate of the relative favorability of profiles with counterfactual levels of a feature. For example, Teele, Kalla, and Rosenbluth (2018) summarize their conjoint on public support "female candidates are favored [over men] by 7.3 percentage points" (6). Similarly, Hainmueller, Hopkins, and Yamamoto (2014) describe some of the results of conjoint on preferences toward political candidates:

> We also see a bias against Mormon candidates, whose estimated level of support is 0.06 (SE = 0.03) lower when compared to a baseline candidate with no stated religion. Support for Evangelical Protestants is also 0.04 percentage points lower (SE = 0.02) than the baseline. (19)

These examples make clear that despite the *causal* inference potentially provided by the AMCE, the quantity of interest is frequently used to provide a characterization of a preferences that has a distinctly descriptive flavor about the relative *levels* of support

[5] In designs that entail constraints between profile features, the average marginal effect is a weighted average of effects across each combination of the constrained features where the weights on the effects are arbitrary but typically uniform. We ignore this distinction in the remainder of this article, as all of our results apply equally to fully randomized and to constrained designs.

across profiles and also across subgroups of respondents. Indeed, this style of description is widespread in conjoint analyses. This use of conjoints to provide descriptive inferences about patterns of preferences is important because AMCEs are defined as *relative* quantities, requiring that patterns of preferences are expressed against a baseline, reference category for each conjoint feature. A positive AMCE is read as higher favorability but it is only higher relative to whatever category serves as the baseline. For example, in the Hainmueller, Hopkins, and Yamamoto candidate example, choosing a non-religious candidate as a baseline and interpreting the resulting AMCES means that the differences between other pairs of marginal means (e.g., evaluations of Mormon and Evangelical candidates) are not obvious. The negative direction, and the size, of the AMCEs for Mormon and Evangelical candidates would be different if the least-liked category of Mormons were the reference group. More trivially, Teele, Kalla, and Rosenbluth (2018) describe their comparisons about public preferences for female candidates relative to male candidates, but could have equivalent described patterns of equal size but opposite sign comparing preferences over male relative to female candidates. Appendix B includes some additional illustrations of this point for interested readers.

Consequences of Arbitrary Reference Category Choice

How do researchers decide which of tens of thousands of possible experimental cells should be selected as the reference category? Examining recently published conjoint analyses, it appears that the choice of reference category is either arbitrary or based upon substantive intuition about the meaning of feature levels. For example, Hainmueller, Hopkins, and Yamamoto (2014) choose female immigrants as a baseline in their immigration experiment, thus providing an estimate of the AMCE of being male, while Teele, Kalla, and Rosenbluth (2018) choose male candidates as a baseline in their conjoint, thus providing an estimate of the AMCE of being female. The choice is seemingly innocuous. Sometimes choices of reference category appear to be driven

by substantive knowledge: on language skills of immigrants in their immigration experiment, Hainmueller, Hopkins, and Yamamoto (2014) choose fluency as a baseline; on the prior trips to the US feature, "never" is chosen as the baseline.

While seemingly arbitrary and innocuous, the choice of reference category can provide highly distorted descriptive interpretations of preferences among subgroups of respondents. This occurs when researchers examine *conditional* AMCEs, wherein AMCEs are calculated separately for subgroups of respondents and those conditional estimates are directly compared (Hainmueller, Hopkins, and Yamamoto, 2014, 13). Conditional AMCEs convey the causal effect of an experimental factor on overall favorability among the subgroup of interest. Consider, for example, a two-condition candidate choice experiment where Democratic and Republican respondents are exposed to either a male or female candidate and opinions toward the candidate serve as the outcome. It is reasonable to imagine that effects of candidate sex might differ for the two groups and therefore to compare the size of treatment between the two groups. Perhaps Democrats are more responsive to candidate sex than are Republicans, making the causal effect larger for Democrats than Republicans. When conjoint analysts engage in subgroup comparisons, they are engaging in this kind of search for heterogeneous treatment effects across subgroups, but across a much larger number of experimental factors.

As Table 1 shows, discussions of conditional AMCEs in conjoint analyses often compare the size, and direction, of subgroup causal effects. Given the common practice of descriptively interpreting conjoint experimental results, such subgroup analyses seem perfectly intuitive. The set of subgroups listed in the last column of Table 1 contains some unsurprising covariates, such as partisanship, that are of obvious theoretical interest in almost any study of individual preferences. If interpreted as a difference in the size of the *causal effect* for two groups, such comparisons are perfectly consistent with more traditional experimental analysis and a perfectly acceptable interpretation of the conjoint results.

Yet, just as analysis of full sample conjoint data is often descriptive in nature, it

Table 1: Uses of Subgroup Analysis Published in Political Science Journals

Paper	Journal	Topic	Subgroup Comparisons
Bechtel and Scheve (2013)	PNAS	Climate agreement preferences	Environmentalism and International Reciprocity Attitudes
Franchino and Zucchini (2014)	PSRM	Candidate preferences	Political Interest, Left-right self-placement
Hainmueller, Hopkins, and Yamamoto (2014)	Political Analysis	Immigration preferences	Ethnocentrism
Hansen, Olsen, and Bech (2014)	Politcal Behavior	Policy preferences	Partisanship
Carlson (2015)	World Politics	Candidate preferences	Co-ethnicity
Bansak, Hainmueller, and Hangartner (2016)	Science	Immigration preferences	Left-right self-placement, age, education, income
Ballard-Rosa, Martin, and Scheve (2016)	JOP	Tax preferences	Various
Campbell et al. (2016)	BJPS	Candidate preferences	Partisanship
Carnes and Lupu (2016)	APSR	Candidate preferences	Partisanship
Mummolo (2016)	JOP	News selection	Various
Vivyan and Wagner (2016)	EJPR	Candidate preferences	Political attitudes
Mummolo and Nall (2017)	JOP	Mobility preferences	Partisanship
Bechtel, Genovese, and Scheve (2017)	BJPS	Climate agreement preferences	Employment sector emissions
Bechtel, Hainmueller, and Margalit (2017)	EJPR	International bailout preferences	Various
Gallego and Marx (2017)	J. European Public Policy	Labor market policy	Left-right self-placement
Kirkland and Coppock (2017)	Political Behavior	Candidate preferences	Partisanship
Sen (2017)	PRQ	Judicial candidate preferences	Partisanship
Sobolewska, Galandini, and Lessard-Phillips (2017)	J. Ethnic & Migration Studies	Immigrant integration	Various
Eggers, Vivyan, and Wagner (2018)	JOP	Candidate preferences	Sex
Hankinson (2018)	APSR	Housing policy preferences	Various
Oliveros and Schuster (2018)	CPS	Bureaucrat candidate preferences	Various
Teele, Kalla, and Rosenbluth (2018)	APSR	Candidate preferences	Sex, Partisanship
Carey et al. (2018)	Politics, Groups, and Identities	Hiring preferences	Various

All articles in this table use subgroup conditional AMCEs to make inferences about differences in preferences between subgroups.

is also the case that conjoint analysts frequently interpret differences in conditional AMCEs descriptively rather than causally. For example, in one analysis Hainmueller, Hopkins, and Yamamoto (2014) visually compare the pattern of AMCEs among high- and low-ethnocentrism respondents and interpret that "the patterns of support are generally similar for respondents irrespective of their level of ethnocentrism" (22). Ballard-Rosa, Martin, and Scheve (2016) make similar comparisons in their tax policy conjoint: "While there are few strong differences in preferences for taxing the lower three income groups (the 'hard work' group has slightly lower elasticities for taxing the poor), there are strong differences in preferences for taxing the rich" (12). In the Bechtel and Scheve (2013) conjoint on support for international climate change agreements in the United States, United Kingdom, Germany, and France, they summarize their results as "We find that individuals in all four countries largely agree on which dimensions are important and to what extent" (13765). In these examples, the differences between conditional AMCEs are used as a way of descriptively characterizing differences in *preferences* (i.e. levels of support) between the groups rather than differences in *causal effects on preferences* in the groups.

The selection of a reference category, while earlier an innocuous analytic decision, becomes substantially consequential for a descriptive reading of conditional AMCEs. Most obviously, using AMCEs descriptively prevents any description of the levels of favorability in the reference category. It can also lead to misinterpretations of patterns in preferences. AMCEs are relative, not absolute, statements about preferences. As such, there is simply no predictable connection between subgroup causal effects and the levels of underlying subgroup preferences. Yet analysts and their readers frequently interpret differences in conditional AMCEs as differences in underlying preferences. AMCEs do provide insight into the descriptive variation in preferences within-group and across-features, and conditional AMCEs do estimate the size of causal effects of features within groups. But AMCEs cannot provide direct insight into the pattern of preferences between groups because they do not provide information about *absolute* levels of favorability toward profiles with each feature (or combination

of features).

This additional information matters. Consider again the simple two-condition experiment in which the effect of a male as opposed to female candidate, $x \in 0, 1$, is compared across a single two-category covariate, $z \in 0, 1$ such as Democratic or Republican self-identification. Subgroup regression equations to estimate effects for each group are:

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon, \quad \forall z = 0$$
$$\hat{y} = \beta_2 + \beta_3 x + \epsilon, \quad \forall z = 1$$

The effect of x when $z = 0$ is given by β_1. The effect of x when $z = 1$ is given by β_3. These are, in essence, the conditional AMCEs in a conjoint analysis. Yet the difference in AMCEs ($\beta_3 - \beta_1$) is not equal to the difference in preferences between the two groups, which is $\bar{y}_{z=1|x=1} - \bar{y}_{z=0|x=1}$ (estimated by $(\beta_2 + \beta_3) - (\beta_0 + \beta_1)$). The difference-in-AMCEs only equals the difference in preferences when $\beta_2 \equiv \beta_0$. Yet the standard AMCE-centric conjoint analysis does not present absolute favorability in the reference category. Similarity of conditional AMCEs only means similarity of the *causal effect* of the feature across groups, not similarity of *preferences* unless preferences toward profiles with the reference category are equivalent in both groups. Given the reference category choice is typically arbitrary or driven by substantive knowledge of the levels, there is never any reason to expect that the reference category satisfies this equality requirement. When using a difference-in-AMCEs comparison to estimate a difference in preferences, the size and direction of the bias is determined by the size of the difference in preferences toward the reference category within each subgroup.

To draw this example out more fully, the upper panel of Figure 2 shows AMCEs for Teele, Kalla, and Rosenbluth's candidate choice experiment for the full sample of respondents. The second panel shows full sample marginal means. Respondents' preference for female candidates is very apparent in both forms of analysis in the upper

Figure 2: Replication of Results for 'Candidate Sex' Feature from Teele et al. (2018) Candidate Experiment using Full Sample AMCEs and MMs and Subgroup AMCEs and MMs for Democrats and Republicans

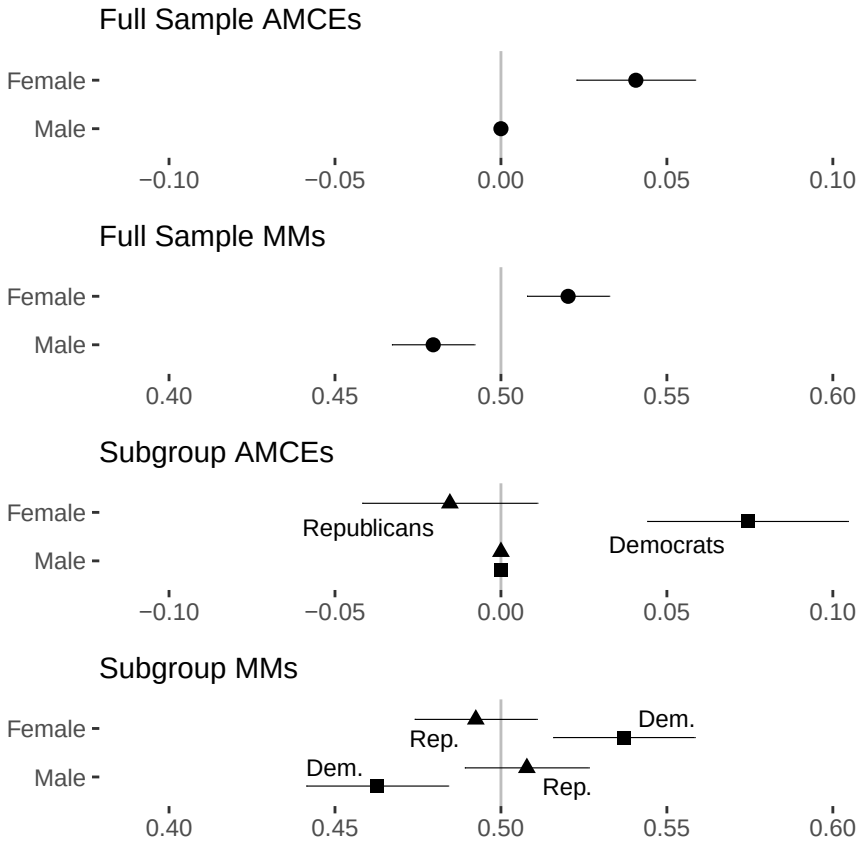
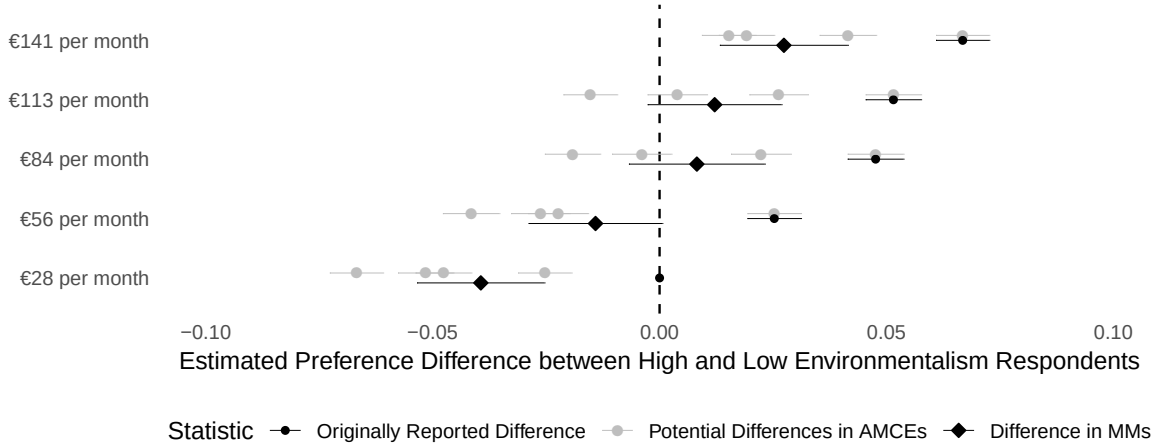


Figure 3: True Difference in Favorability and Implied Preference Differences between High and Low Environmentalism Respondents for 'Monthly Cost' Feature from Bechtel and Scheve (2013) Climate Agreement Experiment for Each Possible Reference Category



two panels because the AMCE definitionally equals the difference in marginal means. But how do Republicans and Democrats differ in their preferences over male and female candidates? The third panel shows conditional AMCEs separately for Democratic and Republican voters, as provided in the original paper and the lower panel shows the results using conditional marginal means for Democratic and Republican voters.[6] By requiring a reference category fixed to zero, the conditional AMCE results in the third panel suggest that there is a very large difference in favorability toward female candidates between Republican and Democratic respondents. In reality, however, the difference in these conditional AMCEs (0.089) reflects the true difference in favorability toward female candidates (difference: 0.045; Democrats: 0.537, Republicans: 0.492) *plus* the difference in favorability toward male candidates (difference: 0.045; Democrats: 0.463, Republicans: 0.508). Because Democrats and Republicans actually differ in their views of profiles containing the reference (male) category, AMCEs sum the true differences in preferences for a given feature level with the difference in preferences toward the reference category.[7]

[6] We opt here for visual presentation of results; tabular presentation of AMCEs, marginal means, and associated standard errors for all examples are included in the Appendix.

[7] Another example that clearly demonstrates the discrepancy between the differences in preferences and the differences in conditional AMCEs can be seen very clearly in the "political experience" feature of this experiment (see Appendix C).

Visual or numerical similarity of subgroup AMCEs is therefore an analytical artefact, not an accurate statement of the similarity of patterns of preferences. We can see this bias in a reanalysis of Bechtel and Scheve's four-country climate change agreement experiment. Figure 3 shows an analysis for the feature capturing the monthly household cost for a potential international climate agreement. This replicates a portion of their results which compare high- and low-environmentalism respondents pooled across countries (Bechtel and Scheve, 2013, 13767 figure 4). The original analysis has conditional AMCEs for the two subgroups with 28 Euro per month as the reference category. Conditional AMCEs for both groups are presented as negative with conditional AMCEs for low-environmentalism respondents being more negative than the conditional AMCEs for high-environmentalism respondents at every feature level. This implies positive differences in favorability toward each monthly cost between high- and low-environmentalism respondents. Figure 3 presents the implied difference-in-AMCEs from the original analysis as black circles, demonstrating the substantial and positive *apparent* differences between the two groups. For example, the difference-in-AMCEs for the 56 Euro per month level (incorrectly) implies that high-environmentalism respondents are *more* favorable toward a 56 Euro per month household cost of an agreement than are low-environmentalism respondents. Yet the opposite is actually true: high environmentalism respondents are less favorable toward this option than low environmentalism respondents. By using the 28 Euro per month level as the reference category, the original analysis implies that preferences are identical between the two groups when in reality high-environmentalism respondents are much less favorable toward a 28 Euro per month cost than low-environmentalism respondents. The black diamonds in Figure 3 show these true differences in favorability as marginal means for the two groups.

Furthermore, the gray dots in Figure 3 represent the alternative differences-in-AMCEs that *could have been generated* from alternative choices of reference category using the same data. Not only is it possible for reference categories choice to significantly color the apparent size of differences between subgroup, that choice can also

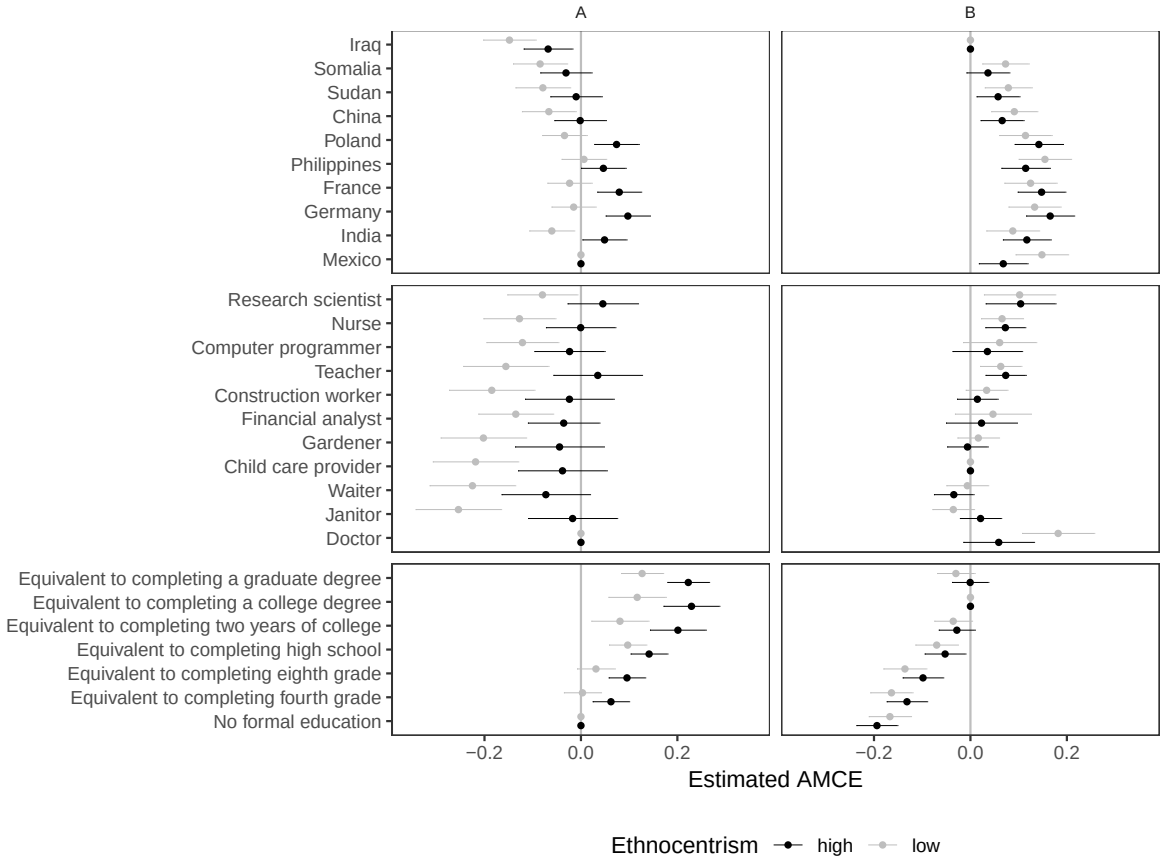
impact the direction and statistical significance of subgroup differences. An analyst could easily choose a reference category that presents differences between these two group as large and positive, small and positive, small and negative, large and negative, or negligible. The original analysis (again, black circles) happens to show large and positive differences between the groups.

It is worth highlighting two further features in Figure 3. First, the alternative differences-in-AMCEs estimates vary mechanically around the difference in marginal means, as the reference category varies. The difference between marginal means for two groups are always fixed in the data, so the differencing of subgroup AMCEs is merely an exercise is centering those differences at arbitrary points along the range of observed differences in marginal means. Second, and more practically, because there is no category for which the preferences of the two subgroups in this example are identical, no choice of reference category would have led to inferences from differences-in-AMCEs that accurately reflect the underlying difference in preferences. Even in the 84 Euro per month level, the difference between the two groups is slightly positive. Were there a category for which subgroup preferences were exactly equal, then we could choose that as the reference category and interpret differences-in-AMCEs as differences in preferences. But there is never any guarantee that such a reference category exists. Thus, there is no way to use conditional AMCEs or differences between those conditional AMCEs to convey the underlying similarity or differences in preferences across sample subgroups.

Improved Subgroup Analyses in Conjoint Designs

Researchers and consumers of conjoints interested in describing levels of respondent favorability toward profiles with varying features can avoid the inferential errors that accompany conditional AMCEs by focusing attention on (subgroup) marginal means, differences between subgroup marginal means to infer subgroup differences in preferences toward particular features, and omnibus nested model comparisons to infer sub-

Figure 4: Comparison of AMCEs for Low- and High-Ethnocentrism Respondents Using Two Alternative Reference Categories Choices for Three Features from Hainmueller et al.'s (2014) Immigration Experiment



group differences across many features. To demonstrate each of these three techniques we provide a complete example based upon Hainmueller, Hopkins, and Yamamoto's analysis of their immigration conjoint by respondent enthnocentrism, which finds that "the patterns of support are generally similar for respondents irrespective of their level of ethnocentrism" (Hainmueller, Hopkins, and Yamamoto, 2014, 22). First, we show how different reference categories could have led to distinctly different conditional AMCEs and, therefore, interpretations of subgroup preference similarity. Second, we show how differences in marginal means clearly convey the similarity of these two subgroups without any sensitivity to reference category. Finally, we show how tested model comparisons would have provided Hainmueller, Hopkins, and Yamamoto with a statistic test of the claimed similarity in levels of support between these two respondent subgroups.

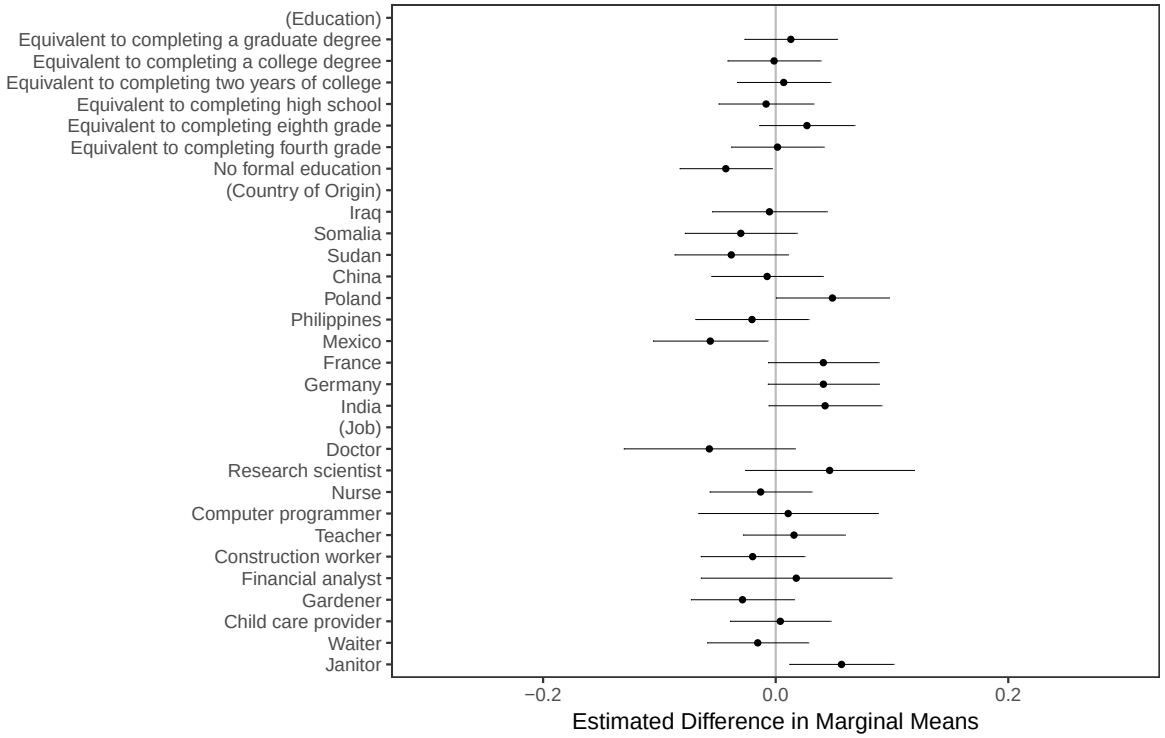
To begin, consider the left and right facets of Figure 4, which shows estimated subgroup AMCEs for three features from the immigration study. In panel "A" (left), all features are configured so that the reference category is the one with the largest difference in levels of support between the two subgroups thus distorting the size of differences at all other levels. In panel "B" (right), all features are configured so that the reference category is the one with the smallest difference in preferences between the two subgroups.

Panel A gives the impression that there are significant differences in preferences between high and low ethnocentrism respondents toward immigrants from different countries of origin, with different careers, and with different educational attainments because the reference category choice cascades the difference in reference category favorability into AMCEs for all other feature levels. By contrast, Panel B gives the impression that these differences are negligible. The experimental data and analytic approach in the two portrayals is identical; the only difference is the choice of reference category. Given what we have shown about the relationship between differences in conditional AMCEs and differences in conditional marginal means, Panel B is a more "truthful" visualization, which Cairo (2016) uses to mean avoidance of self-deception in the presentation of data, and a more "functional" visualization, by which Cairo means choosing graphics based on how they will be interpreted by the visualization's consumers. The differences between subgroup AMCEs there more accurately convey differences in underlying preferences because the reference categories used in Panel B are the most similar between the two groups.

Next, making a comparison of levels of favorability toward different types of immigrants without using AMCEs would have been even more truthful. Figure 5 *directly* shows that comparison of preferences as differences in subgroup marginal means between the two groups for these three features, with 95% confidence intervals for the difference.[8] The two groups indeed have similar preferences, something that would have happened to be clear had the conditional AMCEs in the right panel of Figure 4

[8] A presentation of subgroup marginal means for all features can be found in Appendix E.

Figure 5: Differences in Conditional Marginal Means, by Ethnocentrism, for Three Features From Hainmueller et al.'s (2014) Immigration Experiment



been presented but that would have been far less obvious were the conditional AMCEs in the left panel of that figure presented. Pairwise difference in means tests would provide formal procedures for testing the statistical significance of these differences.

Yet, finally, the similarity of subgroup preferences in conjoints is often characterized in an *omnibus* fashion, as in the quote from Hainmueller, Hopkins, and Yamamoto (2014) describing "patterns of support." An appropriate test in such cases is one that evaluates whether a model of support that accounts for group differences better fits the data than a model of support with only conjoint features as predictors. This type of test is known as a "nested model comparison" which compares the fit of a "restricted" regression (the restriction being that interactions between features and a subgroup identifier are held to be zero) nested within an "unrestricted" regression that allows for arbitrary interactions between conjoint features and the subgroup identifier. Formally, a nested model comparison provides an F-test of the null hypothesis that all interaction terms are equal to zero.[9]

[9]Like any ANOVA this hypothesis test may yield substantively different insight from a series of tests of pairwise mean differences. Figure 5 shows three instances where the 95% confidence intervals for

To make this concrete, for a feature with four levels (one treated as a reference category), the first (restricted) equation would be:

$$Y = \beta_0 + \beta_1 Level_2 + \beta_2 Level_3 + \beta_3 Level_4 + u \tag{1}$$

The second (unrestricted) equation would allow for interactions between feature levels and the subgroup identifier:

$$\begin{aligned} Y = &\beta_0 + \beta_1 Level_2 + \beta_2 Level_3 + \beta_3 Level_4 + \beta_4 Group + \\ &\beta_5 Level_2 * Group + \beta_6 Level_3 * Group + \beta_7 Level_4 * Group + u \end{aligned} \tag{2}$$

While Equation 1 imposes the constraint that $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$, Equation 2 allows for subgroup differences in favorability. Testing this null entails computing an F-statistic comparing the fit of each equation:

$$F = \frac{\frac{SSR_{Restricted} - SSR_{Unrestricted}}{r}}{\frac{SSR_{Unrestricted}}{n-k-1}} \tag{3}$$

where $SSR_{Restricted}$ is the sum of squared residuals for Equation 1, $SSR_{Unrestricted}$ is the sum of squared residuals for Equation 2, where r is the number of restrictions (in the above example, 4), n is the number of cases, and k is the number of feature levels in the unrestricted model.[10]

For the education feature, the resulting F-test for the model comparison in this case again gives us little reason to believe there are subgroup differences: F(7, 11493)=0.68, $p \leq 0.69$. We could repeat such pairwise comparisons or omnibus comparisons for each feature in the design — for country of origin (F(10, 11490)=1.56, $p \leq 0.11$) or job (F(11, 11489)=0.87, $p \leq 0.56$) — or for all features as a whole (F(98, 11402)=1.16, $p \leq 0.14$).

pairwise differences in marginal means do not include zero even though the omnibus test fails to reject the null at $\alpha = 0.05$.

[10]Note that this test is not sensitive to reference category even though it requires specifying a regression equation.

This visual display in Figure 5 and these statistical tests make clear what could not be directly inferred from conditional AMCEs alone: there are indeed no sizeable and only a few statistically apparent differences in preferences between the two groups.

This kind of nested model comparison test can also be used to assess heterogeneity across conjoint features (see also Egami and Imai, 2018). For example, Teele, Kalla, and Rosenbluth (2018) report just such a test for how effects of features other than candidate sex may differ between male and female candidates, finding no such heterogeneity (8–9). Fortunately, the original analysis accurately detected an absence of subgroup differences, yet a subtly different set of analytic decisions about reference categories (as shown in Figure 4) could have led to quite different inferences. As an example, Bechtel and Scheve (2013) argue that their conjoint results show "individuals in all four countries [Germany, France, United States, United Kingdom] largely agree on which dimensions are important and to what extent" (Bechtel and Scheve, 2013, 13765), but a nested model comparison shows the countries do differ in their preferences F(54, 67982)=3.72, $p \leq 0.00$. This cross-country variation is largely driven by differences in sensitivity to monthly household costs feature, F(15, 67995)=3.80, $p \leq 0.00$, with the United Kingdom and United States being more cost sensitive than Germany and France. Visual comparisons of conditional AMCEs can sometimes provide accurate insights into subgroup differences in preferences (as in the Hainmueller, Hopkins, and Yamamoto case), but ultimately there is no guarantee that they do in any particular analysis.

Conclusion

This article has identified several challenges related to the analysis and reporting of conjoint experimental designs, particularly analyses of subgroup differences. We suggest that conjoint analyses should report not only average marginal component effects (AMCEs) but also descriptive quantities about levels of favorability that better convey underlying preferences over profile features and better convey subgroup differences

in those preferences. Marginal means contain all of the information provided by AMCEs and more. Consequently, our intention here is not to substantively undermine any previous set of results, but instead to urge researchers moving forward to demonstrate considerable caution in how they design, analyze, and present the results of these types of descriptive experiments and how they test for differences in preferences between subgroups.

We have relatively straightforward and hopefully uncontroversial advice for how analysts of conjoint experiments should proceed:

1. Always report unadjusted marginal means when attempting to provide a *descriptive* summary of respondent preferences in addition to, or instead of, AMCEs.
2. Exercise caution when explicitly, or implicitly, interpreting differences-in-AMCEs across subgroups. Differences-in-AMCEs are differences in effect sizes for subgroups, not statements about the relative favorability of the subgroups toward profiles with a given feature. Heterogeneous effects do not necessarily mean different underlying preferences. If differences in AMCEs are reported, the choice of reference categories should be discussed explicitly and diagnostics should be provided to justify it.
3. When descriptively characterizing differences in preference level between subgroups, directly estimate the subgroup difference using conditional marginal means and differences between conditional marginal means, rather than relying on the difference-in-AMCEs.
4. To formally test for group differences in preferences, regression with interaction terms between the subgrouping covariate and all feature levels will generate estimates of level-specific differences in preferences via the coefficients on the interaction terms. A nested model comparison between this equation against one without such interactions provides an omnibus test of subgroup differences, which should be reported when characterizing overall patterns of subgroup differences.

Following this advice, we hope, will allow researchers to more clearly and more accu-

rately represent descriptive results of conjoint experiments.

The popularity of conjoint analyses in recent years highlights the power of the design and the important contributions made by Hainmueller, Hopkins, and Yamamoto (2014) in providing a novel causal interpretation of these fully randomized factorial designs. Yet with new tools always come new challenges. The now-common practice of descriptively interpreting conjoints requires more caution than is immediately obvious. To facilitate improved analysis and, especially, to provide easy-to-use tools for calculating marginal means and performing reference category selection diagnostics, we provide software called **cregg** (Leeper, 2018) available from the Comprehensive R Archive Network. Additionally, this manuscript is written as a reproducible knitr document (Xie, 2015) that contains complete code examples that will perform all analyses and visualization used throughout this article. With these resources in-hand, researchers should be well-equipped to analyze subgroup preferences in conjoint designs without running into the analytic challenges discussed here.

References

Ballard-Rosa, Cameron, Lucy Martin, and Kenneth Scheve. 2016. "The Structure of American Income Tax Policy Preferences." *The Journal of Politics* 79(1): 1–16.

Bansak, Kirk, Jens Hainmueller, and Dominik Hangartner. 2016. "How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers." *Science* 354(6309): 217–222.

Bechtel, Michael M., and Kenneth F. Scheve. 2013. "Mass Support for Global Climate Agreements Depends on Institutional Design." *Proceedings of the National Academy of Sciences* 110(34): 13763–13768.

Bechtel, Michael M., Federica Genovese, and Kenneth F. Scheve. 2017. "Interests, Norms and Support for the Provision of Global Public Goods: The Case of Climate Co-operation." *British Journal of Political Science*: Forthcoming.

Bechtel, Michael M., Jens Hainmueller, and Yotam Margalit. 2017. "Policy Design and Domestic Support for International Bailouts." *European Journal of Political Research* 56(4): 864–886.

Cairo, Alberto. 2016. *The Truthful Art*. New Riders.

Campbell, Rosie, Philip Cowley, Nick Vivyan, and Markus Wagner. 2016. "Legislator Dissent as a Valence Signal." *British Journal of Political Science*: Forthcoming.

Carey, John M., Kevin R. Carman, Katherine P. Clayton, Yusaku Horiuchim, Mala Htun, and Brittany Ortiz. 2018. "Who wants to hire a more diverse faculty? A conjoint analysis of faculty and student preferences for gender and racial/ethnic diversity." *Politics, Groups, and Identities*: Forthcoming.

Carlson, Elizabeth. 2015. "Ethnic Voting and Accountability in Africa: A Choice Experiment in Uganda." *World Politics* 67(2): 353–385.

Carnes, Nicholas, and Noam Lupu. 2016. "Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class." *American Political Science Review* 110(04): 832–844.

Clayton, Katherine, Jeremy Ferwerda, and Yusaku Horiuchi. 2018. "Exposure to Immigration and Admission Preferences: Evidence from France." : Forthcoming. Unpublished paper, Dartmouth University.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4): 627–635.

Egami, Naoki, and Kosuke Imai. 2018. "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis." *Journal of the American Statistical Association*: Forthcoming.

Eggers, Andrew C., Nick Vivyan, and Markus Wagner. 2018. "Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life?" *The Journal of Politics* 80(1): 321–326.

Franchino, Fabio, and Francesco Zucchini. 2014. "Voting in a Multi-dimensional Space: A Conjoint Analysis Employing Valence and Ideology Attributes of Candidates." *Political Science Research and Methods* 3(2): 221–241.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1): 1–20.

Gallego, Aina, and Paul Marx. 2017. "Multi-dimensional preferences for labour market reforms: a conjoint experiment." *Journal of European Public Policy* 24(7): 1027–1047.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4): 413–434.

Hainmueller, Jens, and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science*: Forthcoming.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multi-Dimensional Choices via Stated Preference Experiments." *Political Analysis* 22: 1–30.

Hankinson, Michael. 2018. "When Do Renters Behave Like Homeowners? High Rent, Price Anxiety, and NIMBYism." *American Political Science Review* 112(3): 473–493.

Hansen, Kasper M., Asmus L. Olsen, and Mickael Bech. 2014. "Cross-National Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic." *Political Behavior* 37(4): 767–789.

Kirkland, Patricia A., and Alexander Coppock. 2017. "Candidate Choice Without Party Labels." *Political Behavior* 40(3): 571–591.

Leeper, Thomas J. 2018. *cregg: Simple Conjoint Analyses and Visualization*. R package version 0.2.1.

Leeper, Thomas J., Sara B. Hobolt, and James Tilley. 2019. *Replication Data for 'Measuring Subgroup Preferences in Conjoint Experiments'*. doi:10.7910/DVN/ARHZU4.

Mummolo, Jonathan. 2016. "News from the Other Side: How Topic Relevance Limits the Prevalence of Partisan Selective Exposure." *The Journal of Politics* 78(3): 763–773.

Mummolo, Jonathan, and Clayton Nall. 2017. "Why Partisans Do Not Sort: The Constraints on Political Segregation." *The Journal of Politics* 79(1): 45–59.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

Oliveros, Virginia, and Christian Schuster. 2018. "Merit, Tenure, and Bureaucratic Behavior: Evidence From a Conjoint Experiment in the Dominican Republic." *Comparative Political Studies* 51(6): 759–792.

Ratkovic, Marc, and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 25(1): 1–40.

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations." *Political Research Quarterly* 70(2): 374–393.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3): 289–310.

Sniderman, Paul M. 2011. "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press.

Sobolewska, Maria, Silvia Galandini, and Laurence Lessard-Phillips. 2017. "The public view of immigrant integration: multidimensional and consensual: Evidence from survey experiments in the UK and the Netherlands." *Journal of Ethnic and Migration Studies* 43(1): 58–79.

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

Vivyan, Nick, and Markus Wagner. 2016. "House or home? Constituent preferences over legislator effort allocation." *European Journal of Political Research* 55(1): 81–99.

Wright, Matthew, Morris Levy, and Jack Citrin. 2016. "Public Attitudes Toward Immigration Policy Across the Legal/Illegal Divide: The Role of Categorical and Attribute-Based Decision-Making." *Political Behavior* 38(1): 229–253.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1498716963.