Measuring Subgroup Preferences in Conjoint Experiments*

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

April 13, 2019

Abstract

Conjoint analysis is now a common tool for studying political preferences. The method disentangles patterns in respondents' favorability toward complex, multidimensional objects, such as political candidates or public policies. Most conjoint analyses rely upon a fully randomized conjoint 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, people's support for the overall profile relative to a baseline, averaging across all respondents and all other profile 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 preferences or favorability. This often means comparing AMCEs among subgroups of respondents. We show that using conditional AMCEs to describe the degree of agreement among subgroups 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 the published literature and provide suggestions for improved reporting and interpretation using two quantities of interest: the marginal mean and the omnibus F-test. Given the accelerating use of conjoint analyses in political science, we offer advice for best practice in the analysis and presentation of conjoint experiments.

*We thank Benjamin Lauderdale, Jamie Druckman, Yusaku Horiuchi, the editor, and anonymous reviewers for feedback on this manuscript.

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.

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.

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

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 reference 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 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 — is demonstrated throughout. We conclude with advice for best practices in the analysis and presentation of conjoint results.

[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.

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.

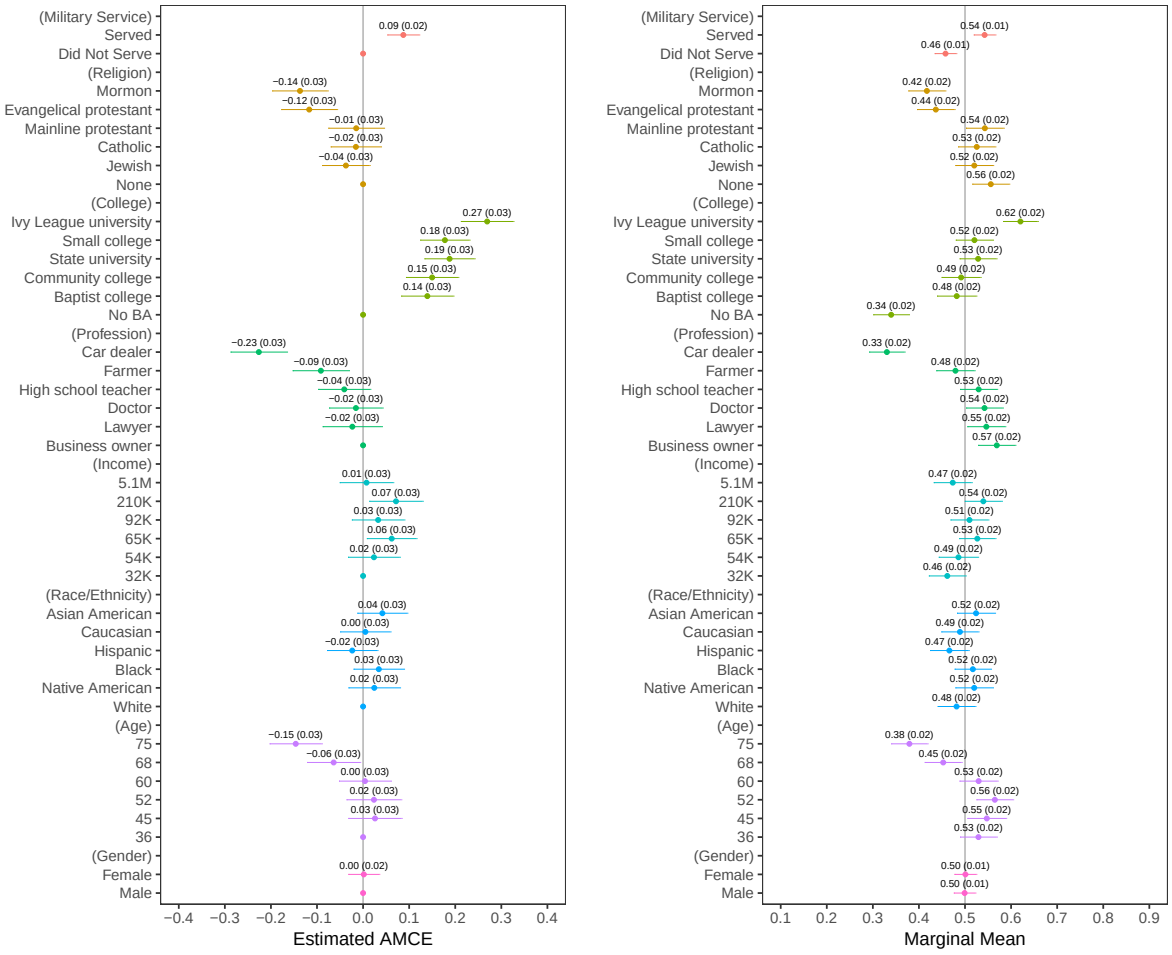
Analytically, a fully randomized conjoint design without constraints between profile features is simply a full-factorial experiment (with some cells possibly, albeit randomly, left unobserved). All quantities of interest relevant to the analysis of conjoint designs therefore derive from combinations 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

[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.

[4]It is not possible for the marginal mean to equal zero or one if pairs of profiles shown together are

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-

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 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 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.

[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.

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 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

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.

the poor), there are strong differences in preferences for taxing the rich" (12). In 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

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

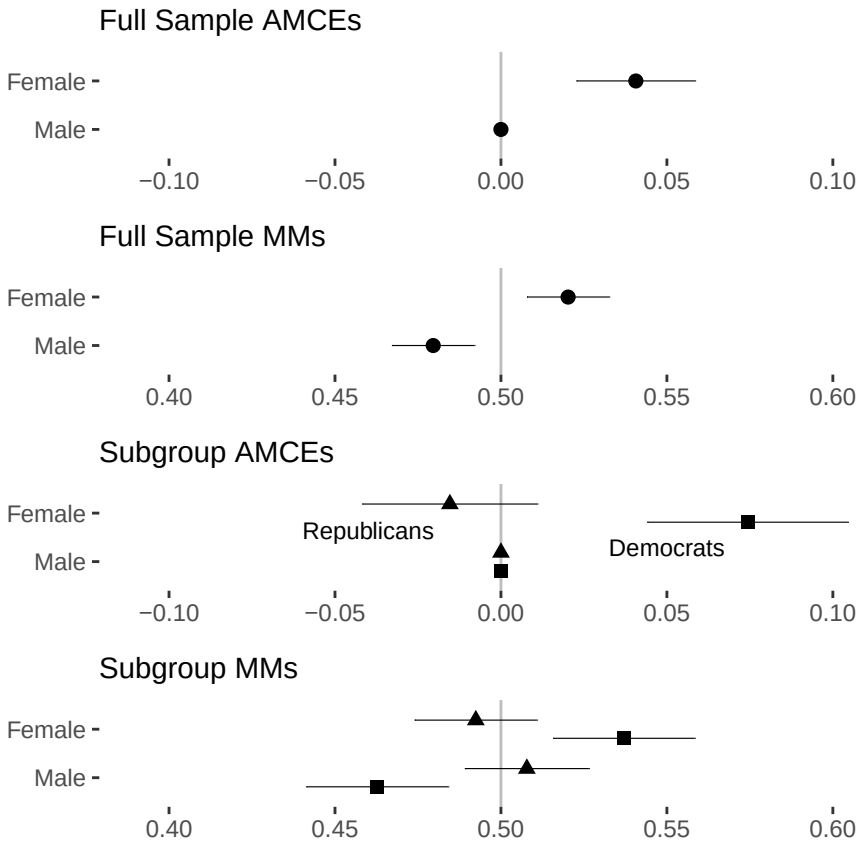
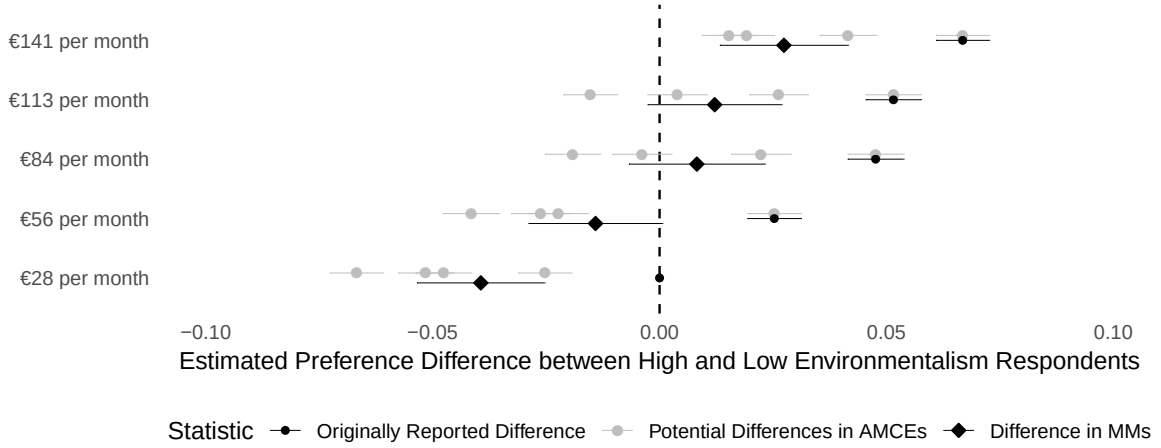


Figure 3: True Difference in Favorability and Implied Preference Differences between High and Low Environmentalism Respondents for 'Monthly Cost' Feature from Bechtel and Scheve (2013) Climate Agreement Experiment for Each Possible Reference Category



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 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]

We can also 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

[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).

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. Differences-in-AMCEs for a given feature level must therefore necessarily sometimes be positive and sometimes be negative, depending on the reference category used in estimating them. The direction of the difference per se conveys no information about underlying pattern of preferences in the two groups.

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. If multiple subgroup analyses are performed, it is unlikely the same reference category would work well across all analyses, making consistent interpretation difficult. And because conjoint analysis generates a sparse feature matrix (where many combinations of feature levels are unobserved in the data), it is rarely possible to empirically select an appropriate set of reference categories using the observed data. It is impossible to know which cell — of the tens of thousands in the design — is the best choice of reference. Indeed

it is possible that there is no such cell for which preferences are identical in the two groups; such a cell may exist, but there is no reason to expect that it should exist in any given application or that it would happen to be observed. 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 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 subgroup differences across many features. To demonstrate each of these three techniques

Here we provide a more complete example, demonstrating the full extent of this problem for interpretation of conjoint results and present alternative forms of analysis that more robustly convey subgroup preferences and the differences (if any) between them. Specifically, we show how reference category changes can lead to visual patterns of differences-in-AMCEs that provide strikingly different interpretations. We then show how to visualize and formally compare subgroup marginal means using simple means comparisons tests to avoid this problem. Finally, we demonstrate how to use nested model comparison tests to formally test for omnibus differences between groups.

Consider the left and right facets of Figure 4, which show the exact same analysis (comparing AMCEs for high and low ethnocentrism respondents) on the same experimental data from Hainmueller, Hopkins, and Yamamoto's immigration experiment. In panel "A" (left), all features are configured so that the reference category is the one with the largest difference in preferences between the two subgroups. 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.[8]

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. By contrast, Panel B gives the impression that these differences — indeed all differences — are negligible. The experimental data and analytic approach in the two portrayals is identical; the only difference is the choice of reference category for the profile features. Given what we have shown about the relationship between differences in conditional AMCEs and differences in conditional marginal means, Panel B is the more truthful visualization (Cairo, 2016). 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. Yet even this may not *perfectly* convey differences because no feature generates perfect agreement between the subgroups.

Alternatively presenting subgroup differences using conditional marginal means (as in Figure 5) provides the intended descriptive comparison of subgroup preferences. Each dot and error bar represents the conditional marginal mean (and its standard

[8]The appendix contains comparable plots for experiments by Bechtel and Scheve (2013) and Teele, Kalla, and Rosenbluth (2018).

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

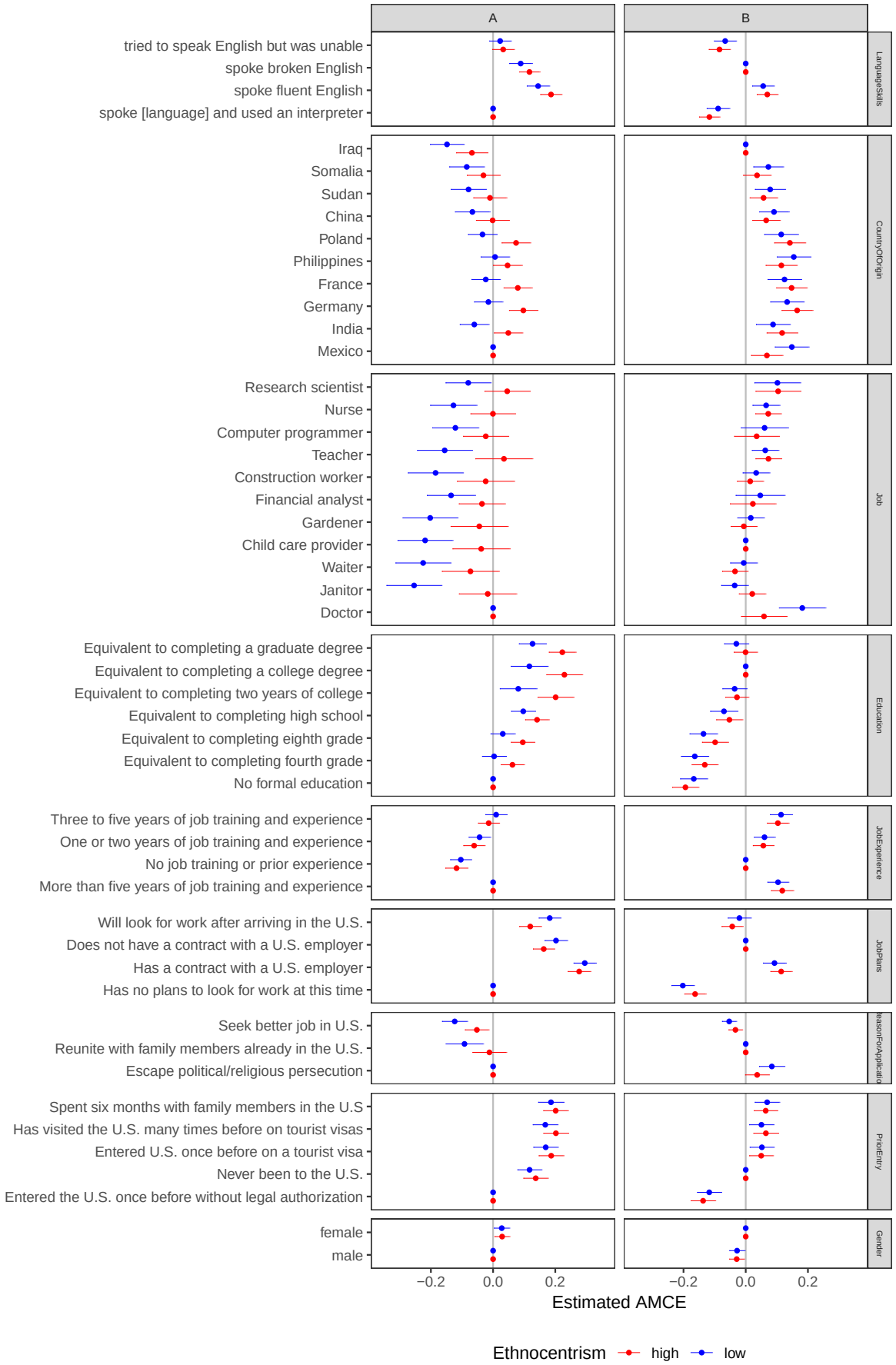
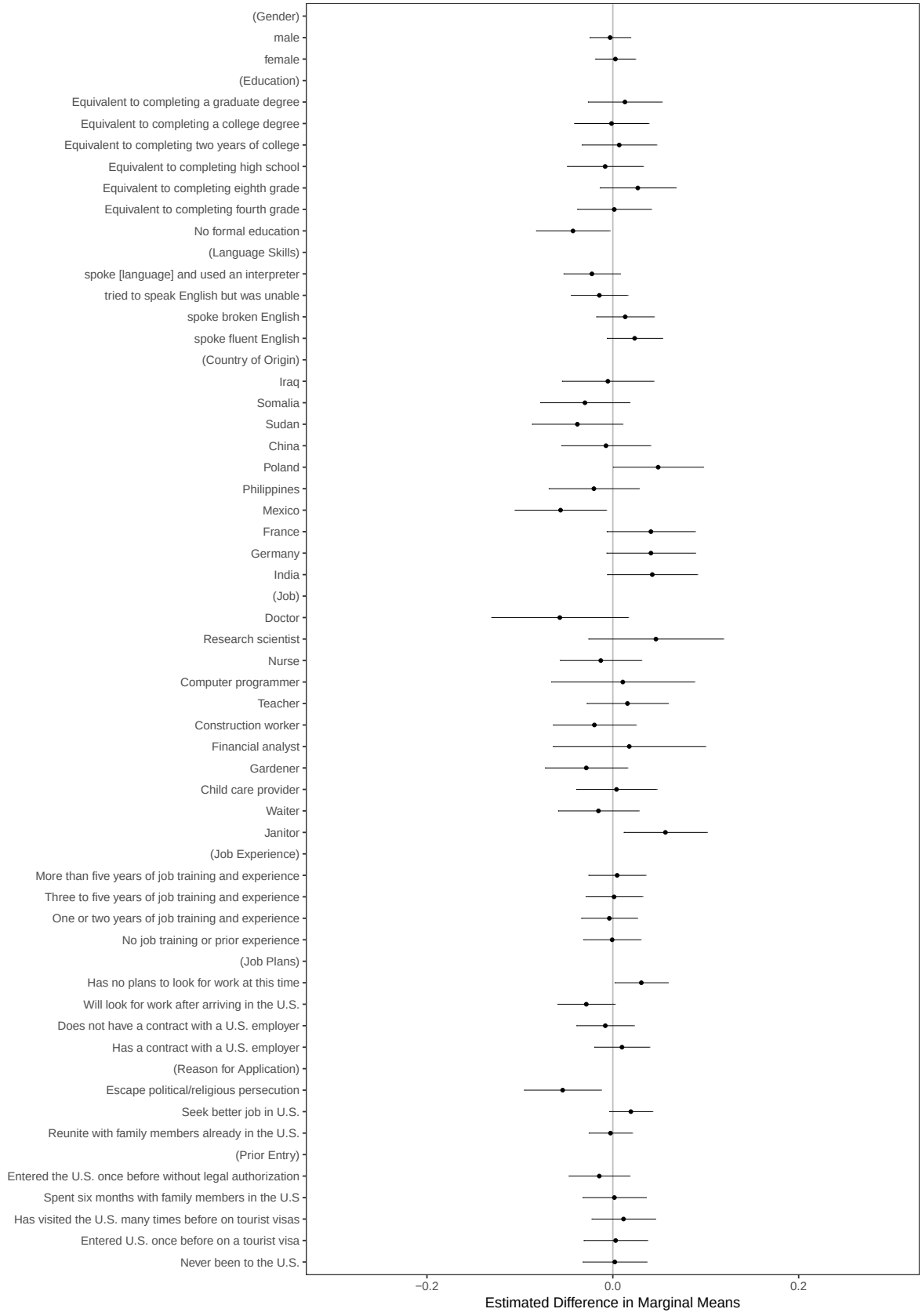


Figure 5: Differences in Conditional Marginal Means, by Ethnocentrism, for Hainmueller et al.'s (2014) Immigration Experiment



error) for high ethnocentrism (in red) and low ethnocentrism (in blue) respondents. This display of conditional marginal means highlights just how similar the preferences are for the two groups. For example, in the first set of estimates, both groups of respondents display minimally positive preferences toward female immigrants and minimally negative preferences toward male immigrants, averaging across those immigrants' other profile features. The second set of estimates shows both group are also more favorable toward higher-educated immigrants and less favorable toward less-education immigrants with no visually apparent differences. The third set of estimates, related to language skills, shows again similar patterns: both groups are more favorable toward immigrants with higher English proficiency than immigrants with lower English proficiency.

These estimates are less obviously identical for the two groups but look quite close. To test for pairwise differences between high and low ethnocentrism respondents, we can calculate differences in conditional conditional marginal means at each feature level, with associated significance tests:

- spoke fluent English: 0.02 (0.02, z_{diff}=1.30, $p \leq 0.20$)
- spoke broken English: 0.01 (0.02, z_{diff}=0.71, $p \leq 0.48$)
- tried to speak English but was unable: -0.01 (0.02, z_{diff}=-0.78, $p \leq 0.43$)
- spoke [language] and used an interpreter: -0.02 (0.02, z_{diff}=-1.22, $p \leq 0.22$)

These pairwise tests show that are our eyes have not deceived us. None of the level-specific differences in conditional marginal means are statistically distinguishable from zero. Were we interested in an omnibus tests of whether any of these differences were non-zero, we could perform a nested model comparison of two equations: (a) one estimating only marginal effects of the "Language Skills" feature levels, and (b) the same model with additional interactions between the subgrouping covariate and the feature levels. The test compares the sum of squared residuals for the two equations.

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

$$Y = \beta_0 + \beta_1 Level_1 + \beta_2 Level_2 + \beta_3 Level_3 + u \tag{1}$$

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

$$\begin{aligned} Y = &\beta_0 + \beta_1 Level_1 + \beta_2 Level_2 + \beta_3 Level_3 + \beta_4 Group + \\ &\beta_5 Level_1 * Group + \beta_6 Level_2 * Group + \beta_7 Level_3 * Group + u \end{aligned} \tag{2}$$

While Equation 1 imposes the constraint that $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$, Equation 2 allows for subgroup differences in favorability. The nested model comparison allows us to test the joint null hypothesis that there are no differences between the subgroups in any of the feature levels. This entails computing an F-statistic:

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

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. Note that this test is not sensitive to reference category used in the regression

For the Hainmueller, Hopkins, and Yamamoto language feature, the resulting F-test for the model comparison in this case again gives us little reason to believe there are subgroup differences: F(4, 11496)=1.06, $p \leq 0.37$. We could repeat such pairwise comparisons or omnibus comparisons for each feature in the design.

Furthermore, we could also directly visualize differences in conditional marginal means for this feature — and all features — as in Figure 5. This provides a more direct presentation of *differences* between subgroup preferences as the vertical line indicates feature levels for which there is no difference between the two groups. Positive values to the right of the line indicate positive differences (high ethnocentrism respondents are more favorable toward immigrants with this feature than low ethnocentrism respondents) and negative value to the left of zero convey the opposite. A further advantage of this plot is that unlike displays of conditional AMCEs, differences in conditional marginal means communicate subgroup differences for all feature levels including the reference categories. This display makes readily clear what was only indirectly apparent in Figure 5: there are indeed no sizeable and only a few statistically apparent differences in preferences between the two groups.

As before, we can perform an omnibus tests for the presence of any subgroup differences across all features, again using nested model comparison of two equations: (a) one estimating only effects of the features, and (b) the same model with additional interactions between the subgrouping covariate and all features. The result of that test for differences by ethnocentrism from the immigration experiment is: F(98, 11402)=1.16, $p \leq 0.14$, which further demonstrates that the substantive interpretation provided by Hainmueller, Hopkins, and Yamamoto (2014) accurately identified a lack of between-group differences. By contrast, Bechtel and Scheve (2013) argue in their cross-country conjoint examining climate change agreements "that individuals in all four countries largely agree on which dimensions are important and to what extent" (Bechtel and Scheve, 2013, 13765). Yet a nested model comparisons 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 as is plainly visible in a plot of marginal means by country (see SI).

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 an quite different conclusion.

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 that better convey underlying preferences over profile features and better convey subgroup differences in those preferences. 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. 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.[9]
2. Exercise caution when explicitly, or implicitly, interpreting differences-in-AMCEs across subgroups. Differences-in-AMCEs are differences in effect sizes for subgroups, not statments 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 preferences 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.[10] 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 accurately 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. This paper has demonstrated several such challenges and hopefully provides useful advice for how researchers should proceed with the analysis of such designs.

[9] Like the presentation of AMCEs, displaying marginal means in constrained conjoint designs may also distort apparent patterns given that not all features can co-occur. Partitioning the design into fractions such that each fraction contains a fully unconstrained design would mitigate any concern with that presentation.

[10] The analysis is slightly more complicated in constrained designs.

To facilitate such 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) that will perform these analyses and also provides the simple-to-use visualization tools used throughout this article. With that resource in-hand, researchers should be well-equipped to analyze 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).

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* , 1–23.

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* , 1–24.

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* , 1–19.

Carlson, Elizabeth. 2015. "Ethnic Voting and Accountability in Africa: A Choice Experiment in Uganda." *World Politics* 67(02): 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.". 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* , 1–34.

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(02): 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(04): 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* .

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. Unpublished paper.

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* .

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

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(01): 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.

Contents

A Definition of Quantities of Interest

A conjoint experiment serves two purposes: (1) description of the conditional distribution of favorability over variations in multiple features, and (2) leveraging the random observation of combinations of features (so-called "profiles") to infer that any differences in favorability over features are causally attributable to the features as opposed to something else. The quantities of interest are therefore functions of the features being randomized as in any factorial experiment. But additionally, conjoints typically involve within-subjects research designs (i.e., multiple, different profile observations per participant) thus necessitating some additional notation to account for the *survey implementation* of the conjoint in addition to the definition of the descriptive and causal parameters of interest.

Ultimately, a conjoint since Hainmueller, Hopkins, and Yamamoto (2014) is a complex survey-experimental design involving multiple observations across a high-dimension factorial experimental space. Specifically, I respondents ($i \in \{1, \ldots, I\}$) are presented with K rating or forced choice decision tasks, each involving J (typically 2) alternative profiles of, for example, candidates or policies. Each profile consists of a vector of F (typically discrete) features or attributes that describe the profile (e.g., age, sex), each composed of D_f alternative levels, a number which can vary across features. The experiment thus generates a dataset with $N = I \times J \times K$ observations of some rating scale or discrete choice outcome, Y, from a random sample of profiles drawn from the $C = \prod_{f=1}^{F} D_f$ population of experimental *cells* in the F-dimension feature space.

The survey implementation of the conjoint therefore generates N observations that can be indexed by i, j, k, forming an $N \times (L + 4)$ dimensional data matrix $\mathbf{M}$ with each row representing the vector of feature levels $\vec{F}$ in each profile j of respondent i's task k, with indicators for i, j, k, and the corresponding outcome $Y_{i,j,k}$.[11]

With no loss of information, we can think of each row in this matrix equivalently as an observation of $Y_{i,\vec{F}}$. This is because Hainmueller, Hopkins, and Yamamoto (2014) make several important assumptions that allow us to interpret these data in a different way than the survey implementation implies. First, they assume no carryover effects (Assumption 1), such that multiple observations from the same respondent can be treated as independent of one another. Second, they assume no profile order effects within-task (Assumption 2), such that profiles within a task can be treated as independent of each other. Assumptions 1 and 2 imply that the survey implementation indices for task, k, and profile-within-task, j, can be ignored. They have no bearing on any quantity of interest, by assumption.

The analyst is therefore left with a dataset of N observations, grouped into i participants, each providing into $Y_{\vec{F}}$. All quantities of interest must therefore be specified over as features of the distribution of Y over the F-dimensional feature space. In what follows, we therefore focus on the experimental features being randomized rather than the survey design factors being assumed away. Hainmueller, Hopkins, and Yamamoto (2014) make a third assumption that profiles are randomly constituted (Assumption 3), which in a fully randomized design, has the effect of meaning that features and feature combinations are randomly sampled for observation. If this randomization is uniform

[11] In typical paired designs (where $J = 2$), this means each task generates two data points: $Y_{i,1,k}$ and $Y_{i,2,k}$. Note, too, that in fully randomized designs, these two profiles can be identical. Furthermore in fully randomized, forced-choice designs this can yield the additional curiosity that $Y_{i,\mathbf{c}} \neq Y_{i,\mathbf{c}}$ for a given respondent, i, and profile, $\mathbf{c}$.

(which it almost always is in applied examples) this means we can additionally ignore the probability of observing any given combination (as all profiles are equally likely to be observed). This is a point we return to in a moment.

The most basic thing that can be estimated about the distribution of Y is the expected value, $E[Y]$, or *grand mean* (in the parlance of factorial experiments). We can think of this quantity in terms of the survey implementation process (namely, respondents, tasks, and profiles) or as a simple function of the resulting data:

$$\bar{Y} = \frac{1}{I \times J \times K} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{i,j,k} = \frac{1}{N} \sum_{n=1}^{N} Y_n \tag{4}$$

The nested summation over i, j, k could be stated explicitly but is unnecessary as the grand mean is simply the mean of all observed Y. A useful check on intuition is that in a forced choice design, where a respondent must choose only one profile, j, of all those presented in each task k, then by design $\bar{Y} = \frac{1}{J}$. For common, two-alternative, forced choice designs, $\bar{Y}$ therefore always equals 0.5. By contrast, in rating scale designs, $\bar{Y}$ can take any value between the lower and upper bounds of the rating scale.

In a *full factorial* experiment where $N > C$ (the number of observations is larger than the number of cells) due to a large sample, or few factors, or levels of each factor, or both (or both of these design characteristics), a sensible next quantity of interest is the *cell mean*: $E[Y|\vec{X} = \vec{x}]$, which in a conjoint simply measures the mean favorability toward a particular profile, $\vec{x}$. An effort to actually estimate this quantity will, however, become obviously intractable when one recognizes that the number of observations in a typical conjoint is much lower the number of feasible profiles ($N \ll C$). The cell mean can be unobserved for many or perhaps most experimental cells.

Therefore quantities of interest that derive from it — such as pairwise differences of means between cells — cannot be estimated for any of the arbitrary $\binom{C}{2}$ pairs of cells. As an example, in the Hainmueller, Hopkins, and Yamamoto (2014) candidate experiment, $C = 6^6 * 2^2 = 186,624$ and $N = 3466$, so less than 2% of experimental cells were observable and a minuscule fraction of the 17.4 billion pairwise cell combinations could have generated estimable effects.

It is at this point that the quantities of interest in a conjoint can become confusing. In a typical experiment where $N > C$, these pairwise differences of means are the standard estimator for a causal effect (the estimand being the causal effect on favorability of changing from one profile to another). For example, we might be interested in the effect on Y of changing the value of one feature to another theoretically interesting value of that feature, holding all other feature values in the profile constant:

$$\tau = E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_f = x_f] - E[Y|X_1 = \neg x_1, X_2 = x_2, \ldots, X_f = x_f] \tag{5}$$

but we have no guarantee that both or, in fact, either of those particular cells are observed. If even this minimal causal quantity cannot be guaranteed to be estimable by design, questions about higher-order interactions across features are even more difficult to estimate as they require observing four or more specific cells, any of which may be missing. Even if we were interested in such quantities, we would be unlikely to be able to estimate them.

Conjoint designs therefore ask us to think about completely different quantities of

interest from typical sentiment measurement or experimentation. Consequently, what quantities might we care about that can be estimated from an L-dimension factorial experimental with considerable sparsity other than the grand mean?

Even though $N \ll C$ in most applied conjoints, $N > F$. This means that even if we probably cannot learn about particular high-dimensional *combinations of features*, we can learn about favorability toward particular features alone. That is, we can learn about conditional expectations over each feature dimension, $E[Y|X_f = x_f]$. In the factorial experiments literature, this conditional mean is called the *marginal mean* (as it lies at the margins of a tabular presentation cell means for the complete design). For example, the following 2x3 factorial design contains 6 cell means ($2 * 3$), 1 grand mean, and five marginal means ($2 + 3$, one for each level of each factor):

	$A = 1$	$A = 2$	
$B = 1$	$\bar{Y}_{A=1,B=1}$	$\bar{Y}_{A=2,B=1}$	$E[Y\|B = 1]$
$B = 2$	$\bar{Y}_{A=1,B=2}$	$\bar{Y}_{A=2,B=2}$	$E[Y\|B = 2]$
$B = 3$	$\bar{Y}_{A=1,B=3}$	$\bar{Y}_{A=2,B=3}$	$E[Y\|B = 3]$
	$E[Y\|A = 1]$	$E[Y\|A = 2]$	$E[Y]$

The uniform sampling of cells in the design means that this is quantity can be estimated by the simple mean of $Y \forall X_f = x_f$.[12]

Were a constrained conjoint design used where some feature combinations were impossible, the marginal means would only be intelligible in the fractions of the design where all cells are observed.[13]

To clarify this point, consider the constrained 2x3 design below where one cell is unobserved by design:

	$A = 1$	$A = 2$	
$B = 1$	$\bar{Y}_{A=1,B=1}$	$\bar{Y}_{A=2,B=1}$	$E[Y\|B = 1]$
$B = 2$	$\bar{Y}_{A=1,B=2}$	$\bar{Y}_{A=2,B=2}$	$E[Y\|B = 2]$
$B = 3$	$\bar{Y}_{A=1,B=3}$	–	$E[Y\|B = 3]$
	$E[Y\|A = 1]$	$E[Y\|A = 2]$	$E[Y]$

Were the lower-right cell ($A = 2, B = 3$) observable by design, then a direct comparison of the marginal means, $E[Y|A = 1]$ and $E[Y|A = 2]$, in the lower table margin would provide direct insight into the relative favorability of respondents to profiles

[12] In unbalanced designs where the probability of being in a given cell is not uniform across cells, there is sometimes a distinction made between *descriptive* marginal means that equally weight observations and *design* marginal means that equally weight cells in the design. Given conjoint designs generally do not allow for the observation of cell means, the distinction is not relevant and we refer to *descriptive* marginal means simply as "marginal means."

[13] Practically, the random sampling of cells does not need to be uniform; over- and under-representation of cells is possible. We focus here on fully randomized designs that draw profiles from the full space with equal probability. A nuance in Hainmueller, Hopkins, and Yamamoto's notation is that their quantities of interest are conditioned on an arbitrary joint distribution of features rather than the particular joint distribution of features that was used to construct design or the joint distribution of features that happens to emerge empirically. In other words, they weight cells by an arbitrary joint probability mass function.

with features $A = 1$ and $A = 2$, marginalized over B. But because this cell is unobserved, these marginal means marginalize over different subsets of the possible values of B making them not obviously comparable. By contrast, the first and second marginal means at the top-right of the table — $E[Y|B = 1]$ and $E[Y|B = 2]$ — provide insight into the favorability of participants toward profiles with features $B = 1$ and with feature $B = 2$ marginalizing over the two possible values of A. A researcher could safely conclude that participants are more (less) favorable toward profiles with feature $B = 1$ than $B = 2$ from this information alone. But they would not be able to so for feature A without either (a) an explicit caveat that the comparison is of dissimilar subsets of profiles along dimension B or (b) calculating marginal means over only the completely observable[14] portion of the feature space due to the curse of dimensionality.

For the common *descriptive* use of conjoint designs to measure preferences over multi-dimensional objects, these marginal means alone are of direct interest. They express favorability on the scale of the outcome over alternative values of each feature independent of the features in the design.[15]

For the *causal* interpretation of conjoint designs, comparisons of these marginal means is required. Comparisons between them provide causal inferences about the effect of changing a focal feature, marginalizing across the distribution of other features. Because feature combinations (i.e., the profiles) are randomly constructed and randomly observed from all possible combinations, the distribution of other non-focal features is, in expectation, is independent of the focal feature, thus identical across all levels of the focal feature, and therefore ignorable.

A typical causal effect of interest is therefore the difference in marginal means across two levels of a feature (i.e., the marginal effect of a change in a feature's levels). For an unconstrained design, this difference is the *average marginal component effect* (AMCE) defined by Hainmueller, Hopkins, and Yamamoto (2014). In this way, an AMCE is simply a marginal effect of the factorial design: the difference of two marginal means.

Unfortunately, this is not a perfectly complete definition, but it covers the vast majority of applied cases. The exceptions are few. First, Hainmueller, Hopkins, and Yamamoto allow the joint distribution of features used in calculating the difference of marginal means to be arbitrary. This is meant to accommodate the weighting of marginal means to reflect the real-world distribution of feature combinations (e.g., down-weighting African American Republican political candidates given their rarity in real-world politics). Their definition of an AMCE allows for arbitrary weighting, but in practice this is uncommon.

Second, in constrained designs where some cells are unobservable, care needs to be taken in both defining and estimating AMCEs. Take, for example, the trivial example

[14]Note that what matters here is *observability*, not whether any given cell is actually observed. We know from above that most cells will be unobserved even in a uniformly sampled, unconstrained design.

[15]They do not necessarily convey favorability in an absolute sense. A high marginal mean for a given feature does not imply that the sample prefers that feature in an absolute sense. Instead, favorability has to be understood in light of the features presented to respondents. This is the innovation in conjoints; rather than asking respondents whether they will support a Mormon candidate (for example), we can infer their favorability toward a Mormon candidate in light of other candidate characteristics they may consider. Still our design may not contain all such features, so caution is needed in drawing typical public opinion inferences from these marginal means.

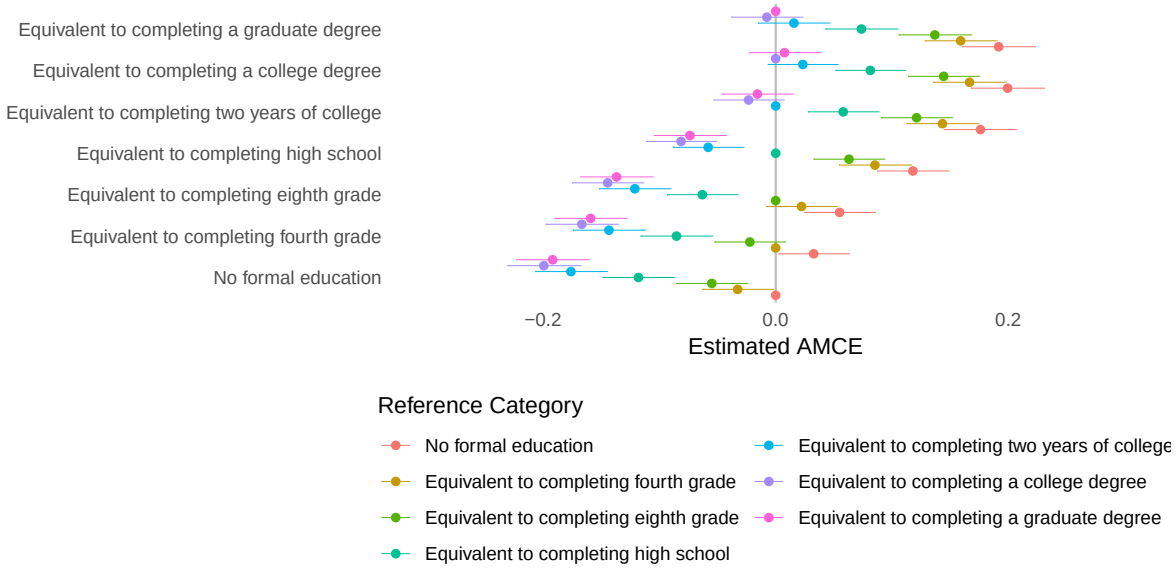
just above. The difference $E[Y|B = 2] - E[Y|B = 1]$ marginalizes over the full set of levels of A but $E[Y|B = 3] - E[Y|B = 1]$ marginalizes only over case where $A = 1$. Thus these two marginal effects reflect different subsets of the data and different combinations of values of A.

Hainmueller, Hopkins, and Yamamoto allow for these two differences to be presented as the AMCE despite the fact that the quantities marginalize over distinct subsets of the design. Indeed, their definition of AMCE for constrained designs diverges from the intuitive marginal effect to instead define the AMCEs for levels of B as an average of marginal effects of B over subsets of A and the AMCEs for levels of A as averages of the marginal effects of A over subsets of B (again, weighting these marginal effects arbitrarily). For example, if feature A is race $Caucasian, AfricanAmerican$ and feature B is religion $Evangelical, Catholic, Jewish$. In Hainmueller, Hopkins, and Yamamoto's notation, the AMCE of a candidate being Jewish relative to being Evangelical Christian is defined only for Caucasian candidates, while the AMCE of being Catholic is defined for both African American and Caucasian candidates. They present these subset marginal effects as the sample AMCEs even though they are not defined for the whole sample. There is nothing inherently problematic about that but, as noted earlier, it requires either being clear about what features are being marginalized over for each AMCE or an analysis of only the complete and comparable subset of the design (i.e., partitioning the design to form two complete, overlapping experimental designs). So, the researcher in this example may prefer to not present the AMCE of being Jewish together with the other results as it does not draw upon the complete set of feature combinations used in other portions of the analysis.

B Impact of Reference Category Choice on AMCEs

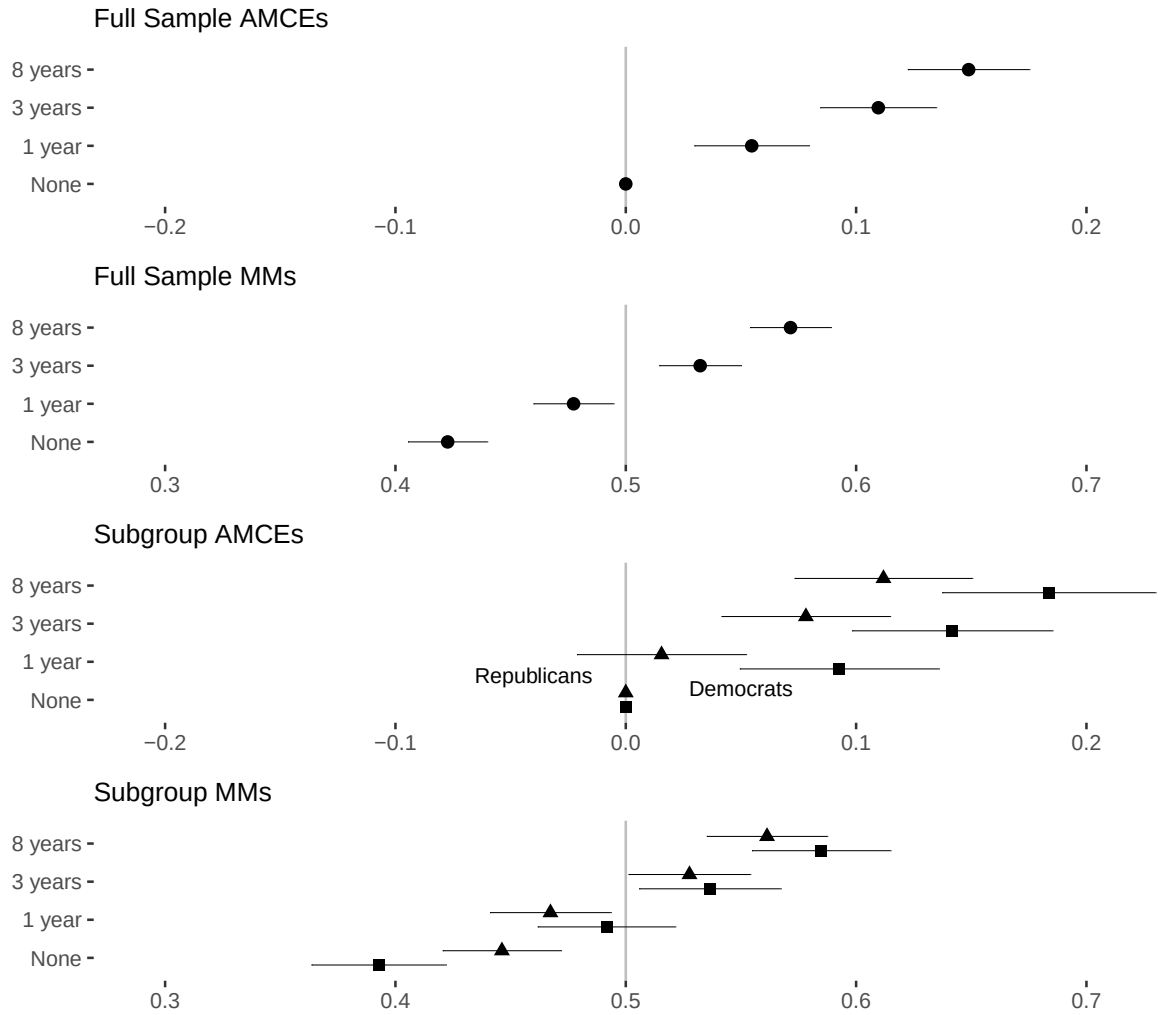
Though seemingly arbitrary, the choice of reference category for estimating AMCEs can be quite consequential. For example, in Hainmueller, Hopkins, and Yamamoto's candidate experiment (again, see 1), the least liked education level ("no formal education") is chosen as a reference category, but the authors could have presented the results using any of the categories as the baseline.

The figure below shows how the estimated AMCEs for each level of the education feature would have differed depending on that choice. Selecting a reference category that receives middling support (i.e., more favorability than some other feature levels but less favorability than others), makes some AMCEs positive and others negative but all AMCEs can be made positive (or negative) simply by choosing a different baseline. The results would be numerically equivalent — the alternative linear models used to the estimate the AMCEs have a mathematical equivalence — but the choice has sizeable consequences for the interpretation of conjoint analyses, as we discuss below.



In *constrained* conjoint designs, the choice of reference category is even more important. Consider, for example, the design of Hainmueller, Hopkins, and Yamamoto's immigration experiment, which constrains the "Country of Origin" feature so that levels 'India,' 'Germany,' 'France,' 'Mexico,' 'Philippines,' and 'Poland' cannot co-occur with the 'Escape Persecution' level of the "Reason for Application" feature. Consequently, the AMCE for the "Escape Persecution" level (relative to the "Reunite with family" reference category) is only defined for the subset of the design involving countries 'China,' 'Sudan,' 'Somalia,' and 'Iraq.' The AMCEs for those four countries (relative to India as a baseline) marginalize across all reasons for application, but the AMCEs for the first six countries marginalize only across the latter two reasons. Thus the interpretation of AMCEs — and the basic ability to estimate them in constrained designs — depends entirely upon the selection of a reference category where all feature levels can co-occur. In a design where *all* features are constrained, then AMCEs are undefined for the design as a whole and only estimable for subsets of the design that are *conditionally* unconstrained.

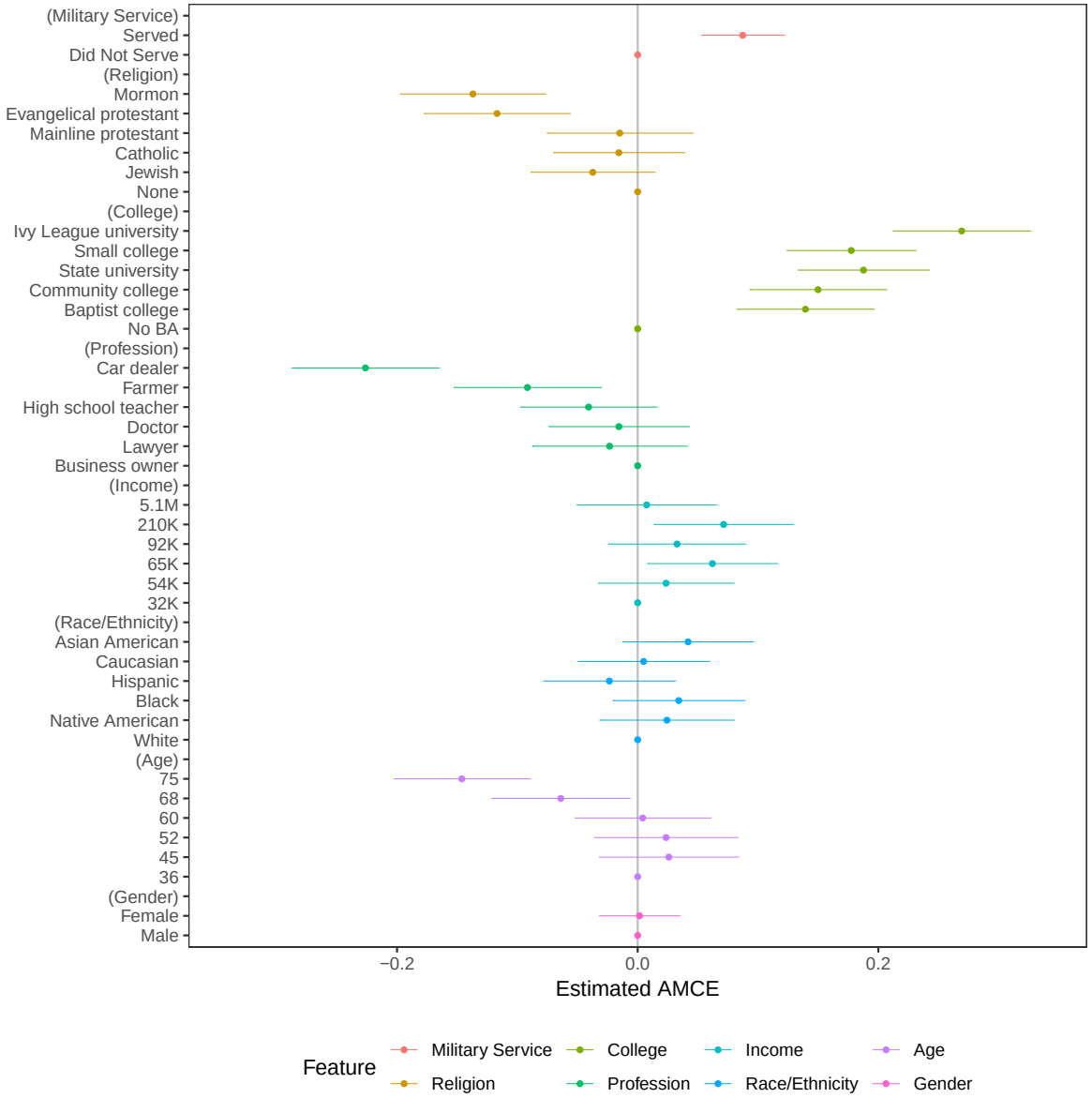
C Re-analysis of 'Political Experience' Feature from Teele et al. (2018)



Conditional AMCEs in this experiment (see 3rd panel, above) correctly convey that both Democrats and Republicans are more likely to favor experienced than inexperienced candidates. Reading the AMCEs descriptively, however, would suggest that Democratic voters are more favorable toward candidates with all levels of experience compared to Republican voters (i.e., Republicans and Democrats differ in their preferences over experienced candidates). Yet the conditional marginal means (4th panel, above) reveal that Democrats and Republicans have very similar preferences toward candidates with 1 or 3 years of experience, but differ dramatically in their preferences over candidates with no experience (the reference category) and those with 8 years experience. Democrats are much more sensitive to experience than are Republicans and important differences in preferences between the groups are apparent for very high and very low experience, but the conditional AMCEs suggest that preferences differ at all levels of experience, when in reality they do not.

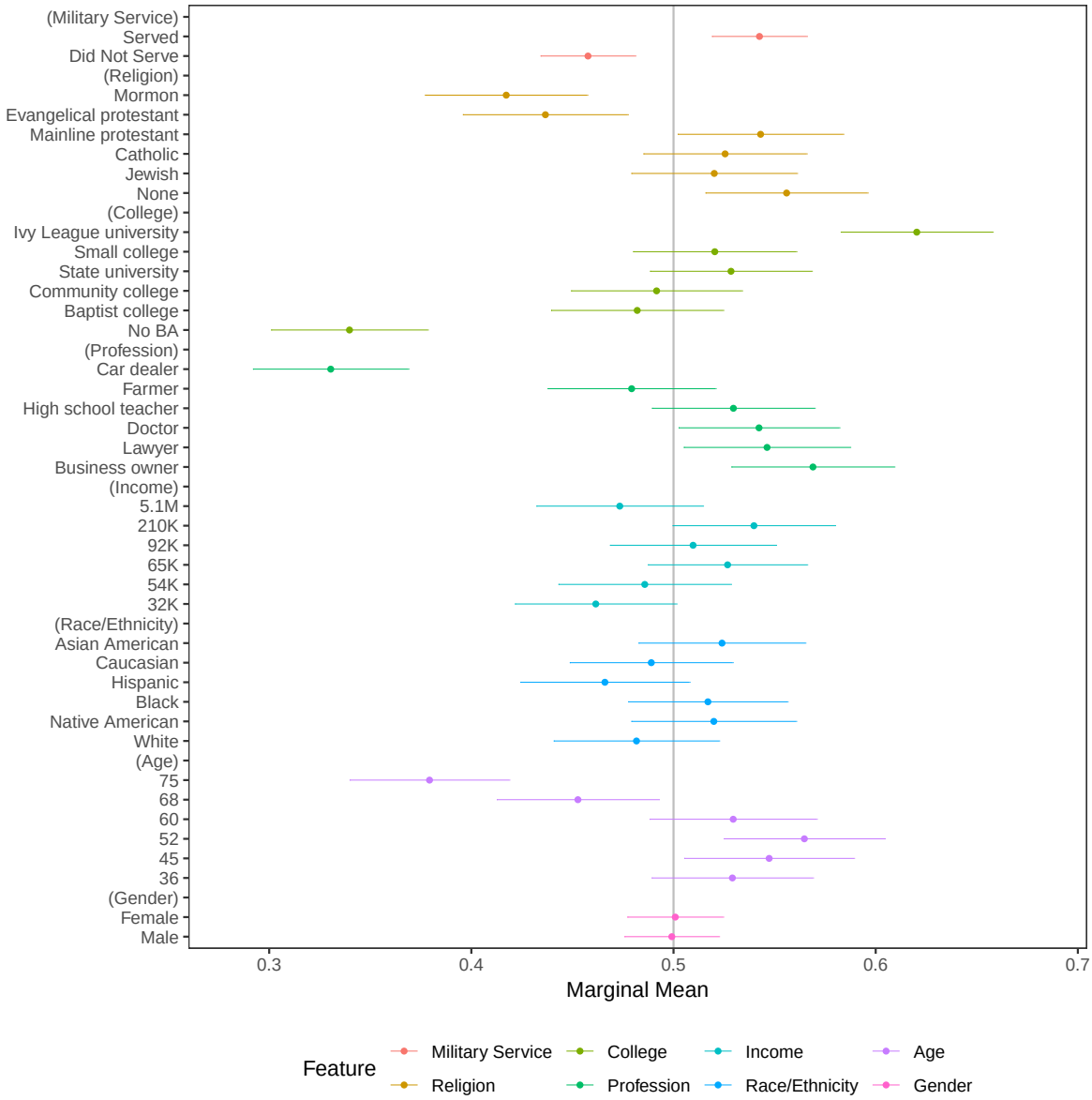
D Hainmueller et al. (2014) Candidate Experiment

D.1 Replication using AMCEs



feature	level	estimate	std.error	z
Military Service	Did Not Serve	0.00		
Military Service	Served	0.09	0.02	4.95
Religion	None	0.00		
Religion	Jewish	-0.04	0.03	-1.42
Religion	Catholic	-0.02	0.03	-0.56
Religion	Mainline protestant	-0.01	0.03	-0.48
Religion	Evangelical protestant	-0.12	0.03	-3.78
Religion	Mormon	-0.14	0.03	-4.46
College	No BA	0.00		
College	Baptist college	0.14	0.03	4.82
College	Community college	0.15	0.03	5.17
College	State university	0.19	0.03	6.77
College	Small college	0.18	0.03	6.50
College	Ivy League university	0.27	0.03	9.26
Profession	Business owner	0.00		
Profession	Lawyer	-0.02	0.03	-0.71
Profession	Doctor	-0.02	0.03	-0.53
Profession	High school teacher	-0.04	0.03	-1.42
Profession	Farmer	-0.09	0.03	-2.94
Profession	Car dealer	-0.23	0.03	-7.24
Income	32K	0.00		
Income	54K	0.02	0.03	0.82
Income	65K	0.06	0.03	2.26
Income	92K	0.03	0.03	1.12
Income	210K	0.07	0.03	2.41
Income	5.1M	0.01	0.03	0.25
Race/Ethnicity	White	0.00		
Race/Ethnicity	Native American	0.02	0.03	0.85
Race/Ethnicity	Black	0.03	0.03	1.22
Race/Ethnicity	Hispanic	-0.02	0.03	-0.84
Race/Ethnicity	Caucasian	0.00	0.03	0.18
Race/Ethnicity	Asian American	0.04	0.03	1.51
Age	36	0.00		
Age	45	0.03	0.03	0.88
Age	52	0.02	0.03	0.78
Age	60	0.00	0.03	0.14
Age	68	-0.06	0.03	-2.17
Age	75	-0.15	0.03	-5.06
Gender	Male	0.00		
Gender	Female	0.00	0.02	0.09

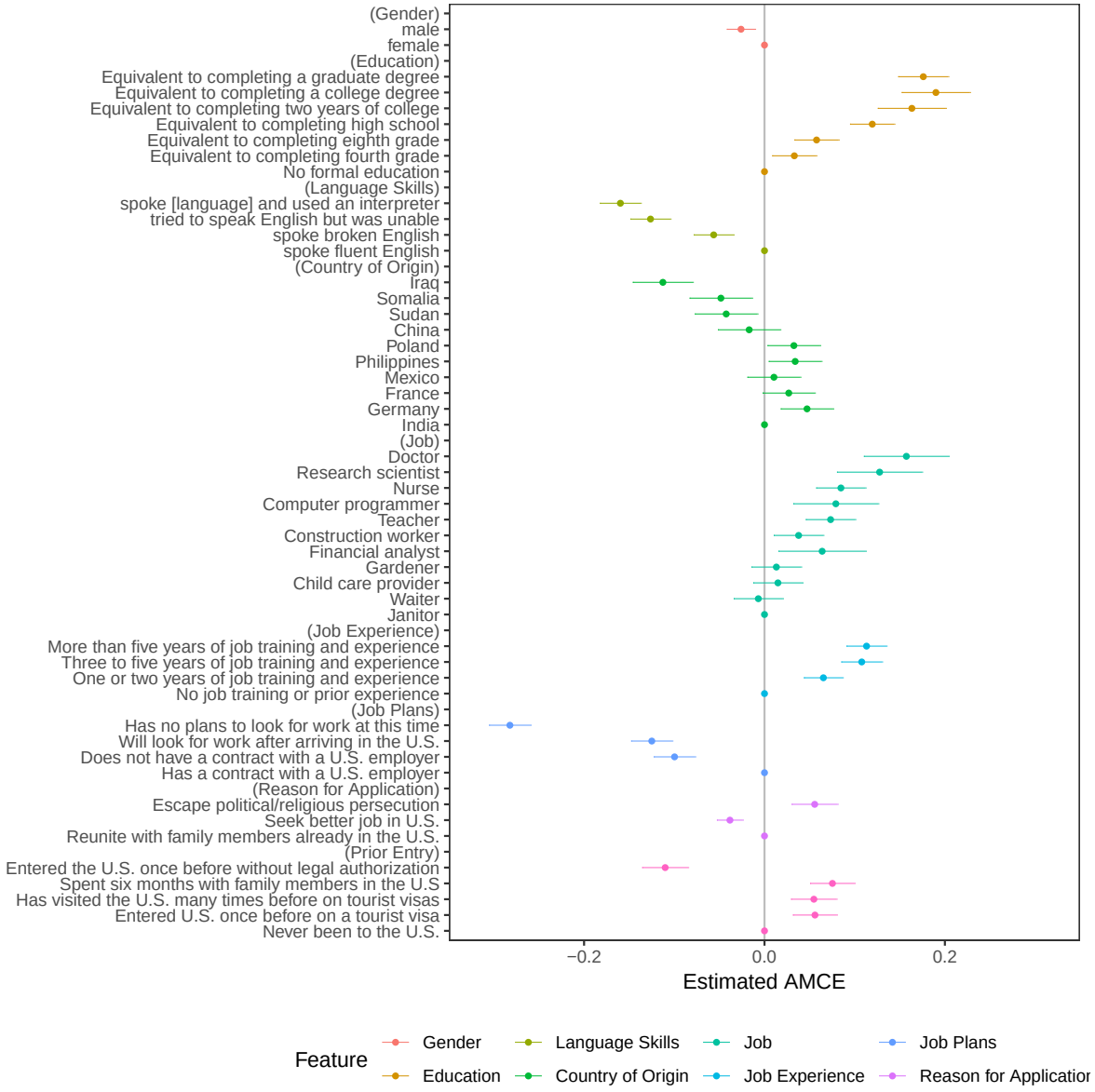
D.2 Replication using MMs

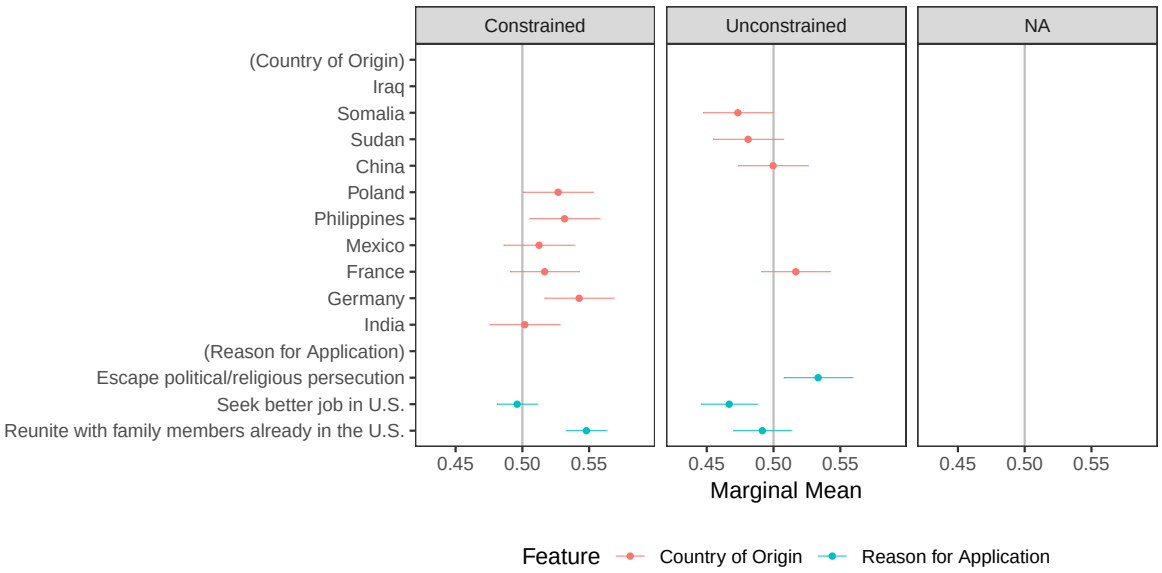


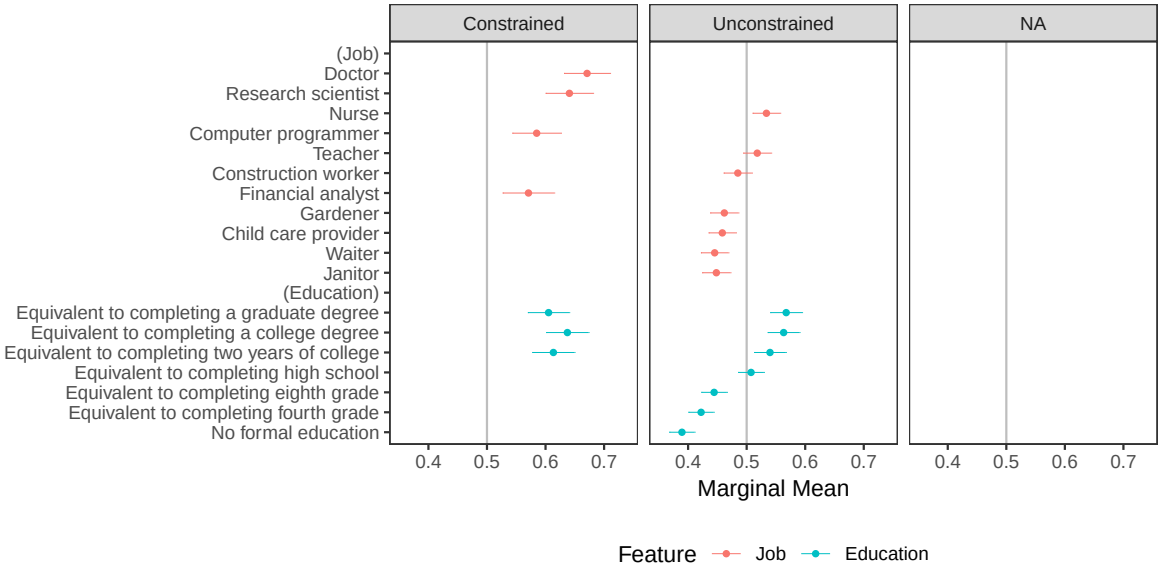
feature	level	estimate	std.error	z
Military Service	Did Not Serve	0.46	0.01	-3.54
Military Service	Served	0.54	0.01	3.55
Religion	None	0.56	0.02	2.73
Religion	Jewish	0.52	0.02	0.96
Religion	Catholic	0.53	0.02	1.24
Religion	Mainline protestant	0.54	0.02	2.06
Religion	Evangelical protestant	0.44	0.02	-3.05
Religion	Mormon	0.42	0.02	-4.04
College	No BA	0.34	0.02	-8.11
College	Baptist college	0.48	0.02	-0.83
College	Community college	0.49	0.02	-0.39
College	State university	0.53	0.02	1.39
College	Small college	0.52	0.02	0.99
College	Ivy League university	0.62	0.02	6.27
Profession	Business owner	0.57	0.02	3.35
Profession	Lawyer	0.55	0.02	2.20
Profession	Doctor	0.54	0.02	2.08
Profession	High school teacher	0.53	0.02	1.44
Profession	Farmer	0.48	0.02	-0.98
Profession	Car dealer	0.33	0.02	-8.64
Income	32K	0.46	0.02	-1.89
Income	54K	0.49	0.02	-0.65
Income	65K	0.53	0.02	1.33
Income	92K	0.51	0.02	0.46
Income	210K	0.54	0.02	1.94
Income	5.1M	0.47	0.02	-1.26
Race/Ethnicity	White	0.48	0.02	-0.88
Race/Ethnicity	Native American	0.52	0.02	0.96
Race/Ethnicity	Black	0.52	0.02	0.85
Race/Ethnicity	Hispanic	0.47	0.02	-1.59
Race/Ethnicity	Caucasian	0.49	0.02	-0.53
Race/Ethnicity	Asian American	0.52	0.02	1.14
Age	36	0.53	0.02	1.43
Age	45	0.55	0.02	2.21
Age	52	0.56	0.02	3.18
Age	60	0.53	0.02	1.40
Age	68	0.45	0.02	-2.31
Age	75	0.38	0.02	-5.99
Gender	Male	0.50	0.01	-0.07
Gender	Female	0.50	0.01	0.07

E Hainmueller et al. (2014) Immigration Experiment

E.1 Replication using AMCEs

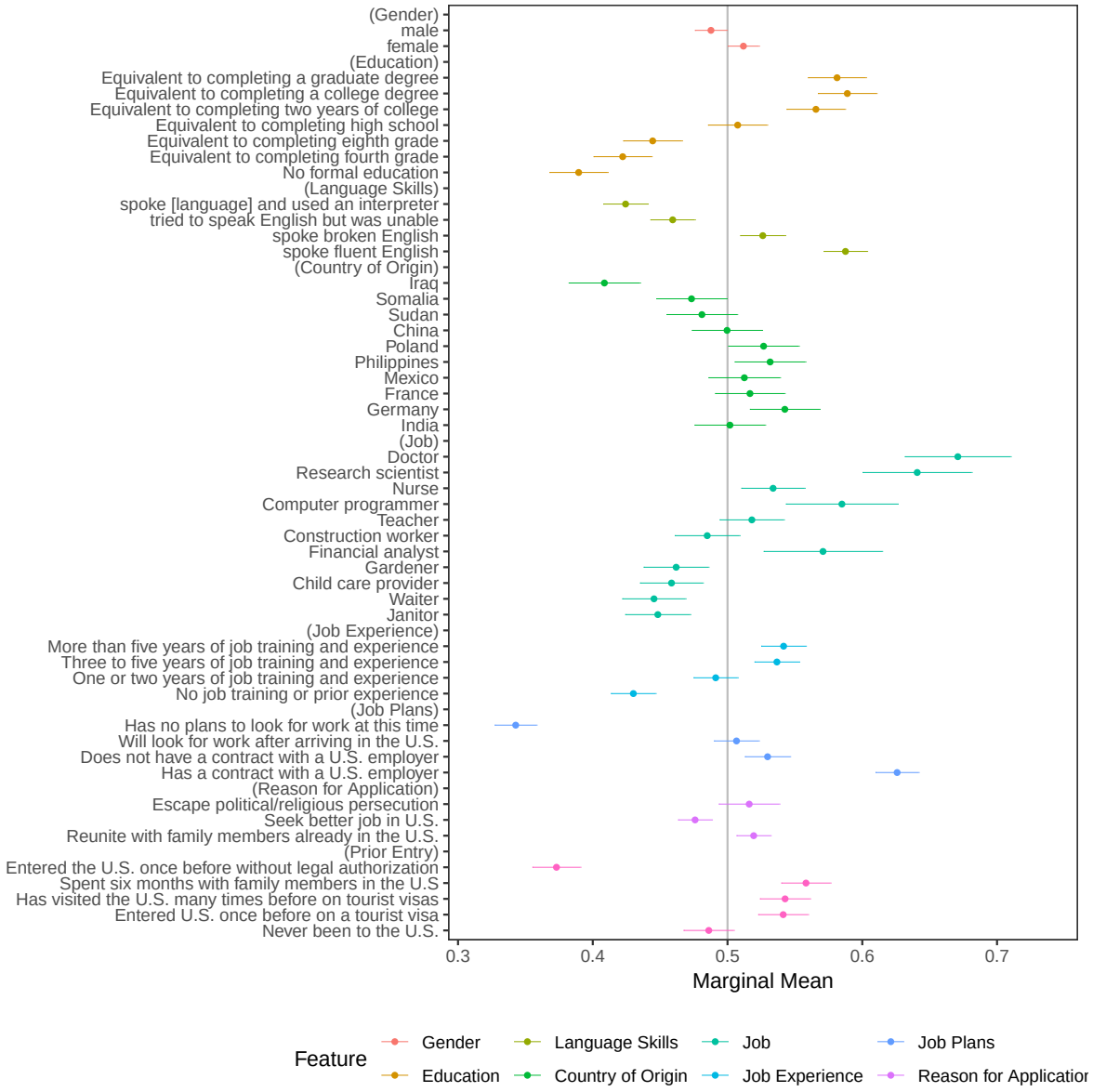



Constrained
Unconstrained
NA
(Country of Origin)
Iraq
Somalia
Sudan
China
Poland
Philippines
Mexico
France
Germany
India
(Reason for Application)
Escape political/religious persecution
Seek better job in U.S.
Reunite with family members already in the U.S.
0.45
0.50
0.55
Marginal Mean
Feature
Country of Origin
Reason for Application


Constrained
Unconstrained
NA
(Job)
Doctor
Research scientist
Nurse
Computer programmer
Teacher
Construction worker
Financial analyst
Gardener
Child care provider
Waiter
Janitor
(Education)
Equivalent to completing a graduate degree
Equivalent to completing a college degree
Equivalent to completing two years of college
Equivalent to completing high school
Equivalent to completing eighth grade
Equivalent to completing fourth grade
No formal education
0.4
0.5
0.6
0.7
Marginal Mean
Feature
Job
Education

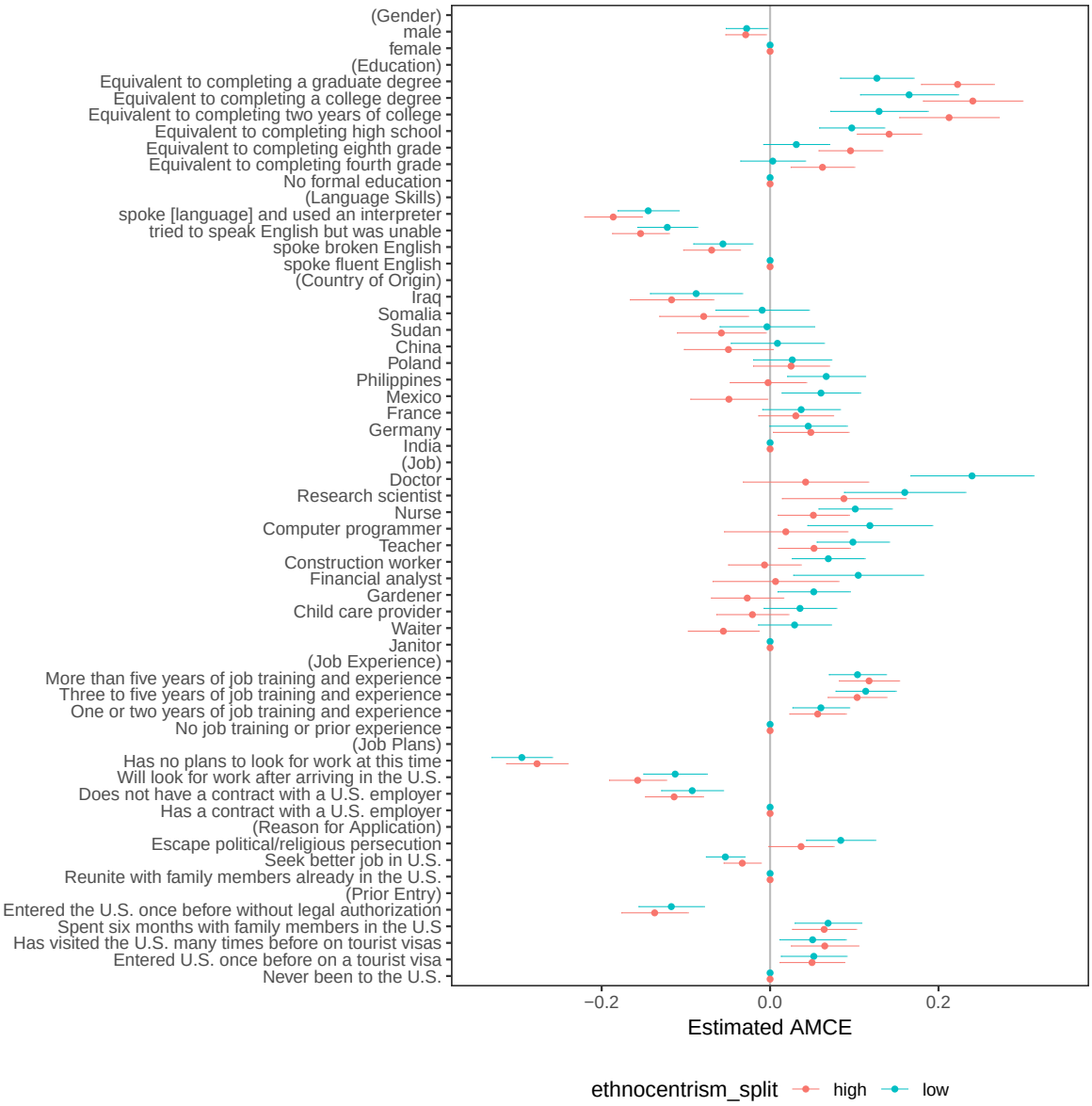
feature	level	estimate	std.error	z
Gender	female	0.00		
Gender	male	-0.03	0.01	-3.25
Education	No formal education	0.00		
Education	Equivalent to completing fourth grade	0.03	0.01	2.22
Education	Equivalent to completing eighth grade	0.06	0.01	3.86
Education	Equivalent to completing high school	0.12	0.01	7.98
Education	Equivalent to completing two years of college	0.16	0.02	7.12
Education	Equivalent to completing a college degree	0.19	0.02	8.26
Education	Equivalent to completing a graduate degree	0.18	0.02	10.41
Language Skills	spoke fluent English	0.00		
Language Skills	spoke broken English	-0.06	0.01	-4.98
Language Skills	tried to speak English but was unable	-0.13	0.01	-11.11
Language Skills	spoke [language] and used an interpreter	-0.16	0.01	-13.78
Country of Origin	India	0.00		
Country of Origin	Germany	0.05	0.02	2.66
Country of Origin	France	0.03	0.02	1.53
Country of Origin	Mexico	0.01	0.02	0.59
Country of Origin	Philippines	0.03	0.02	1.91
Country of Origin	Poland	0.03	0.02	1.83
Country of Origin	China	-0.02	0.02	-0.81
Country of Origin	Sudan	-0.04	0.02	-2.01
Country of Origin	Somalia	-0.05	0.02	-2.29
Country of Origin	Iraq	-0.11	0.02	-5.56
Job	Janitor	0.00		
Job	Waiter	-0.01	0.02	-0.41
Job	Child care provider	0.01	0.02	0.89
Job	Gardener	0.01	0.02	0.78
Job	Financial analyst	0.06	0.03	2.17
Job	Construction worker	0.04	0.02	2.26
Job	Teacher	0.07	0.02	4.39
Job	Computer programmer	0.08	0.03	2.76
Job	Nurse	0.08	0.02	5.08
Job	Research scientist	0.13	0.03	4.44
Job	Doctor	0.16	0.03	5.49
Job Experience	No job training or prior experience	0.00		
Job Experience	One or two years of job training and experience	0.07	0.01	5.92
Job Experience	Three to five years of job training and experience	0.11	0.01	9.32
Job Experience	More than five years of job training and experience	0.11	0.01	9.96
Job Plans	Has a contract with a U.S. employer	0.00		
Job Plans	Does not have a contract with a U.S. employer	-0.10	0.01	-8.50
Job Plans	Will look for work after arriving in the U.S.	-0.12	0.01	-10.69
Job Plans	Has no plans to look for work at this time	-0.28	0.01	-23.91
Reason for Application	Reunite with family members already in the U.S.	0.00		
Reason for Application	Seek better job in U.S.	-0.04	0.01	-4.37
Reason for Application	Escape political/religious persecution	0.06	0.02	3.58
Prior Entry	Never been to the U.S.	0.00		
Prior Entry	Entered U.S. once before on a tourist visa	0.06	0.01	4.49
Prior Entry	Has visited the U.S. many times before on tourist visas	0.05	0.01	4.24
Prior Entry	Spent six months with family members in the U.S	0.08	0.01	5.98
Prior Entry	Entered the U.S. once before without legal authorization	-0.11	0.01	-8.45

E.2 Replication using MMs

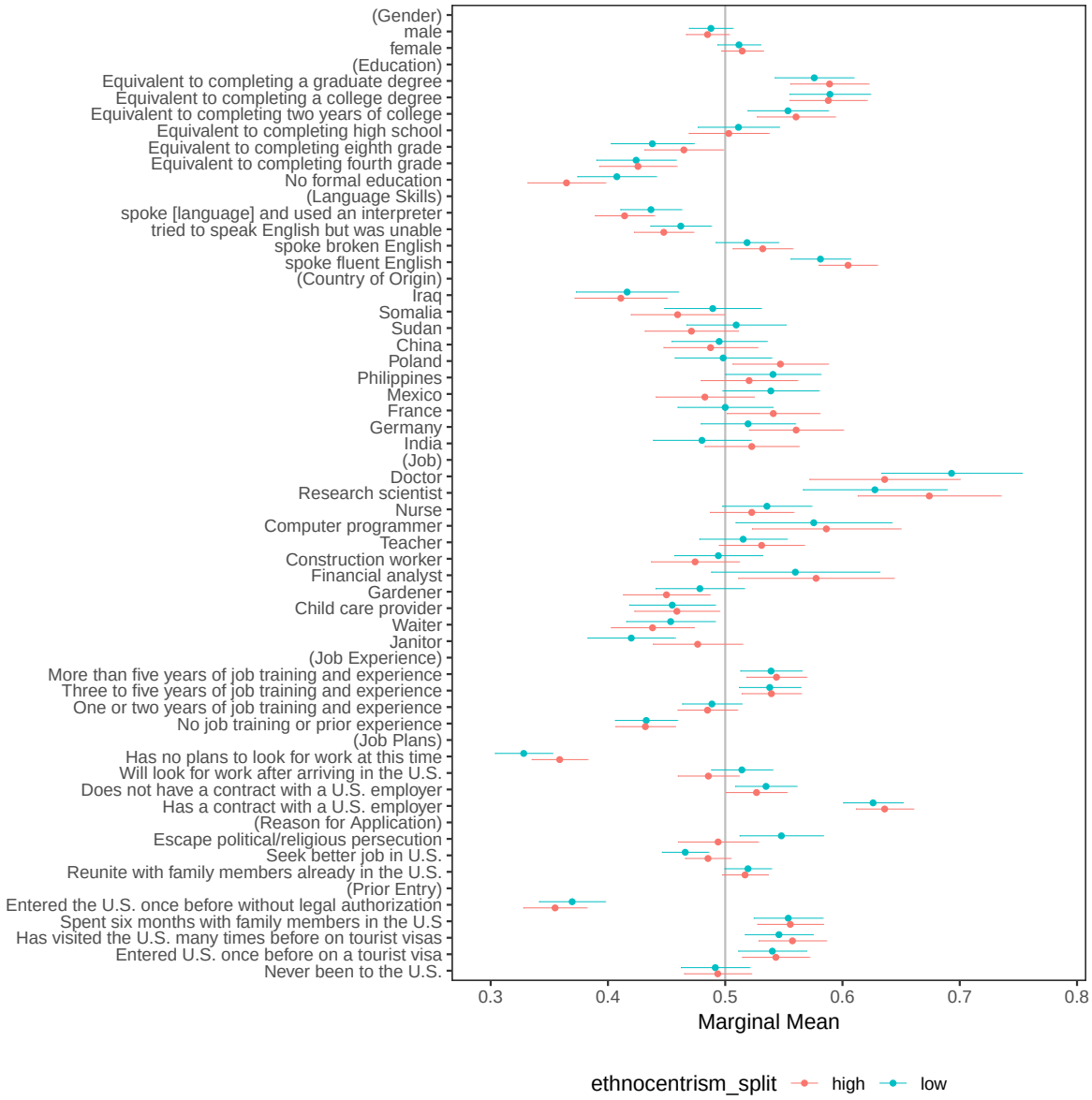


feature	level	estimate	std.error	z
Gender	female	0.51	0.01	1.99
Gender	male	0.49	0.01	-2.03
Education	No formal education	0.39	0.01	-10.04
Education	Equivalent to completing fourth grade	0.42	0.01	-7.08
Education	Equivalent to completing eighth grade	0.44	0.01	-5.00
Education	Equivalent to completing high school	0.51	0.01	0.67
Education	Equivalent to completing two years of college	0.57	0.01	5.92
Education	Equivalent to completing a college degree	0.59	0.01	8.00
Education	Equivalent to completing a graduate degree	0.58	0.01	7.40
Language Skills	spoke fluent English	0.59	0.01	10.63
Language Skills	spoke broken English	0.53	0.01	3.07
Language Skills	tried to speak English but was unable	0.46	0.01	-4.83
Language Skills	spoke [language] and used an interpreter	0.42	0.01	-8.98
Country of Origin	India	0.50	0.01	0.13
Country of Origin	Germany	0.54	0.01	3.22
Country of Origin	France	0.52	0.01	1.26
Country of Origin	Mexico	0.51	0.01	0.92
Country of Origin	Philippines	0.53	0.01	2.36
Country of Origin	Poland	0.53	0.01	2.01
Country of Origin	China	0.50	0.01	-0.03
Country of Origin	Sudan	0.48	0.01	-1.42
Country of Origin	Somalia	0.47	0.01	-2.01
Country of Origin	Iraq	0.41	0.01	-6.76
Job	Janitor	0.45	0.01	-4.20
Job	Waiter	0.45	0.01	-4.56
Job	Child care provider	0.46	0.01	-3.50
Job	Gardener	0.46	0.01	-3.11
Job	Financial analyst	0.57	0.02	3.16
Job	Construction worker	0.48	0.01	-1.23
Job	Teacher	0.52	0.01	1.49
Job	Computer programmer	0.58	0.02	4.01
Job	Nurse	0.53	0.01	2.82
Job	Research scientist	0.64	0.02	6.82
Job	Doctor	0.67	0.02	8.53
Job Experience	No job training or prior experience	0.43	0.01	-8.27
Job Experience	One or two years of job training and experience	0.49	0.01	-1.05
Job Experience	Three to five years of job training and experience	0.54	0.01	4.33
Job Experience	More than five years of job training and experience	0.54	0.01	4.92
Job Plans	Has a contract with a U.S. employer	0.63	0.01	15.40
Job Plans	Does not have a contract with a U.S. employer	0.53	0.01	3.47
Job Plans	Will look for work after arriving in the U.S.	0.51	0.01	0.78
Job Plans	Has no plans to look for work at this time	0.34	0.01	-19.86
Reason for Application	Reunite with family members already in the U.S.	0.52	0.01	3.00
Reason for Application	Seek better job in U.S.	0.48	0.01	-3.76
Reason for Application	Escape political/religious persecution	0.52	0.01	1.40
Prior Entry	Never been to the U.S.	0.49	0.01	-1.47
Prior Entry	Entered U.S. once before on a tourist visa	0.54	0.01	4.37
Prior Entry	Has visited the U.S. many times before on tourist visas	0.54	0.01	4.50
Prior Entry	Spent six months with family members in the U.S	0.56	0.01	6.24
Prior Entry	Entered the U.S. once before without legal authorization	0.37	0.01	-13.96

E.3 Subgroup Analysis using AMCEs



E.4 Subgroup Analysis using MMs

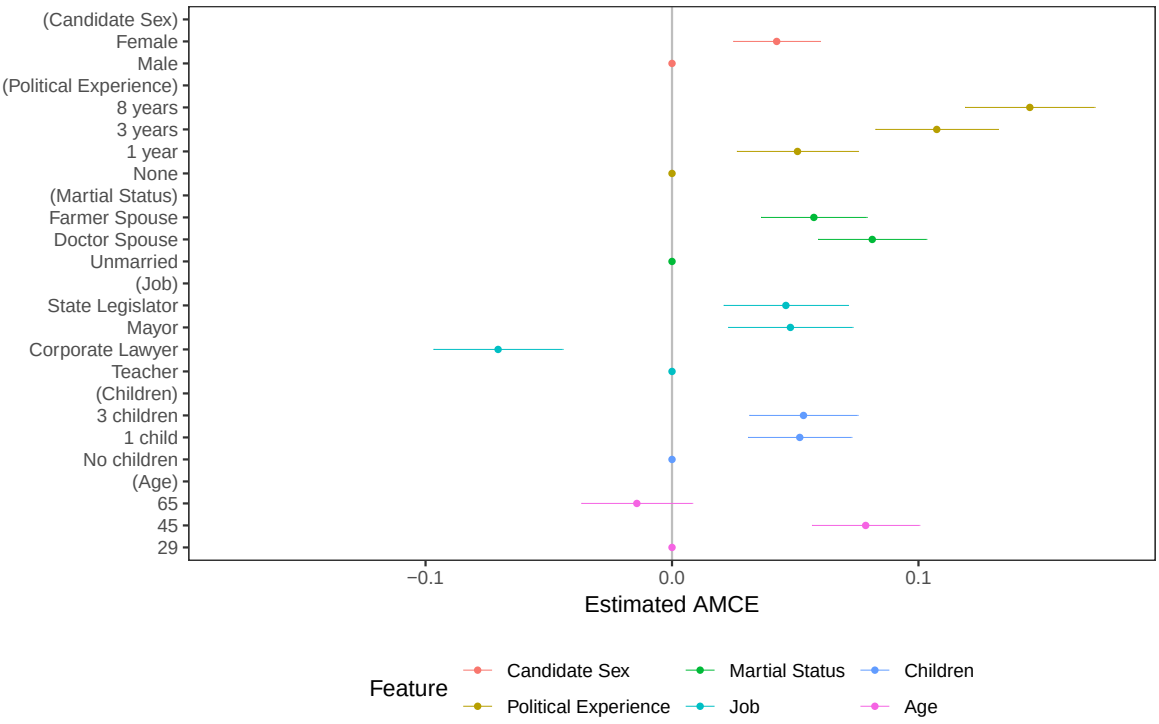


E.5 Nested Model Comparison

```
## Analysis of Deviance Table
##
## Model 1: ChosenImmigrant ~ Gender + Education + LanguageSkills + CountryOfOrigin +
##     Job + JobExperience + JobPlans + ReasonForApplication + PriorEntry +
##     Education:Job + CountryOfOrigin:ReasonForApplication
## Model 2: ChosenImmigrant ~ Gender + Education + LanguageSkills + CountryOfOrigin +
##     Job + JobExperience + JobPlans + ReasonForApplication + PriorEntry +
##     ethnocentrism_split + Education:Job + CountryOfOrigin:ReasonForApplication +
##     Gender:ethnocentrism_split + Education:ethnocentrism_split +
##     LanguageSkills:ethnocentrism_split + CountryOfOrigin:ethnocentrism_split +
##     Job:ethnocentrism_split + JobExperience:ethnocentrism_split +
##     JobPlans:ethnocentrism_split + ReasonForApplication:ethnocentrism_split +
##     PriorEntry:ethnocentrism_split + Education:Job:ethnocentrism_split +
##     CountryOfOrigin:ReasonForApplication:ethnocentrism_split
##   Resid. Df Resid. Dev Df Deviance     F Pr(>F)
## 1     11402     2500.7
## 2     11304     2475.8 98   24.834 1.157 0.1384
```

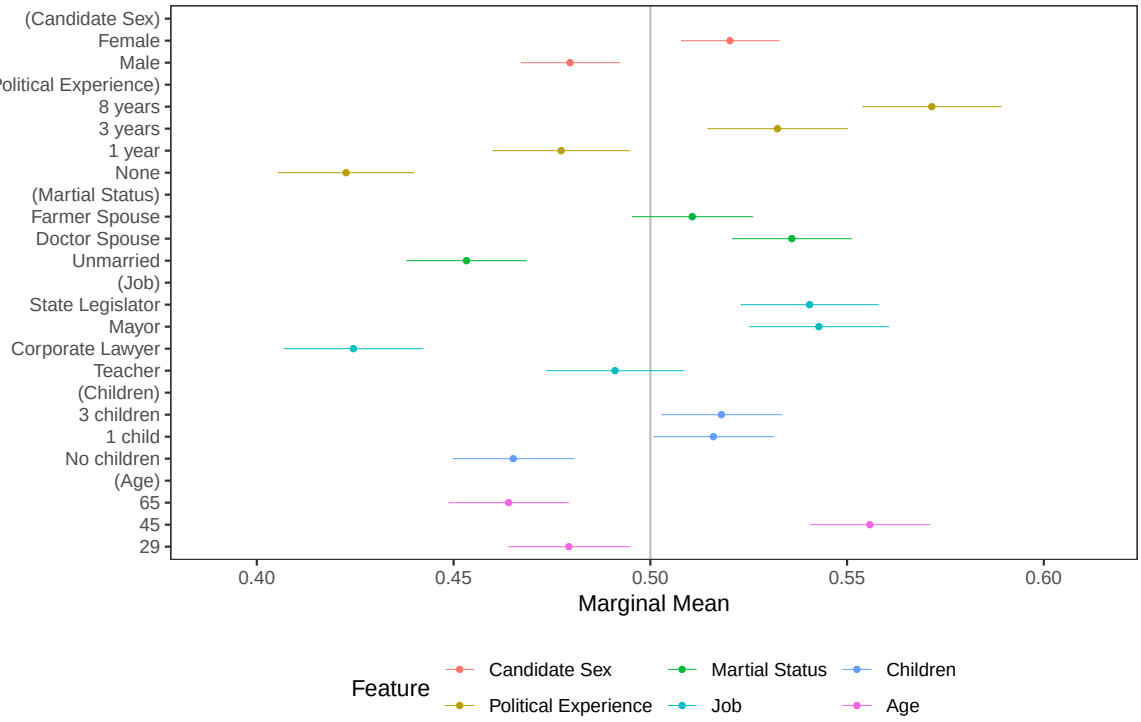
F Teele et al. (2018) Candidate Experiment

F.1 Replication using AMCEs



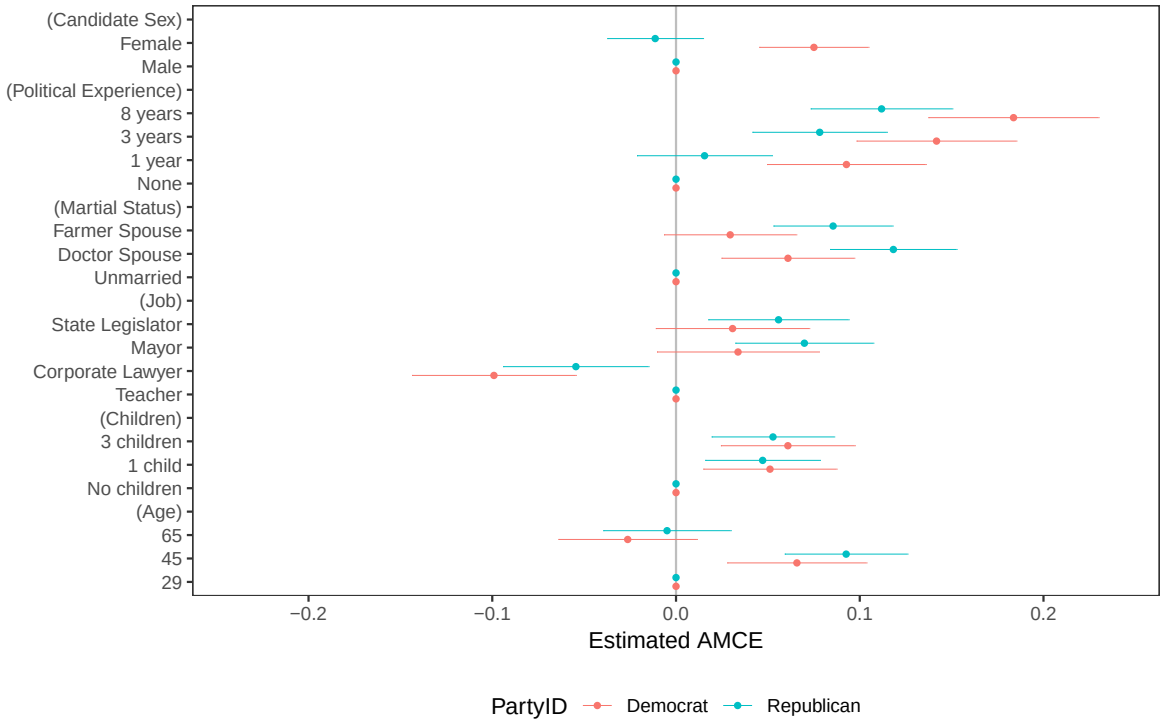
feature	level	estimate	std.error	z
Candidate Sex	Male	0.00		
Candidate Sex	Female	0.04	0.01	4.74
Political Experience	None	0.00		
Political Experience	1 year	0.05	0.01	4.06
Political Experience	3 years	0.11	0.01	8.47
Political Experience	8 years	0.15	0.01	10.83
Martial Status	Unmarried	0.00		
Martial Status	Doctor Spouse	0.08	0.01	7.25
Martial Status	Farmer Spouse	0.06	0.01	5.26
Job	Teacher	0.00		
Job	Corporate Lawyer	-0.07	0.01	-5.29
Job	Mayor	0.05	0.01	3.74
Job	State Legislator	0.05	0.01	3.59
Children	No children	0.00		
Children	1 child	0.05	0.01	4.84
Children	3 children	0.05	0.01	4.77
Age	29	0.00		
Age	45	0.08	0.01	7.07
Age	65	-0.01	0.01	-1.24

F.2 Replication using MMs

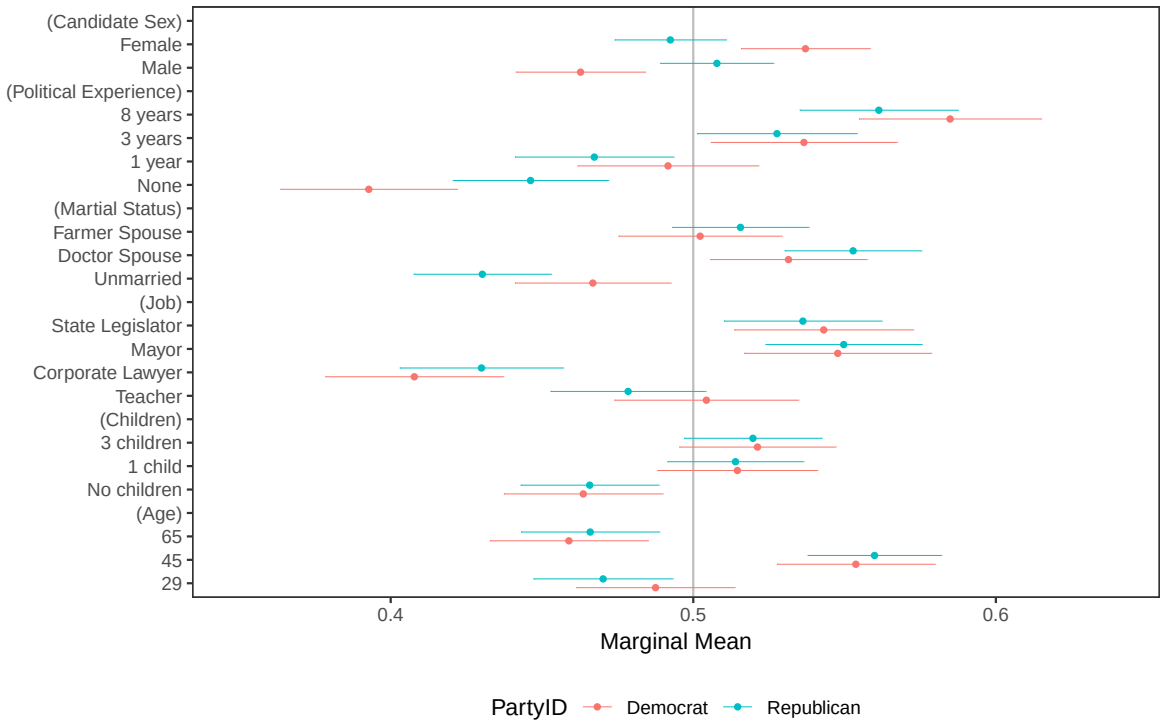


feature	level	estimate	std.error	z
Candidate Sex	Male	0.48	0.01	-3.22
Candidate Sex	Female	0.52	0.01	3.20
Political Experience	None	0.42	0.01	-8.81
Political Experience	1 year	0.48	0.01	-2.56
Political Experience	3 years	0.53	0.01	3.58
Political Experience	8 years	0.57	0.01	7.99
Martial Status	Unmarried	0.45	0.01	-6.05
Martial Status	Doctor Spouse	0.54	0.01	4.66
Martial Status	Farmer Spouse	0.51	0.01	1.37
Job	Teacher	0.49	0.01	-1.01
Job	Corporate Lawyer	0.42	0.01	-8.44
Job	Mayor	0.54	0.01	4.77
Job	State Legislator	0.54	0.01	4.55
Children	No children	0.47	0.01	-4.47
Children	1 child	0.52	0.01	2.07
Children	3 children	0.52	0.01	2.34
Age	29	0.48	0.01	-2.65
Age	45	0.56	0.01	7.28
Age	65	0.46	0.01	-4.66

F.3 Subgroup Analysis using AMCEs



F.4 Subgroup Analysis using MMs



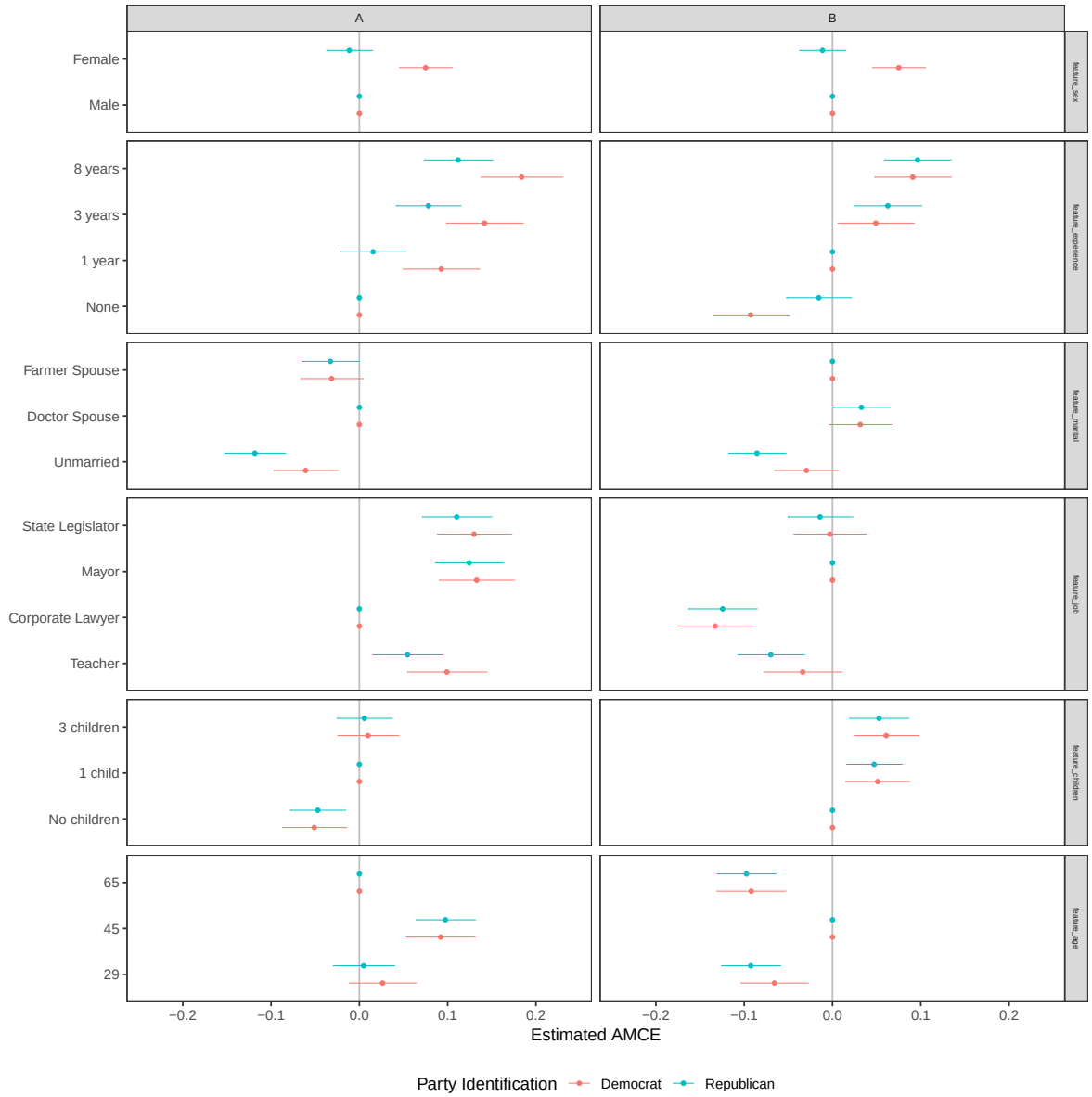
F.5 Nested Model Comparison: Male/Female Voters

```
## Analysis of Deviance Table
##
## Model 1: winner ~ feature_sex + feature_experience + feature_marital +
##     feature_job + feature_children + feature_age
## Model 2: winner ~ feature_sex + feature_experience + feature_marital +
##     feature_job + feature_children + feature_age + Sex + feature_sex:Sex +
##     feature_experience:Sex + feature_marital:Sex + feature_job:Sex +
##     feature_children:Sex + feature_age:Sex
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1     12436     2997.2
## 2     12422     2992.2 14   5.0738 1.5046 0.1002
```

F.6 Nested Model Comparison: Democratic/Republican Voters

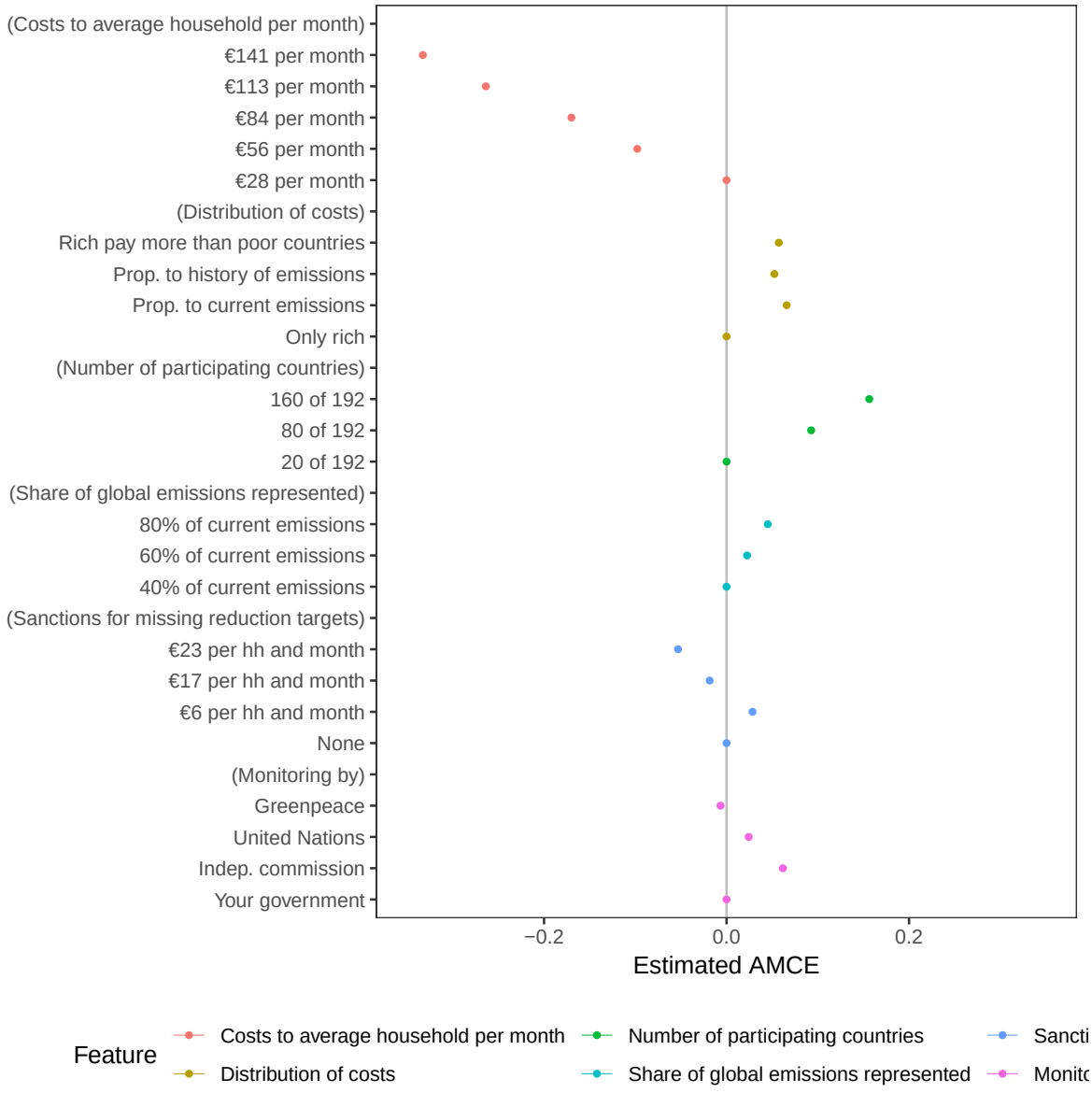
```
## Analysis of Deviance Table
##
## Model 1: winner ~ feature_sex + feature_experience + feature_marital +
##     feature_job + feature_children + feature_age
## Model 2: winner ~ feature_sex + feature_experience + feature_marital +
##     feature_job + feature_children + feature_age + PartyID +
##     feature_sex:PartyID + feature_experience:PartyID + feature_marital:PartyID +
##     feature_job:PartyID + feature_children:PartyID + feature_age:PartyID
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      9776     2352.8
## 2      9762     2342.9 14   9.8697 2.9373 0.0001739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F.7 Comparison of Alternative Reference Categories

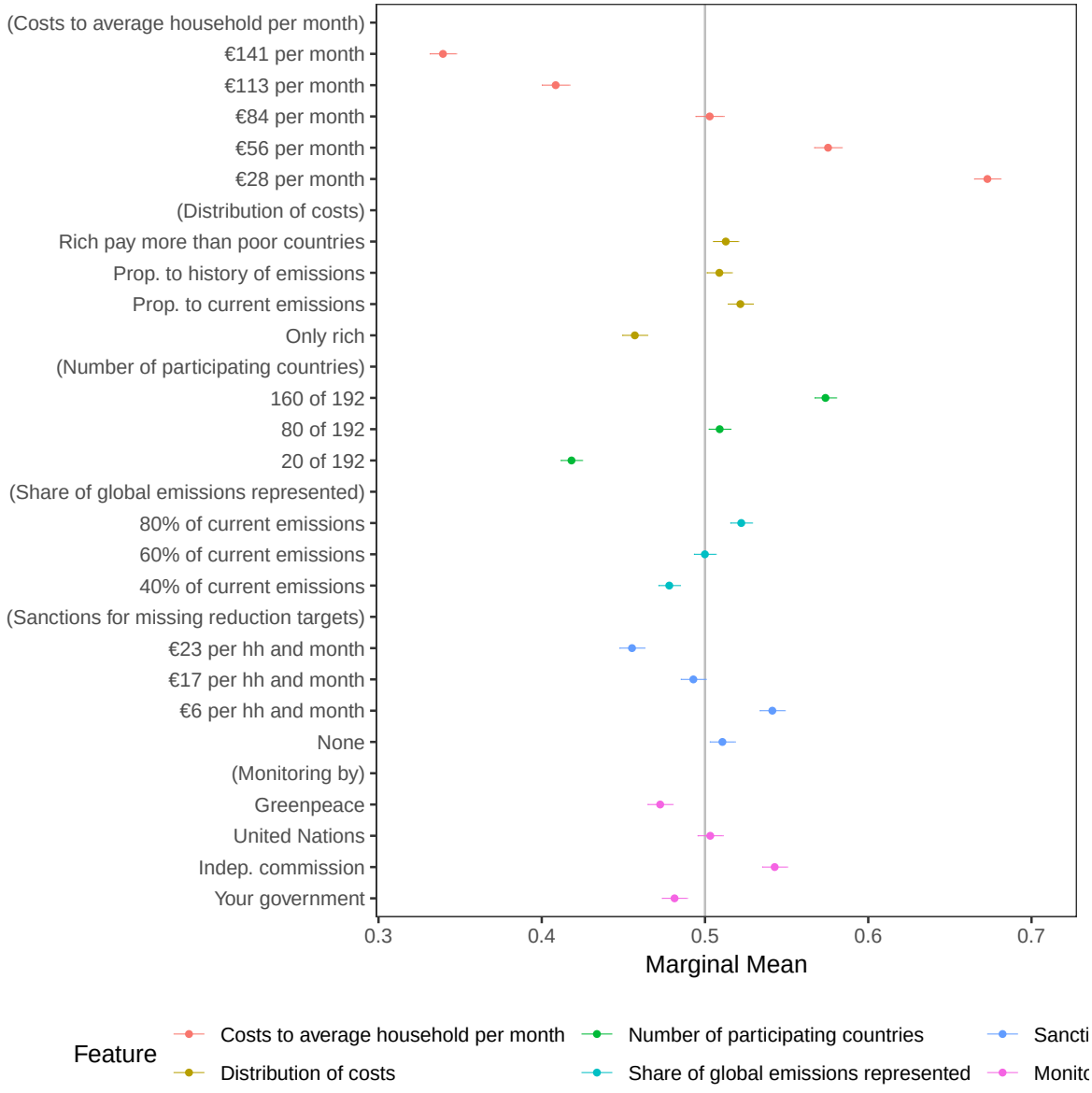


G Bechtel and Scheve (2013) Climate Agreement Experiment

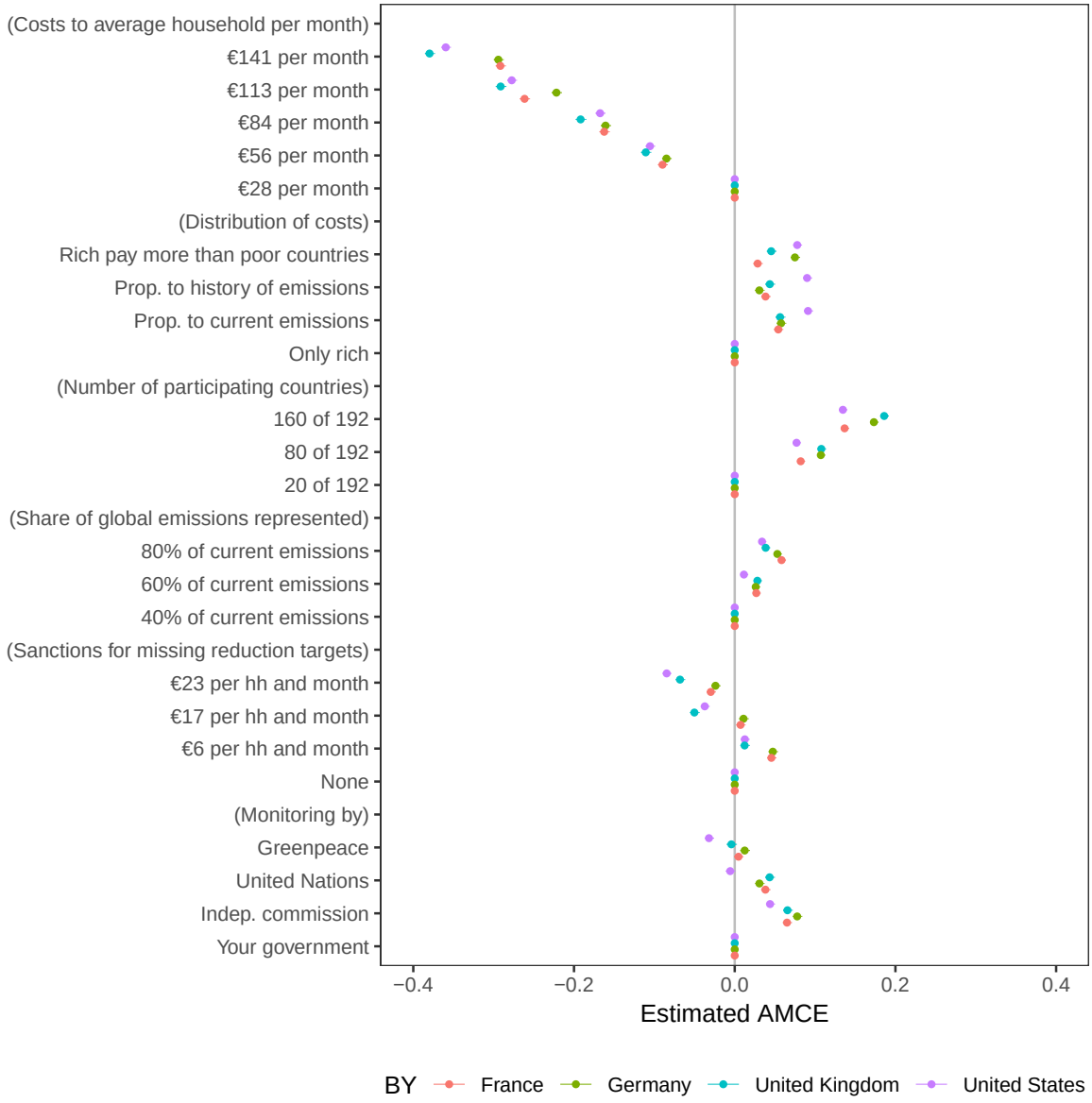
G.1 Replication using AMCEs



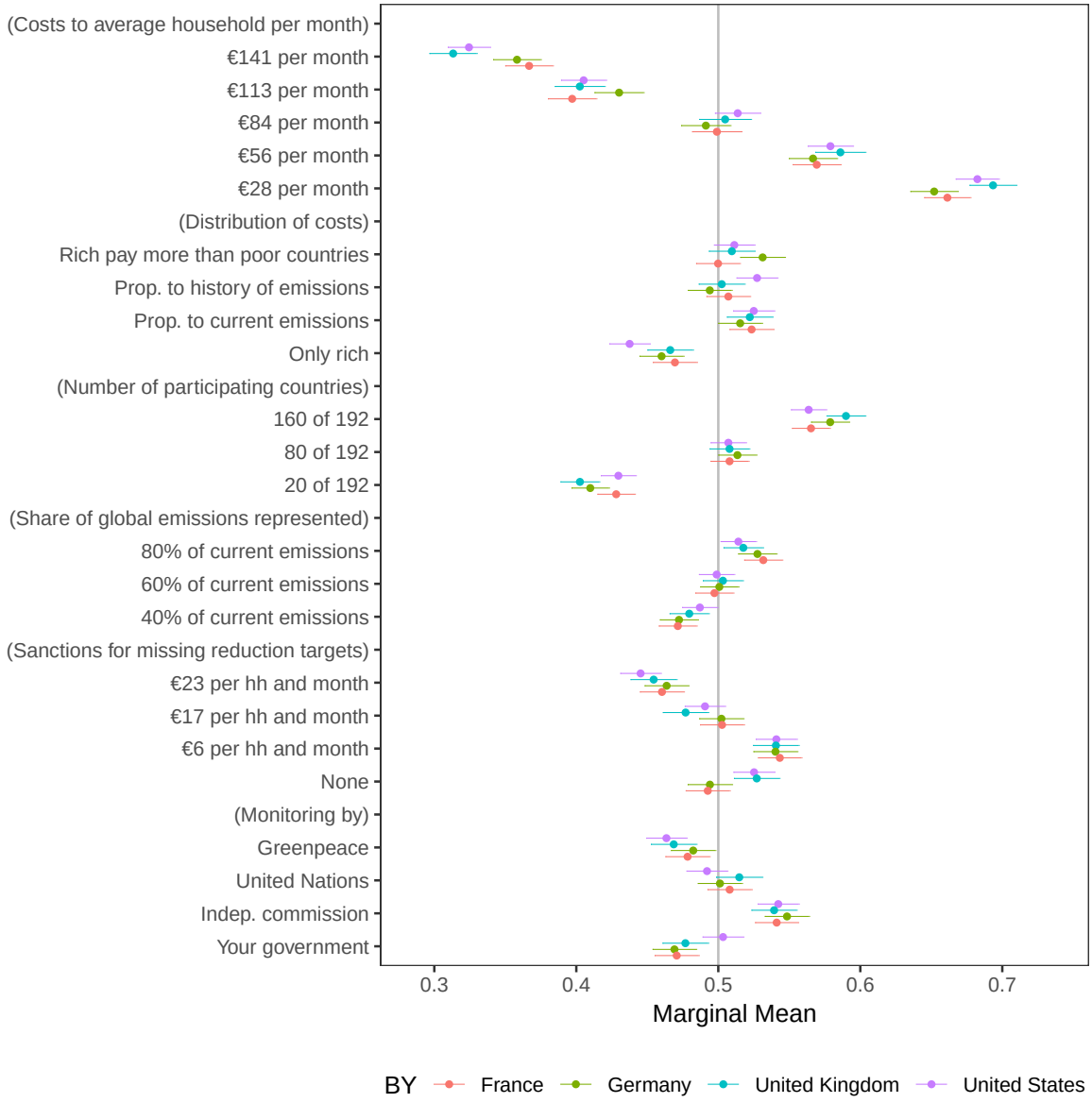
G.2 Replication using MMs



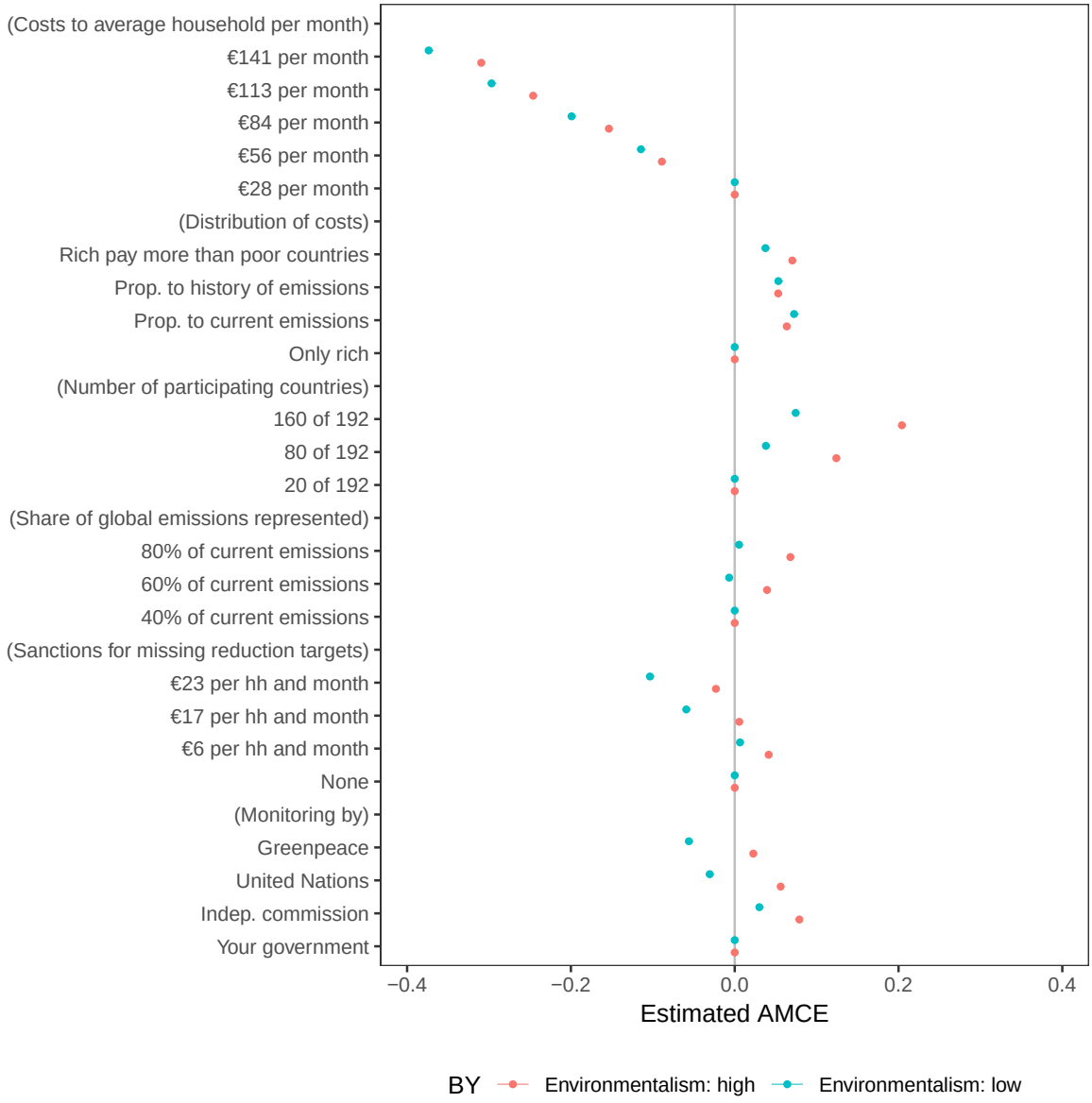
G.3 Subgroup Analysis using AMCEs: Country

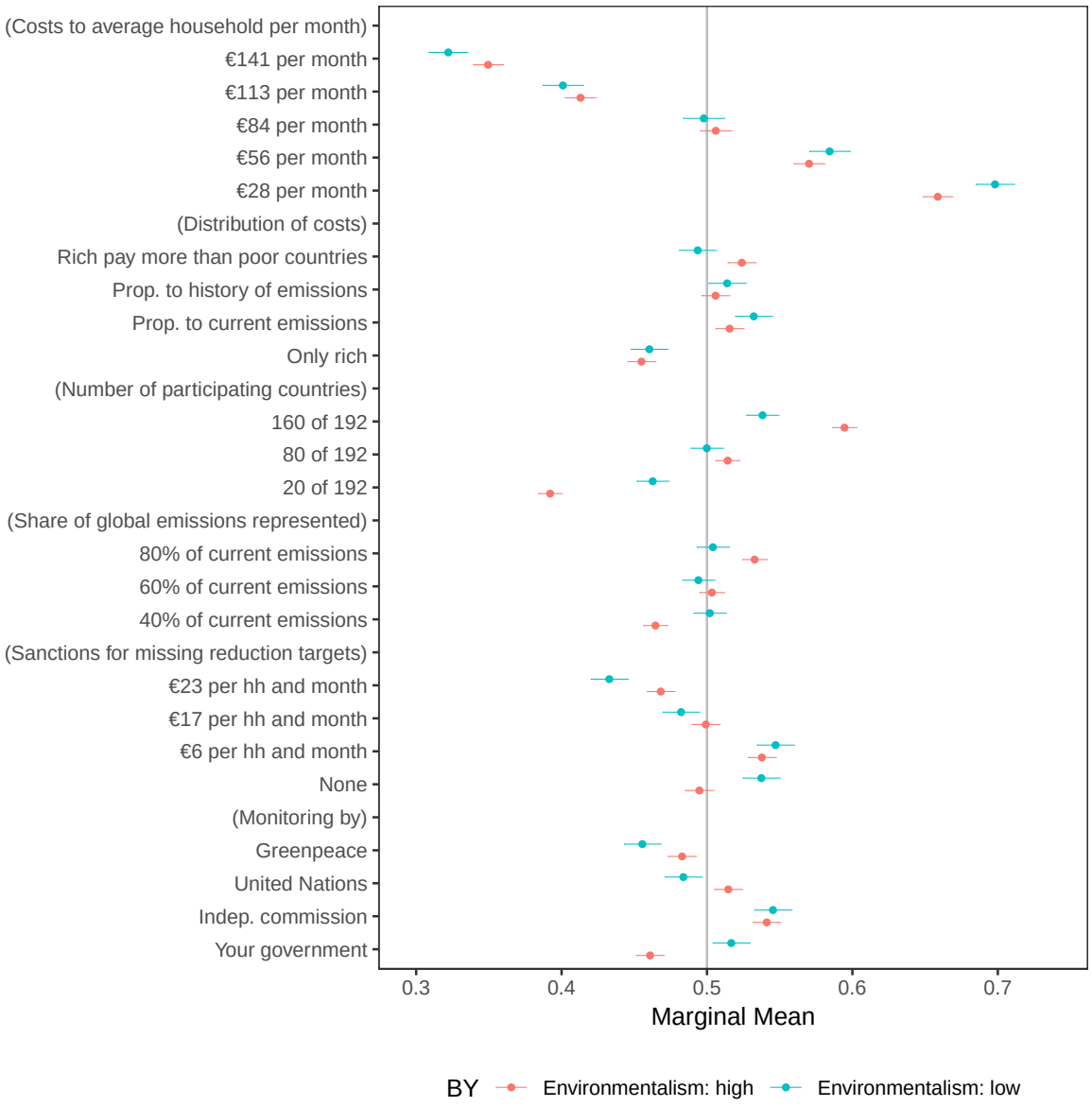


G.4 Subgroup Analysis using MMs: Country

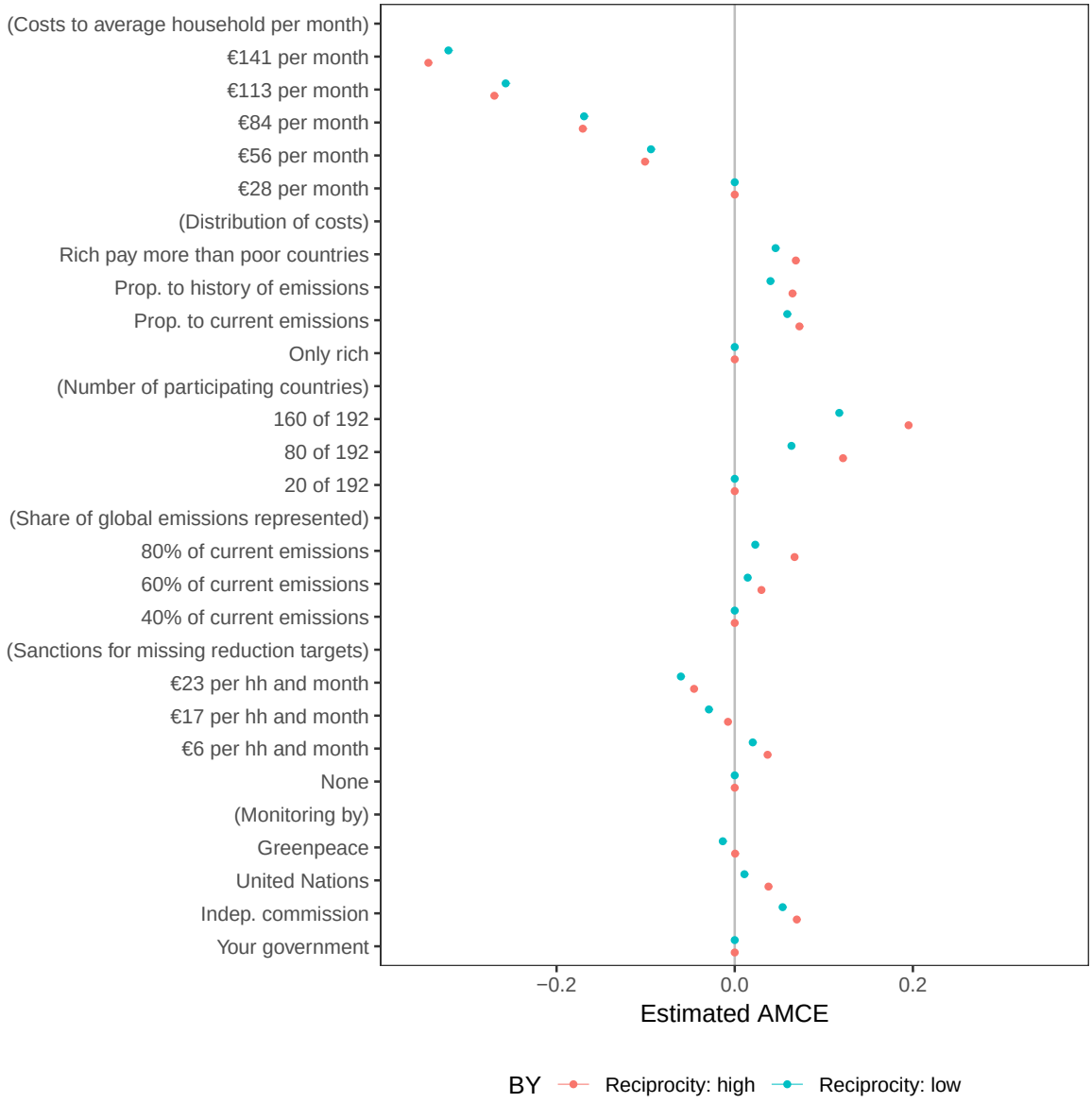


G.5 Subgroup Analysis using AMCEs: Environmentalism

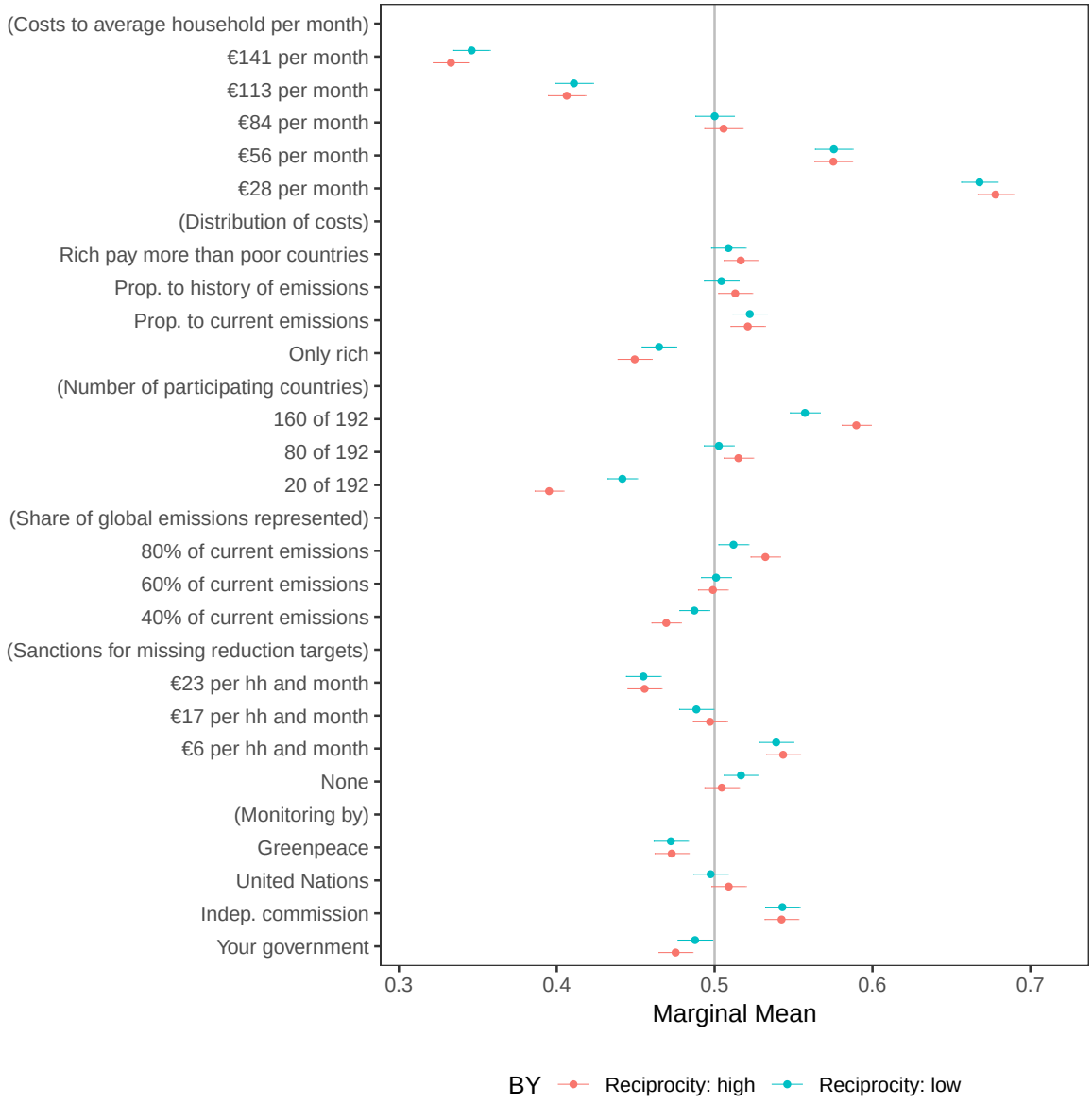



(Costs to average household per month)
€141 per month
€113 per month
€84 per month
€56 per month
€28 per month
(Distribution of costs)
Rich pay more than poor countries
Prop. to history of emissions
Prop. to current emissions
Only rich
(Number of participating countries)
160 of 192
80 of 192
20 of 192
(Share of global emissions represented)
80% of current emissions
60% of current emissions
40% of current emissions
(Sanctions for missing reduction targets)
€23 per hh and month
€17 per hh and month
€6 per hh and month
None
(Monitoring by)
Greenpeace
United Nations
Indep. commission
Your government
0.3
0.4
0.5
0.6
0.7
Marginal Mean
BY
Environmentalism: high
Environmentalism: low

G.6 Subgroup Analysis using AMCEs: Reciprocity



G.7 Subgroup Analysis using MMs: Reciprocity



G.8 Nested Model Comparison: Country

```
## Analysis of Deviance Table
##
## Model 1: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj
## Model 2: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj + country + cost_cj:country +
##     distrib_cj:country + ctries_cj:country + emissions_cj:country +
##     sanctions_cj:country + monitoring_cj:country
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1     67982      15601
## 2     67928      15555 54   45.983 3.7187 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

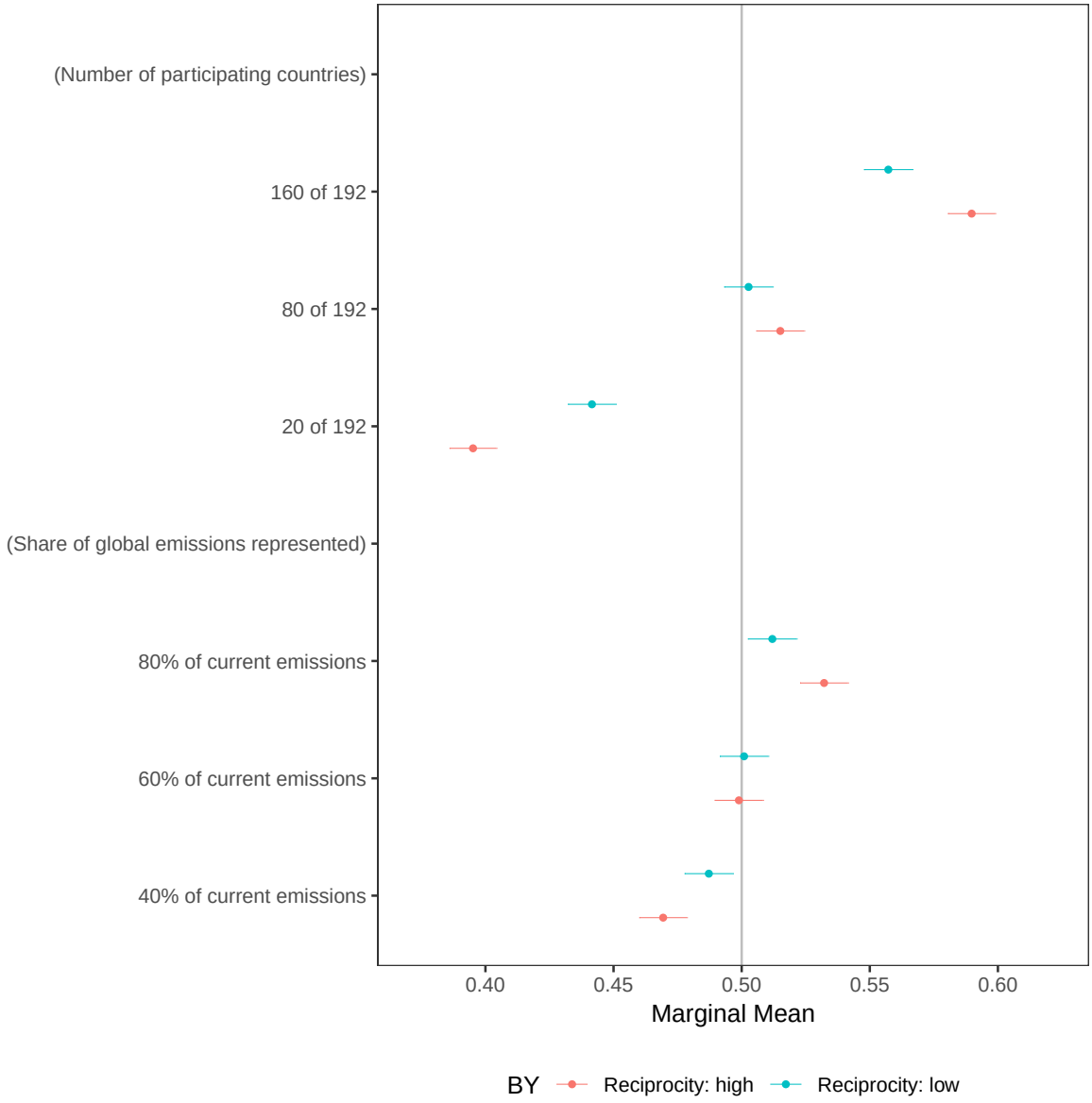
G.9 Nested Model Comparison: Environmentalism

```
## Analysis of Deviance Table
##
## Model 1: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj
## Model 2: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj + environmentalism + cost_cj:environmentalism +
##     distrib_cj:environmentalism + ctries_cj:environmentalism +
##     emissions_cj:environmentalism + sanctions_cj:environmentalism +
##     monitoring_cj:environmentalism
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1     67974      15599
## 2     67956      15491 18   107.83 26.279 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

G.10 Nested Model Comparison: Reciprocity

```
## Analysis of Deviance Table
##
## Model 1: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj
## Model 2: choice_cj ~ cost_cj + distrib_cj + ctries_cj + emissions_cj +
##     sanctions_cj + monitoring_cj + reciprocity + cost_cj:reciprocity +
##     distrib_cj:reciprocity + ctries_cj:reciprocity + emissions_cj:reciprocity +
##     sanctions_cj:reciprocity + monitoring_cj:reciprocity
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1     67982      15601
## 2     67964      15570 18   30.831 7.4767 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

G.11 Comparison of Alternative Reference Categories



This paper was built using `knitr::knit2pdf()` under the following environment:

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gridExtra_2.3 ggplot2_3.1.1 cregg_0.3.1   rio_0.5.16
##
## loaded via a namespace (and not attached):
##  [1] zip_2.0.1         Rcpp_1.0.1        cellranger_1.1.0
##  [4] pillar_1.3.1      compiler_3.5.3    plyr_1.8.4
##  [7] forcats_0.4.0     tools_3.5.3       digest_0.6.18
## [10] lattice_0.20-38   evaluate_0.13     tibble_2.1.1
## [13] gtable_0.3.0      ggstance_0.3.1    pkgconfig_2.0.2
## [16] rlang_0.3.4       Matrix_1.2-17     openxlsx_4.1.0
## [19] curl_3.3          haven_2.1.0       xfun_0.6
## [22] withr_2.1.2       stringr_1.4.0     knitr_1.22
## [25] hms_0.4.2         lmtest_0.9-36     grid_3.5.3
## [28] data.table_1.12.2 R6_2.4.0          survival_2.44-1.1
## [31] readxl_1.3.1      foreign_0.8-71    magrittr_1.5
## [34] codetools_0.2-16  splines_3.5.3     scales_1.0.0
## [37] assertthat_0.2.1  xtable_1.8-3      colorspace_1.4-1
## [40] sandwich_2.5-1    survey_3.35-1     stringi_1.4.3
## [43] lazyeval_0.2.2    munsell_0.5.0     crayon_1.3.4
## [46] zoo_1.8-5
```