Measuring Subgroup Preferences in Conjoint Experiments

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

December 12, 2018

Abstract

Conjoint analysis is an increasingly prominent tool for studying political preferences. The method powerfully 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), which measure the degree to which a given value of a conjoint profile feature increases or decreases respondents' support for the overall profile relative to a baseline, averaging across all respondents and all other profile features. While the AMCE has a clear causal interpretation, most published conjoint analyses also use AMCEs to simply describe preferences, often including comparisons of AMCEs between subgroups of respondents. We show how this descriptive use of conditional AMCEs can be substantially misleading about the degree of agreement or disagreement between subgroups due the simple, but often forgotten, property that interactions are sensitive to the reference category used in regression 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 rapidly accelerating use of conjoint analyses, this paper makes an important contribution by highlighting pitfalls and presenting advice for best practice in the analysis and presentation of conjoint experiments.

Amidst the dramatically increased use of experiments within political science (Druckman et al., 2006; Mutz, 2011), conjoint experimental designs have recently become a prominent methodological tool in political science. While traditional survey experiments tend to examine 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 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), or public policies (Gallego and Marx, 2017; Hankinson, 2018). The driving force behind this use of conjoint analysis has been the introduction by Hainmueller, Hopkins, and Yamamoto (2014) of a fully randomized conjoint design and an associated analytic approach that emphasizes a single quantity of interest: namely, 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, giving the AMCE a clear causal interpretation: the degree to which a given value of a feature increases or decreases respondents' favorability toward a packaged conjoint profile relative to a baseline, averaging across all respondents and all other profile features.

While randomization of profile features gives the AMCE a causal interpretation, most published conjoint analyses in political science use AMCEs for *descriptive* purposes: that is, to map variation in favorability toward a multidimensional object across its various features. This is particularly the case when researchers engage in subgroup analyses of conjoint experiments in search of preference heterogeneity. For example, Hainmueller, Hopkins, and Yamamoto (2014) perform a subgroup analysis on their original immigration experiment in which they perform a median split on a measure of ethnocentrism and then compare 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. In a different domain, Ballard-Rosa, Martin, and Scheve (2016) compare preferences over tax policies across a number of subgroups defined by demographics and political orientations. Teele, Kalla, and Rosenbluth (2018) compare AMCEs for features of male and female political candidates among male and female respondents. Kirkland and Coppock (2017) do a similar comparison between Democrats and Republicans in hypothetical elections.

In these and many other articles, interpretation of these subgroup analyses focus not just on the *causal effects* of profile features within each subgroup (what Hainmueller et al. term "conditional AMCEs"; 13) but also on an implied quantity of interest: the *difference* between two conditional AMCEs across subgroups. Searching for causal effect heterogeneity is an increasingly common feature of experimental analysis (Green and Kern, 2012; Ratkovic and Tingley, 2017; Grimmer, Messing, and Westwood, 2017), yet most conjoint analyses use this difference-in-AMCEs instead to *descriptively* interpret apparent differences in favorability toward objects with a given feature (e.g., immigrants from Syria) between the two groups (e.g., low and high ethnocentrism respondents).

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 differences-in-AMCEs. Yet authors frequently use visual or more formal comparisons of conditional AMCEs to make descriptive claims about such differences, leading themselves and readers astray. Differences in AMCEs do not provide inference into difference between subgroups' favorability toward a conjoint feature. In

this paper, we show that a difference in underlying subgroup preferences — like a difference in willingness to support a Syrian immigrant between high and low ethnocentrism respondents — is only reflected in the difference-in-AMCEs under particular preference configurations and analytic choices. The underlying cause of this error is simple and familiar to any applied researcher but appears to be forgotten in most applied conjoint work.

As we will show, where preferences in subgroups toward the experimental reference category are similar, the difference-in-AMCEs conveys preferences reasonably well but where preferences between subgroups diverge in the reference category, the difference-in-AMCEs is a misleading representation of underlying patterns of favorability. Yet most published conjoint studies appear to 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. Ultimately AMCEs are relative, not absolute, statements about preferences so subgroup differences are also relative not absolute statements about preference heterogeneity.[1] There is simply no predictable inference to be drawn from subgroup causal effects to the levels of underlying subgroup opinion. This inferential error — interpreting differences in the size of causal effects as descriptive differences in preferences — appears to be widespread in published conjoint analyses. 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 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.

In what follows, we demonstrate the challenges of conjoint analysis and reminder readers of how reference category choice for profile features creates significant problems for comparing conditional AMCEs across respondent subgroups. We show how the use of an arbitrary reference category means the size and the direction of differences-in-AMCEs have little relationship to the underlying degree of favorability of the subgroups toward profiles with particular features and that reference category choices can make similar preferences look dissimilar and dissimilar preferences look similar, using examples drawn from the published political science literature (namely experiments by Hainmueller, Hopkins, and Yamamoto 2014; Ballard-Rosa, Martin, and Scheve 2016; 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: (1) unadjusted marginal means, a quantity measuring favorability toward a given feature, and (b) an omnibus F-test, measuring differences therein. Newly developed 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.

[1] For example, in a comparison of subgroup effects for Democrats and Republicans, Republicans might display a smaller effect because their preferences in the reference category are already very positive, such that a large positive effect for Democrats occurs despite Democrats being less supportive than Republicans in either experimental condition.

Quantities of Interest in Conjoint Experiments

Conjoint analysis serves two purposes. One is to assess causal effects. Another is preference description.[2] In causal inference, 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.

Importantly, 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 derive from combinations of cell means, marginal means, and the grand mean, as is common in the traditional analysis of factorial experiments. In a forced choice 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 so 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.

Average marginal component effects (AMCEs) 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, marginalizing across 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 $Pr(Y = 1|X = x) = 0$ while a marginal mean of 1 indicates respondents select profiles with that feature level with probability ($Pr(Y = 1|X = x) = 1$.[3] With rating scale outcomes, marginal means can vary arbitrarily along

[2]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.

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

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 multicollinearity.[4] 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 (upper panel) and an analogous examination of the results using marginal means (lower panel). Here and throughout we use visual presentation of results, but full numerical estimates including appropriate standard errors are presented in the Appendix. 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.

The AMCE is often described as an estimate of the relative favorability of profiles with counterfactual levels of a feature; for example, "male candidates are preferred to female candidates" (Teele, Kalla, and Rosenbluth, 2018, 6). Hainmueller, Hopkins, and Yamamoto (2014) similarly describe some of the results of conjoint on preferences toward Congressional 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)

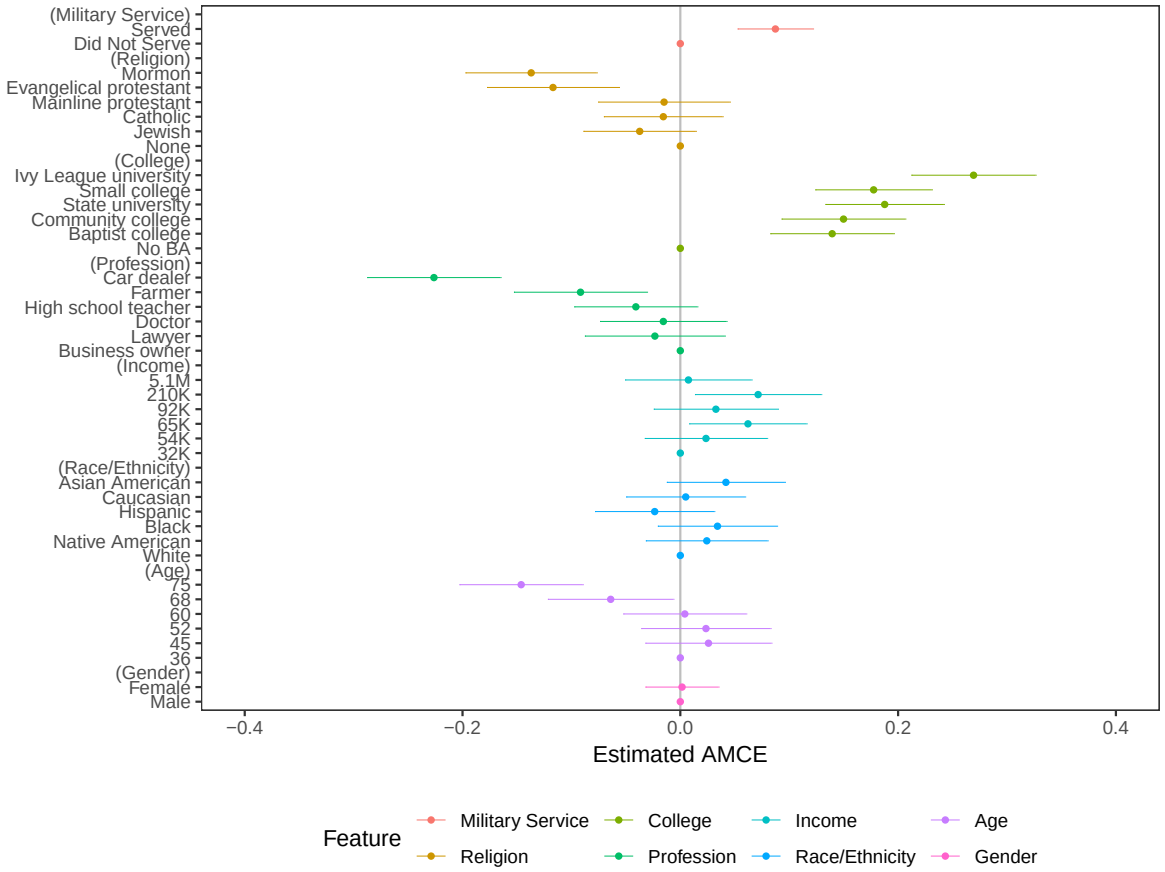
These examples make clear that despite the *causal* inference potentially provided by the AMCE, the quantity of interest is frequently used to provide a characterization of a preferences that has a distinctly descriptive flavor. Indeed, this style of description is widespread in conjoint analyses. Ballard-Rosa, Martin, and Scheve (2016) interpretation their tax preference conjoint:

> we find strong support for progressive preferences over federal income taxes among the American public [...] respondents are less likely to support a given tax plan as the tax rate on the poorest three groups increases but more likely

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). 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.

[4] 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.

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



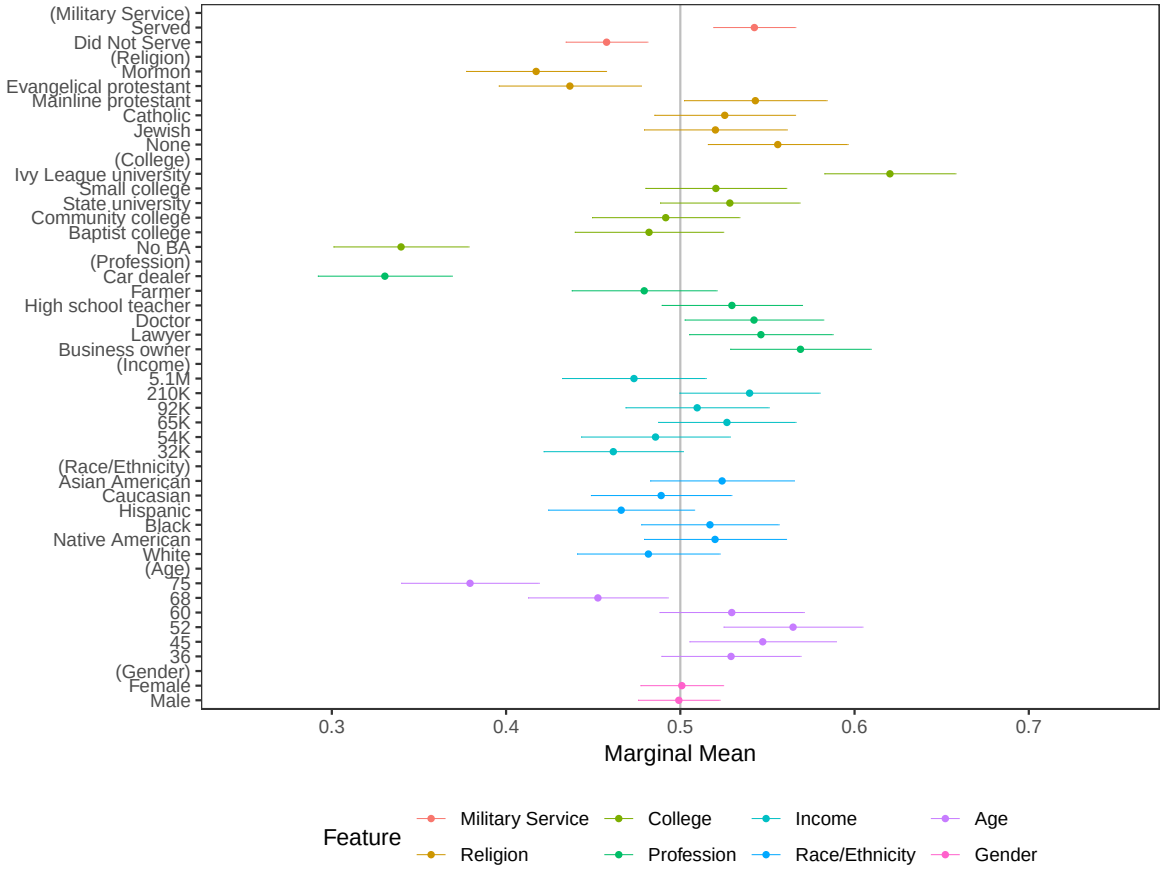
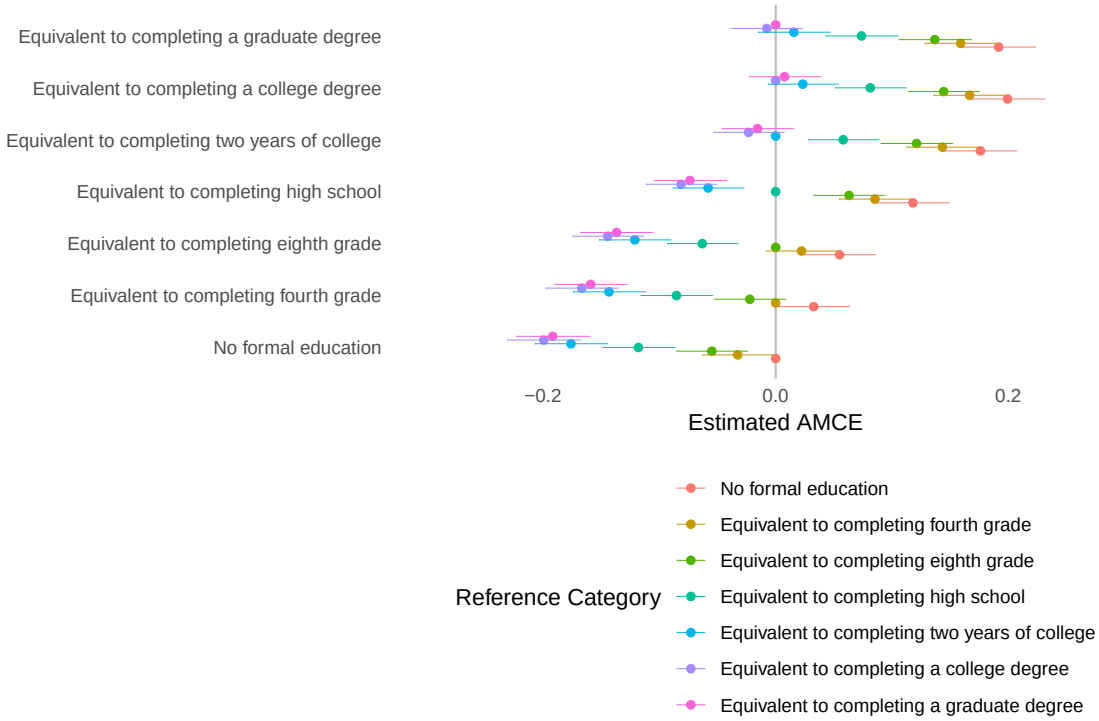


Figure 2: Reference Category Diagnostic for the 'Education' Feature from Hainmueller et al.'s (2014) Immigration Experiment



to support an income tax policy when the tax rate on the richest two groups increases, at least to a point.

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 (negative) AMCE is read as higher (lower) favorability but it is only higher (lower) 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 means the AMCEs in the candidate experiment are all expressed relative to this non-religious baseline; the difference (if any) between other pairs of marginal means (e.g., evaluations of Mormon and Evangelical candidates) is not obvious. Relatedly, the negative direction (and the size) of the AMCEs for Mormon and Evangelical candidates would be different if the least-liked category (Mormon candidates) were the reference group. In that case, the AMCE for Evangelicals would be small and positive and the AMCEs for all other categories (including the presented reference category, "none") would be large and positive.

Being a familiar analytic problem in any regression context, this choice of reference category for estimating AMCEs can seem trivial but is quite consequential. For example, in Hainmueller, Hopkins, and Yamamoto's candidate experiment, 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. Figure 2 shows how the estimated AMCEs for each level of education 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.[5] 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.[6]

Consequences of Arbitrary Reference Category Choice

Given the need to choose a reference category for every feature in order to estimate AMCEs, an important question is: 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, Hainmueller, Hopkins, and Yamamoto (2014) choose fluency as a baseline; on the prior trips to the US feature, "never" is chosen as the baseline. These seem sensible on face value — it might seem less useful to define all effects relative to "having visited many times as a tourist" as the AMCEs then lose an immediately intuitive interpretation.

Yet the choice is consequential. A possibly surprising consequence of the seemingly arbitrary selection of a reference category — and the resulting arbitrariness of both the size and direction of AMCEs — is that it can provide highly distorted descriptive interpretation 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). Table 1 reports a list of recently published articles in political science that engage in this form of subgroup analysis.[7] Given the commonly

[5] As another example, in Ballard-Rosa, Martin, and Scheve's tax preference experiment, the lowest level of taxation is chosen as the reference category for each feature for reasons of substantive interpretation, yet despite the substantive intuitiveness of this, favorability toward the lowest level of taxation is not necessarily higher or lower than preferences for alternative tax rates.

[6] 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.

[7] Ratkovic and Tingley (2017) considered efficient methods for performing subgroup analyses in conjoint designs. Our focus here is on the narrower problem of interpreting subgroup analyses as traditionally performed using subsetting or interaction terms in a regression framework.

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

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

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

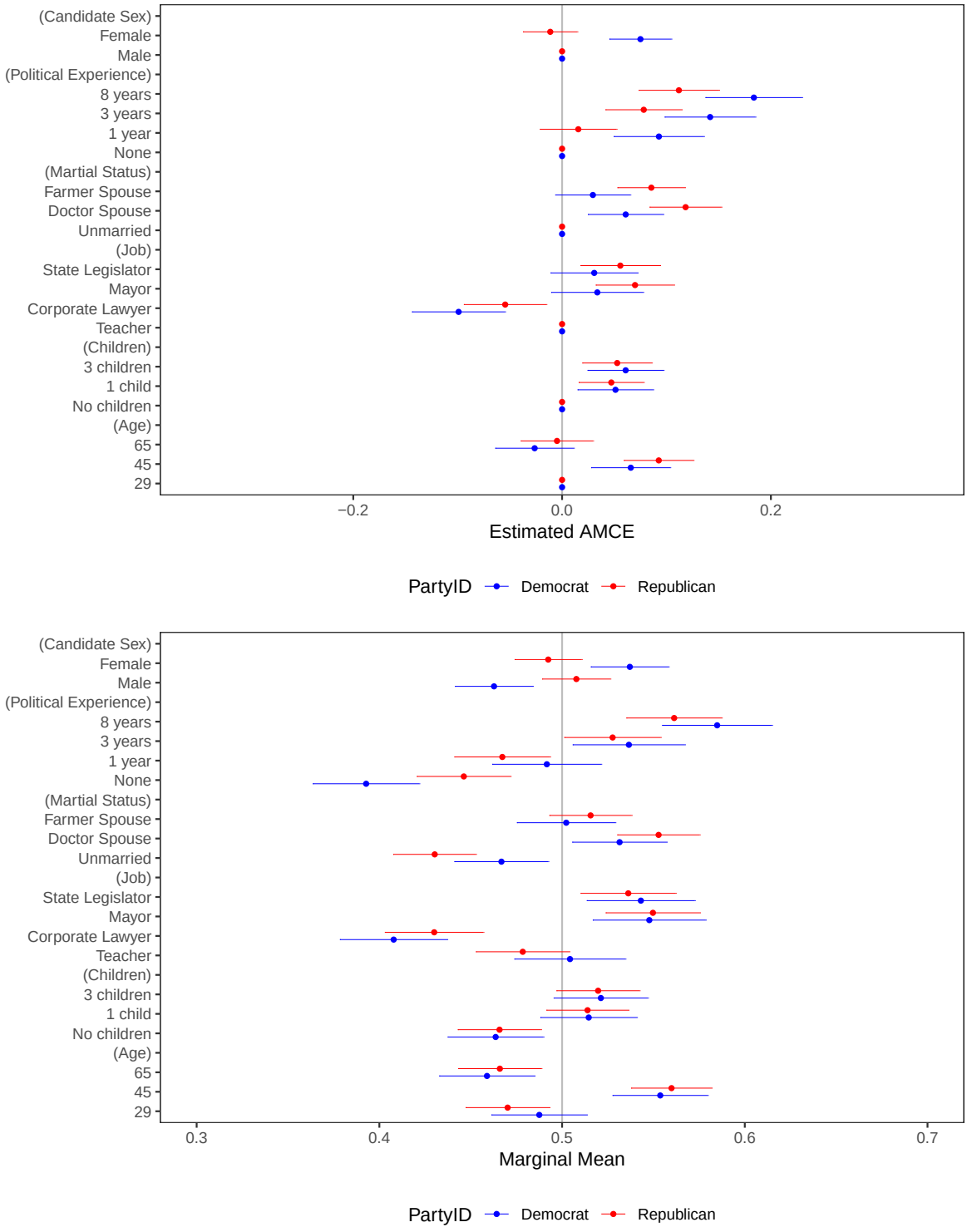
descriptive interpretations of conjoint experimental results, such subgroup analyses seem perfectly intuitive and the set of subgroups listed in the last column of Table 1 contains some unsurprising covariates, such as partisanship, that are of obvious theoretical interest in almost any study of individual preferences. Analytically, these conditional AMCEs can be obtained either from regression estimates on respondent subgroups or through interactions between conjoint features and respondent characteristics, the details of which are unimportant for our purposes.

Conditional AMCEs are not per se a problematic quantity of inference. Like subgroup analysis of any experiment, they convey the causal effect of an experimental factor on overall favorability among the subgroup of interest. Consider, for example, a two-condition party cue experiment where Democrats and Republicans are exposed to an endorsement cue from the Democratic party or no cue and opinions toward the policy serve as the outcome. It is sensible to imagine that effects of the cue might differ for the two groups and therefore to compare the size of cue effect among the two groups. Perhaps Democrats are more responsive to the Democratic party cue than are Republicans, making the causal effect larger for Democrats than Republicans. Discussions of conditional AMCEs in conjoint analyses often explicitly or implicitly engage in this kind of discussion comparing the size and direction of subgroup causal effects. 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, so too do conjoint analysts frequently interpret differences in conditional AMCEs descriptively rather than causally. For example, in one analysis Hainmueller, Hopkins, and Yamamoto (2014) visually compare the pattern of AMCEs among high- and low-ethnocentrism respondents and interpret that "the patterns of support are generally similar for respondents irrespective of their level of ethnocentrism" (22). Ballard-Rosa, Martin, and Scheve (2016) make similar comparisons in their tax policy conjoint: "While there are few strong differences in preferences for taxing the lower three income groups (the 'hard work' group has slightly lower elasticities for taxing the poor), there are strong differences in preferences for taxing the rich. Respondents who believe luck plays a role in economic success are more strongly progressive, although preferences over taxing the \$175K–\$375K bracket are relatively flat" (12). In these examples, the differences between conditional AMCEs are used as a way of descriptively characterizing differences in *preferences* between the groups rather than differences in *causal effects on preferences* in the groups.

As a more complete example, the upper panel of Figure 3 shows AMCEs for Teele, Kalla, and Rosenbluth's candidate choice experiment separately for Democratic and Republican voters, as provided in the original paper, and the lower panel shows the results using conditional marginal means. Again, we opt for visual presentation of results; tabular presentation of AMCEs, marginal means, and associated standard errors for all examples are included in the Appendix. Respondents' preference for female candidates is very apparent in both forms of analysis. Yet the discrepancy between the differences in preferences (i.e., conditional marginal means) and the differences in conditional AMCEs can be seen very clearly in the "political experience" feature in Figure 3 (the second set of estimates from the top in both panels). The conditional AMCEs in the upper panel 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 ex-

Figure 3: Replication of Teele et al. (2018) Candidate Experiment using AMCEs and MMs



perience compared to Republican voters (i.e., Republicans and Democrats differ in their preferences over experienced candidates). In reality, however, the conditional marginal means shown in the lower panel demonstrate that actually 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 are apparent for extreme categories in the visualization of conditional marginal means, but the conditional AMCEs suggest that preferences differ at all levels of experience, when in reality they do not. The selection of a reference category is therefore hugely consequential for a descriptive reading of the AMCE results.

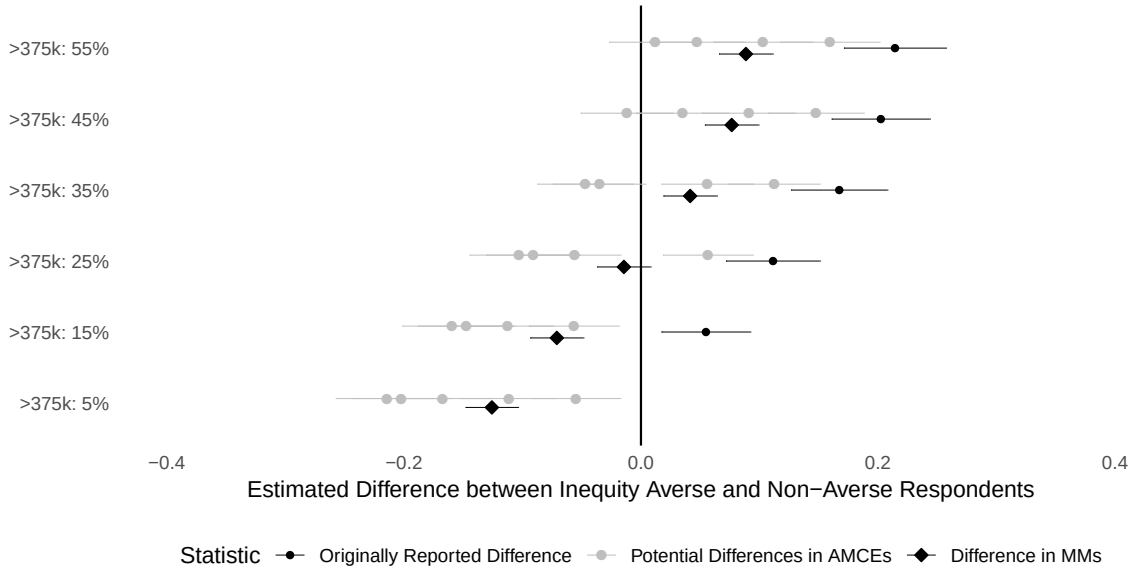
Interpreting conjoint AMCEs as measures of preferences is an inferential error. In a simple experiment like the party cue example just given, this kind of interpretation would be obviously flawed. A larger causal effect of the party cue for Democrats than Republicans does not necessarily mean that Democrats are more supportive than Republicans on average or in either condition. Effects are relative, not absolute, statements about preferences. Republicans, for example, might experience a smaller effect because their preferences in the control group are already very supportive, such that a large positive effect for Democrats occurs despite Democrats being less supportive than Republicans in either experimental condition. There is simply no predictable connection between subgroup causal effects and the levels of underlying subgroup preferences. This inferential error — interpreting differences in the size of causal effects as descriptive differences in preferences — appears to be widespread in published conjoint analyses. While 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, 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 cue treatment, $x \in 0, 1$, 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:

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

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 or characterize either of these quantities. Similarity of conditional AMCEs therefore only convey similarity of the *causal effect* of the feature, but do not convey similarity of *preferences* unless preferences toward profiles with the reference category are equivalent across 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 an arbitrarily selected reference category satisfies that equality assumption. When using a difference-in-AMCEs comparison to estimate a difference in

Figure 4: Estimated Preference Differences between Inequity Averse and Non-Averse Respondents from Ballard-Rosa et al. (2016) Tax Preference Experiment for Each Possible Reference Category



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.

We can see this bias clearly in a reanalysis of Ballard-Rosa, Martin, and Scheve's tax preference experiment. Figure 4 shows an analysis for the feature capturing the tax rate for the highest earners (those over $375,000 per year) replicating a portion of the results they present comparing inequity averse and non-averse respondents Ballard-Rosa, Martin, and Scheve (2016, 9 figure 2). The original analysis was presented as conditional AMCEs for the two subgroups with inequity averse respondents having positive AMCEs for all tax levels (relative to 5% as the reference category) and AMCEs for non-averse respondents being largely indistinguishable from zero. Figure 4 presents the implied difference-in-AMCEs from the original analysis as round black dots, demonstrating the substantial and positive *apparent* differences between the two groups. The black diamonds show the true differences in marginal means between the two groups. The gray dots represent the alternative differences-in-AMCEs that could have been generated from alternative choices of reference category using the same data. Because respondents in the two groups actually have substantially different preferences over the reference category 5% tax rate (inequity averse respondents are much less favorable toward this rate than non-averse respondents) differences-in-AMCEs make it seem like inequity averse respondents are much *more* favorable toward 15% and 25% tax rates than non-averse respondents, when in actuality averse respondents are *less* favorable toward the 15% tax rate than non-averse respondents and the two groups have largely similar views of a 25% tax rate. Not only does the difference-in-AMCEs overestimate group differences for very high tax rates (35%, 45%, 55%) but the difference-in-AMCEs flips the true direction of group differences for lower rates. The authors correctly read their data as showing "support for more progressive preferences is correlated with concern over societal inequality" (Ballard-Rosa, Martin, and Scheve, 2016, 9) but for the wrong reason: inequity averse and non-averse respondents are similarly favorable toward middling tax rates and diverge in their views of very high and very low rates for high earners.

It is worth highlighting two further features in Figure 4. 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 are therefore necessarily sometimes positive and sometimes 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. Yet the choice of reference category — likely unintentionally — resulted in the most extreme, positive difference-in-AMCEs that could be estimated from the data but alternative reference categories could have conveyed different, equally incorrect insights.

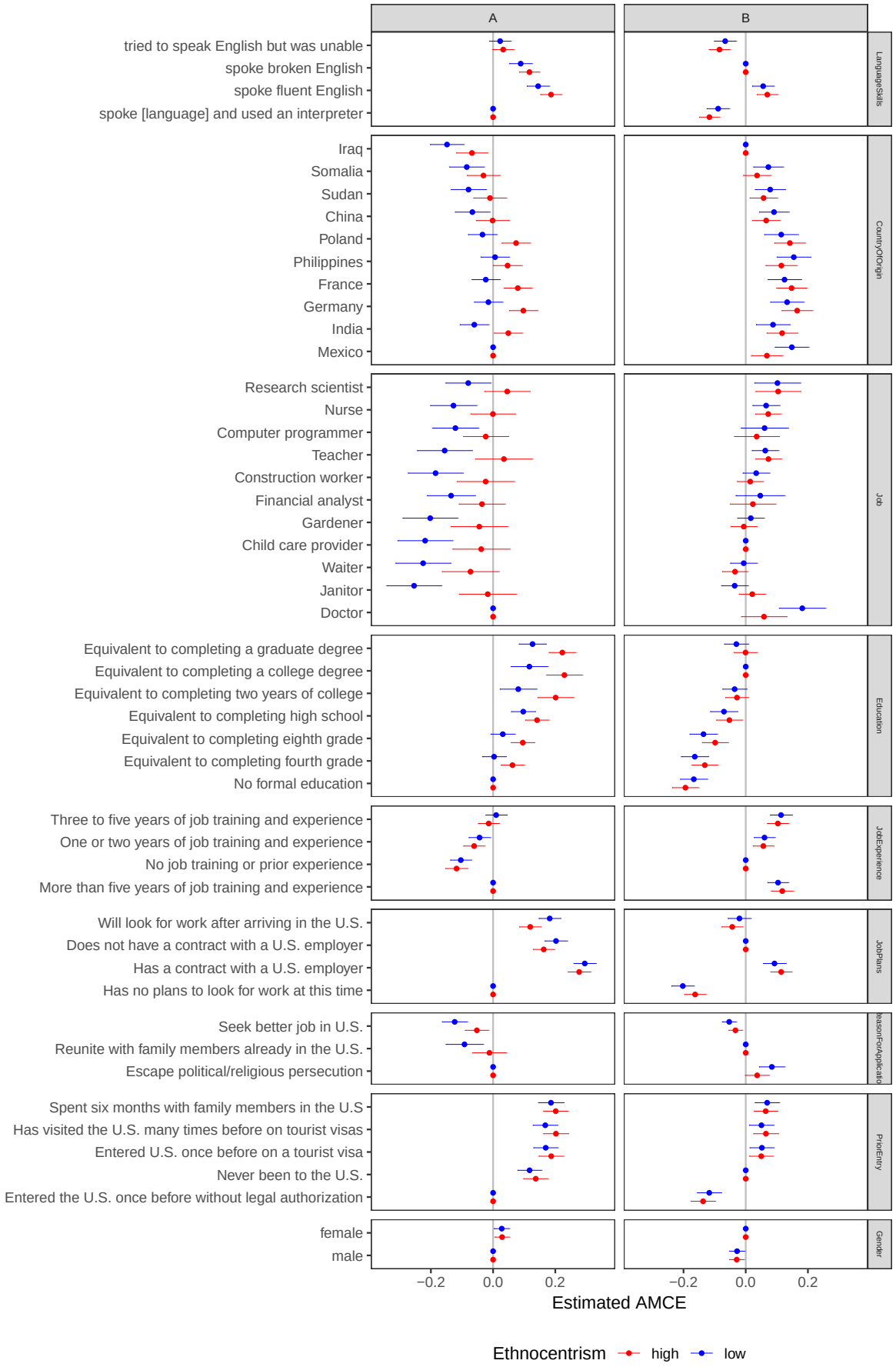
Second, and more practically, because there is no category for which the preferences of the two subgroups in this example are identical, no choice of reference category would have led to inferences from differences-in-AMCEs that accurately reflect the underlying difference in preferences. Even in the 25% tax rate category, the difference between the two groups is slightly negative. Were there a category for which preferences were equivalent, that could be sensibly chosen as the reference category in order to be able to interpret differences-in-AMCEs as differences in preferences. There is never any guarantee, however, that such a reference category exists in any given experimental dataset. If multiple subgroup analyses are performed, it is unlikely the same reference category would work well across all analyses, making consistent interpretation difficult. As such, even if the AMCEs are similar, it does not necessarily mean that *preferences* are similar; if AMCEs are dissimilar, it does not necessarily mean that preferences differ.

Ultimately, it is important to note that because conjoint analysis generates a sparse feature matrix (where there is never any guarantee that a particular combination of feature levels is observed in the data), it is also not possible to empirically select an appropriate set of reference categories using the data. It is impossible to know which cell — of the tens of thousands in the design — is the best choice of reference. This is because while it might seem possible to select a *marginally* appropriate reference category (i.e., one where preferences are similar with respect to a given feature), preferences toward that reference category may differ across other dimensions in the analysis. And it is furthermore 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. 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 Analysis of Subgroup Preferences in Conjoint Designs

We have shown that subgroup analyses of conjoint designs frequently entail the use of difference-in-AMCE comparisons and we have also shown that such analyses, counter-intuitively, do not demonstrate differences in preferences between groups due the near-impossibility of selecting a non-arbitrary reference category against which to estimate AMCEs. Thus the choice of reference category — while seemingly irrelevant — has dramatic inferential consequences in conjoint analyses. Here we provide a more complete

Figure 5: Comparison of AMCEs for Low- and High-Ethnocentrism Respondents Using Two Alternative Reference Categories Choices for Hainmueller et al. (2014) Immigration Experiment



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. Consider the left and right facets of Figure 5, 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 6) provides the intended descriptive comparison of subgroup preferences. Each dot and error bar represents the conditional marginal mean (and its standard 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, $p \leq 0.22$)

[8]The appendix contains comparable plots for experiments by Ballard-Rosa, Martin, and Scheve (2016) and Teele, Kalla, and Rosenbluth (2018).

Figure 6: Conditional Marginal Means, by Ethnocentrism, for Hainmueller et al.'s (2014) Immigration Experiment

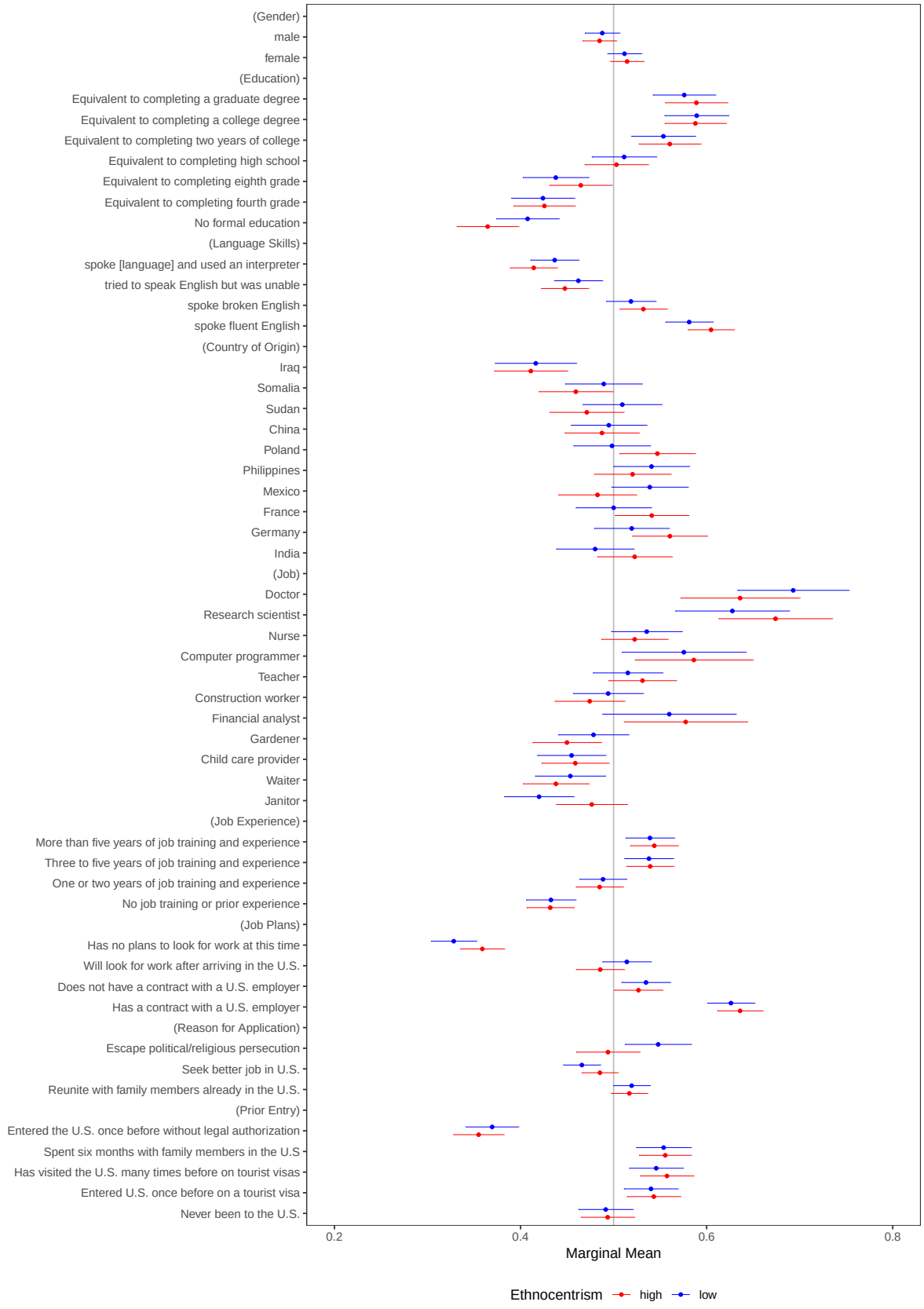
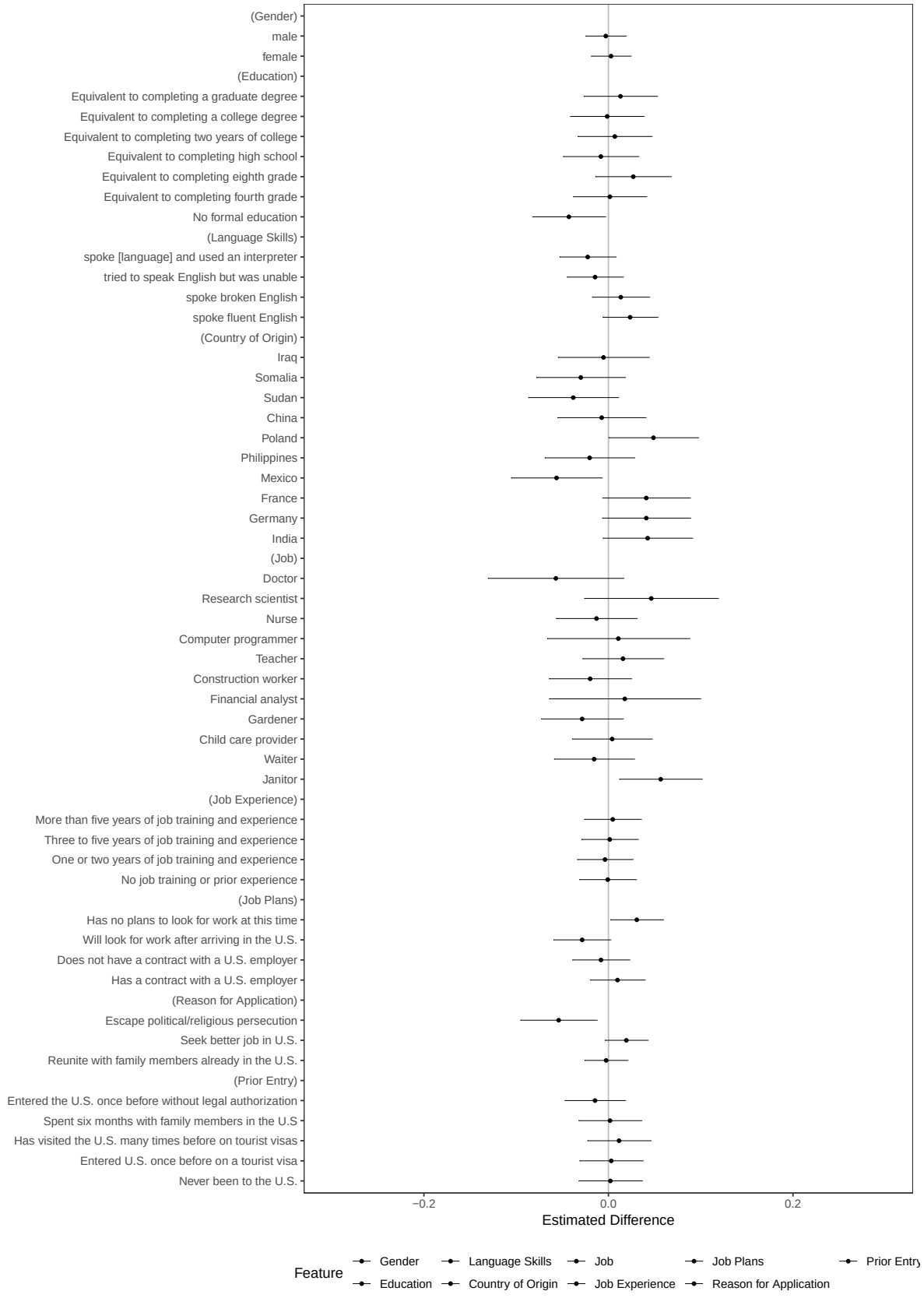


Figure 7: Differences in Conditional Marginal Means, by Ethnocentrism, for Hainmueller et al.'s (2014) Immigration Experiment



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, and (b) the same model with additional interactions between the subgrouping covariate and the features. The resulting F-test for the model comparison in this case again gives us little reason to believe there are subgroup differences: $F(4)=1.06$, $p \leq 0.37$. 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 7. 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 6: 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: $F(98)=1.16$, $p \leq 0.14$, which further demonstrates that the substantive interpretation provided by Hainmueller, Hopkins, and Yamamoto (2014) accurately identified a lack of between-group differences. This kind of test can also be used to assess heterogeneity across conjoint features. 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 5) 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 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. While that quantity conveys the difference in effects of a change in a given feature, heterogeneous effects do not necessarily mean different underlying preferences. Differences-in-AMCEs are almost always a biased estimate (of unknown sign and direction) of the difference in 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.

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.

[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.

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 Cooperation." *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.

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.

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 (which it almost always is in applied examples) this means we can additionally ignore the probability of

[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, $\mathbf{c}$.

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 learned 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:

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

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 an experiment where $N > C$ (the number of observations is large 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 characteristic), 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. 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:

$$\tau = E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_f = x_f] - E[Y|X_1 = \neg x_1, X_2 = x_2, \ldots, X_f = x_f] \tag{2}$$

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

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

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

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 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[13] portion of the feature space.

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.

For the *causal* interpretation of conjoint designs, comparisons of these marginal means is required. Comparisons between them provide causal inferences about the effect of

[12]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.

[13]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.

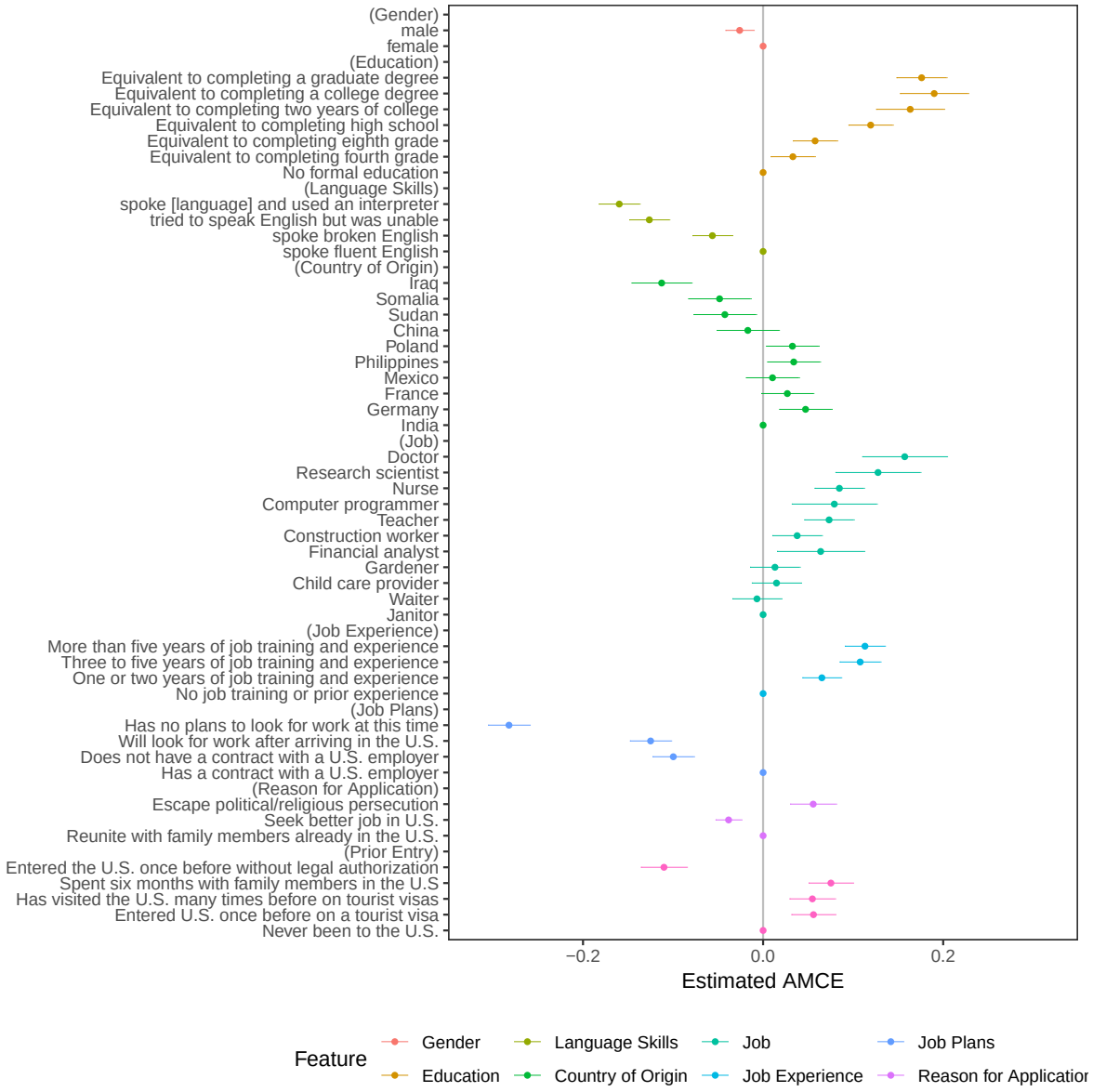
changing a focal feature, marginalizing across the distribution of other features. Because feature combinations (i.e., the profiles) are randomly constructed and randomly observed from all possible combinations, the distribution of other non-focal features is, in expectation, is independent of the focal 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. 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 effects 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) 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 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 $A = 1$. 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. 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. There is nothing inherently problematic about that but, as noted earlier, it requires either being clear about what features are being marginalized over or an analysis of only the complete and comparable subset of the design. So, researchers using such designs 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.

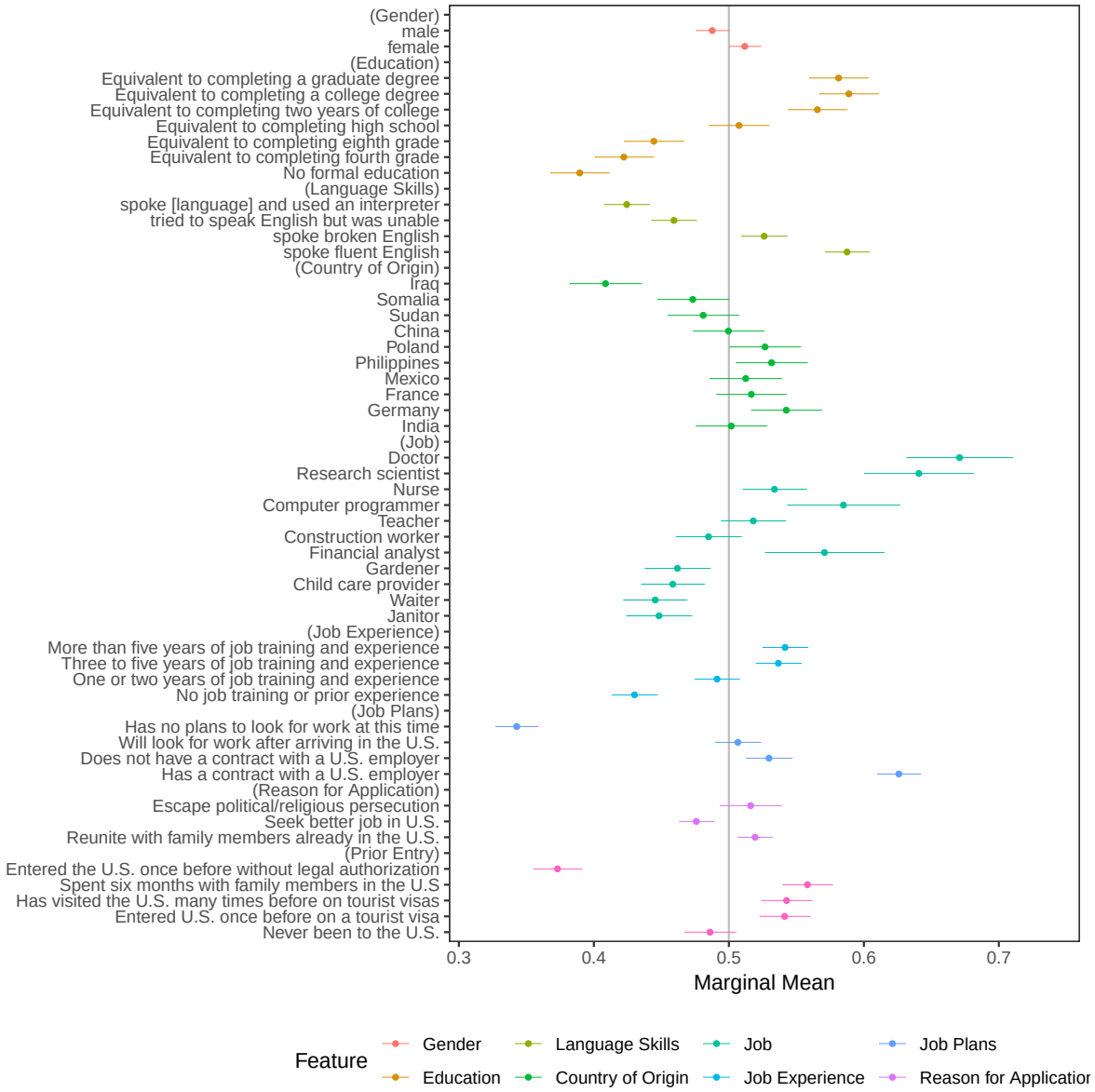
B Hainmueller et al. (2014) Immigration Experiment

B.1 Replication using AMCEs



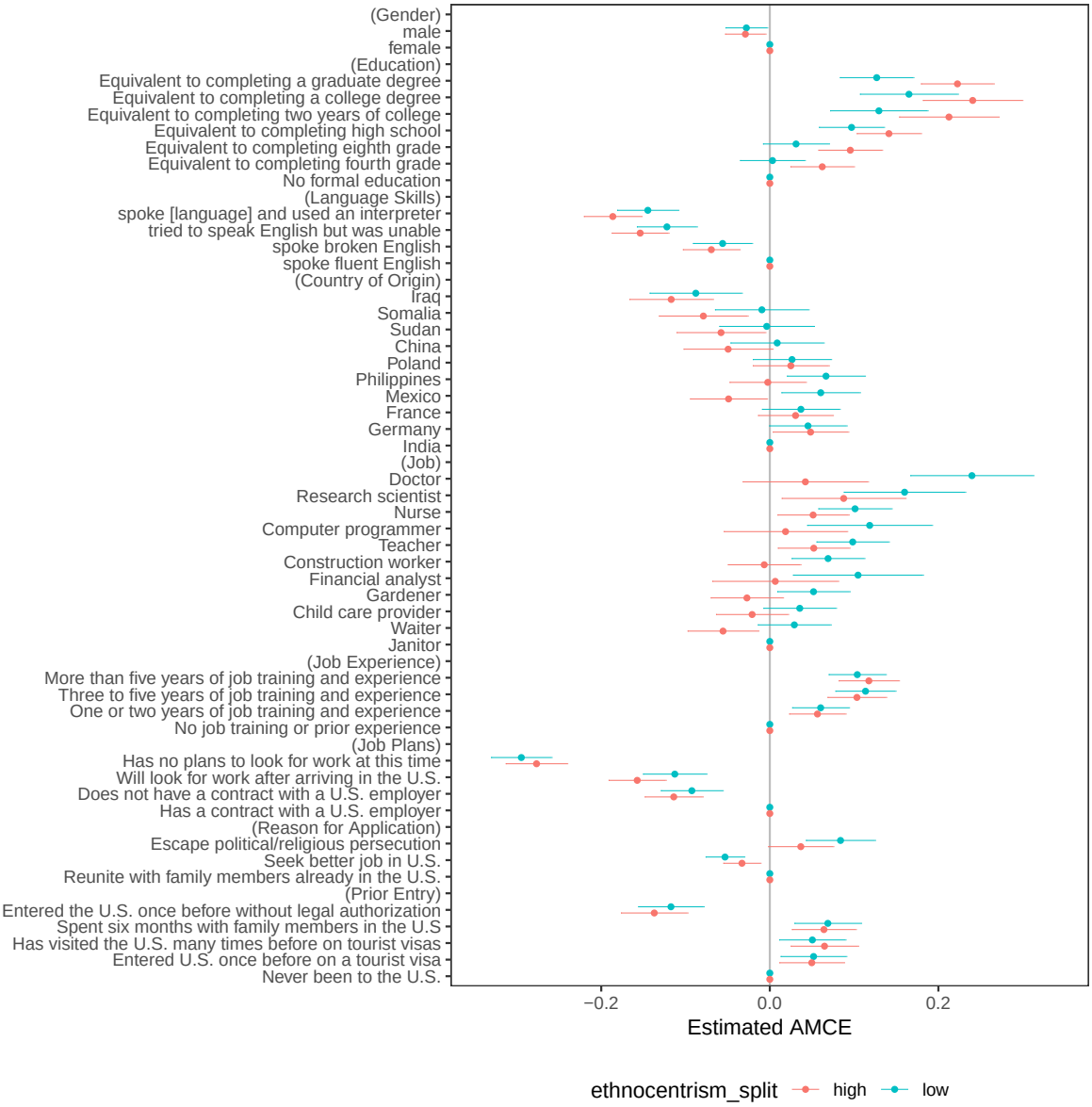
feature	level	estimate	std.error	z
Gender	female	0.00		
Gender	male	-0.03	0.01	-3.25
Education	No formal education	0.00		
Education	Equivalent to completing fourth grade	0.03	0.01	2.22
Education	Equivalent to completing eighth grade	0.06	0.01	3.86
Education	Equivalent to completing high school	0.12	0.01	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

B.2 Replication using MMs

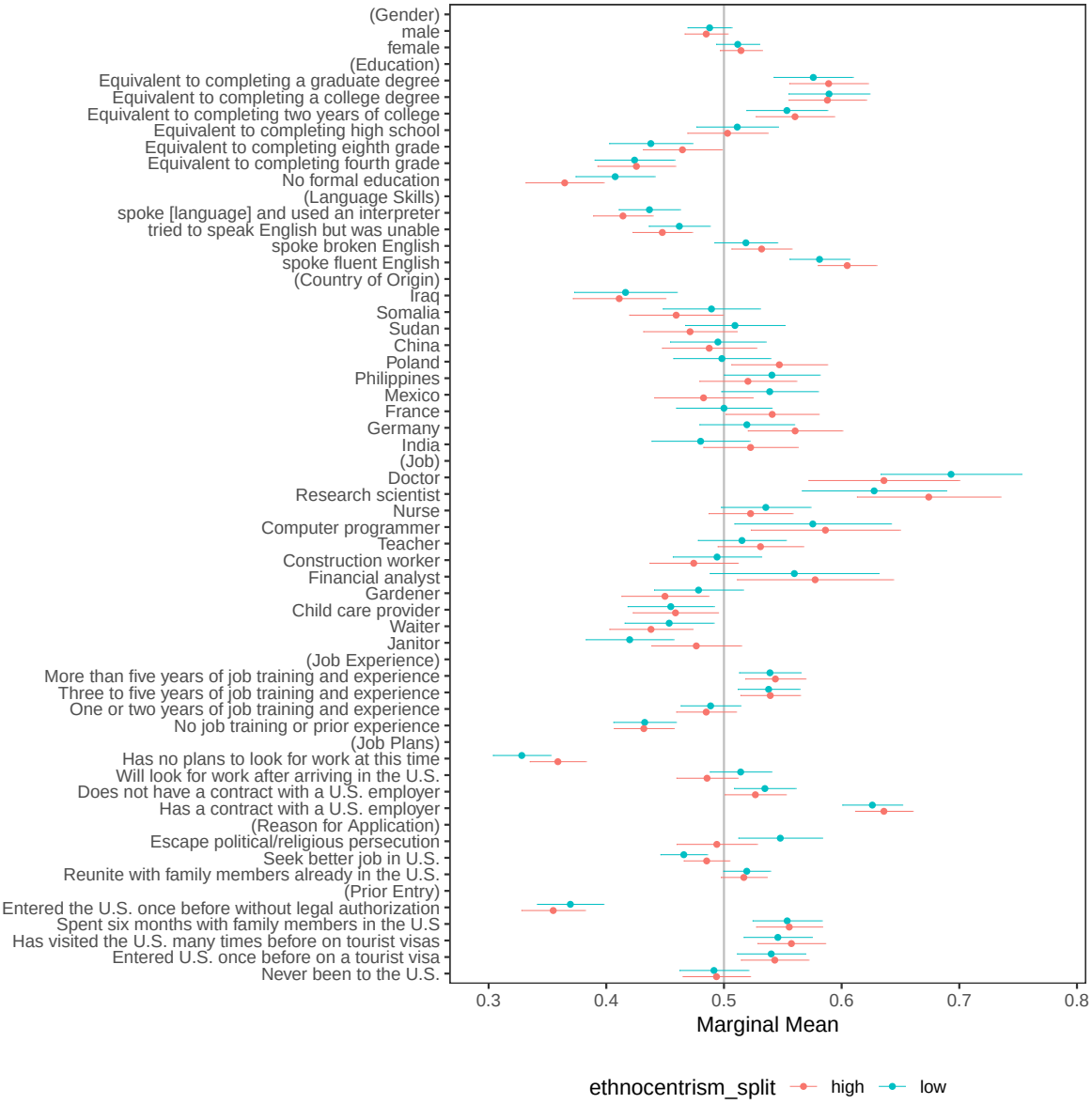


feature	level	estimate	std.error	z
Gender	female	0.51	0.01	1.99
Gender	male	0.49	0.01	-2.03
Education	No formal education	0.39	0.01	-10.04
Education	Equivalent to completing fourth grade	0.42	0.01	-7.08
Education	Equivalent to completing eighth grade	0.44	0.01	-5.00
Education	Equivalent to completing 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

B.3 Subgroup Analysis for Hainmueller et al. (2014) Immigration Experiment using AMCEs

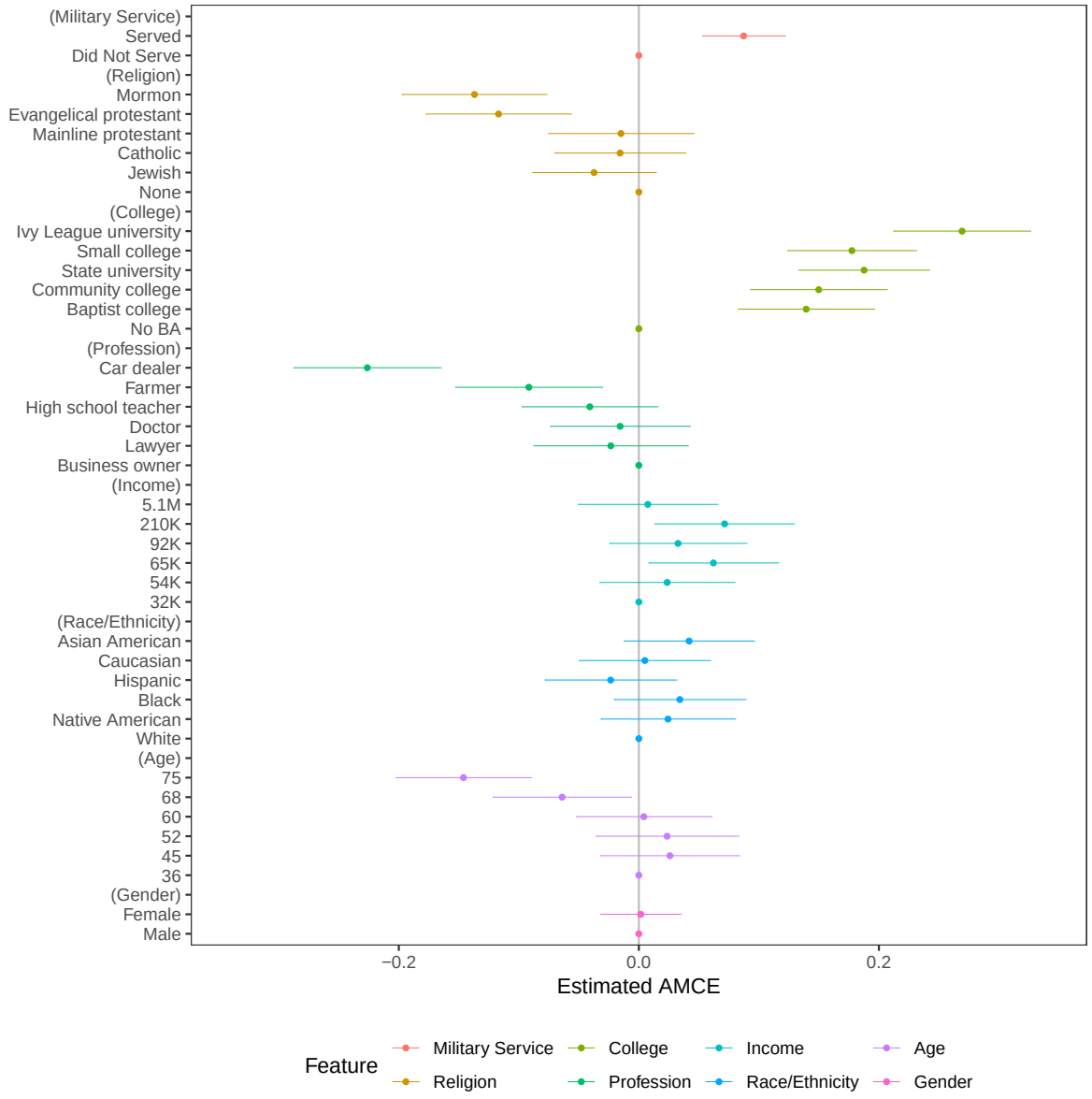


B.4 Subgroup Analysis for Hainmueller et al. (2014) Immigration Experiment using MMs



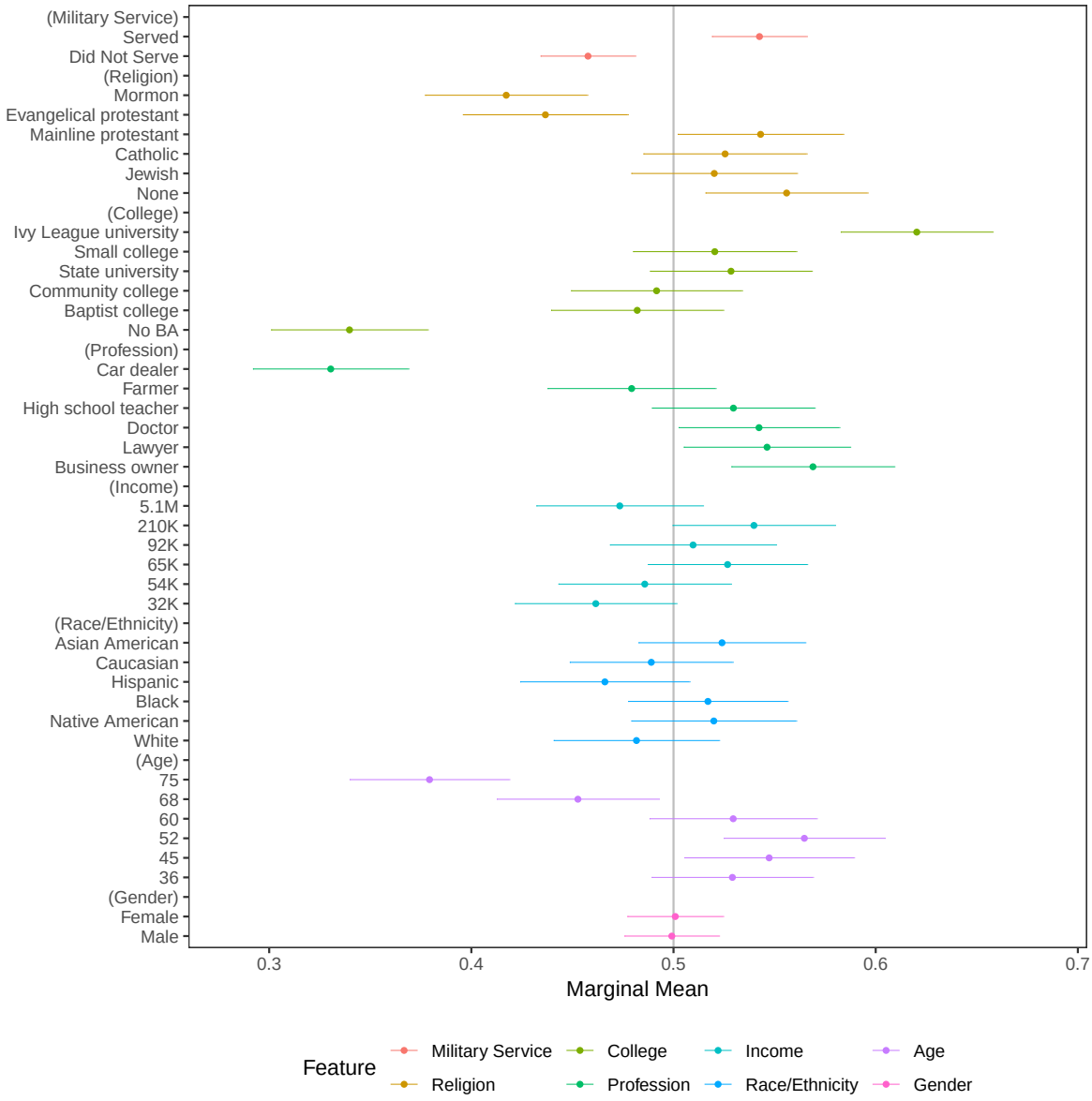
C Hainmueller et al. (2014) Candidate Experiment

C.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

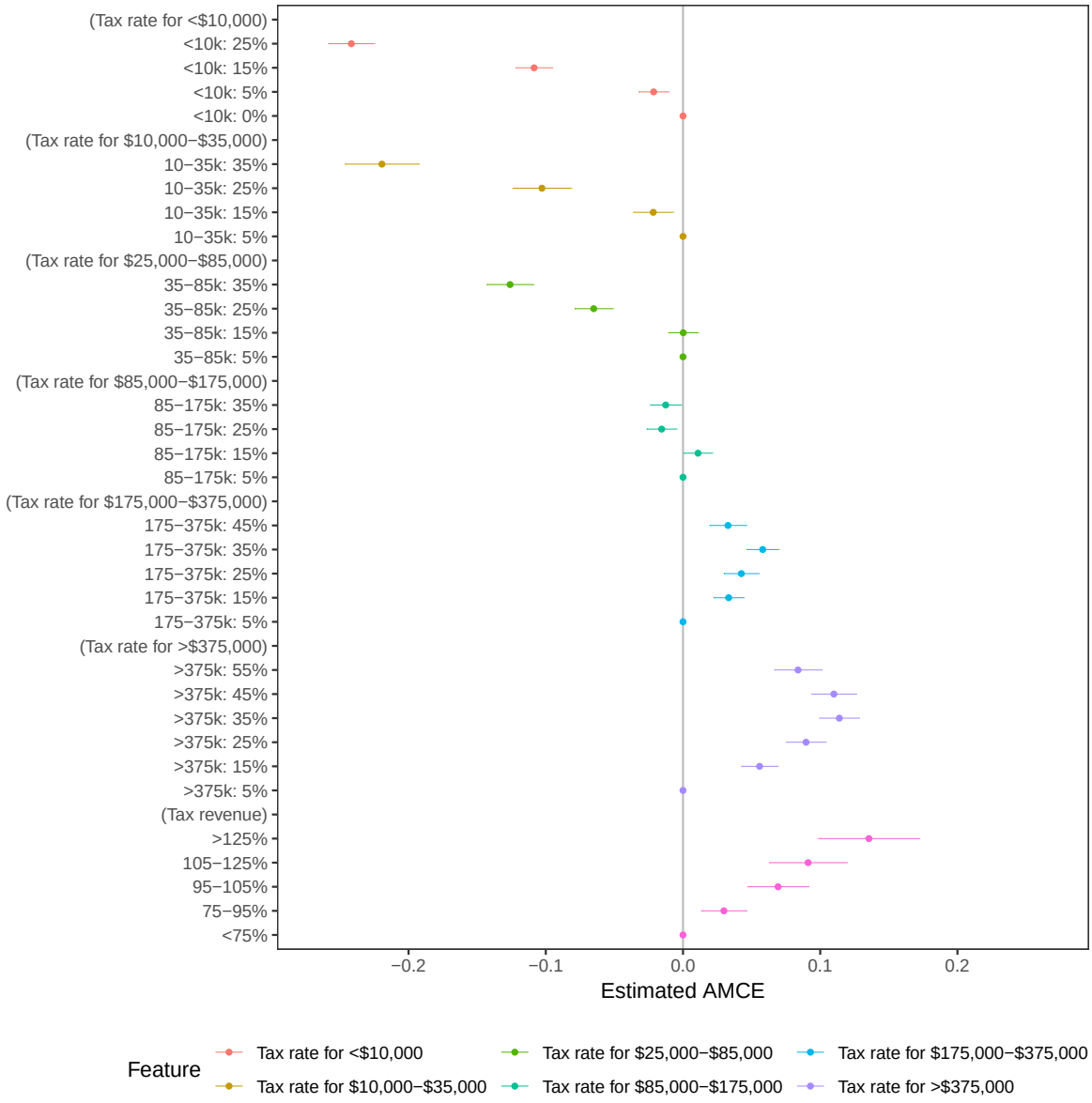
C.2 Replication using MMs



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

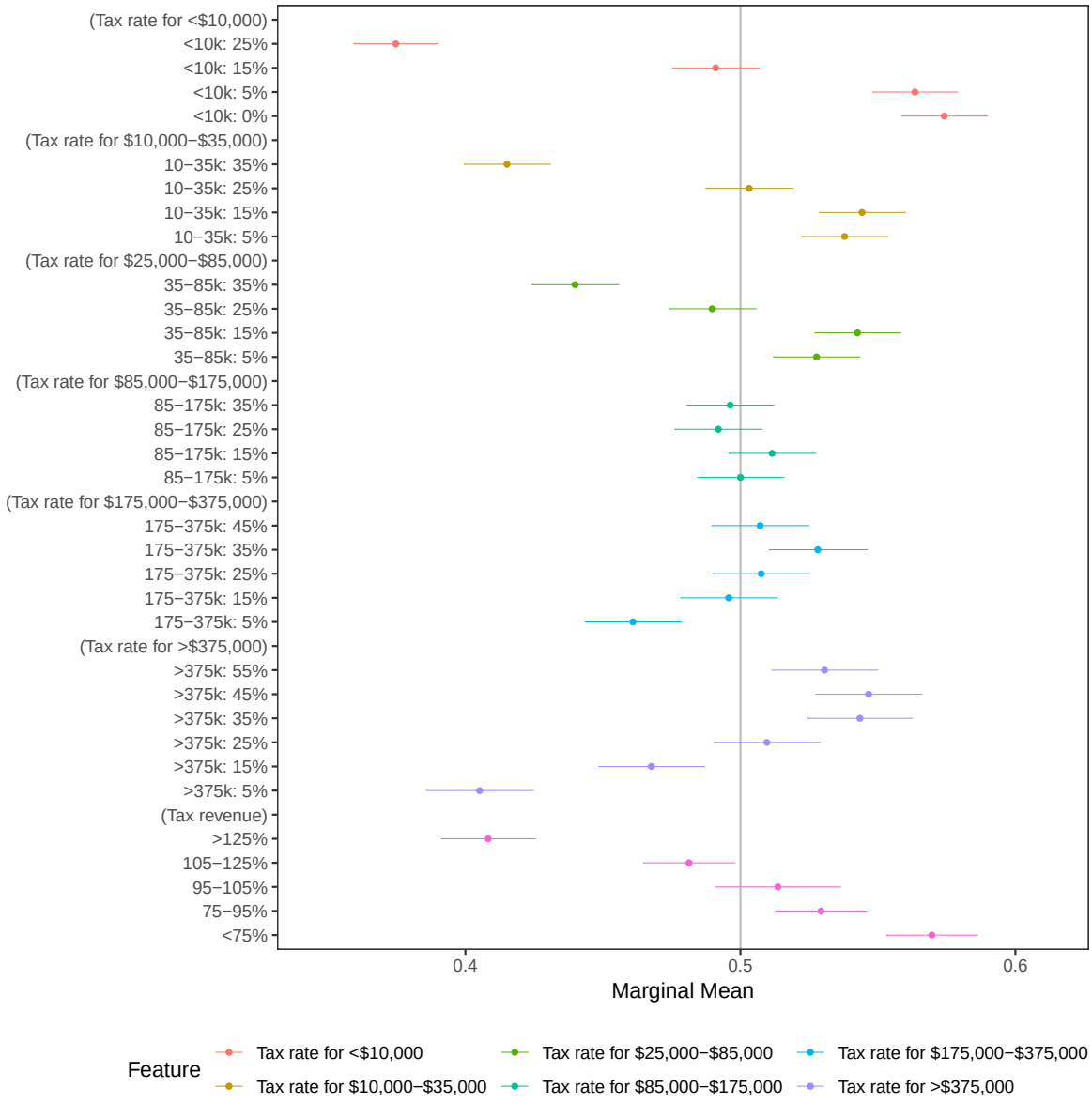
D Ballard-Rosa et al. (2016) Tax Preference Experiment

D.1 Replication using AMCEs



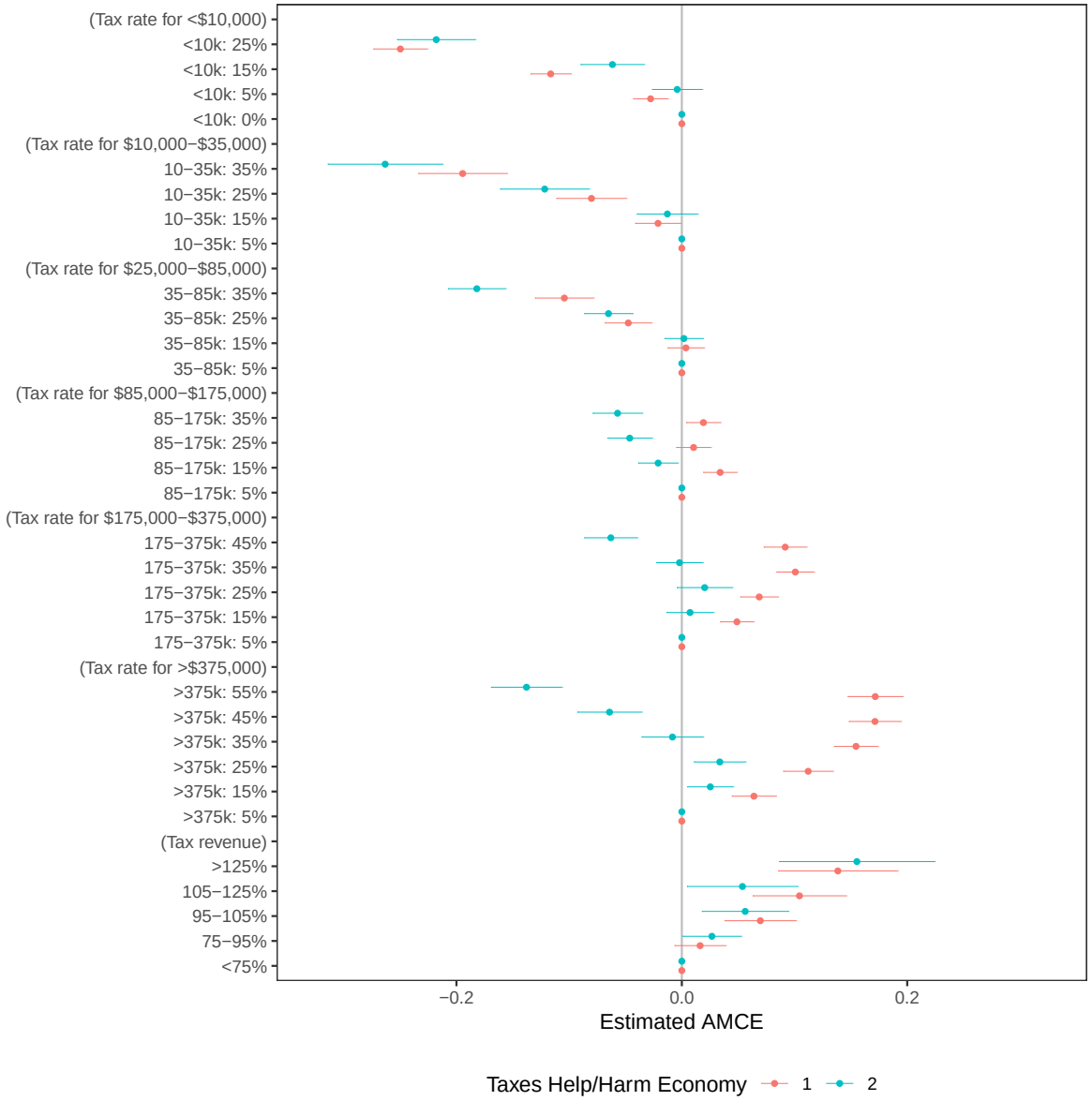
feature	level	estimate	std.error	z
Tax rate for <$10,000	<10k: 0%	0.00		
Tax rate for <$10,000	<10k: 5%	-0.02	0.01	-3.81
Tax rate for <$10,000	<10k: 15%	-0.11	0.01	-15.97
Tax rate for <$10,000	<10k: 25%	-0.24	0.01	-28.33
Tax rate for $10,000-$35,000	10-35k: 5%	0.00		
Tax rate for $10,000-$35,000	10-35k: 15%	-0.02	0.01	-2.94
Tax rate for $10,000-$35,000	10-35k: 25%	-0.10	0.01	-9.42
Tax rate for $10,000-$35,000	10-35k: 35%	-0.22	0.01	-15.96
Tax rate for $25,000-$85,000	35-85k: 5%	0.00		
Tax rate for $25,000-$85,000	35-85k: 15%	0.00	0.01	0.03
Tax rate for $25,000-$85,000	35-85k: 25%	-0.07	0.01	-9.18
Tax rate for $25,000-$85,000	35-85k: 35%	-0.13	0.01	-14.55
Tax rate for $85,000-$175,000	85-175k: 5%	0.00		
Tax rate for $85,000-$175,000	85-175k: 15%	0.01	0.01	2.06
Tax rate for $85,000-$175,000	85-175k: 25%	-0.02	0.01	-2.80
Tax rate for $85,000-$175,000	85-175k: 35%	-0.01	0.01	-2.19
Tax rate for $175,000-$375,000	175-375k: 5%	0.00		
Tax rate for $175,000-$375,000	175-375k: 15%	0.03	0.01	5.97
Tax rate for $175,000-$375,000	175-375k: 25%	0.04	0.01	6.53
Tax rate for $175,000-$375,000	175-375k: 35%	0.06	0.01	9.77
Tax rate for $175,000-$375,000	175-375k: 45%	0.03	0.01	4.80
Tax rate for >$375,000	>375k: 5%	0.00		
Tax rate for >$375,000	>375k: 15%	0.06	0.01	8.22
Tax rate for >$375,000	>375k: 25%	0.09	0.01	12.18
Tax rate for >$375,000	>375k: 35%	0.11	0.01	15.44
Tax rate for >$375,000	>375k: 45%	0.11	0.01	13.23
Tax rate for >$375,000	>375k: 55%	0.08	0.01	9.55
Tax revenue	<75%	0.00		
Tax revenue	75-95%	0.03	0.01	3.56
Tax revenue	95-105%	0.07	0.01	6.11
Tax revenue	105-125%	0.09	0.01	6.33
Tax revenue	>125%	0.14	0.02	7.22

D.2 Replication using MMs

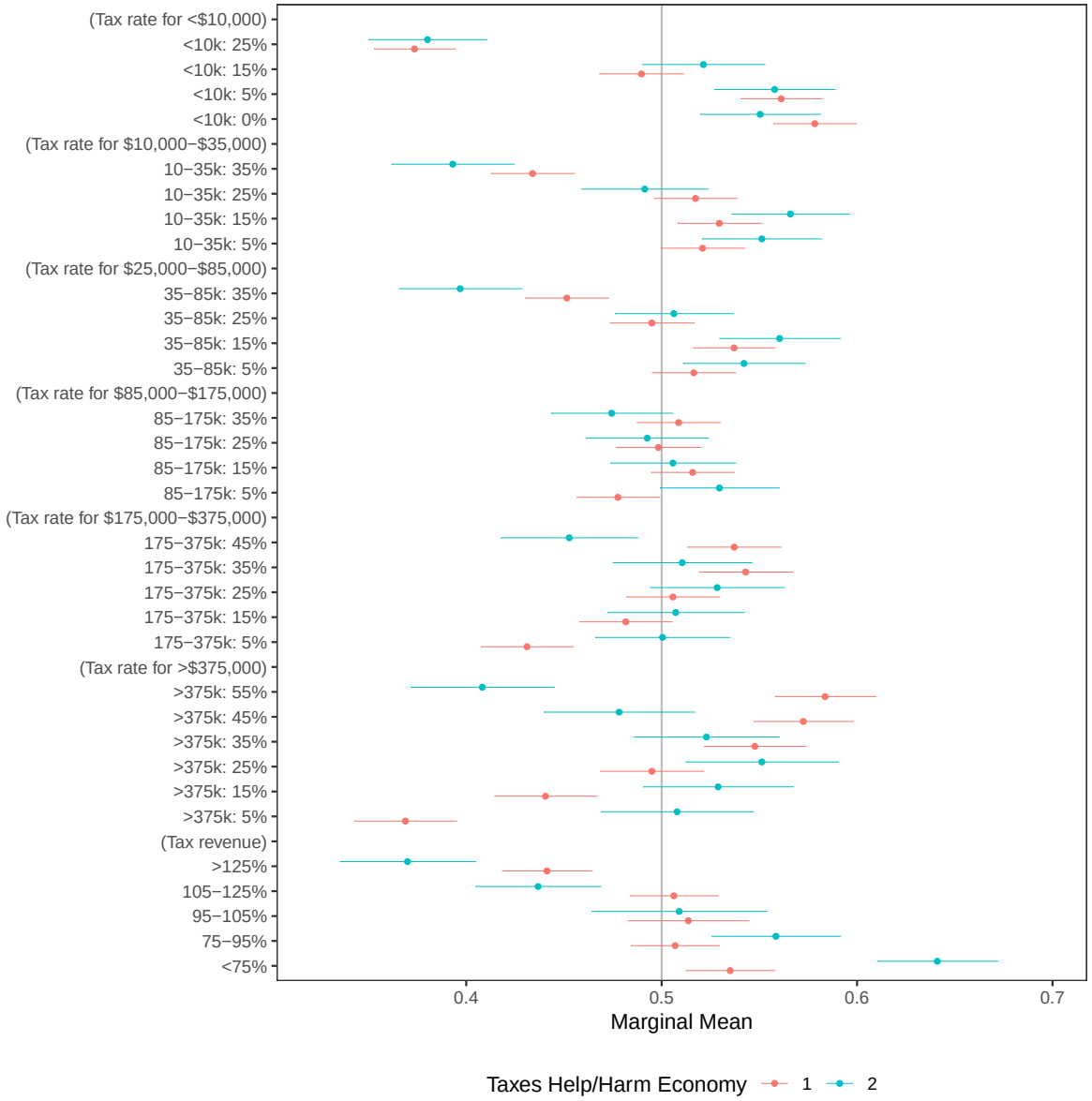


feature	level	estimate	std.error	z
Tax rate for <$10,000	<10k: 0%	0.57	0.01	9.37
Tax rate for <$10,000	<10k: 5%	0.56	0.01	8.06
Tax rate for <$10,000	<10k: 15%	0.49	0.01	-1.11
Tax rate for <$10,000	<10k: 25%	0.37	0.01	-16.07
Tax rate for $10,000-$35,000	10-35k: 5%	0.54	0.01	4.75
Tax rate for $10,000-$35,000	10-35k: 15%	0.54	0.01	5.55
Tax rate for $10,000-$35,000	10-35k: 25%	0.50	0.01	0.39
Tax rate for $10,000-$35,000	10-35k: 35%	0.42	0.01	-10.62
Tax rate for $25,000-$85,000	35-85k: 5%	0.53	0.01	3.45
Tax rate for $25,000-$85,000	35-85k: 15%	0.54	0.01	5.37
Tax rate for $25,000-$85,000	35-85k: 25%	0.49	0.01	-1.27
Tax rate for $25,000-$85,000	35-85k: 35%	0.44	0.01	-7.50
Tax rate for $85,000-$175,000	85-175k: 5%	0.50	0.01	0.00
Tax rate for $85,000-$175,000	85-175k: 15%	0.51	0.01	1.42
Tax rate for $85,000-$175,000	85-175k: 25%	0.49	0.01	-1.00
Tax rate for $85,000-$175,000	85-175k: 35%	0.50	0.01	-0.46
Tax rate for $175,000-$375,000	175-375k: 5%	0.46	0.01	-4.37
Tax rate for $175,000-$375,000	175-375k: 15%	0.50	0.01	-0.47
Tax rate for $175,000-$375,000	175-375k: 25%	0.51	0.01	0.84
Tax rate for $175,000-$375,000	175-375k: 35%	0.53	0.01	3.11
Tax rate for $175,000-$375,000	175-375k: 45%	0.51	0.01	0.80
Tax rate for >$375,000	>375k: 5%	0.41	0.01	-9.57
Tax rate for >$375,000	>375k: 15%	0.47	0.01	-3.30
Tax rate for >$375,000	>375k: 25%	0.51	0.01	0.97
Tax rate for >$375,000	>375k: 35%	0.54	0.01	4.48
Tax rate for >$375,000	>375k: 45%	0.55	0.01	4.75
Tax rate for >$375,000	>375k: 55%	0.53	0.01	3.12
Tax revenue	<75%	0.57	0.01	8.22
Tax revenue	75-95%	0.53	0.01	3.47
Tax revenue	95-105%	0.51	0.01	1.18
Tax revenue	105-125%	0.48	0.01	-2.21
Tax revenue	>125%	0.41	0.01	-10.54

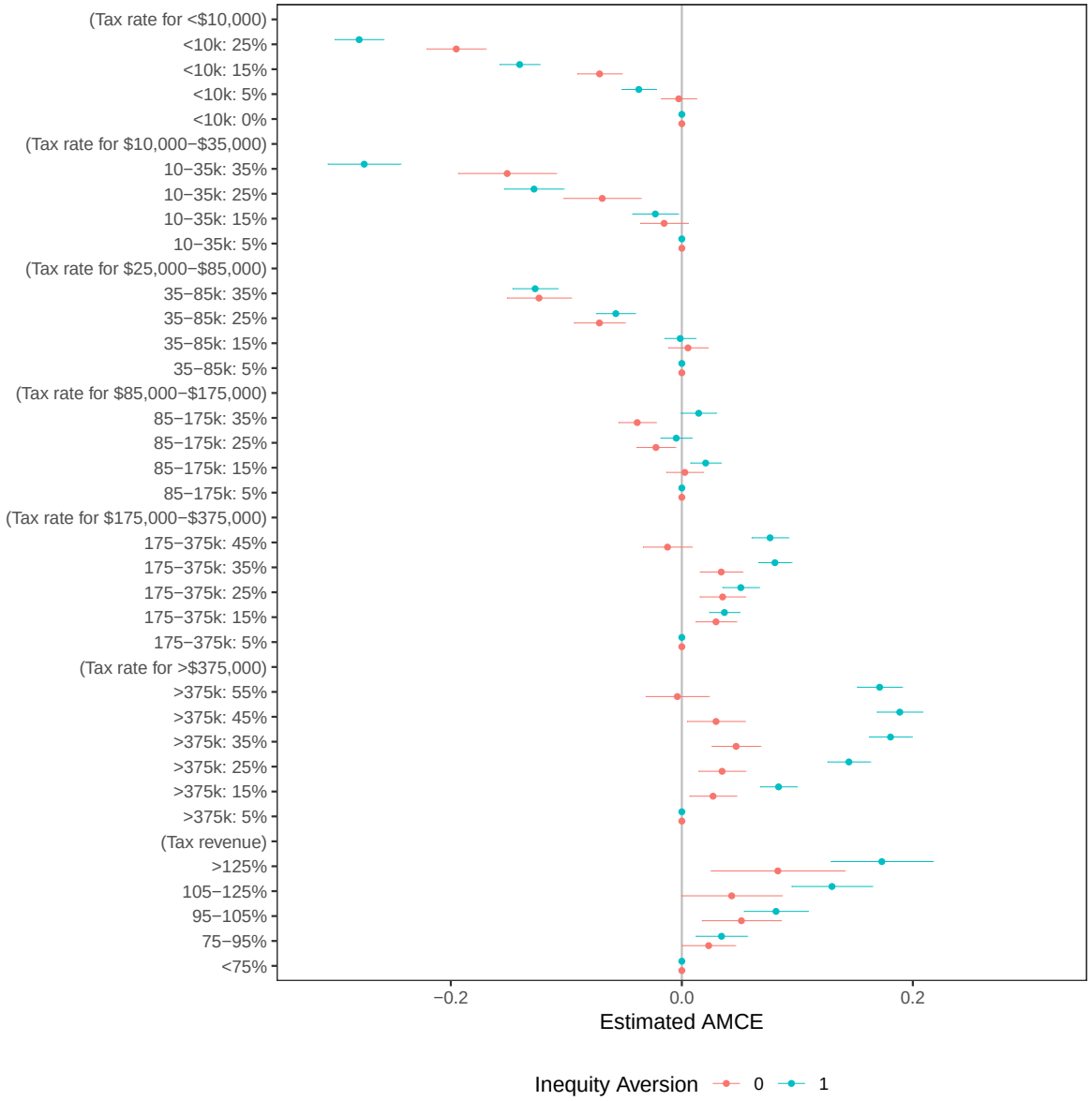
D.3 Subgroup Analysis for Ballard-Rosa et al. (2016), by "Taxes Harm Economy" Split using AMCEs



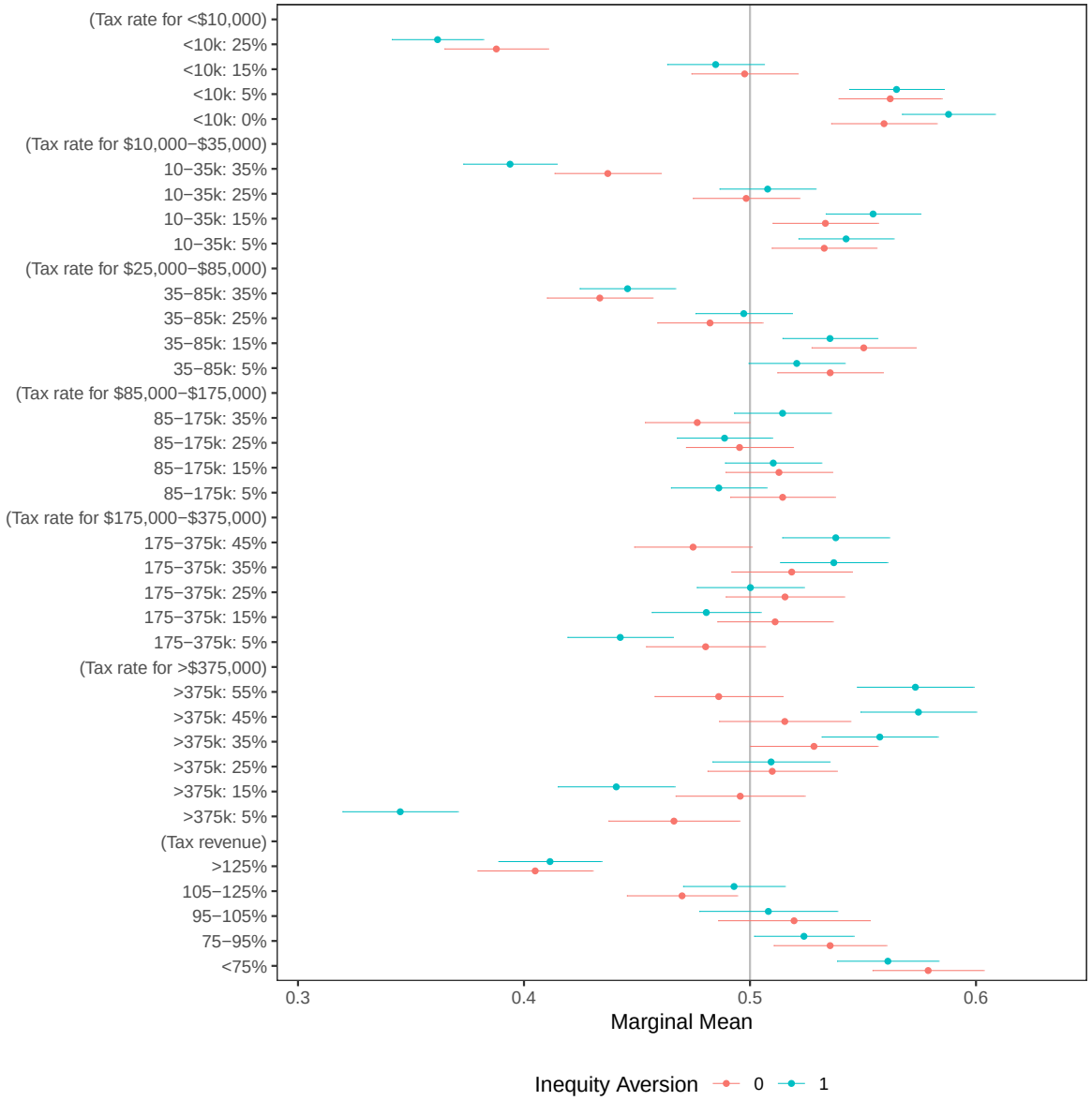
D.4 Subgroup Analysis for Ballard-Rosa et al. (2016), by "Taxes Harm Economy" Split using MMs



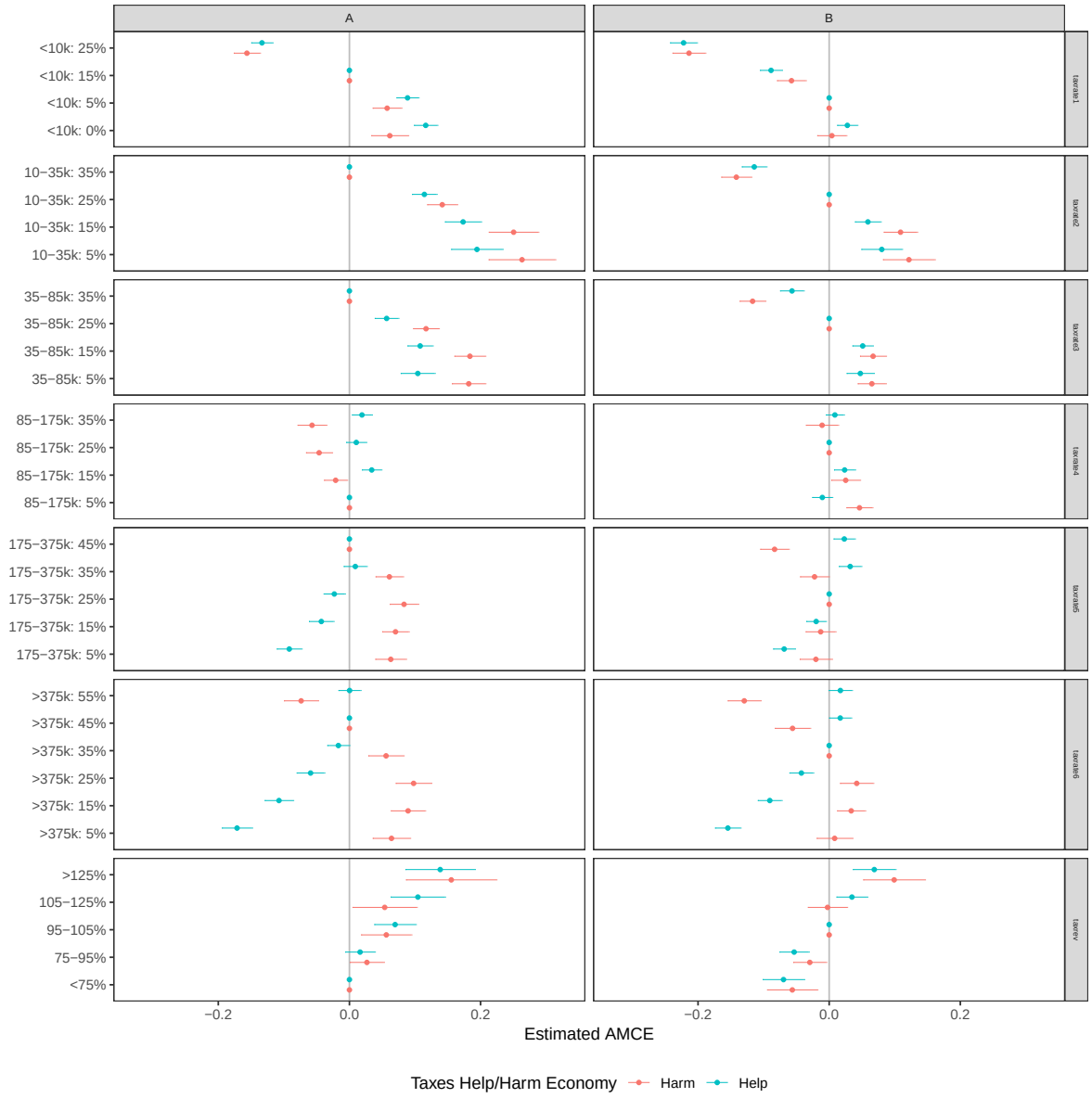
D.5 Subgroup Analysis for Ballard-Rosa et al. (2016), by Inequity Aversion using AMCEs



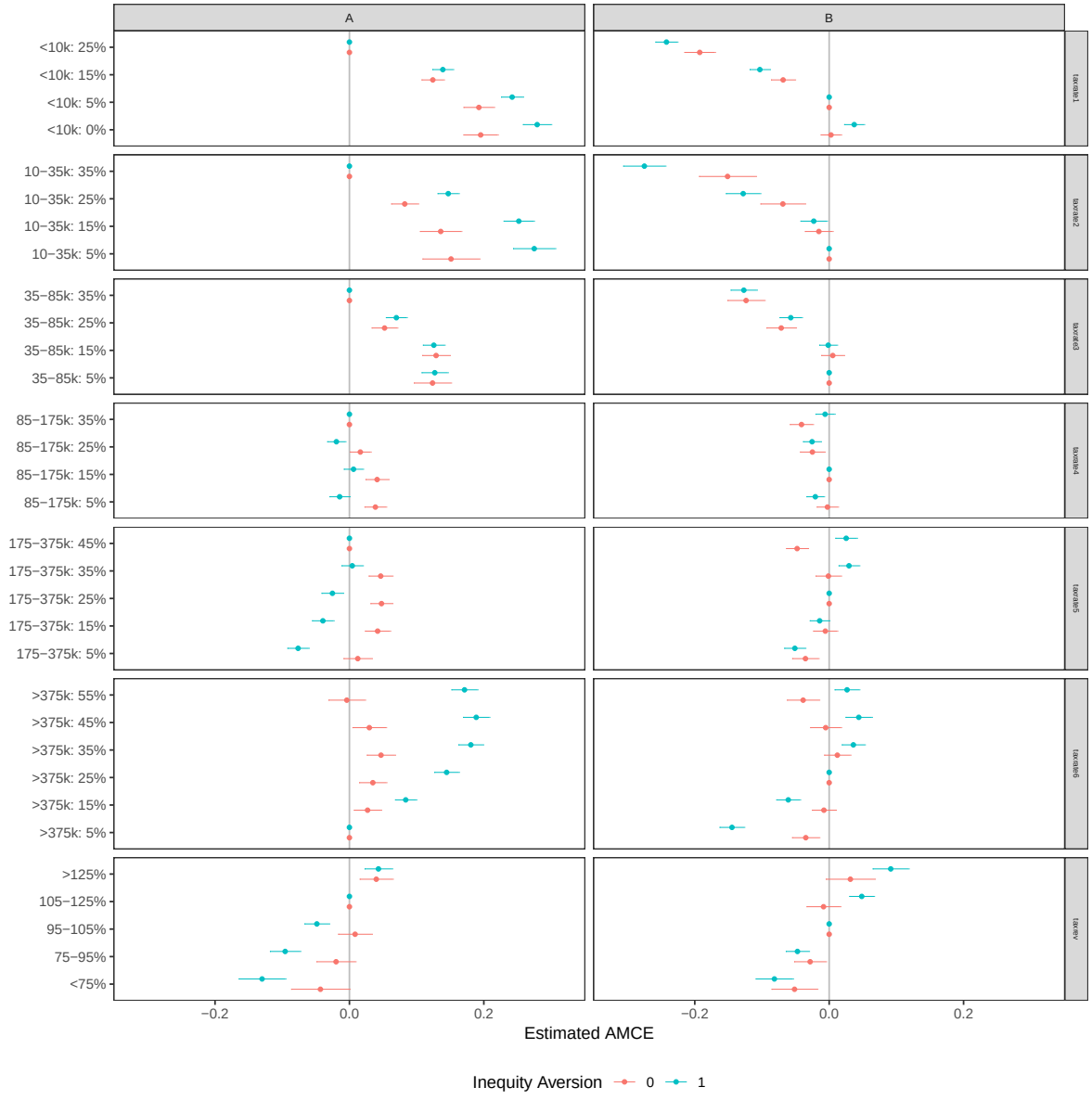
D.6 Subgroup Analysis for Ballard-Rosa et al. (2016), by Inequity Aversion using MMs



D.7 Comparison of Alternative Reference Categories for Ballard-Rosa et al. (2016) Tax Preference Experiment, by "Taxes Harm Economy" Split

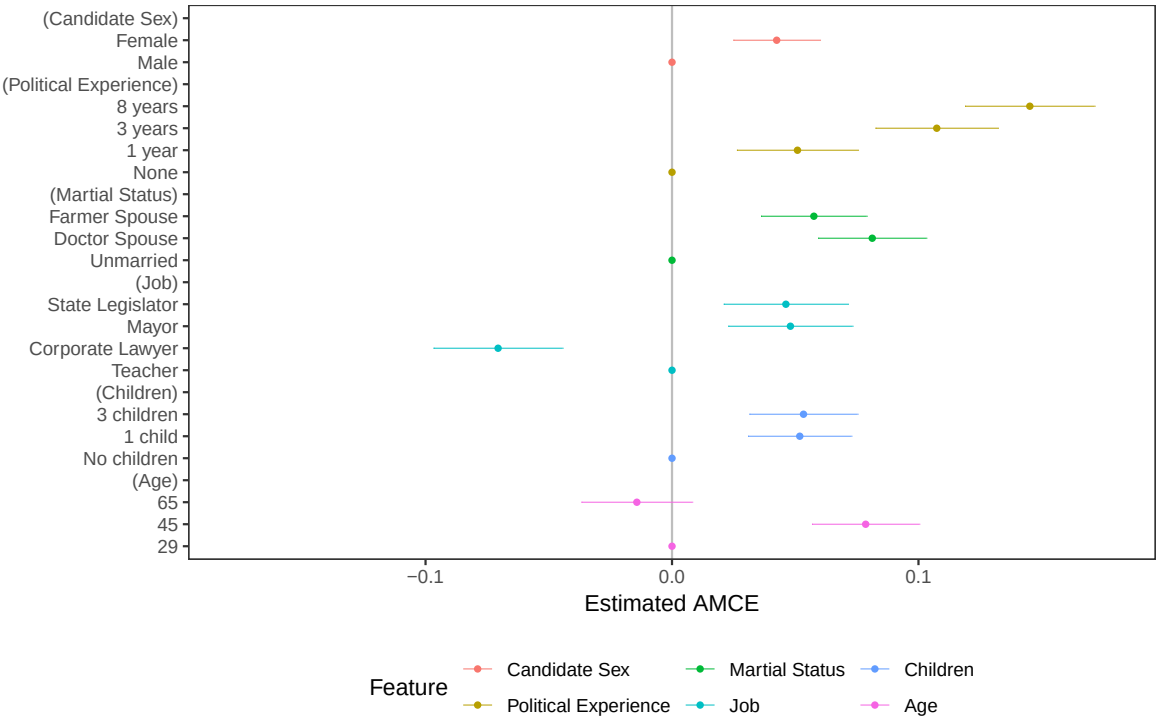


D.8 Comparison of Alternative Reference Categories for Ballard-Rosa et al. (2016) Tax Preference Experiment, by Inequity Aversion



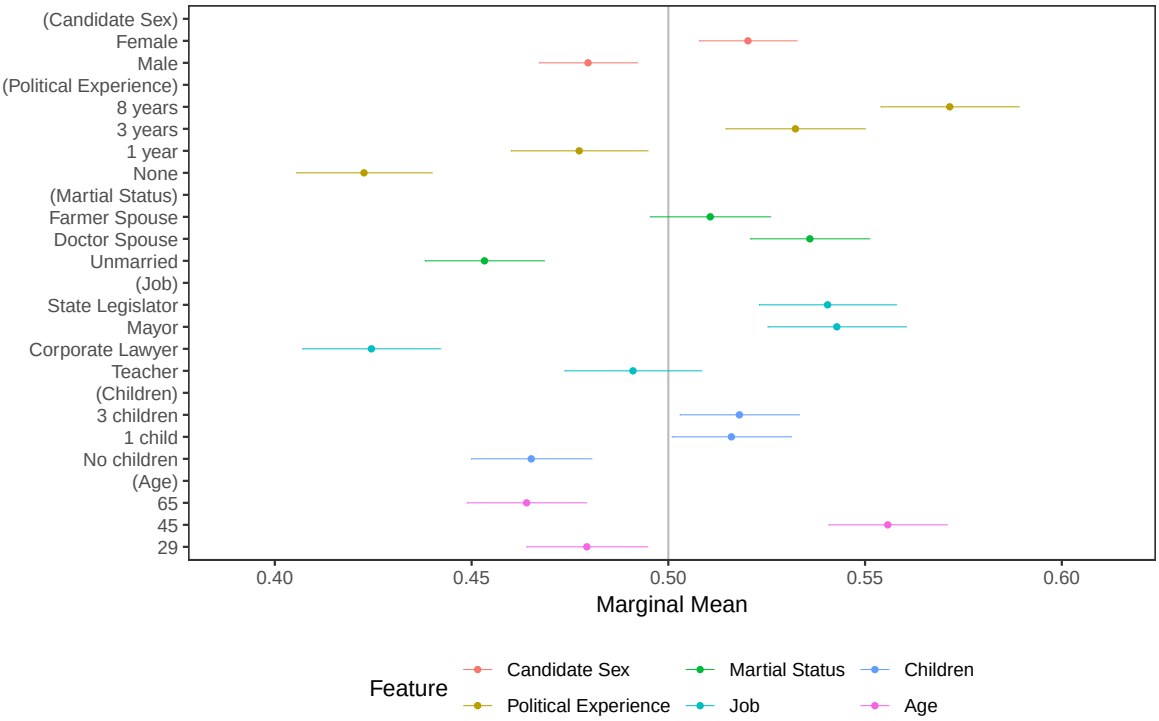
E Teele et al. (2018) Candidate Experiment

E.1 Replication using AMCEs



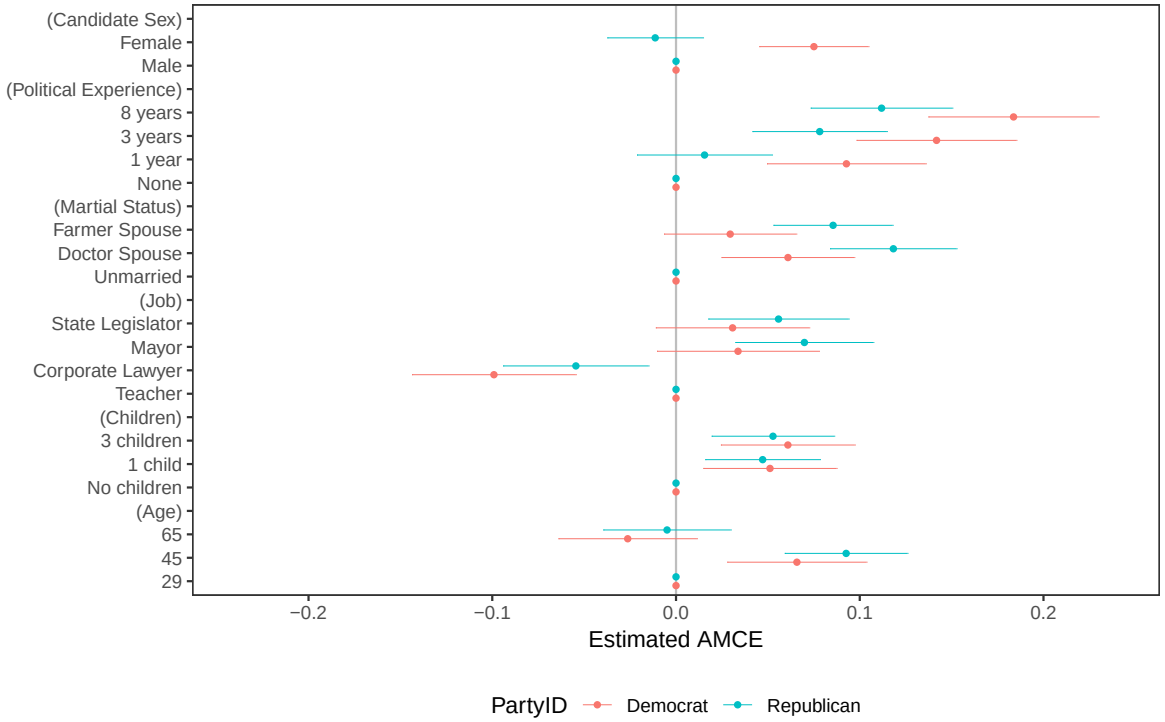
feature	level	estimate	std.error	z
Candidate Sex	Male	0.00		
Candidate Sex	Female	0.04	0.01	4.74
Political Experience	None	0.00		
Political Experience	1 year	0.05	0.01	4.06
Political Experience	3 years	0.11	0.01	8.47
Political Experience	8 years	0.15	0.01	10.83
Martial Status	Unmarried	0.00		
Martial 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

E.2 Replication using MMs

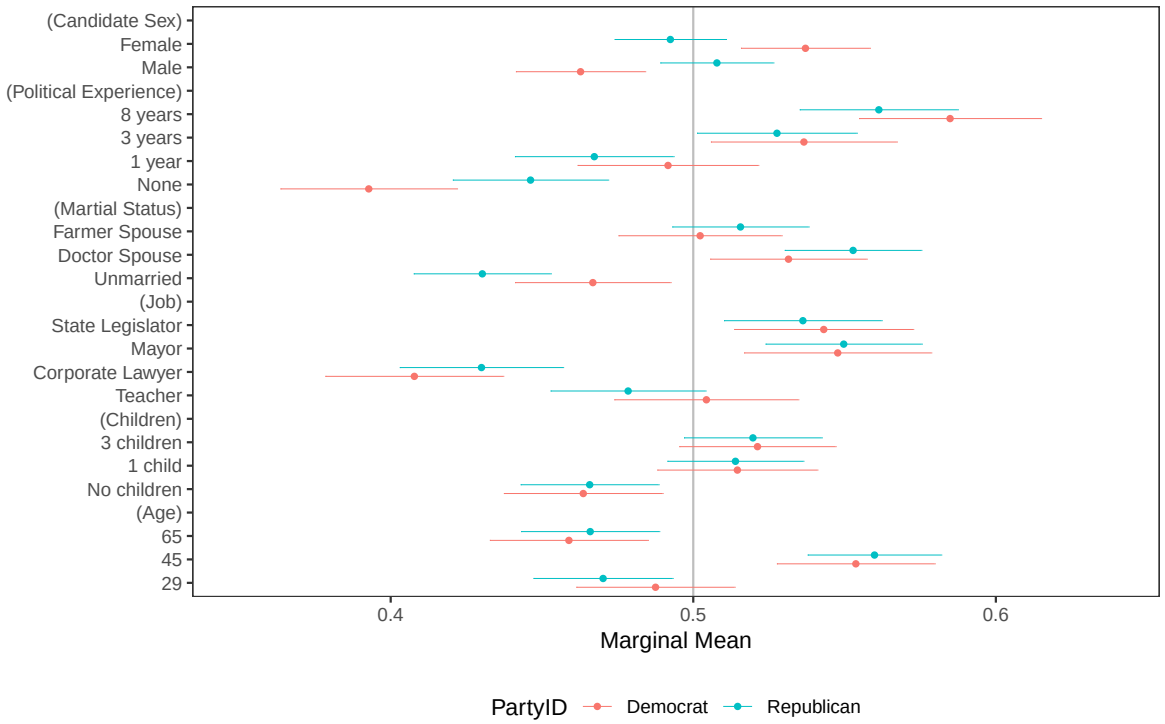


feature	level	estimate	std.error	z
Candidate Sex	Male	0.48	0.01	-3.22
Candidate Sex	Female	0.52	0.01	3.20
Political Experience	None	0.42	0.01	-8.81
Political Experience	1 year	0.48	0.01	-2.56
Political Experience	3 years	0.53	0.01	3.58
Political Experience	8 years	0.57	0.01	7.99
Martial Status	Unmarried	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

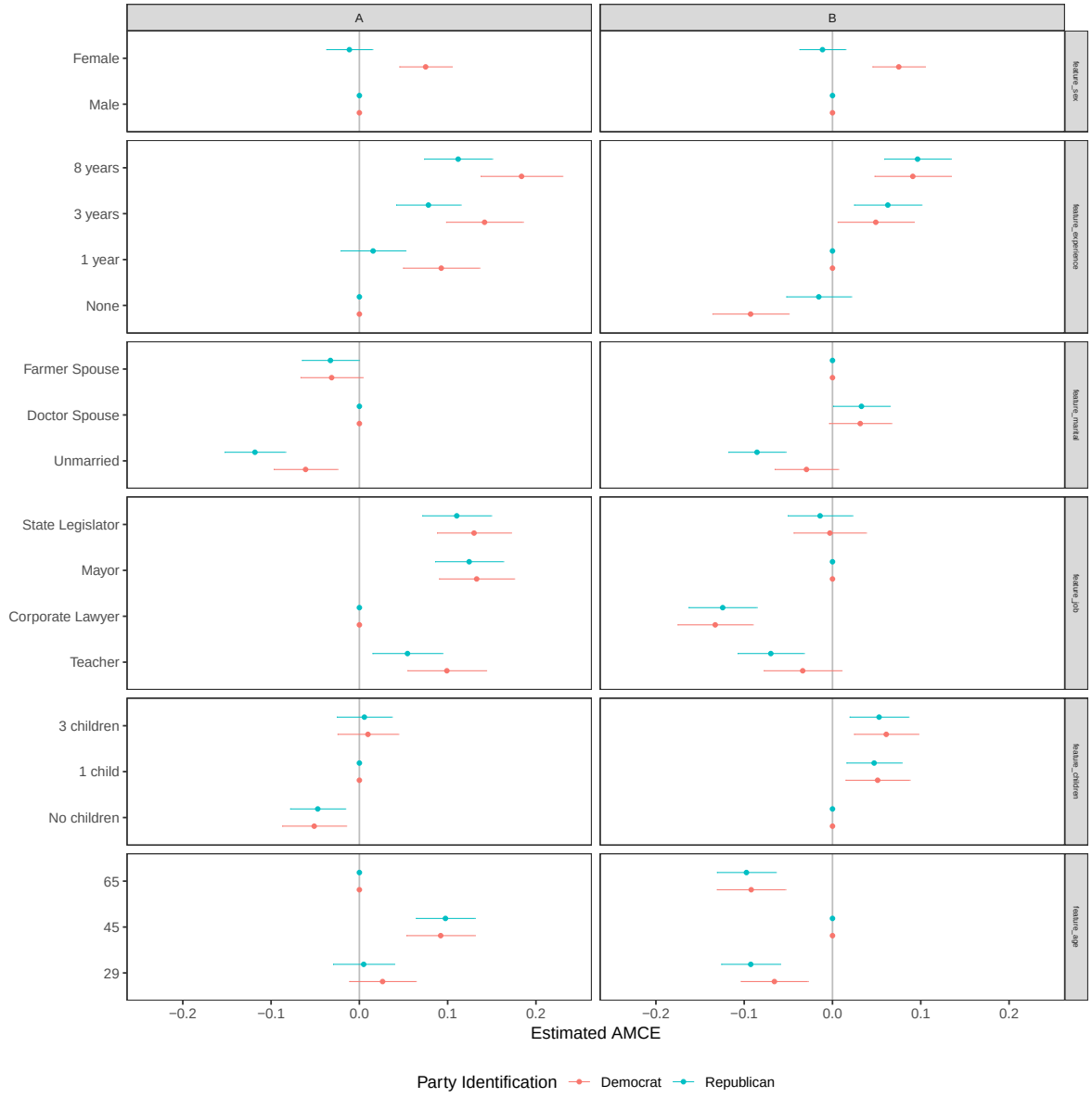
E.3 Subgroup Analysis for Teele et al. (2018) Candidate Experiment using AMCEs



E.4 Subgroup Analysis for Teele et al. (2018) Candidate Experiment using MMs



E.5 Comparison of Alternative Reference Categories for Teele et al. (2018) Candidate Experiment



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

```
## R version 3.5.1 (2018-07-02)
## 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] ggplot2_3.1.0 cregg_0.3.0   rio_0.5.16
##
## loaded via a namespace (and not attached):
##  [1] zip_1.0.0         Rcpp_1.0.0        cellranger_1.1.0
##  [4] pillar_1.3.0.9000 compiler_3.5.1    plyr_1.8.4
##  [7] bindr_0.1.1       forcats_0.3.0     tools_3.5.1
## [10] digest_0.6.18     lattice_0.20-38   ggstance_0.3.1
## [13] evaluate_0.12     tibble_1.4.2      gtable_0.2.0
## [16] pkgconfig_2.0.2   rlang_0.3.0.1     Matrix_1.2-15
## [19] openxlsx_4.1.0    curl_3.2          haven_2.0.0
## [22] bindrcpp_0.2.2    withr_2.1.2       stringr_1.3.1
## [25] dplyr_0.7.8       knitr_1.20        hms_0.4.2
## [28] lmtest_0.9-36     tidyselect_0.2.5  grid_3.5.1
## [31] glue_1.3.0        data.table_1.11.8 R6_2.3.0
## [34] survival_2.43-3   readxl_1.1.0      foreign_0.8-71
## [37] purrr_0.2.5       magrittr_1.5      splines_3.5.1
## [40] scales_1.0.0      assertthat_0.2.0  xtable_1.8-3
## [43] colorspace_1.3-2  sandwich_2.5-0    survey_3.34
## [46] stringi_1.2.4     lazyeval_0.2.1    munsell_0.5.0
## [49] crayon_1.3.4      zoo_1.8-4
```