Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution*

Brandon de la Cuesta† Naoki Egami‡ Kosuke Imai§

Forthcoming in *Political Analysis*

Abstract

Conjoint analysis has become popular among social scientists for measuring multidimensional preferences. When analyzing such experiments, researchers often focus on the average marginal component effect (AMCE), which represents the causal effect of a single profile attribute while averaging over the remaining attributes. What has been overlooked, however, is the fact that the AMCE critically relies upon the distribution of the other attributes used for the averaging. Although most experiments employ the uniform distribution, which equally weights each profile, both the actual distribution of profiles in the real world and the distribution of theoretical interest are often far from uniform. This mismatch can severely compromise the external validity of conjoint analysis. We empirically demonstrate that estimates of the AMCE can be substantially different when averaging over the target profile distribution instead of uniform. We propose new experimental designs and estimation methods that incorporate substantive knowledge about the profile distribution. We illustrate our methodology through two empirical applications, one using a real-world distribution and the other based on a counterfactual distribution motivated by a theoretical consideration. The proposed methodology is implemented through an open-source software package.

*The proposed methodology is implemented via an open-source software `R` package `factorEx`, available through the Comprehensive R Archive Network (`https://cran.r-project.org/package=factorEx`). We thank Jens Hainmueller, Dan Hopkins, Dean Knox, Shiro Kuriwaki, Thomas Leavitt, Erik Peterson, and Teppei Yamamoto for helpful comments and conversations. The replication materials are available as de la Cuesta *et al.* (2020).

†Postdoctoral Research Fellow, King Center on Global Development, Stanford University, Palo Alto CA 94305. Email: brandon.delacuesta@stanford.edu, URL: `https://brandondelacuesta.com`

‡Assistant Professor, Department of Political Science, Columbia University, New York NY 10027. Email: ne2312@columbia.edu, URL: `https://naokiegami.com`

§Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu, URL: `https://imai.fas.harvard.edu`

1 Introduction

Conjoint analysis is a factorial survey experiment that is designed to measure multidimensional preferences. In a typical application, respondents are presented with a pair of hypothetical profiles whose attributes are randomly selected, and are then asked to choose their preferred profile. Examples of such profiles include political candidates (e.g., Teele *et al.*, 2018), immigrants (e.g., Hainmueller and Hopkins, 2015), and public policies (e.g., Ballard-Rosa *et al.*, 2017). Although it has been extensively used in marketing research (e.g., Green *et al.*, 2001; Marshall and Bradlow, 2002), conjoint analysis has quickly gained popularity in political science due to its wide applicability and relative simplicity (Hainmueller *et al.*, 2014). Indeed, as shown in Figure 1, the number of major political science journal articles that utilize conjoint analysis has increased dramatically over the last five years.

The most commonly used quantity of interest in conjoint analysis is the average marginal component effect (AMCE), which represents the causal effect of changing one attribute of a profile while averaging over the distribution of the remaining profile attributes (Hainmueller *et al.*, 2014). Because conjoint analysis often involves many attributes, averaging over their distribution makes the interpretation of causal effects simpler and more practical than conditioning on their specific values. For example, a researcher may be interested in the AMCE of candidate's gender that averages over the distribution of other candidate characteristics such as age, education, race, and policy positions. Thus, the definition of the AMCE critically depends on the distribution used to average over profile attributes.

Unfortunately, while this point is theoretically understood, in practice little attention has been paid to the choice of this distribution. As Figure 1 demonstrates, nearly 90% of the existing conjoint analyses use the uniform distribution. The problem is that the resulting estimate of the AMCE, which we call the *uniform* AMCE (uAMCE), gives equal weights to all conjoint profiles even when some of them are unrealistic from a substantive point of view. Ignoring the distribution of profiles fundamentally contradicts the key promise of conjoint analysis that the provision of information about several profile attributes makes the choice task realistic for respondents (Hainmueller *et al.*, 2015). In fact, if other attributes do not systematically affect respondents' evaluation of the main attribute of interest, then one could simply elicit preferences over each attribute separately, making a conjoint experiment unnecessary. Therefore, conjoint analysis is beneficial precisely when we expect multiple attributes to jointly affect human decision making, and this is also the exact setting where the choice of profile distribution

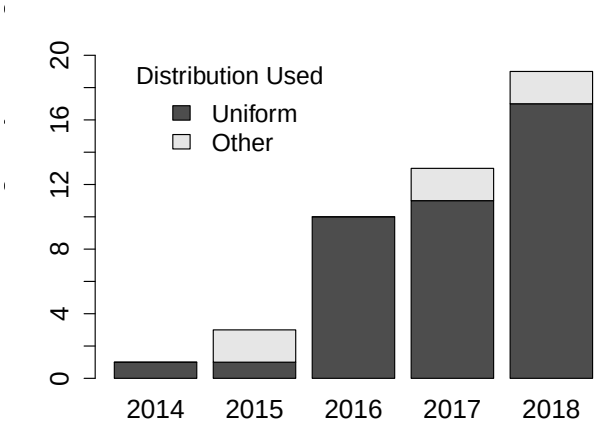

Figure 1: Recent Growth of Conjoint Analysis and Use of the Uniform Distribution for Randomization in Political Science Journal Articles. Darker (lighter) fill represents the proportion of articles in which all the factors are randomized with the uniform distribution. 88% of all reviewed articles use the uniform distribution. The plot is based on a review of articles published in political science journals from 2014 to 2018. See Appendix A for the information about how the review was conducted.

affects estimates of the AMCE the most.

In this paper, we study how the choice of profile distribution affects the conclusions of conjoint analysis. We define the *population* AMCE (pAMCE), which averages over the distribution of profile attributes in a target population of interest. Unlike the uAMCE, which is based on the uniform distribution, the pAMCE accounts for the relative frequency with which each profile occurs in the target population. This target profile distribution should be chosen according to substantive interests of each study, similar to the choice of a target population of respondents in traditional survey sampling. The choice of distribution may be based on (1) real-world data, such as the characteristics and policy positions of actual politicians, or (2) a counterfactual distribution of theoretical interest. For each of the two scenarios, we provide empirical applications. We show that the difference between the uAMCE and pAMCE is large when the target profile distribution differs from uniform and when there exists interaction between the main attribute of interest and other attributes.

We propose two new strategies to estimate the pAMCE. The first approach, which we call *design-based confirmatory analysis*, incorporates the target profile distribution in the design stage (Section 4.1). We introduce three experimental designs that differ in terms of data requirements and necessary assumptions. In the most natural design, which we term joint population randomization, we propose randomizing conjoint profiles according to their target profile distribution rather than the uniform. We then use a nonparametric estimator of the pAMCE, which can be computed using a weighted linear regression. This is a straightforward

generalization of a widely-used regression estimator (Hainmueller *et al.*, 2014).

Our second approach, *model-based exploratory analysis*, takes into account the target profile distribution at the analysis stage, after randomizing profiles and collecting data (Section 4.2). This approach is useful in estimating the pAMCE when researchers have to randomize profiles based on distributions different from the target profile distribution, such as uniform. We propose fitting a flexible two-way interaction model and estimating the pAMCE as a weighted average of coefficients. Although this approach yields less precise estimates than the design-based confirmatory analysis, we discuss how to use regularization methods to partially recoup the loss of statistical efficiency (Egami and Imai, 2019).

One potential challenge of incorporating the target profile distribution is that the joint distribution of all attributes is difficult to obtain in some applications. For example, in a conjoint experiment of immigrant profiles, it may not be feasible to obtain the joint distribution of the (potentially many) attributes of immigrants that researchers wish to study. Recognizing this practical data constraint, we propose the marginal population randomization design, which only requires the knowledge of each factor's marginal distribution. Here, researchers randomize each factor independently with its marginal distribution. While this design requires a stronger assumption of no three-way or higher-order interactions, we provide a method to test this assumption empirically. We also discuss how researchers can combine marginal distributions and partial joint distributions among several factors to relax this assumption.

The concern for unrealistic profiles is not new. In fact, researchers often remove a set of unusual profile combinations (e.g., doctors without college degree). Unfortunately, avoiding extreme cases is not sufficient for estimating the pAMCE. While some have begun to use unequal probabilities when randomizing profiles to partially address this concern (e.g., Huff and Kertzer, 2018; Hainmueller *et al.*, 2015; Leeper and Robison, 2018),[1] an overwhelming majority of researchers still use the uniform distribution without theoretically motivating it.[2] The substantive implication of this choice is that the resulting estimates of the AMCE are externally valid only when there is no interaction between attributes or when the uniform is the theoretically relevant profile distribution. Even though scholars have clearly discussed the importance

[1] See also Barnes *et al.* (2019) who point out that traditional conjoint experiments fail to generate realistic budget tradeoffs when studying public attitudes towards government spending.

[2] Less than 4% of existing conjoint studies theoretically motivate distributions used for randomization in the article's main text. See Appendix A, for additional information and a description of how these values were calculated.

of distributions used to randomize profiles (Hainmueller *et al.*, 2014),[3] there currently exists no systematic way to incorporate the target profile distribution into the estimation of the AMCE. The proposed methodology directly addresses this problem by developing new experimental designs and estimation strategies. We note that our focus is on the external validity of conjoint *profiles*, and is distinct from another important issue of representativeness of *respondents* in survey experiments (see e.g., Mutz, 2011; Mullinix *et al.*, 2015; Coppock *et al.*, 2018; Miratrix *et al.*, 2018).

We illustrate the proposed methodology using two empirical applications. First, we reanalyze a conjoint experiment of political candidates by Ono and Burden (2019). The primary goal of the study is to estimate the effect of a candidate's gender on voter choice. The original study estimates the uAMCE of being female and finds that women candidates face discrimination in presidential but not in congressional elections. Specifying the target profile distribution to be the 115th U.S. Congress, we estimate the pAMCEs separately for Republican and Democratic legislators. We show that the null effect of gender found in the original analysis for Congressional candidates is due to the large number of unrealistic profiles produced by the uniform distribution. Once we average profiles according to their real-world distributions, we recover a different result: women face a disadvantage when they run for Congress as Republicans but have an advantage when they run as Democrats. We also demonstrate that the uAMCE and pAMCE are similar for Presidential candidates because there exists little interaction between the main attribute of interest and other attributes within this subgroup.

As is the case for our first application, for many conjoint analyses, there exist natural target profile distributions, for which we can collect relevant data. In some cases, however, it might be impractical to gather corresponding real-world distributions (e.g., conjoint analysis of refugee profiles in Bansak *et al.*, 2016). Alternatively, researchers may be interested in counterfactual profiles of theoretical interest, which may be rare or even absent in the real world. For example, Ballard-Rosa *et al.* (2017) examines a variety of hypothetical tax policy proposals that are infeasible in the real world politics, but are nonetheless essential in testing the authors' theoretical argument. Importantly, even in these scenarios, the AMCE estimates

[3]Hainmueller *et al.* (2014) write "the choice of [population distribution] is important. It should always be made clear what weighting distribution of the treatment components was used in calculating the AMCE, and the choice should be convincingly justified. In practice, we suggest that the uniform distribution over all possible attribute combinations be used as a default, unless there is a strong substantive reason to prefer other distributions." (p. 12)

do depend on the choice of profile distribution. Thus, it is essential to use the proposed pAMCE framework to systematically investigate the sensitivity of the AMCE estimates to alternative theoretically relevant profile distributions.

Our second application, which is based on Peterson (2017), considers precisely these research settings, where no natural target population exists or counterfactual profiles are of theoretical interest. Peterson (2017) examines how the amount of information about candidates alters the importance of copartisanship. By randomizing how much information voters receive, the author finds that the copartisan effect is weaker when they are shown additional information on policy positions and candidate attributes. We revisit this finding by applying the proposed methodology. We build three theoretically relevant counterfactual distributions that simulate high, medium, and low-information environments. We then show that the reduction in the effect of copartisanship is driven by the outsized influence of candidates' positions on abortion and deficit spending. While the original findings are based on a specific information environment, the proposed pAMCE framework enables the systematic investigation of their robustness.

2 Motivating Empirical Applications

Before presenting the proposed methodology, we describe a conjoint analysis that will motivate and illustrate the methodology proposed in this paper. We provide two empirical applications. The first application (Ono and Burden, 2019) is a common type of conjoint analysis based on profiles of politicians, which we use to demonstrate how to incorporate a real-world distribution of politicians' characteristics. In the second application (Peterson, 2017), we illustrate the importance of considering alternative profile distributions even in settings where no natural real-world distribution exists. We show how to systematically examine counterfactual profile distributions motivated by theoretical considerations.

2.1 The Effect of Candidates Gender on Voter Choice

Scholars have long been interested in the conditions under which female candidates face obstacles to being elected (McDermott, 1997). A primary focus of the literature has been on whether a bias against female candidate is the result of taste-based or statistical discrimination (see e.g. Arrow, 1998). While the taste-based discrimination argument implies that voters dislike the idea of having female candidates in office *per se*, the statistical discrimination hypothesis contends that voters, rightly or wrongly, associate female politicians with certain

political backgrounds and policy preferences, and this association in turn shapes their vote choice. Under the statistical discrimination hypothesis, the provision of sufficient information about politicians beyond their gender should eliminate the bias against female politicians. If, on the other hand, voters are engaging in taste-based discrimination, they will disfavor female candidates even when other attributes are known.

In a recent study, Ono and Burden (2019) uses a conjoint analysis to study the effects of candidate's gender on vote choice. The authors test the aforementioned hypotheses by varying the gender of candidates and other factors such as partisanship. As in a typical conjoint analysis, respondents were asked to choose one of the two hypothetical political candidates, each of whom has the following factors: three demographic characteristics (age, race, gender), six political background (family life, years in office, area of expertise, partisanship, favorability rating, character trait), and four policy preferences (positions on abortion, immigration, national security and deficit reduction). In addition to attributes of the candidates, the original authors also randomly vary the office being sought at the candidate pair level; whether they run for President or Congress.

In Table 1, we summarize the levels of each factor used in this study. Each of 1,583 respondents evaluates 10 pairs of candidate profiles, indicating which one of the two profiles they prefer. Following the conventional conjoint analysis, all factors are independently randomized according to the uniform distribution so that each profile is equally likely. Under this uniform randomization design, the authors estimate the AMCE of candidate being female relative to male, marginalizing all other attributes, to be -1.25 percentage points (95% CI $= [-2.36, -0.19]$). This result implies that that female candidates suffer from a small disadvantage. The authors suggest that, because the conjoint analysis also presents other relevant information about politicians, this negative estimate represents evidence of taste-based rather than statistical discrimination. Importantly, Ono and Burden (2019) finds that the overall effect is driven by presidential candidates and there is little gender effect on congressional candidates. In particular, the estimated AMCE of being female is only -0.09 percentage points ($[-1.71, 1.48]$) for congressional candidates. On the other hand, the authors find a large negative effect of -2.42 percentage points ($[-3.96, -0.88]$) for presidential candidates. These findings led to the conclusion that discrimination against female candidates exists mostly in presidential elections rather than congressional elections.

Factors	Levels
`Gender`	Male, Female
`Race`	Asian, Black, Hispanic, White
`Age`	36, 44, 52, 60, 68, 76
`Family`	Divorced, Never married, Married (no children), Married (2 children)
`Experience`	None, 4 years, 8 years, 12 years
`Expertise`	Economic policy, Education, Environmental issues, Foreign policy, Health care, Public safety
`Character Trait`	Compassionate, Honest, Intelligent, Knowledgeable, Leadership, Empathetic
`Party`	Republican, Democrat
`Immigration Policy`	Favors guest worker program, Opposes guest worker program
`Security Policy`	Strong military, Cut defense spending
`Abortion Policy`	Pro-choice, Neutral, Pro-life
`Deficit Policy`	Increase taxes, Take no action, Reduce spending
`Favorability Rating`	34, 43, 52, 61, 70%

Table 1: Factors and Levels Used in Ono and Burden (2019). All factors are independently and uniformly randomized with levels in each factor shown with equal probability.

2.2 The Effect of Information Environment on Partisan Voting

The study of copartisanship in the United States has long shown that voters demonstrate a strong preference for candidates of their own party (Campbell *et al.*, 1960). Although the importance of copartisanship is widely accepted, researchers disagree about its underlying mechanisms. Some argue that voters' support for parties is deeply rooted (Bartels, 2000). As a result, voters may use partisan motivated reasoning when making decisions about which candidates to support, assessing information as favorable as possible given their partisan attachments (Bolsen *et al.*, 2014; Druckman, 2014). Others argue that partisan cues mainly serve as substitutes for relevant information such as political background and policy preferences (Bullock, 2011; Lau and Redlawsk, 2001).

To adjudicate between these two theories, Peterson (2017) uses a conjoint analysis to estimate the extent to which the amount of information presented to voters conditions the importance of partisan cues. Respondents are asked to choose one of the two hypothetical candidates that vary along ten dimensions such as age, gender, race, and policy positions.

These factors and their levels are given in Table 2.

A key feature of this study is that the randomization occurs in three steps. First, the author randomly selects the number of attributes to be presented to a respondent. The primary factor of interest, candidate party, is always shown, but the remaining nine factors are randomized to be shown or not shown. In particular, the number of additional factors is randomized to be 1, 3, 5, 7, or 9. In the second step, he then randomly chooses the selected number of factors from the remaining nine attributes. Finally, as in a typical conjoint analysis, levels are randomly chosen within each selected factor.

Under this design, Peterson (2017) examines how the effect of copartisanship changes with the amount of information about candidates respondents possess. The original analysis finds that showing more information greatly reduces the effect of copartisanship, suggesting that partisanship partially serves as substitutes for other relevant information. The author also extends this analysis by investigating which factor plays an outsized role in reducing the effect of copartisanship. This analysis shows that the information about a candidate's position on abortion policy and deficit spending diminish the effect of copartisanship more than demographic features such as race and gender.

3 Causal Quantities of Interest

In this section, we consider causal quantities of interest in conjoint analysis. We first show that most existing conjoint analyses implicitly estimate the *uniform* average marginal component effect (uAMCE) that gives equal weights to all conjoint profiles. Unfortunately, the profile distribution in the real world is likely to be far from uniform. Therefore, we consider an alternative quantity, the population average marginal component effect (pAMCE) that directly incorporates the knowledge about the target profile distribution. We discuss the conditions under which the pAMCE differs from the uAMCE.

3.1 The Setup

Following the setup of Hainmueller *et al.* (2014), consider a conjoint analysis with a total of N respondents. In the experiment, each respondent, indexed by $i \in \{1, \ldots, N\}$, completes K choice (or rating) tasks, and for a given task, a respondent chooses one of J profiles (or rate each of them). A conjoint profile is composed of L attributes represented by the corresponding L factors, where each factor ℓ has a total of D_ℓ levels. For example, the conjoint analysis of Ono and Burden (2019) has $N = 1,583$ respondents who are assigned to $K = 10$ tasks of choosing

Factor	Levels
`Age`	28, 34, 40, 40, 46, 52, 58, 62, 68, 74
`Gender`	Male, Female
`Race`	Asian American, Black, Hispanic, White
`Education`	No college degree, AA (community college), BA from state university/small college/Ivy league
`Profession`	Business owner, Car dealer, Doctor, Farmer, High school teacher, Lawyer
`Family`	Never married, Married with 0/1/2 children, Divorced with 0/1/2 children
`Military service`	Served in U.S. military, No military service
`Party`	Democrat, Republican
`Abortion stance`	Never permissible, Permissible only when mother in danger, Always permissible
`Spending stance`	Large decrease, Small decrease, No change, Small increase, Large increase

Table 2: Factors Used in Peterson (2017). Each respondent completed 3 choice tasks with each task containing two profiles. The full sample includes 1,059 respondents and 6,354 profiles. The design randomizes the number of factors shown to the respondent, which factors are shown, and the levels of each selected factor. The candidate's partisanship is always shown.

one of $J = 2$ candidates. Candidates differ in $L = 13$ factors and the levels of each factor are given in Table 1; e.g., $D_1 = 2$ and $D_2 = 6$ where the first and second factors represent gender and age, respectively.

We denote the jth profile presented to respondent i in the kth task by a profile vector $\mathbf{T}_{ijk}$ of length L. The ℓ th element of this vector represents the ℓth factor of the profile, which takes one of D_ℓ levels, i.e., $T_{ijk\ell} \in \{0, 1, \ldots, D_\ell - 1\}$. For example, if for the first respondent, the first attribute of the first profile in the first task is male, then we have $T_{1111} = 0$.

Using the potential outcomes framework (Neyman, 1923; Rubin, 1974), let $Y_{ijk}(\mathbf{t})$ represent the potential outcome for respondent i when the stacked vector of J profiles $\mathbf{T}_{ik} = \mathbf{t}$ are presented to respondent i as the kth task. When the outcome is choice-based, only one of J potential outcomes for task k by respondent i takes the value of one whereas the other $J - 1$ potential outcomes are equal to zero. In contrast, when the outcome is rating-based, each outcome Y_{ijk} corresponds to the rating of profile j given by respondent i in the kth task.

This notation is based on the stable unit treatment value assumption (Cox, 1958; Rubin, 1990). In particular, we assume no carryover effect, implying that the outcome of a task is not

affected by the same respondent's previous tasks (Hainmueller *et al.*, 2014). In addition, it is often assumed that the position of profiles does not affect the outcome (Hainmueller *et al.*, 2014). Under these assumptions, the potential outcome $Y_{ijk}(\mathbf{t})$ can be simplified as $Y_{ik}(\mathbf{t})$ because respondents would reveal the same outcomes regardless of positions of profiles j.

Under this framework, we review the definition of the AMCE originally proposed by Hainmueller *et al.* (2014). The AMCE represents the average causal effect of changing levels within each factor while averaging over other factors. For example, we might be interested in estimating the effect of a candidate's gender, averaging over the distribution of the other candidate characteristics such as age, ideology, and policy positions.

Definition 1 (Average Marginal Component Effect (Hainmueller *et al.*, 2014)) The average causal effect of changing factor ℓ from level t_0 to t_1 for a given profile while averaging over the other factors is given by,

$$\tau_\ell(t_1, t_0; \Pr(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})) = \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) \in \mathcal{T}} \mathbb{E}\left[Y_{ik}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) - Y_{ik}(t_0, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})\right] \times \Pr(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}),$$

where $\mathbf{t}_{ijk,-\ell}$ represents an $(L-1)$ dimensional vector representing the levels of all factors except for factor ℓ of the jth profile in the kth task completed by respondent i, $\mathbf{t}_{i,-j,k}$ denotes the levels of all factors for the remaining profiles other than profile j, and $\mathcal{T}$ is the support of $\Pr(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})$. Finally, the expectation is over a random sample of the respondents and task positions.

In its core, the AMCE averages not only across respondents but also across conjoint profiles, such as political candidates. We show below that this marginalizing distribution over profiles plays an essential role in conjoint analysis.

3.2 The Uniform Average Marginal Component Effect

The definition of the AMCE clearly shows that the use of different profile distributions $\Pr(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})$ can lead to substantively different conclusions (Hainmueller *et al.*, 2014). Nevertheless, in practice, little attention is paid to the choice of this profile distribution. In particular, most existing conjoint analyses use the uniform distribution, in which each factor is independently and uniformly randomized, making each conjoint profile equally likely. We call the resulting quantity as the *uniform* average marginal component effect (the uAMCE).

Definition 2 (Uniform Average Marginal Component Effect) The uniform average causal effect of changing factor ℓ from level t_0 to t_1 for a given profile while marginalizing the other factors is given by,

$$\tau_\ell^{\mathtt{U}}(t_1, t_0) = \tau_\ell(t_1, t_0; \Pr{}^{\mathtt{U}}(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}))$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) \in \mathcal{T}^{\mathtt{U}}} \mathbb{E}\left[Y_{ik}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) - Y_{ik}(t_0, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})\right] \Pr^{\mathtt{U}}(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}),$$

where $\Pr^{\mathtt{U}}(\cdot)$ denotes the uniform distribution and $\mathcal{T}^{\mathtt{U}}$ is the support of $\Pr^{\mathtt{U}}(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})$.

The central problem of the uAMCE is that it equally weights all profiles regardless of how realistic they are. Because any AMCE represents a weighted average of causal effects across all profiles used in the experiment, the estimates partially based on unrealistic profiles may yield misleading findings. The problem is not entirely new. In fact, users of conjoint experiments are often concerned about unrealistic profiles and remove highly unlikely profiles (e.g., Hainmueller *et al.*, 2014). Although this restricted randomization can eliminate extreme cases (e.g., doctors without college degree), the overall distribution of profiles may still be far away from a target profile distribution. Given that one of the core advantages of conjoint experiments is to mimic real-world decision making process (Hainmueller *et al.*, 2015), it is critical to define causal quantity of interest that reflects a target population.

3.3 The Population Average Marginal Component Effect

To improve the external validity of conjoint analysis, we consider the *population* AMCE (pAMCE), which marginalizes factors over the target population distribution of profiles rather than the uniform distribution. This target population of profiles depends on the substantive context of each application, similarly to survey research where a target population of respondents must be specified. This can be obtained from a real world data set on the attributes of actual politicians as in the case of Ono and Burden (2019) study (our first application). Alternatively, it can be a counterfactual distribution of theoretical interest that, for example, represents a different information environment for voters as in the Peterson (2017) study (our second application). Formally, we define the pAMCE as follows.

Definition 3 (Population Average Marginal Component Effect) The population average causal effect of changing factor ℓ from level t_0 to t_1 for a given profile while marginalizing the other factors is given by,

$$\begin{aligned}\tau^*_\ell(t_1, t_0) &= \tau_\ell(t_1, t_0; \Pr^*(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})) \\ &= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) \in \mathcal{T}^*} \mathbb{E}\left[Y_{ik}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) - Y_{ik}(t_0, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})\right] \Pr^*(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}),\end{aligned}$$

where $\Pr^*(\cdot)$ denotes the target population distribution and $\mathcal{T}^*$ is the support of $\Pr^*(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})$.

The distinction between the uAMCE and the pAMCE is simple and yet important. While the uAMCE marginalizes other factors over the uniform distribution, the pAMCE averages

them over the target population distribution of profiles. Therefore, the pAMCE appropriately weights profiles according to the frequency with which they occur in the target distribution. Formally, we can characterize the difference between these two quantities as follows,

$$
\begin{aligned}
&\tau^*_\ell(t_1, t_0) - \tau^{\mathtt{U}}_\ell(t_1, t_0) \\
&= \sum_{\substack{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) \\ \in \mathcal{T}^* \cup \mathcal{T}^{\mathtt{U}}}} \mathbb{E}\big[\{Y_{ik}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) - Y_{ik}(t_0, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})\} - \{Y_{ik}(t_1, \mathbf{t}'_{ijk,-\ell}, \mathbf{t}'_{i,-j,k}) - Y_{ik}(t_0, \mathbf{t}'_{ijk,-\ell}, \mathbf{t}'_{i,-j,k})\}\big] \\
&\qquad \times \{\Pr^*(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) - \Pr^{\mathtt{U}}(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})\}. \qquad (1)
\end{aligned}
$$

This difference between the uAMCE and the pAMCE has two components. The first term quantifies the average causal interaction effect between the factor of interest and all the other factors including those of other profiles (Egami and Imai, 2019). For example, the effect of being female relative to male might be larger for white candidates than black candidates. The second term represents the difference between the uniform and the target profile distributions. Therefore, the difference between the uAMCE and the pAMCE is large when the causal effect of factor ℓ interacts with other factors and when the target profile distribution is far away from the uniform distribution.

3.4 Empirical Illustrations

Using the two studies introduced in Section 2, we empirically illustrate the importance of target profile distributions. For the first application, there exists a natural real-world profile distribution that can be used to estimate the pAMCE. Using data on the characteristics of actual politicians, we construct a distribution of profiles that more accurately reflects what real-world politicians look like. We show that this distribution is strikingly different from uniform. In our second application, we demonstrate how the pAMCE can be useful even when there exists no natural real-world profile distribution for which data can be collected. Specifically, we analyze theoretically relevant counterfactual distributions and systematically investigate how empirical findings change according to the choice of profile distributions.

3.4.1 The Use of Real-world Distributions

As in the vast majority of conjoint analyses, Ono and Burden (2019) randomizes factors independently by choosing each level with equal probability. This produces a uniform distribution in which all attribute combinations are equally likely. While the uniform distribution is commonly used in applications of the conjoint analysis, the corresponding real-world distribution of attributes are rarely uniform.

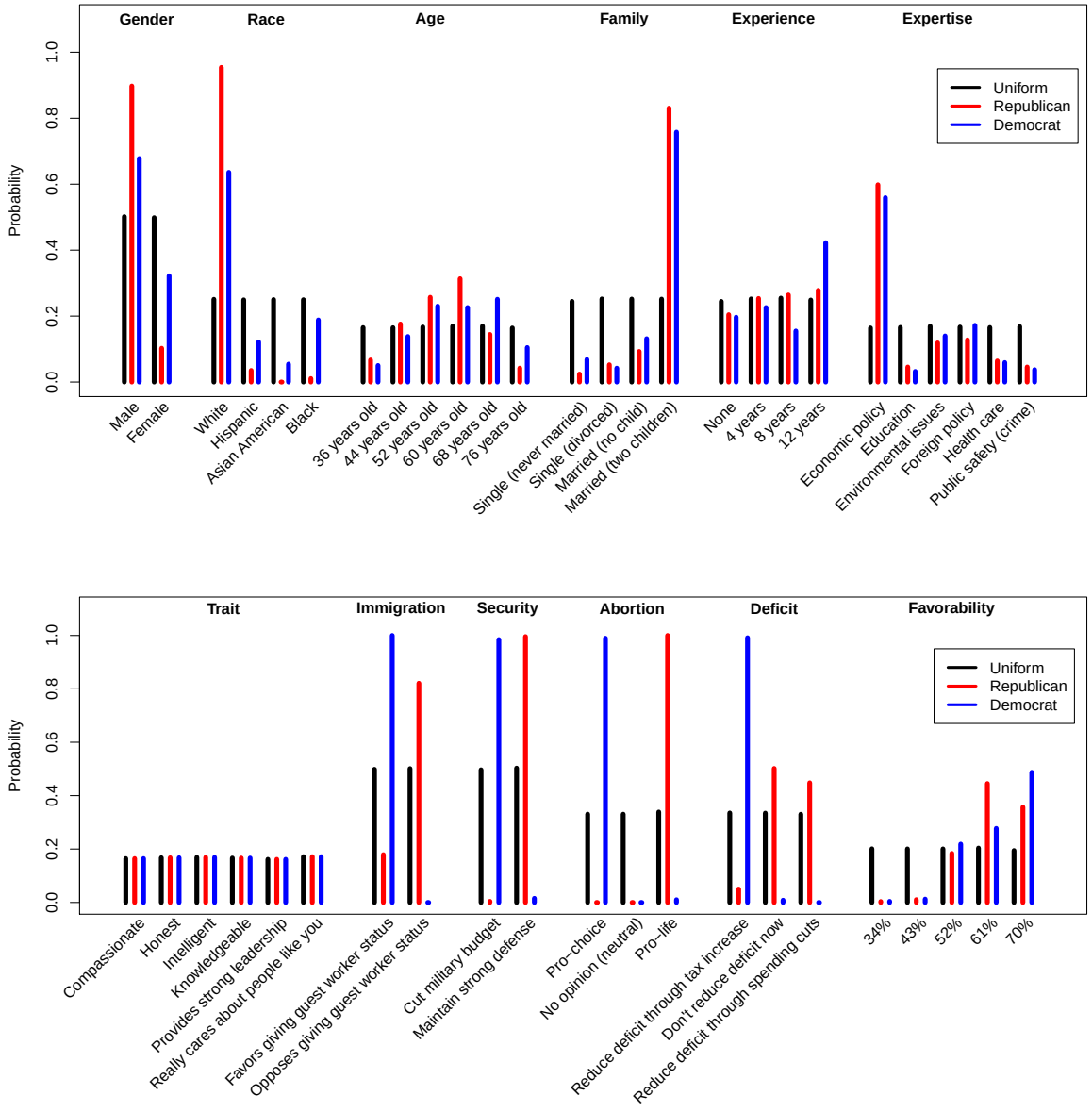

Figure 2: Experimental and Target Profile Distributions of Factors in Ono and Burden (2019). We compare the uniform distribution used in the original experiment and two real-world distributions of politicians' characteristics and policy positions; Republican and Democrat legislators.

Indeed, the uniform randomization produces highly unusual profiles. For example, two-thirds of Republican candidate profiles will have abortion positions of "neutral" or "pro-choice." The difference between this distribution and the one that of actual Republican politicians is stark. Using a legislator scorecard produced by the National Right to Life Council, a conservative non-profit that advocates for pro-life policies, only 1 of the 296 Republican legislators (0.33%) could be classified as pro-choice and only 2 (0.67%) as neutral. A similar pattern emerges for Democrats. Two-thirds of presented candidate profiles take a value of neutral or

pro-life, yet similarly low percentages of Democratic politicians hold those positions.

The case of abortion position may be especially dramatic, but the real-world distributions of nearly all of the attributes presented in Table 1 differ markedly from the uniform distribution. As a target profile distribution, we use data of actual legislators in the 115th Congress and compute the real-world joint distribution of 12 of the 13 attributes examined in Ono and Burden (2019). We do not produce the distribution of the `Trait` attribute due to its highly subjective quality and thus keep the uniform distribution for it. Because party is strongly correlated with nearly all remaining attributes, we consider the target profile distributions of Republican and Democrat politicians separately. Appendix B includes details about the construction of this joint distribution.

Figure 2 shows that the marginal distributions of actual politicians' characteristics (blue bars for Democrats, red for Republicans) differ substantially from the uniform distribution (black bars). In the case of the gender, which is the focus of the original analysis, neither the Republican nor Democratic distributions resemble the uniform: only 10.2% of Republicans and 32.2% of Democratic legislators are female. We find a similar pattern for the remaining attributes. The difference is most pronounced for the attributes that are likely to be salient to subjects, such as race and major policy positions. This suggests that the uAMCE may significantly differ from the pAMCE.

Finally, we note that the original experiment considers hypothetical political candidates. Thus, the ideal target profile distribution would be the real-world distribution of the attributes for all candidates, not only for elected legislators. Unfortunately, because the original conjoint experiment was not designed with fidelity to the real-world distribution in mind, there are many factors for which it is not possible to gather corresponding real-world distributions using data from all candidates. As a result, we use politicians in the 115th Congress as our main target profile distribution, for whom we were able to collect real-world distributions for most factors (as visualized in Figure 2).

In Section 5.1, we consider the robustness of the pAMCE estimates by replacing profile distributions of race, gender, and experience, based on publicly available candidate-level datasets. We also consider different theoretically relevant profile distributions on policy dimensions. Even when it is infeasible to collect the real-world distribution of all factors for all candidates, it is critical to take into account more realistic profile distributions and improve the external validity of conjoint analysis.

3.4.2 The Use of Counterfactual Distributions

Peterson (2017) is primarily interested in how the effect of copartisanship changes according to the amount of other relevant information about candidates. Therefore, our analysis focuses on the first two steps of the original randomization — randomizing the number of factors to show and then randomly selecting which factor to present given the selected number of factors to be shown. Because each randomization uses the uniform distribution, every factor is equally likely to be shown. In particular, the marginal probability of each factor being shown is a little above 50% (see Figure 3). If researchers use the widely-used linear regression estimator (Hainmueller *et al.*, 2014), the resulting estimate of the AMCE represents the causal effect of copartisanship while averaging over low, medium, and high information environments.

Rather than averaging over different information environments that have distinct substantive meanings, we may be interested in investigating how the pAMCE depends on different information environments. In particular, we consider two counterfactual distributions: a low information environment in which subjects observe each factor (other than copartisanship) only 20% of the time, and a high information environment in which each factor is observed 80% of the time. Figure 3 compares these low- and high-information counterfactual distributions to the one used in the original analysis. As the figure demonstrates, these low and high-information environments differ substantially from the medium-information environment produced by the original design. This suggests that the AMCE estimate based on the conventional regression estimator may differ from the pAMCEs based on the two counterfactual distributions representing specific information environments of theoretical interest. The framework of the pAMCE is essential to systematically assess how the AMCE estimates might change under different profile distributions.

4 The Proposed Methodology

In this section, we propose two new approaches to estimate the pAMCE. First, we show how to conduct a *design-based confirmatory analysis*, in which we incorporate target profile distributions when designing experiments. In contrast, the second approach — a *model-based exploratory analysis* — takes into account target distributions after randomizing profiles. This latter approach is useful in estimating the pAMCE from existing conjoint experiments that have randomized profiles with distributions different from the target population.

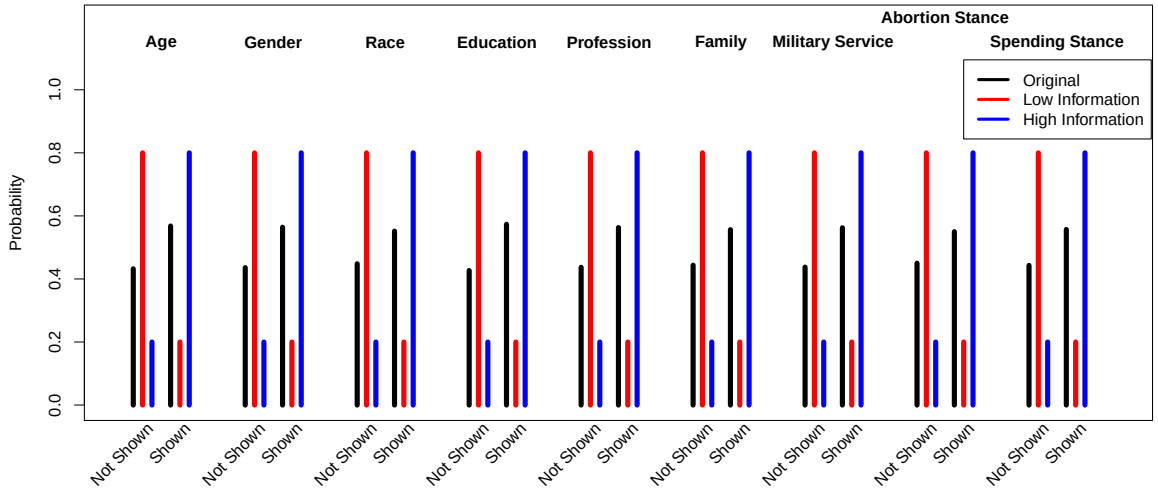

Figure 3: Original and Counterfactual Distributions of Factors in the Information Experiment (Peterson, 2017). We compare the distribution used in the experiment and two counterfactual distributions of information environment.

4.1 Design-based Confirmatory Analysis

The proposed design-based confirmatory analysis consists of new experimental designs and their associated estimators of the pAMCE. We describe each in turn.

4.1.1 Experimental Designs

We introduce three experimental designs; the joint population randomization design, the marginal population randomization design, and the mixed randomization design. While all experimental designs allow for the consistent estimation of the pAMCE, they differ in terms of data requirements and assumptions.

We begin with the *joint population randomization design*. In this design, researchers randomize profiles according to their target profile distribution.

Definition 4 (Joint Population Randomization Design)

$$\Pr^{\mathtt{R}}(\mathbf{T}_{ik} = \mathbf{t}) = \Pr^{*}(\mathbf{T}_{ik} = \mathbf{t}) \quad \text{for all } \mathbf{t} \in \text{support of } \mathbf{T}_{ik} \text{ and for all } i \text{ and } k, \tag{2}$$

where $\Pr^{\mathtt{R}}(\cdot)$ denotes the distribution used for randomization and $\Pr^{*}(\cdot)$ represents the target profile distribution.

This design is simple and intuitive since it directly incorporates the target profile distribution into randomization. The main advantage is that the design allows for nonparametric estimation of the pAMCE using a weighted difference-in-means estimator described in the next section.

While the joint population randomization design enables nonparametric estimation, it requires the knowledge of the joint distribution of profile attributes. In practice, this requirement

might be difficult to satisfy for many applications. An alternative design that relaxes this stringent data requirement is the *marginal population randomization design*. Under this design, researchers randomize each factor independently according to its marginal profile distribution of the target population.

Definition 5 (Marginal Population Randomization Design)

$$\Pr^{\mathsf{R}}(T_{ijk\ell} = t) = \Pr^{*}(T_{ijk\ell} = t) \quad \text{for all levels } t \text{ and for all } i, j, k, \ell. \tag{3}$$

For example, we randomize three factors {`Gender`, `Race`, `Education`} independently with each marginal distribution, $\Pr^{*}(\texttt{Gender}), \Pr^{*}(\texttt{Race})$, and $\Pr^{*}(\texttt{Education})$, respectively, rather than using the joint distribution $\Pr^{*}(\texttt{Gender}, \texttt{Race}, \texttt{Education})$.

The main advantage of this approach is that it only requires information about separate marginals of the target profile distribution. Gathering data on marginal distributions is likely to be easier in most contexts. In fact, some researchers have begun to incorporate marginal distributions of the target profile population in their research (see Leeper and Robison, 2018). Another significant benefit is that we can estimate the pAMCE using simple difference-in-means under this design. In practice, this means that researchers can estimate the pAMCE using a linear regression because factors are independent of each other.

The marginal population randomization design is not free of limitations. In particular, without further assumptions, this design estimates the approximate pAMCE where we only partially capture the target profile distribution. Nevertheless, compared to the uAMCE, this design already greatly improves the external validity of conjoint analysis. Indeed, a similar approximation strategy is often used in other contexts, including survey research, in which sampling weights are computed using population marginals, and causal inference with observational data, in which observed covariates are balanced only with respect to their marginal means.

What assumption is required for the consistent estimation of the pAMCE only with separate marginal distributions rather than the joint distribution of profile attributes? It turns out that we only need to assume the absence of three-way or higher order interactions among factors. Suppose that there are three factors `Gender`, `Race`, and `Education`, and they have two-way interactions; `Gender` $\times$ `Race`, `Gender` $\times$ `Education`, and `Race` $\times$ `Education`. In this case, a simple difference-in-means estimator is still consistent for the pAMCE so long as there exists no three-way or higher order interaction such as `Gender` $\times$ `Race` $\times$ `Education`. It is important to emphasize that the marginal population randomization design allows for the existence of

any two-way interaction, which often captures the strongest interaction in many applications.

There are several ways to address concerns about the assumption of no three-way or higher-order interaction. First, researchers can extend this marginal population randomization design by incorporating the partial joint distributions. Suppose that the joint distribution $\Pr^*(\texttt{Race}, \texttt{Education})$ is available while all other factors are randomized independently according to their separate marginal distributions. In this case, we can consistently estimate the pAMCE of `Gender` via a weighted difference-in-means (see Section 4.1.2) even when there exists the three-way interaction `Gender` $\times$ `Race` $\times$ `Education` if the joint distribution of `Race` and `Education` is incorporated into randomization. In general, if we incorporate the joint distributions of M factors, the consistent estimation of the pAMCE is possible even if there exist $(M+1)$-way interactions involving these factors. Finally, we can test the assumption of no three-way and higher-order interactions using the standard F-test (see Section 4.2.4).

As the final design, we introduce the *mixed randomization design*, which can yield more efficient estimates when researchers are interested in only a small number of factors (e.g., one or two) and view the remaining factors as background information they control for. For this design, we first separate L factors into two types $\mathbf{T} = \{\mathbf{T}^{\mathcal{M}}, \mathbf{T}^{\mathcal{C}}\}$; (1) *main factors* of interest $\mathbf{T}^{\mathcal{M}}$, for which researchers wish to estimate the pAMCE, and (2) *control factors* $\mathbf{T}^{\mathcal{C}}$, which are included as the background information. The distinction between the main and control factors is essential because there is a statistical tradeoff; as the number of the main factors increases, the estimation of the pAMCE becomes less precise. Under the mixed randomization design, we randomize the main factors of interest based on the uniform distribution and the control factors based on their target profile distribution.

Definition 6 (Mixed Randomization Design)

$$\begin{aligned} \text{Main factors:} \quad & \Pr^{\mathtt{R}}(T_{ijk\ell} = t) = \frac{1}{D_\ell} \quad \text{for all levels } t \text{ in factor } \ell \in \mathcal{M} \text{ and for all } i, j, k \\ \text{Control factors:} \quad & \Pr^{\mathtt{R}}(\mathbf{T}^{\mathcal{C}}_{ik} = \mathbf{t}) = \Pr^*(\mathbf{T}^{\mathcal{C}}_{ik} = \mathbf{t}) \quad \text{for all } i \text{ and } k. \end{aligned} \tag{4}$$

For example, as in the original study (Ono and Burden, 2019), suppose researchers are primarily interested in estimating the pAMCEs of factor `Gender` and use the other 12 factors as control factors. Under the mixed design, we randomize `Gender` using uniform while randomizing the remaining factors based on their target profile distribution.

This design has two primary advantages. First, by pre-specifying a small number of main factors at the design stage, researchers can increase the research transparency and credibility in the same way that pre-registration does (Blair *et al.*, 2019). Second, under the mixed

randomization design, we can often estimate the pAMCEs of the main factors more efficiently than under the two alternative designs. In fact, we show that when researchers have a single main factor, the mixed randomization design is optimal under the assumption of no cross-profile interaction effects (see Appendix D.3). In contrast, when multiple factors are of interest, the comparison of statistical efficiency across the three designs gives an inconclusive answer (see Section 4.1.3 for the sample size formula).

4.1.2 The Weighted Difference-in-Means Estimator

We introduce a general weighted difference-in-estimator that is consistent for the pAMCE under all three experimental designs described above. We then show how this general estimator can be simplified under some designs.

Formally, the weighted difference-in-means estimator of the pAMCE can be written as follows (Hájek, 1971),

$$\widehat{\tau}^{*}_{\ell}(t_1, t_0) \;=\; \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}w_{ijk\ell}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}w_{ijk\ell}}, \tag{5}$$

where the weights are defined as,

$$w_{ijk\ell} \;=\; \frac{1}{\Pr^{\mathtt{R}}(T_{ijk\ell} \mid \mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})} \times \frac{\Pr^{*}(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})}{\Pr^{\mathtt{R}}(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})}. \tag{6}$$

The weights equal the product of two terms. The first term represents the randomization distribution of $T_{ijk\ell}$ given all the other factors $\{\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k}\}$, whereas the second term is the ratio between the target profile distribution of $\{\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k}\}$ and their randomization distribution. Therefore, the weights are greater for observations that are more prevalent in the target profile distribution than in the randomization distribution. We prove the consistency of this estimator in Appendix D.1.

Under the joint population randomization design, the second term of the weights is equal to one and thus, weights are simplified as follows,

$$w^{\mathtt{Joint}}_{ijk\ell} \;=\; \frac{1}{\Pr^{*}(T_{ijk\ell} \mid \mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})}.$$

Under the marginal population randomization design, both the first and second terms are canceled out and hence, weights are equal to one for all observations. Therefore, simple difference-in-means is consistent for the pAMCE under the assumption of no three-way or higher-order interaction.

Result 1 (Estimation under Marginal Population Randomization Design) Under the assumption of no three-way or higher-order interaction, the following simple difference-in-means estimator is consistent for the pAMCE after randomizing profiles according to the marginal population randomization design (equation (3)).

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}} \xrightarrow{p} \tau_\ell^*(t_1,t_0) \qquad (7)$$

This difference-in-means estimator can be computed by regressing Y_{ijk} on an intercept and $\mathbf{X}_{ijk\ell}$ with regression where $\mathbf{X}_{ijk\ell}$ is a vector of $(D_\ell - 1)$ dummy variables for the levels of $T_{ijk\ell}$ excluding the baseline level t_0. Then, this difference-in-means estimator equals the estimated coefficient on the dummy variable for the level t_1 of factor ℓ (Greene, 2011; Hainmueller *et al.*, 2014). We provide the proof in Appendix D.2.

Finally, under the mixed randomization design, while weights do not have a simple expression, we can use the general weighted difference-in-means estimator given in equation (5).

In practice, the proposed weighted difference-in-means estimator can be computed via a weighted linear regression model.[4] Since the weighted linear regression is used only to compute the nonparametric weighted difference-in-means estimator, no additional modeling assumption is imposed. This weighted regression estimator generalizes the regression estimator proposed in Hainmueller *et al.* (2014).

4.1.3 Effective Sample Size

When using the proposed weighting estimator, it is important to compute the effective sample size (ESS) to determine the statistical efficiency of each design prior to conducting an experiment. We use Monte Carlo simulation by randomizing profiles according to a specific design and then compute the ESS as follows (Kish, 1965),

$$\text{ESS} \;=\; \frac{(\sum_{ijk} w_{ijk\ell})^2}{\sum_{ijk} w_{ijk\ell}^2}. \qquad (8)$$

When weights are equal to one for every observation, the ESS is equal to the total sample size NJK. As weights diverge from one, the ESS becomes smaller. Using ESS, we can easily compute the following standard error multiplier between any two designs,

$$\sqrt{\frac{\text{ESS under one design}}{\text{ESS under another design}}}, \qquad (9)$$

[4] As before, the weighted difference-in-means estimator defined in equation (5) can be computed by regressing Y_{ijk} on an intercept and $\mathbf{X}_{ijk\ell}$ with weights $w_{ijk\ell}$ where $\mathbf{X}_{ijk\ell}$ is a vector of $(D_\ell - 1)$ dummy variables for the levels of $T_{ijk\ell}$ excluding the baseline level t_0. Then, the weighted difference-in-means estimator equals the estimated coefficient on the dummy variable for the level t_1 of factor ℓ.

which quantifies the expected ratio of standard error that would result under one design over that under another design. By computing the ESS and the standard error multiplier at the design stage, researchers can choose an experimental design that most efficiently estimates the pAMCEs. Note that since weights are different for each pAMCE, we must compute these statistics separately.

4.2 Model-based Exploratory Analysis

When researchers incorporate the target profile distribution at the design stage, the above approach estimates the pAMCEs without bias. In some cases, however, we may wish to explore the pAMCEs of various factors using a conjoint experiment that has been fielded using profile distributions different from the target population. This is especially important when there exists no natural target profile distribution, leading to the use of the uniform randomization. Even in such cases, it is essential to examine the robustness of the AMCE estimates to alternative profile distributions that are of theoretical interest. To do so, we introduce a model-based estimator. While it requires additional modeling assumptions, this approach is useful for exploratory and sensitivity analyses. We also provide diagnostic tools for relevant modeling assumptions in Appendix E.

4.2.1 Latent Utility Model

We begin by introducing a latent utility model that allows all two-way interactions. Specifically, we assume that the latent utility for each profile is a function of the main effect of each factor, the two-way interactions between all the factors, and the two-way interaction of the same factor between the two profiles within a given pair (e.g., the effect of age of one profile may depend on the age of the other profile). The modeling assumption is violated if three-way or higher order interaction effects exist. Although we believe that in most practical settings this assumption approximately holds, we offer a simple model specification test in Section 4.2.4.

Formally, our latent utility model of respondent i for profile j when compared against profile j' in the kth task is defined as follows,

$$\begin{aligned}\widetilde{Y}_{ijk}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k}) \;=\; & \widetilde{\alpha} + \sum_{\ell=1}^{L} \mathbf{X}_{ijk\ell}^{\top}\widetilde{\beta}_{\ell} + \sum_{\ell=1}^{L}\sum_{\ell'\neq\ell}(\mathbf{X}_{ijk\ell} \bigtimes \mathbf{X}_{ijk\ell'})^{\top}\widetilde{\gamma}_{\ell\ell'} - \sum_{\ell=1}^{L} \mathbf{X}_{ij'k\ell}^{\top}\widetilde{\beta}_{\ell} \\ & - \sum_{\ell=1}^{L}\sum_{\ell'\neq\ell}(\mathbf{X}_{ij'k\ell} \bigtimes \mathbf{X}_{ij'k\ell'})^{\top}\widetilde{\gamma}_{\ell\ell'} + \sum_{\ell=1}^{L}(\mathbf{X}_{ijk\ell} \bigtimes \mathbf{X}_{ij'k\ell})^{\top}\widetilde{\delta}_{\ell\ell} + \widetilde{\epsilon}_{ijk}, \quad (10)\end{aligned}$$

where $\mathbf{X}_{ijk\ell}$ is a vector of $(D_\ell - 1)$ dummy variables for the levels of $T_{ijk\ell}$ excluding the baseline level and $\times$ represents the cartesian product operator, e.g., $(\mathbf{X}_{ijk\ell} \times \mathbf{X}_{ijk\ell'})^{\top}\widetilde{\gamma}_{\ell\ell'} =$

$\sum_{d=1}^{D_\ell-1}\sum_{d'=1}^{D_{\ell'}-1} X_{ijk\ell d}X_{ijk\ell' d'}\widetilde{\gamma}_{\ell d\ell' d'}$. The coefficients $\widetilde{\beta}_\ell$ denote the main effects of factor ℓ, while the coefficients $\widetilde{\gamma}_{\ell\ell'}$ indicate two-way interactions between the two factors ℓ and ℓ'. Finally, the coefficients $\widetilde{\delta}_{\ell\ell}$ represent two-way interactions between factor ℓ across the two profiles j and j'. Under the assumption of no profile-order effects, the effects of factors in profile j and those in profile j' are symmetric. This is why the effect of $\mathbf{X}_{ijk\ell}$ is $\widetilde{\beta}_\ell$ and that of $\mathbf{X}_{ij'k\ell}$ is $-\widetilde{\beta}_\ell$. Similarly, the effect of $\mathbf{X}_{ijk\ell} \times \mathbf{X}_{ijk\ell'}$ is $\widetilde{\gamma}_{\ell\ell'}$ while that of $\mathbf{X}_{ij'k\ell} \times \mathbf{X}_{ij'k\ell'}$ is $-\widetilde{\gamma}_{\ell\ell'}$.

As in the conventional latent utility model, we do not directly observe the latent utility. Instead, we observe the choices made by respondents. Each respondent is assumed to choose profile j when its latent utility is higher than the latent utility of the other profile j', i.e.,

$$Y_{ik}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k}) \;=\; \begin{cases} 1 & \text{if} \quad \widetilde{Y}_{ijk}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k}) > \widetilde{Y}_{ij'k}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k}), \\ 0 & \text{otherwise} \end{cases}$$

There are many ways to connect the latent utility model to the choice outcome model. For example, when we assume the error term follows the type I extreme value distribution, we obtain the well-known conditional logit model (McFadden, 1974). For the ease of interpretation, we rely on the following linear probability model (Egami and Imai, 2019),

$$\begin{aligned} &\Pr(Y_{ik} = 1 \mid \mathbf{T}_{ijk}, \mathbf{T}_{ij'k}) \\ &= \left\{\widetilde{Y}_{ijk}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k}) - \widetilde{Y}_{ij'k}(\mathbf{T}_{ijk}, \mathbf{T}_{ij'k})\right\} + 0.5 \\ &= \alpha + \sum_{\ell=1}^{L}(\mathbf{X}_{ijk\ell} - \mathbf{X}_{ij'k\ell})^\top \beta_\ell + \sum_{\ell=1}^{L}\sum_{\ell'\neq\ell}(\mathbf{X}_{ijk\ell} \bigtimes \mathbf{X}_{ijk\ell'} - \mathbf{X}_{ij'k\ell} \bigtimes \mathbf{X}_{ij'k\ell'})^\top \gamma_{\ell\ell'} + \sum_{\ell=1}^{L}(\mathbf{X}_{ijk\ell} \bigtimes \mathbf{X}_{ij'k\ell})^\top \delta_{\ell\ell} \end{aligned} \tag{11}$$

where the coefficients have direct connections to the latent utility model given in equation (10), i.e., $\beta_\ell = 2\widetilde{\beta}_\ell$, $\gamma_{\ell\ell'} = 2\widetilde{\gamma}_{\ell\ell'}$ and $\delta_{\ell\ell} = 2\widetilde{\delta}_{\ell\ell}$. We estimate this linear probability model via ordinary least squares by regressing Y_{ijk} on an intercept, the difference in the main terms for all the factors, the difference in the interaction terms for all the two-way interactions, and the interaction terms across profiles for all the factors.

This model does not impose the linearity assumption because each level of a given factor enters the model as a separate dummy variable. The model also allows for all two-way interactions between and across profiles. Therefore, the key assumption is the absence of three-way or higher order interactions, which can be easily relaxed at the expense of statistical efficiency.

4.2.2 Estimation of the Population AMCE

Using the above linear probability model, we can estimate the pAMCE as a weighted average of the estimated coefficients,

$$\widehat{\tau}^*_\ell(t_1, t_0) \;=\; \widehat{\beta}_{\ell 1} + \sum_{\ell' \neq \ell} \sum_{d=1}^{D_{\ell'}-1} \widehat{\gamma}_{\ell 1 \ell' d} \Pr^*(T_{ijk\ell'} = d) + \sum_{d=1}^{D_\ell - 1} \widehat{\delta}_{\ell 1 \ell d} \Pr^*(T_{ijk\ell} = d). \tag{12}$$

where the marginal distributions are used as weights. Thus, under the two-way interactive linear probability model, we only need to collect the marginal distributions of the target profile population $\Pr^*(T_{ijk\ell} = d)$. This greatly relaxes data requirements in practice.

As we saw earlier, when there is no interaction between or across factors, the uAMCE equals the pAMCE. That is, when $\widehat{\gamma}_{\ell 1 \ell' d} = \widehat{\delta}_{\ell 1 \ell d} = 0$, we have $\widehat{\tau}^*_\ell(t_1, t_0) = \widehat{\tau}^{\mathtt{U}}_\ell(t_1, t_0) = \widehat{\beta}_{\ell 1}$. In addition, it is straightforward to estimate the difference between the uAMCE and the pAMCE,

$$\begin{aligned}
\widehat{\text{Diff}} \;&=\; \widehat{\tau}^*_\ell(t_1, t_0) - \widehat{\tau}^{\mathtt{U}}_\ell(t_1, t_0) \\
&=\; \sum_{\ell' \neq \ell} \sum_{d=1}^{D_{\ell'}-1} \widehat{\gamma}_{\ell 1 \ell' d} \{\Pr^*(T_{ijk\ell'} = d) - \Pr^{\mathtt{U}}(T_{ijk\ell'} = d)\} + \sum_{d=1}^{D_\ell - 1} \widehat{\delta}_{\ell 1 \ell d} \{\Pr^*(T_{ijk\ell} = d) - \Pr^{\mathtt{U}}(T_{ijk\ell} = d)\}.
\end{aligned}$$

Thus, as mentioned earlier, the difference is large when there exist significant interactions and when the target profile distribution is far away from the uniform distribution. Finally, we can decompose this difference as the sum of components due to different factors,

$$\widehat{\text{Diff}} \;=\; \sum_{\ell'=1}^{L} \widehat{\text{Diff}}_{\ell'} \;=\; \sum_{\ell'=1}^{L} \sum_{d=1}^{D_{\ell'}-1} \widehat{\gamma}_{\ell 1 \ell' d} \{\Pr^*(T_{ijk\ell'} = d) - \Pr^{\mathtt{U}}(T_{ijk\ell'} = d)\}. \tag{13}$$

Through this decomposition, researchers can unpack the origin of the difference between the uAMCE and the pAMCE.

4.2.3 Regularization

The main drawback of the model-based exploratory analysis is its large estimation uncertainty. When there are many factors and each factor has several levels, the model with all two-way interaction effects can produce large standard errors. We consider regularization as a way to partially recoup this loss of statistical efficiency relative to the design-based confirmatory analysis. For example, the conjoint analysis of Ono and Burden (2019) contains 13 factors with a total of 49 levels. This means that the estimated pAMCE will be the weighted average of a large number of interaction terms. In such cases, a regularized regression approach can be effective in reducing estimation uncertainty.

In particular, we follow Egami and Imai (2019) and collapse levels within factors using a regularized regression. For instance, even though Ono and Burden (2019) use six levels for factor `Age` (36, 44, 52, 60, 68, 76 years old), not all the differences between the six levels may be relevant. It may be, for example, that the effects for the first three levels are indistinguishable from each other and can be collapsed into fewer levels (e.g., 36/44/52, 60/68, 76 years old). We can use a regularized regression to identify such coarsening patterns. By reducing the number of levels, the proposed regularized regression can improve efficiency and estimate the pAMCE more precisely.

Specifically, we estimate the pAMCE by collapsing levels while avoiding regularization bias through cross fitting (Chernozhukov *et al.*, 2018). We begin by randomly splitting data into two parts, training and test data. Using the training data, we first collapse levels within factors via the generalized lasso (Tibshirani and Taylor, 2011),

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ijk} - \beta_0 - \sum_{\ell=1}^{L}\mathbf{X}_{ijk\ell}^{\top}\beta_\ell)^2 + \lambda \sum_{\ell=1}^{L}\sum_{d=1}^{D_\ell - 1}\pi_{\ell d}|\beta_{\ell d} - \beta_{\ell,d-1}|, \tag{14}$$

where we select tuning parameter λ using cross validation. By weighting according to effect size, the adaptive weights help regularize smaller effects more and larger effects less.[5] Importantly, we do not shrink the coefficients $\beta_{\ell d}$ themselves and instead regularize their differences $|\beta_{\ell d} - \beta_{\ell,d-1}|$ so that we can collapse unnecessary levels (Egami and Imai, 2019). When levels are unordered, researchers can use an alternative penalty that regularizes all pairwise differences, i.e., $\sum_{\ell=1}^{L}\sum_{d=0}^{D_\ell-1}\sum_{d'\neq d}|\beta_{\ell d} - \beta_{\ell,d'}|$.

Second, using the separate test data, we fit the proposed linear probability model with collapsed levels and then estimate the pAMCE based on the weighted average expression given in equation (12). Because unnecessary levels are removed in the previous step, we can estimate the pAMCE more precisely. It is important that we collapse levels with the training data and estimate the pAMCE with the separate test data to remove bias due to regularization.

Finally, we flip the role of training and test data and repeat the two steps described above. We average the two estimates from each test data as the estimate of the pAMCE. For uncertainty estimates, we use the block bootstrap by sampling respondents with replacement. We implement the cross-fitting for each bootstrap replicate. Uncertainty estimates are calculated based on the empirical distribution of the estimated pAMCE over the bootstrap sample. In

[5] Adaptive weights are defined as $\pi_{\ell d} = \sqrt{N_{\ell d} + N_{\ell,d-1}}/|\hat{\beta}_{\ell d}^{OLS} - \hat{\beta}_{\ell,d-1}^{OLS}|$ where $N_{\ell d}$ is the number of observations with $T_{ijk\ell} = t_d$ and $\hat{\beta}_{\ell d}^{OLS}$ is the OLS estimate of $\beta_{\ell d}$ (Gertheiss *et al.*, 2010).

Approach	Data Requirement	Assumption	Note
Design-based Confirmatory Analysis			
• Joint Population Randomization	Joint distribution over all profile attributes	None	
• Marginal Population Randomization	Marginal distributions of each profile attribute	Absence of three-way or higher order interaction	Relax the assumption with partial joint distributions
• Mixed Randomization	Joint distribution over control factors	None	Efficient when focus on one or two main factors
Model-based Exploratory Analysis			
• Linear Probability Model	Marginal distributions of each profile attribute	Absence of three-way or higher order interaction	Relax the assumption with partial joint distributions

Table 3: Data Requirements and Assumptions of Design-based and Model-based Approaches.

Appendix F, we provide simulation studies to show how much the proposed regularization method can improve efficiency without inducing bias.

4.2.4 Assessing the Absence of Higher-order Interaction

The model introduced above (equation (11)) as well as the marginal population randomization design (equation (3)) assumes the absence of three-way or higher order interaction. We can directly test the assumed absence of three-way interaction by conducting the standard F-test. Specifically, we incorporate three-way interactions between three factors ℓ, ℓ', and ℓ'' by adding $(\mathbf{X}_{ijk\ell} \times \mathbf{X}_{ijk\ell'} \times \mathbf{X}_{ijk\ell''})^\top \zeta_{\ell\ell'\ell''}$ to the two-way interactive model of equation (11) where $\zeta_{\ell\ell'\ell''}$ is a vector of coefficients for the three-way interactions. Then, we test the existence of this three-way interaction via F-test with the null hypothesis, $H_0 : \zeta_{\ell\ell'\ell''} = \mathbf{0}$. When the statistical power of detecting three-way interaction effects is of concern, it is recommended to rely on the regularization approach described above.

4.3 Summary

Table 3 summarizes the methodologies introduced in this section in terms of required data and assumptions. Several points are worth emphasizing. First, if researchers expect the target profile distribution to differ from the uniform distribution and factors to interact with one another, we recommend that they use one of the proposed experimental designs. The design-based approach is considerably more efficient than the model-based approach.

Second, the choice of experimental designs largely depends on the availability of data about the target profile distribution although the sample size calculation can be conducted to compare the statistical efficiency of these designs. Ideally, researchers have the joint distribution, and hence are able to use the joint population randomization design. If only the marginal distributions are available, the marginal population randomization can be used at the cost of making an additional assumption about the absence of third or higher-order interactions. If large higher-order interaction effects are expected, incorporating partial joint distributions can relax the assumption. In addition, the mixed-randomization design is available if researchers are interested in testing hypotheses about one or two factors while controlling for other factors.

Finally, even when there exists no natural target profile distribution for which data can be collected, it is important to conduct the model-based approach to explore the robustness of the AMCE estimates to the choice of profile distributions. We recommend researchers systematically examine different counterfactual profile distributions motivated by a theoretical consideration (see our second example based on Peterson (2017) in Section 5.2).

5 Empirical Applications

We apply the proposed methodology to the empirical applications described in Section 2. For the first application, we find that two key conclusions regarding the effect of gender are due to the uniform distribution used in the original study. Estimating the pAMCE using the real-world profile distribution instead, we find that the effect of being a female candidate varies according to party and office they seek for. For the second application, we show how to systematically explore the pAMCE based on counterfactual distributions of theoretical interest.

5.1 The Effect of Candidate's Gender on Voter Choice

In the first application, the primary quantity of interest is the AMCE of candidate's gender on voter choice. Instead of the uniform distribution used in the original analysis, we estimate the pAMCE using the real-world distribution of elected politicians in the 115th Congress, as described in Section 3.4. In particular, we estimate this quantity separately using the distributions of Democratic and Republican politicians' characteristics (see Figure 4).

5.1.1 Design-based Analysis

We begin by performing the design-based confirmatory analysis proposed in Section 4.1. Because in the original study each attribute is randomized according to the uniform distribution,

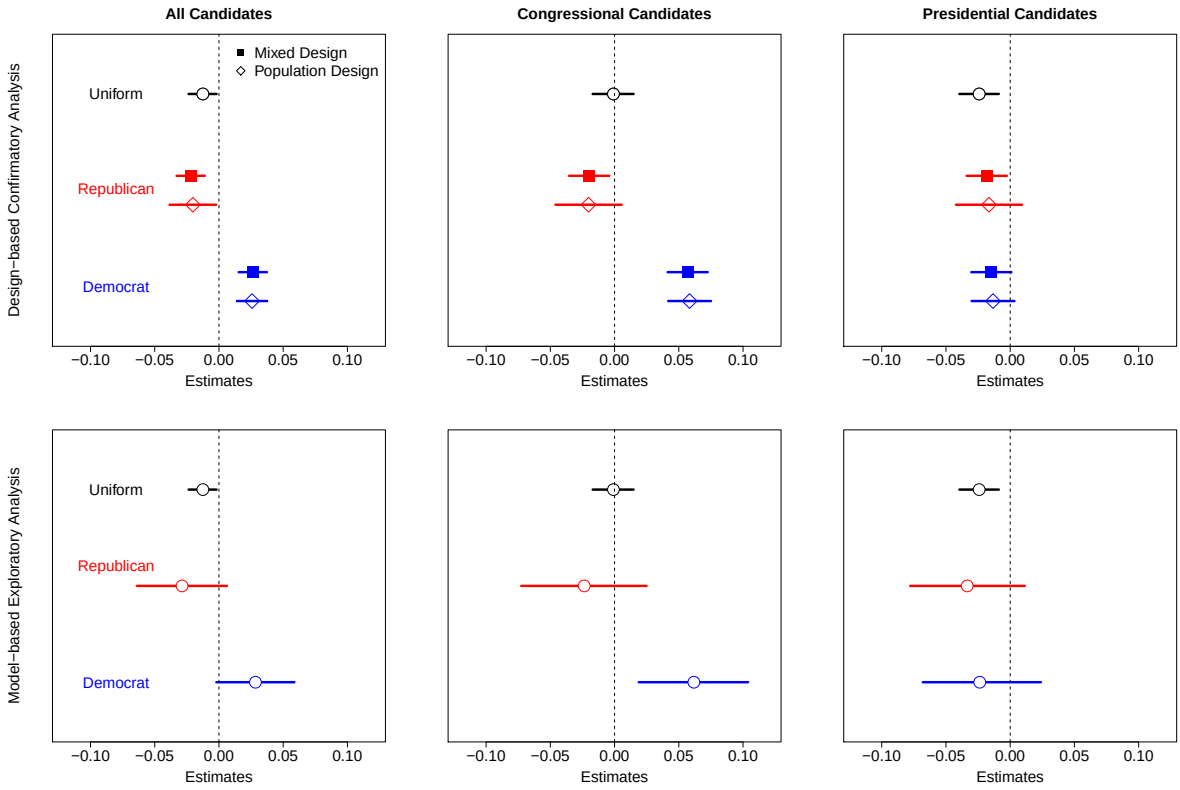

Figure 4: Estimates of the pAMCEs of Being Female in Ono and Burden (2019). We estimate the pAMCE for Republican (red) and Democrat politicians (blue). Even though an estimate of uAMCE is close to zero for congressional candidates, the pAMCE for congressional candidates under the Democrat distribution is large and positive.

we conduct a simulation study to assess the performance of the marginal population randomization and mixed randomization designs.[6] To do this, we first fit a linear regression model with all two-way interactions between the thirteen factors summarized in Table 1 and use this estimated model as the true data generating process. For the marginal population randomization design, we randomize each factor independently based on a marginal distribution of the target population. For the mixed randomization design, the primary factor of interest, i.e., gender, is randomized according to the uniform distribution and the remaining factors are randomized using their target population distribution. We estimate the pAMCE via the simple difference-in-means under the marginal population design and the weighted difference-in-means under the mixed design. Standard errors are clustered by respondents. We repeat the same procedure 100 times and average over point estimates and standard errors.

The left plot in the top row of Figure 4 presents the results. First, we focus on the re-

[6] As we propose in Section 4.1, researchers can directly conduct the design-based confirmatory analysis when researchers can incorporate target profile distributions in the design stage.

sults based on the mixed randomization design. In contrast to the estimates of the uAMCE, we find that the pAMCE is estimated to be -2.20 percentage points (95% confidence interval = $[-3.30, -1.10]$) when using the distribution of Republicans and 2.64 percentage points ($[1.53, 3.74]$) for Democrats. We obtain similar results under the marginal population randomization design although the standard errors are slightly larger. Recall that under the uniform distribution, the estimated uAMCE of gender on vote choice is small and negative. This demonstrates that the estimated AMCE critically depends on the target distribution of candidates' attributes.

One key conclusion of the original study is that the negative effect of being female is found only for presidential candidates but not for congressional candidates. We revisit this finding by using the real-world politicians as the target profile distributions. In particular, we now conduct the design-based confirmatory analysis separately for congressional and presidential candidates. These results are presented in the top row of the second and third columns of Figure 4. Consistent with the original analysis, the estimated uAMCE of being female is -0.09 percentage points ($[-1.71, 1.48]$) for congressional candidates and -2.42 percentage points ($[-3.96, -0.88]$) for presidential candidates. For presidential candidates (the third plot in the top row), the pAMCE of being female is similar to the corresponding uAMCE for both Democratic and Republican distributions. Female presidential candidates face barriers compared to male candidates regardless of party. This result shows that the pAMCE and the uAMCE estimates can be similar even when the target profile distribution is far from uniform. This is because there exists little interaction between gender and other factors within this subgroup (see formal discussions in Section 3.3).

Interestingly, for congressional candidates (the second plot in the top row), the results of the uAMCE and pAMCEs diverge. The uAMCE implies that gender has little effect in congressional races. Yet a more realistic profile distribution suggests a more nuanced finding: women are disadvantaged when they run as Republicans and advantaged when they run as Democrats. Under the mixed randomization design, female Republican candidates are 1.98 percentage points ($[0.42, 3.54]$) less likely to be chosen than their male counterparts, while female Democratic candidates are 5.69 percentage points ($[4.13, 7.25]$) more likely to be chosen. The latter effect is large in substantive terms, equaling the effect of candidates' experience in office and their position on deficit reduction.

5.1.2 Model-based Analysis

Now, we illustrate the model-based exploratory analysis introduced in Section 4.2. This approach is useful especially when researchers are interested in exploring the pAMCE with conjoint experiments that have already been conducted using the uniform or any distributions different from the target distribution. First, we focus on estimating the pAMCE for both presidential and congressional candidates together as done in the original analysis. As explained in Section 4.2.3, we incorporate all two-way interactions among all the thirteen factors in Table 1 except for `Office` and then collapse levels within factors using the generalized lasso. Standard errors are calculated with 2,000 block bootstraps clustered by respondents.

As expected, the results are similar to those from the design-based analysis but with larger standard errors (see the left plot in the bottom row of Figure 4). The estimated pAMCE is -2.87 percentage points ($[-6.39, 0.63]$) when using the distribution of Republican politicians and 2.84 percentage points ($[-0.20, 5.87]$) for Democratic politicians. We also repeat the same analysis for presidential and congressional candidates separately (the second and third plots in the second row). As in the design-based confirmatory analysis, we find that female congressional candidates have a disadvantage when they run as Republicans and have an advantage when they run as Democrats.

The standard errors are much larger than those in the design-based confirmatory analysis because the uniform distribution used in the experiment is markedly different from the target profile distribution. This post-adjustment of the large differences reduces the effective sample size. Since the model-based analysis marginalizes all the two-way interactions, the efficiency loss is especially severe when the number of factors and levels within each factor are large, as in this example. Although regularization recoups some of this efficiency loss, the design-based analysis yields smaller standard errors in such high dimensional designs.

To investigate the sources of the difference between the uAMCE and pAMCE, we apply the decomposition formula given in equation (13). For the sake of illustration, we focus on the difference between the estimated uAMCE and pAMCE for congressional Democratic candidates (the second plot in the bottom row). The first plot of Figure 5 shows the results of this decomposition, with each row representing the difference attributable to a single factor. We find, for example, that the difference due to factor `Security Policy` is about 1.6 percentage points ($[0.14, 3.12]$), implying that the estimated AMCE increases by 1.6 percentage points when we use the distribution of Democratic politicians for `Security Policy` rather than the

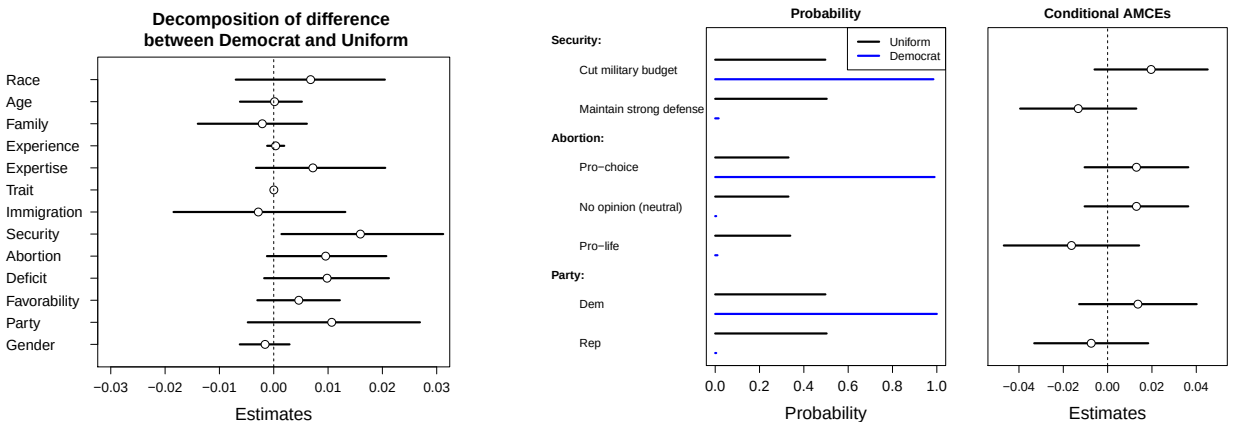

Figure 5: Decomposing the Difference between the Estimated uAMCE and pAMCE of Being a Female Candidate. For congressional candidates, we compare the uAMCE and the pAMCE based on the Democrat distribution. The first plot decomposes the overall difference into each factor. The second and third factors investigates how effect heterogeneity and the difference in the profile distributions result in the difference in the uAMCE and the pAMCE.

uniform distribution. Importantly, less than 20% of the overall difference is attributable to `Party`, meaning that we cannot estimate the pAMCE just by considering the interaction between `Gender` and `Party`. The results show that the difference between the uAMCE (-0.09 percentage points) and pAMCE (6.17 percentage points) can be attributed to a combination of many factors even though the contribution of each factor is not necessarily precisely estimated. Even if the difference due to each factor is small, when aggregated across many factors, the overall difference between the uAMCE and the pAMCE can be substantial. This result underscores the need to collect relevant data for as many factors as possible when building the target distribution.

To further unpack the source of this difference, we examine the *conditional* average marginal component effect (cAMCE). The cAMCE is the AMCE of the factor of interest — in this case, gender — conditional on the level of another factor.[7] A difference in the cAMCEs across the levels of the second factor implies an interaction with the factor of interest. For example, a difference in the cAMCEs of `Gender` conditional on `Security Policy` would indicate an interaction between `Gender` and `Security Policy`. We can use the cAMCE to determine whether each factor's contribution to the difference between the uAMCE and the pAMCE (the first plot of Figure 5) is due to large interactions, large changes in distribution, or a combination of the two. If interactions between the primary factor of interest and secondary factors are responsible for most of the difference between the uAMCE and the pAMCE, even small changes

[7] Within each factor, the weighted sum of the cAMCEs — with weights equal to the population probabilities of each level — is equal to the pAMCE of interest.

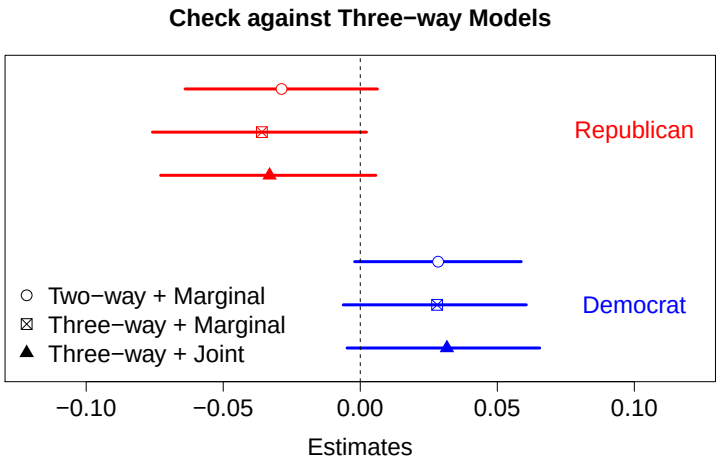

Figure 6: Assess the Existence of Three-way Interactions. We compare the pAMCE estimates from models that assume two-way interactions and that incorporate three-way interactions.

in distribution will make the uAMCE different from the pAMCE.

The right two plots of Figure 5 visualize the distributions of three factors `Security Policy`, `Abortion Policy`, and `Party` alongside the cAMCEs of `Gender` conditional on each of the three factors. For example, the first row in the third plot presents the estimated cAMCE of being female relative to male, conditional on having the `Security Policy` factor equal to `Cut military budget`. Focusing on the `Security Policy` factor, we observe that although its cAMCEs are modest in size, the distribution for Democratic politicians differs substantially from the uniform distribution. Thus, the difference induced by the `Security Policy` factor is being driven primarily by distributional differences. Repeating this approach for each factor tells us whether the difference between the uAMCE and the pAMCE is a function of distributional changes or causal interactions.

As an important diagnostic, we evaluate the assumption of no three-way interactions. In particular, we incorporate three-way interactions between `Gender`, `Party`, and each of the four policy positions given that the difference between the uAMCE and the pAMCE is mostly attributable to those factors. Because we have information about the joint distribution of politicians' attributes, we use them when we marginalize over three-way interactions to estimate the pAMCE. Figure 6 shows that the pAMCE estimates based on the three-way interaction model are similar to those based on the two-way model both in terms of point estimates and standard errors. This result demonstrates that, even when researchers have access only to marginal distributions of the target profile population, it is possible to consistently estimate the pAMCEs by using the marginal population randomization design.

Finally, we examine the robustness of the pAMCE estimates based on the 115th Congress

to alternative profile distributions based on the available candidate-level data rather than the data on elected politicians. Although these candidate-level data do not contain information for all factors, we can take into account a number of important candidates' characteristics. In particular, we use DIME data set (Bonica, 2015) and the Reflective Democracy (RefDem) dataset[8] to obtain the profile distributions of three key variables (race, gender, and experience). We also use our substantive knowledge to investigate different theoretically relevant profile distributions on policy dimensions. We provide details of these alternative profile distributions in Appendix C. We show that the pAMCE estimates are robust to these different profile distributions that more accurately reflect the real-world distribution of political candidates (see Figure A3 in Appendix C). These results imply that the difference between the pAMCE and uAMCE is mainly driven by the fact that the uniform distribution is far away from the actual distribution of politicians' characteristics. In contrast, the difference between the distribution of attributes for elected politicians and that for candidates is relatively minor and has little impact on the empirical findings.

5.2 The Effect of Information Environment on Partisan Voting

In this section, we revisit a major finding of the original study that the importance of copartisanship declines as voters are given more information about candidates.

5.2.1 Design-based Analysis

We begin with the design-based confirmatory analysis. To estimate the pAMCE, we use the marginal population randomization design by randomizing each factor according to the counterfactual distributions of interest. We also employ the mixed randomization design, retaining the uniform distribution for a primary factor of interest — copartisanship, in this case — and using the counterfactual distributions for all remaining factors. We rely on a weighted difference-in-means estimator (equation (5)) and cluster standard errors by respondents.

The left plot of Figure 7 presents results of this analysis. Consistent with the original finding, the pAMCE of copartisanship is estimated to be the largest under the low information distribution (61.84 percentage points, $[59.06, 64.62]$) while the effect is the smallest under the high information distribution (38.39 percentage points, $[35.13, 41.65]$) using the mixed randomization design. Results are similar under the marginal population randomization design. Thus, the importance of copartisanship in subjects' voting decisions can vary by more than

[8]This dataset is available at `https://wholeads.us/resources/for-researchers/`

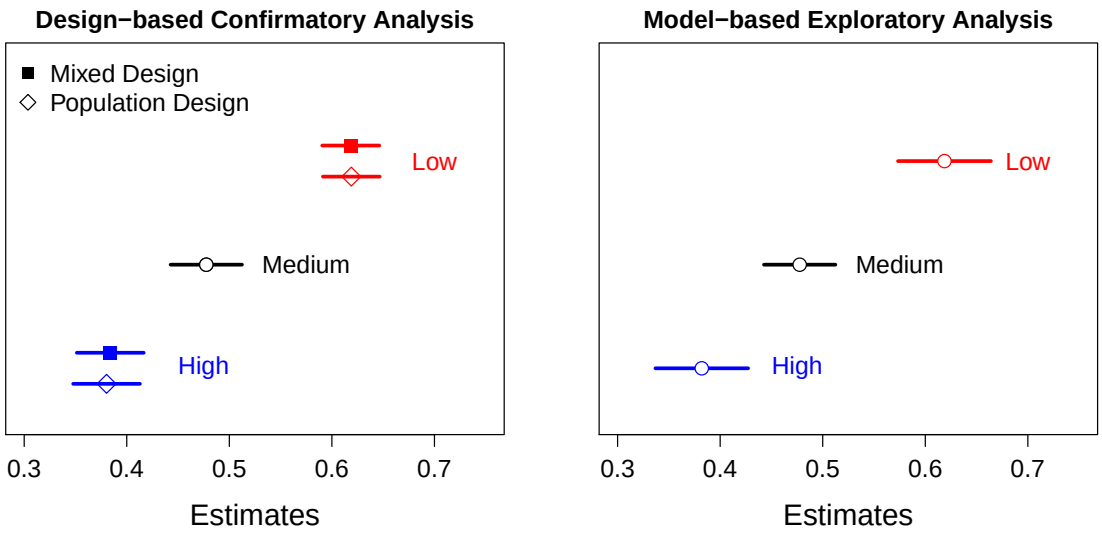

Figure 7: The Estimated Population AMCEs of Copartisanship in Peterson (2017). We estimate the pAMCE of being copartisan under three different distributions – a medium information distribution and the low and high information distributions.

20 percentage points depending on the information environment.

5.2.2 Model-based Analysis

Next, we estimate the same three quantities using model-based exploratory analysis. To do so, we run an unregularized linear regression using all two-way interactions between the ten factors described in Table 2. While regularization is generally preferred, the factors here are binary. Since the goal of regularization is to improve efficiency by collapsing levels of a factor that have similar effects, regularization is not needed in this case. Standard errors are based on 2,000 block bootstraps clustered by respondents.

The second plot in Figure 7 presents these results. As in the design-based confirmatory analysis, the pAMCE of copartisanship is the largest under the low information distribution (61.87 percentage points, $[57.34, 66.40]$) and the effect is the smallest under the high information distribution (38.21 percentage points, $[33.69, 42.73]$). Although standard errors for the model-based exploratory analysis are larger than those of the design-based confirmatory analysis, the difference between them in this application is relatively small. This is due to the fact that the design in Peterson (2017) is low-dimensional, comprised only of binary factors.

After showing that copartisanship effects are indeed smaller when a larger number of candidate characteristics are shown, the author conducts the second analysis to unpack the mechanism by identifying which information is responsible for reducing the effect of copartisanship. To answer this question, he considers an extreme counterfactual distribution, in which only

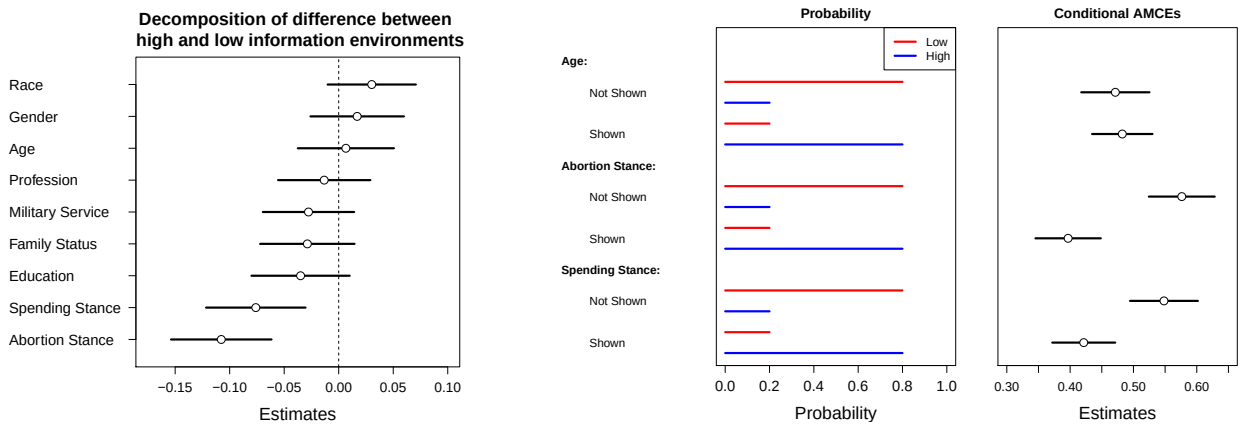

Figure 8: Decomposition of the Difference between the two pAMCEs of Copartisanship between High and Low Information Environment. The left plot decomposes the overall difference into each factor. The difference is mainly due to the two factors, `Spending Stance` and `Abortion Stance`. The second and third plots investigate how the conditional AMCE and the difference in the profile distributions contribute to the difference.

one factor (in addition to copartisanship) is shown to respondents and examines the difference in the pAMCE of copartisanship with and without this additional factor. The author repeats this analysis separately for each of the nine factors and finds that policy positions on spending and abortion result in the largest differences.

In our pAMCE framework, there is no need to consider each factor in isolation. Instead, we directly examine the sources of the difference in the pAMCE of copartisanship between the low and high information environments. To do so, we use the decomposition formula. The left plot of Figure 8 shows that the difference observed in Figure 7 is mainly driven by two factors, `Spending Stance` (-7.60 percentage points, $[-12.16, -3.04]$) and `Abortion Stance` (-10.77 percentage points, $[-15.38, -6.16]$). This result suggests that respondents use copartisanship mainly as a cue for policy stances on spending and abortion, consistent with the original findings.

Finally, we examine why these factors drive the difference in the copartisanship effect between the low and high information environments. The second and third plots of Figure 8 present the distribution and the cAMCEs of each factor. Taking `Spending Stance` as an illustration, we find that the cAMCE for `Shown` (the bottom estimate) is much smaller than for `Not Shown` (the second estimate from the bottom). There is a strong interaction between factors `Party` and `Spending Stance`, yielding the large difference of the pAMCE between the high and low information environments. In contrast, little difference exists in the cAMCEs of copartisanship conditional on `Age` (see the first and second estimates in the third plot). This

is why the difference of the pAMCE due to `Age` is small (third estimate in the first plot).

6 Concluding Remarks

Over the last several years, conjoint analysis has become increasingly popular in political science. One advantage of conjoint analysis is its unique ability to help researchers systematically examine various decision making processes faced by individuals in the real world. This attractive feature has boosted the external validity of empirical conclusions based on conjoint analysis.

Yet, little attention has been paid to the choice of the profile distribution used for randomization. While most researchers use the uniform distribution for convenience, this leads to a causal quantity — the *uniform* average marginal component (uAMCE) effect — that gives equal weights to all possible profiles, including those that rarely occur in the real world.

We address this problem by defining an alternative quantity of interest, the *population* average marginal component effect (pAMCE), using the target profile distribution based on substantive knowledge. We propose new experimental designs and estimation methods for inferring the pAMCE. We then illustrate their use with two empirical applications, one using a real-world distribution and the other based on a counterfactual distribution motivated by a theoretical consideration.

While we focus on the issues related to the distribution of profiles in conjoint analysis, our proposed methodology applies to any factorial experiments with many factors. Moreover, the importance of designing realistic interventions goes beyond conjoint analysis and survey experiments. Indeed, unlike the widely recognized issues related to the representativeness of the experimental sample, the realism of treatments is an essential yet under-appreciated element of external validity. We thus believe that the use of realistic treatments is essential in ensuring the theoretical and practical relevance of any experimental research.

References

Arrow, K. J. (1998). What has economics to say about racial discrimination? *The Journal of Economic Perspectives*, **12**(2), 91–100.

Ballard-Rosa, C., Martin, L., and Scheve, K. (2017). The Structure of American Income Tax Policy Preferences. *The Journal of Politics*, **79**(1), 1–16.

Bansak, K., Hainmueller, J., and Hangartner, D. (2016). How Economic, Humanitarian, and Religious Concerns Shape European Attitudes Toward Asylum Seekers. *Science*, **354**(6309), 217–222.

Barnes, L., Blumenau, J., and Lauderdale, B. (2019). Measuring Attitudes towards Public Spending using a Multivariate Tax Summary Experiment. Technical report, University College London.

Bartels, L. M. (2000). Partisanship and voting behavior, 1952-1996. *American Journal of Political Science*, **44**(1), 35–50.

Blair, G., Cooper, J., Coppock, A., and Humphreys, M. (2019). Declaring and Diagnosing Research Designs. *American Political Science Review*, **113**(3), 838–859.

Bolsen, T., Druckman, J. N., and Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, **36**(2), 235–262.

Bonica, A. (2015). Database on Ideology, Money in Politics, and Elections (DIME).

Bullock, J. G. (2011). Elite influence on public opinion in an informed electorate. *The American Political Science Review*, **105**(3), 496–515.

Campbell, A., Converse, P., Miller, W., and Stokes, D. (1960). *The American voter*. Chicago University Press, Hoboken, NJ.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double Machine Learning for Treatment and Structural Parameters. *Econometrics Journal*, **21**, C1 – C68.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of Heterogeneous Treatment Effect Estimates Across Samples. *Proceedings of the National Academy of Sciences*, **115**(49), 12441–12446.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

de la Cuesta, B., Egami, N., and Imai, K. (2020). Replication Data for: Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution. `https://doi.org/10.7910/DVN/IJXFQF`, Harvard Dataverse.

Druckman, J. N. (2014). Pathologies of studying public opinion, political communication, and democratic re sponsiveness. *Political Communication*, **31**(3), 467–492.

Egami, N. and Imai, K. (2019). Causal Interaction in Factorial Experiments: Application to Conjoint Analysis. *Journal of the American Statistical Association*, **114**(526), 529–540.

Enamorado, T., Fifield, B., and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, **113**(2), 353–371.

Gertheiss, J., Tutz, G., *et al.* (2010). Sparse Modeling of Categorial Explanatory Variables. *The Annals of Applied Statistics*, **4**(4), 2150–2180.

Gobel, S. and Munzert, S. (2019). legislator: political, sociodemographic, and wikipedia-related data on political elites. *R package version 0.2.0*.

Green, P. E., Krieger, A. M., and Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, **31**(3_supplement), 56–73.

Greene, W. H. (2011). *Econometric Analysis*. Pearson.

Hainmueller, J. and Hopkins, D. J. (2015). The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Toward Immigrants. *American Journal of Political Science*, **59**(3), 529–548.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, **22**(1), 1–30.

Hainmueller, J., Hangartner, D., and Yamamoto, T. (2015). Validating Vignette and Conjoint Survey Experiments against Real-World Behavior. *Proceedings of the National Academy of Sciences*, **112**(8), 2395–2400.

Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one,". *The Foundations of Survey Sampling*, **236**.

Huff, C. and Kertzer, J. D. (2018). How the public defines terrorism. *American Journal of Political Science*, **62**(1), 55–71.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Lau, R. R. and Redlawsk, D. P. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*, **45**(4), 951–971.

Leeper, T. J. and Robison, J. (2018). More Important, but for What Exactly? The Insignificant Role of Subjective Issue Importance in Vote Decisions. *Political Behavior*.

Marshall, P. and Bradlow, E. T. (2002). A unified approach to conjoint analysis models. *Journal of the American Statistical Association*, **97**(459), 674–682.

McDermott, M. (1997). Voting cues in low-information elections: candidate gender as a social information variable in contemporary United States elections. *American Journal of Political Science*, **41**(1), 270–283.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, editor, *Frontiers in Econometrics*. Academic Press.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis*, **26**(3), 275–291.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., and Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, **2**(2), 109–138.

Mutz, D. C. (2011). *Population-based Survey Experiments*. Princeton University Press.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). *Statistical Science*, **5**(4), 465–472.

Ono, Y. and Burden, B. C. (2019). The Contingent Effects of Candidate Sex on Voter Choice. *Political Behavior*, pages 1–25.

Peterson, E. (2017). The Role of the Information Environment in Partisan Voting. *The Journal of Politics*, **79**(4), 1191–1204.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**(5), 688.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**(4), 472–480.

Stewart, C. and Woon, J. (2017). Congressional committee assignments, 103rd to 114th congresses, 1993-2017. *House and Senate*.

Teele, D. L., Kalla, J., and Rosenbluth, F. (2018). The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics. *American Political Science Review*, **112**(3), 525–541.

Tibshirani, R. J. and Taylor, J. (2011). The Solution Path of The Generalized Lasso. *The Annals of Statistics*, **39**(3), 1335 – 1371.

Wickham, H. (2019). rvest: easily harvest (scrape) web pages. *R package version 0.3.4*.

Supplementary Appendix for: de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis*

A Review of Conjoint Literature

A review of the literature was conducted to assess several features of current best practices. In order to gather a sufficiently large number of articles, we selected 10 journals for a keyword search ("conjoint"): *The American Journal of Political Science, The American Political Science Review, The British Journal of Political Science, Journal of Experimental Political Science, Journal of Politics, Political Analysis, Political Behavior, Political Science Research and Methods, Research and Politics*, and *the Review of International Organizations*. This search criteria resulted in a total of 40 articles. We then augmented this list by examining articles that cited Hainmueller *et al.* (2014) using Google's "cited by" feature to obtain articles from additional journals or articles from the list above that were missed in the keyword search. This resulted in an additional 25 articles. We removed from the list any article whose contribution was primarily or completely methodological. This procedure left us with a total of 59 articles from 2014 to 2019. This list is not meant to be exhaustive but rather to be broad enough to give an overview of current practice.

Each article was then examined and classified along several dimensions. First, we coded the randomization distribution used in the design, characterizing each article by the number of factors used in the fielded design and the number that were randomized according to the uniform. In many cases, authors made no mention of the exact randomization probabilities or simply noted that their designs were "fully randomized". In cases where there was ambiguity about the distribution used, we consulted the appendix material to determine the distribution. If the appendix did not contain information sufficient to determine the distribution, we examined the uniformity of the standard errors of reported estimates and counted a factor as being randomized according to the uniform if the standard errors of all of that factor's levels were indistinguishable from each other.

We then examined the main text to establish whether the authors justified the distribution they chose on theoretical grounds. Justifications that would yield an affirmative coding include explicit discussion of the desire to match population distributions or the desire for statistical efficiency. An affirmative coding was given even if the discussion was relegated to a footnote and concerned only a single factor. Discussion of the constraints placed on unrealistic factor combinations was not part of the criteria used. As such, some papers that use such constraints were nonetheless considered as not invoking a substantive or theoretical justification for their chosen distribution.

Finally, for each paper we examined all factors that were a part of the design and de-

termined whether it was feasible to collect data that would allow the approximation of the population distribution for that factor. Designs in which population data could be feasibly collected for most or all factors were considered to be amenable to the use of population data in the design or analysis stage. The articles are given below in Table A1 .

Author	**Year**	**Title**
Atkeson and Hamel	2018	Fit for the job: candidate qualifications and vote choice in low information elections
Auerbach and Thachil	2018	How clients select brokers: competition and choice in India's slums
Ballard-Rosa et al	2017	The structure of American income tax policy preferences
Bansak et al	2016	How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers
Bechtel and Scheve	2013	Mass support for global climate agreements depends on institutional design
Bechtel et al	2017	Interests, norms and support for the provision of global public goods: The case of climate co-operation
Bechtel et al	2017	Policy design and domestic support for international bailouts
Berinksy et al	2018	Attribute affinity: U.S. natives' attitudes towards immigrants
Bernauer et al	2019	Do citizens evaluate international cooperation based on information about procedural and outcome equality?
Breitensten	2019	Choosing the crook: a conjoint experiment on voting for corrupt politicians
Bueno	2017	Bypassing the enemy: distributive politics, credit claiming, and nonstate organizations in Brazil
Campbell et al	2016	Legislator dissent as a valence signal
Carnes and Lupu	2016	Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class
Chauchard	2016	Unpacking ethnic preferences: theory and micro-Level evidence from north India
Chilton et al	2017	Reciprocity and public opposition to foreign direct investment
Clayton et al	2019	Exposure to immigration and admission preferences: evidence from France
Crowder-Meyer et al	2018	A different kind of disadvantage: candidate race, cognitive complexity, and voter choice
Eggers et al	2017	Corruption, accountability and gender: do female politicians face higher standards in public life
Franchino and Segatti	2019	Public opinion on the Eurozone fiscal union: evidence from survey experiments in Italy
Franchino and Zucchini	2014	Voting in a multidimensional space: a conjoint analysis employing valence and ideology attributes of candidates

Gallego and Marx	2016	Multi-dimensional preferences for labour market reforms
Goggin et al	2019	What goes with red and blue? Mapping partisan and ideological associations in the minds of voters
Hainmueller and Hopkins	2015	The hidden American immigration consensus: a conjoint analysis of attitudes towards immigrants
Hainmueller et al	2015	Validating vignette and conjoint survey experiments against real-world behavior
Hankinson	2018	When do renters behave like homeowners? High rent, price, anxiety, and NIMBYism
Hartman and Morse	2018	Violence, empathy and altruism: evidence from the Ivorian refugee crisis in Liberia
Hausermann et al	2019	The politics of trade-offs: studying the dynamics of welfare state reform with conjoint experiments
Heinric and Kobayashi	2017	Sanction consequences and citizen support: a survey experiment
Heinric and Kobayashi	2018	How do people evaluate foreign aid to nasty regimes?
Hemker and Rink	2017	Multiple dimensions of bureaucratic discrimination: evidence from German welfare offices
Horiuchi et al	2018	Measuring voters multidimensional policy preferences with conjoint analysis: application to Japans 2014 election
Horiuchi et al	2018	Identifying voter preferences for politicians' personal attributes: a conjoint experiment in Japan
Huff and Kertzer	2017	How the public defines terrorism
Iyengar and Westood	2015	Fear and loathing across party lines: new evidence on group polarization
Karpowitz et al	2017	How to elect more women: gender and candidate success in a field experiment
Kertzer et al	2019	How do observers assess resolve?
Kirkland, Coppock	2018	Candidate choice without party labels
Leeper, Robison	2018	More important, but for what exactly? The insignificant role of subjective issue importance in vote decisions
Li and Zeng	2017	Individual preferences for FDI in developing countries: experimental evidence from China
Liu	2018	The logic of authoritarian political selection: evidence from a conjoint experiment in China
Malhotra and Newman	2019	Explaining immigration preferences: disentangling skill and prevalence
Mares and Visconti	2019	Voting for the lesser evil: evidence from a conjoint experiment in Romania
Mummolo	2016	News from the other side: how topic relevance limits the prevalence of partisan selective exposure
Mummolo and Nall	2016	Why partisans do not sort: the constraints on political segregation

Newman and Malhotra	2018	Economic reasoning with a racial hue: is the immigration consensus purely race neutral?
Oliveros and Schuster	2018	Merit, tenure, and bureaucratic behavior: evidence from a conjoint experiment in the Dominican Republic
Ono and Burden	2018	The contingent effects of candidate sex on voter choice
Ono and Yamada	2018	Do voters prefer gender stereotypic candidates? Evidence from a conjoint survey experiment in Japan
Peterson	2017	The role of the information environment in partisan voting
Peterson and Simonovitis	2018	The electoral consequences of issue frames
Sances	2018	Ideology and vote choice in U.S. mayoral elections: evidence from Facebook surveys
Scheider	2019	Euroscepticism and government accountability in the European Union
Sen	2017	How political signals affect public support for judicial nominations: evidence from a conjoint experiment
Shafranek	2019	Political considerations in nonpolitical decisions: a conjoint analysis of roommate choice
Spilker et al	2016	Selecting partner countries for preferential trade agreements: experimental evidence from Costa Rica, Nicaragua, and Vietnam
Teele et al	2018	The ties that double bind: social roles and women's underrepresentation in politics
Vivyan and Wagner	2016	House or home? Constituent preferences over legislator effort allocation
Ward	2019	Public attitudes towards young immigrant men
Write et al	2016	Mass opinion and immigration policy in the United States: reassessing clientelist and elitist perspectives

Table A1: Conjoint Articles Published From 2014-2019.

B Constructing the Target Profile Distribution

We utilize several sources of data to construct the population distribution used in Section 2. We emphasize that ideally researchers should construct the population profile distribution before designing conjoint analysis in order to match the attributes of the population distribution with those of conjoint analysis. In the current application, we construct the population profile distribution after the conjoint analysis was conducted by Ono and Burden (2019). As a result, for almost all factors, additional *ex post* coding was needed to match the empirical data to the categories chosen by the original authors.

Here, we discuss the data source and the procedure used to produce categories matching those used in the original experiment. We use the legislators in the 115th Congress as the

Factors	Levels	Population Data Source
Age	36, 44, 52, 60, 68, 76	Daily Kos Biographical Database
Gender	Male, Female	
Race	Asian, Black, Hispanic, White	
Family	Divorced, Never married, Married (no children) Married (2 children)	The Hill People Directory
Experience	None, 4 years, 8 years, 12 years	Daily Kos Biographical Database
Expertise	Economic policy, Education, Environmental issues, Foreign policy, Health care, Public safety (crime)	Congressional Committee Assignments
Character Trait	Compassionate, Honest, Intelligent, Knowledgeable, Leadership, Empathetic	None
Immigration Policy	Favors guest worker program, opposes guest worker program	Secure America's Future Act (SAFA) Roll Call Votes
Security Policy	Strong military, Cut defense spending	Center for Security Policy Legislator Scorecard
Abortion Policy	Pro-choice, Neutral, Pro-life	National Right to Life Council Legislator Scorecard
Deficit Policy	Increase taxes, Take no action, Reduce spending	Club for Growth Legislator Scorecard

Table A2: Levels and Data Sources Used to Construct the Population Profile Distribution.

target population distribution in order to maximize our ability to reverse engineer the original factor levels. To merge disparate data sources, we use a probabilistic record linkage method, implemented via the `R` package `fastLink` (Enamorado *et al.*, 2019), with partial matching on the first and last name. Table A2 lists the data source used to build the empirical distribution for each factor. In calculating these factors, we considered only legislators who were seated via popular vote; those who were named to a seat due to a vacancy are omitted.

B.1 Demographic Factors

Age, gender and race — the three demographic factors used in the original study — were obtained from the Daily Kos 115th Congress Members Guide. The dataset contains both biographical and electoral information. Biographical information on legislators is sourced from Pew, Roll Call, news stories and Wikipedia. Data was also checked against a similar dataset available through `legislatoR` (Gobel and Munzert, 2019), an `R` package that allows queries to a database of biographical and political information on legislators from multiple countries. The details about each demographic factor is presented below.

Age. The age factor was produced by binning legislators' ages into the same age ranges as the original categories.

Gender. Legislator gender was taken directly from the data and unaltered.

Race. Racial categories closely matched those of the experiment with some notable exceptions. First, legislators that were coded as identifying as both white and Hispanic were coded as Hispanic in the joint data. Two legislators who identified as white-Portuguese American were coded as White. All Asian-American legislators were coded as Asian-American regardless of their nationality. For example, an Indian-American and Japanese American legislator would both be coded as Asian-American.

B.2 Background Factors

Background factors were constructed from four sources: the Daily Kos 115th Congress Members Guide, legislators' official Wikipedia page; the Congressional Committees dataset (Stewart and Woon, 2017); and biographical information from the People directory of TheHill.com, a digital news site.

Experience. The experience measure was created by first subtracting the first year a legislator was elected to higher office from the most recent election year, resulting in a measure of the total number of years spent in office. To calculate years served, the election year for all House members was taken as 2016—the year of the most recent House elections present in the data—while for Senators the election year in which they won current office was used. If a legislator had served previously, this interval was added to the more recent tenure. In a limited number of cases, a legislator was seated as a result of a special election. In these cases, the year of the special election is used. To map this measure onto the categories of the factor used in the original experiment (0 years, 4 years, 8 years, 12 years), we use the midpoints between each category to determine into which bin each observation falls. For example, a legislator with 1 year experience would fall to the left of the midpoint between the two nearest categories (0 and 4 years, respectively) and be assigned to the 0 years category.

Policy Expertise. The policy expertise factor is difficult to approximate with real-world data because the expertise that legislators claim during campaigns may be a matter of political expedience and may not correspond to their actual expertise. To overcome this difficulty, we used legislators' committee assignments as the basis for producing the joint distribution. Our motivation for using committee assignments is straightforward: legislators are strategic in their choice of committee assignments—or at least in their attempts to obtain assignments that would allow them to claim expertise in politically salient areas. Using the Congressional Committees dataset, we attempted to map each committee—in both the House and Senate—to a corresponding category in the original experiment. Where the committee was a poor match for all of the original categories, such as for the Ways and Means Committee, we coded a legislator's expertise as missing.

Because each legislator serves on multiple committees and our joint distribution is constructed at the legislator-level, there are multiple values possible for each legislator. To overcome this problem, for each legislator we compared the seniority rankings of each committee and assigned to that legislator the committee on which they were the most senior. Because not all committees are the same size, it is possible that a legislator could be assigned a small committee in which they were a higher rank in absolute terms but lower as a percentage of total seats. We allow for such cases because a high absolute ranking on a small committee may be used as the basis for a claim of expertise as easily as a lower ranking in a larger committee.

Party. Party was taken directly from the Daily Kos dataset and then binned into three categories: Democrat, Republican and Independent.

Favorability Rating. Because we were interested in a large pool of legislators, it was not possible to obtain favorability ratings drawn from a sufficiently large survey sample for the majority of our legislators. To overcome this, we used the vote margin in the previous election as a proxy for legislators' approval ratings in their constituency. Due to the mechanics of first-past-the-post elections, this means that the lowest level of favorability rating possible in the experiment (34%) occurs only twice and the next highest rating (43%) occurs only six times. This right-skewed distribution is a good approximation of the true favorability rating for two reasons. First, viable candidates in competitive districts must have reasonably high approval ratings with the general electorate. Second, legislators in stronghold districts are likely to have high approval ratings due to their copartisanship with the majority of their constituents.

Family Status. The original "family" factor included information on marital status and legislators' number of children. Data on both were harvested from legislators' Wikipedia pages using the `rvest` package in `R` (Wickham, 2019) wherever such fields were available. Because legislators may not wish to publicize that they are divorced or unmarried, it is possible that missingness is correlated with legislators' marital status. We attempted to address this problem by augmenting the Wikipedia data with data harvested from the People Directory of TheHill.com. For both the Wikipedia and TheHill.com fields, the names of spouses and number of children were given. In the case of multiple marriages, we coded legislators' marital status based on the status of their most recent marriage. Thus, a legislator who is currently married but was divorced in the past would be coded as married. To use this data to reconstruct the categories used in the conjoint experiment, it was necessary to bin the number of children into the original categories. Legislators with 1 or more children were binned into the "2 children" category. The number of children was disregarded if the legislator was divorced or never married because the original categories contained no information on the number of children for those marital statuses. Legislators for whom the number of children field was missing in both TheHill and Wikipedia datasets were coded as having zero children.

B.3 Policy Positions

The original data contained information on four policy dimensions: abortion (pro-life/pro-choice/neutral); immigration (in favor of/against a guest worker program); security (favors strong military/favors defense spending cuts); and deficit reduction (wants to reduce deficit through tax increases/wants to maintain current deficit/wants to reduce deficit through spending cuts). These factors were difficult to approximate with real-world data for several reasons. First, they correspond to broad issue areas, such as a legislators' stance on abortion. In such cases, real-world legislators' policy positions are likely to be driven by one or more latent dimensions that can only be estimated from voting behavior across many bills. Second, estimates of this latent dimension via voting behavior are complicated by the fact that we are restricted to considering only bills introduced in and voted on during the 115th Congress, resulting in relatively few bills that correspond to the original levels. Third, for positions that are subsets of a broader policy space—such as the desired means of deficit reduction—a bill with a proposal similar to the original levels will often include statutes related to other, similar issues. Special care thus needs to be taken to ensure that legislators' voting behavior was driven at least in part by the statute corresponding to the original levels.

Finally, while policy think tanks often provide legislator scorecards, the score space may not correspond neatly to the levels of the original data. For example, someone who is considered moderate on fiscal issues may not necessarily advocate for no deficit reduction, which is the middle category of the spending policy factor. Given these considerations, we aimed to build a reasonable first approximation using a combination of actual voting behavior. In cases where legislators' voting behavior was not available or driven by other statutes included in a bill, legislator scorecards produced by advocacy organizations and partisan policy institutes were consulted.

To facilitate consistency across the four policy factors in the original data, we used a general heuristic in deciding whether to use a bill or a legislator scorecard to approximate the experimental categories. We began by identifying legislative scorecards whose policy space closely matched the policy referenced in the original factor. If none were available, a scorecard for a more broad issue area could also be used.

Legislator scorecards are typically constructed by "scoring" legislators' votes on bills that are considered important in a particular policy space. The think tank producing the scorecard then rates each legislator according to how closely their voting behavior matches the position favored or advocated by the organization. While such scorecards are available from both conservative and liberal policy institutes, we chose only from scorecards issued by conservative organizations. This was done to ensure that a higher score was always associated with a more conservative policy position. Once the scorecard was obtained, we examined the bills used to produce each legislators' score. If a bill closely matched the original categories and was voted on in the 115th Congress, each legislators' vote was used to assign him or her a policy position.

To ensure that the policy position distribution was not driven by only a few legislators, bills meeting these criteria were only used if they were considered in both the House and Senate in similar form or the House alone.

If there existed no bill that was suitable for approximating actual legislators' values on the original factor, the legislator scorecards were used directly. To do so, legislators were binned into categories according to their numerical score and normalized to range from 0 to 1. A score of 1 was given to legislators considered strong proponents of a given policy position. If the original factor had only two categories — as in the case of the national security factor, for example — legislators with a value at or below the midpoint (a score of 0.5) were given the value corresponding to the liberal position, while those above were assigned to the more conservative category. In cases where the original factor had three categories, legislators with scores from 0 to 0.4 were given the liberal position, those with scores between 0.4 and 0.6 were given the moderate position, and those at 0.6 or above were given the conservative position.

Given these decision rules, we selected legislator scorecards for the abortion, deficit spending and national security factors and a single bill for the immigration factor. Below we describe the data source and coding rules used to produce categories matching those of the original study.

Abortion. Data for abortion position was based on the National Right to Life Council (NRLC) legislator scorecard. While the NRLC is a conservative, pro-life organization, similar scorecards from left-leaning organizations produce similar distributions owing to the highly polarized nature of abortion policy in the United States. The NRLC score is a 0-1, with 1 corresponding to a strongly pro-life legislator and a zero a strongly pro-choice legislator. Legislators between 0 and 0.4 were coded as pro-choice; those between 0.4 and 0.6 were coded as neutral; and those with a score greater than 0.6 were coded as pro-life. Predictably, there are only three neutral legislators according to this criteria, and the distribution is almost perfectly correlated with partisanship.

Immigration. Using the decision rules above, we chose a bill rather than a legislator scorecard, selecting the Secure America's Future Act (SAFA) to serve as our proxy for legislators' positions on the guest worker program. SAFA included in it a provision to abolish the existing H-2B guest worker program with a less generous and more restricted policy dubbed H-2C. While other provisions of the bill were politically important—such as protection for so-called Dreamers—the guest worker program was a prominent feature of the bill. This is the only case for the 115th Congress where an immigration bill including a program closely corresponding to the Ono and Burden levels was both introduced and voted on. Because a vote in favor of SAFA was a vote for the H-2C guest worker program (and thus against the existing H-2B system), a vote of yes was coded as opposition to a guest worker program and a vote of no as favoring a guest worker system. Legislators who did not vote were marked as missing.

National Security. Data for the military spending factor was based on the Center for Security Policy (CfSP) legislator scorecard. The CfSP scored a total of 19 bills to produce a 0-1 score where higher values represents more "pro-security" voting behavior. Legislators with a score less than 0.5 were binned into the "cut military spending" category while those above the cutoff were binned into the "maintain strong defense" category.

Budget. Data for the budget position was based on the Club for Growth's legislator scorecard. The Club for Growth, a conservative organization, rated legislators according to the organization's pro-deficit-reduction position, producing a scorecard with a range of 0-1 where higher values indicate more support for deficit reduction. Legislators with a score between 0 and 0.4 were given a value of "reduce deficit through tax increases", a liberal position; those with a score between 0.4 and 0.6 were given the moderate position "do not reduce deficit now"; and those with a score greater than 0.6 were given the conservative position "reduce deficit through spending cuts".

C Robustness to the Choice of Profile Distribution

The original experiment design of Ono and Burden (2019) considers hypothetical political candidates. Thus, the ideal target profile distribution would be the real-world distribution of the attributes summarized in Table 1 for all candidates, not only sitting legislators. Unfortunately, because the original experiment was not designed with fidelity to the real-world distribution in mind, there are many factors for which it is practically impossible to gather corresponding real-world distributions of all candidates (e.g., character traits of candidates and favorability rating). As a result, in Section 3.4.1, we set our main target profile distribution to be politicians in the 115th Congress, for whom we were able to collect real-world distributions for most factors (as described in detail in Section B).

In this section, we use the model-based approach to investigate the robustness of the pAMCE estimates based on the 115th Congress (reported in Section 5.1) to alternative profile distributions based on candidate-level data. Although these candidate-level data do not include information for all factors used in the original experiment, we can incorporate a number of relevant candidates' characteristics. In particular, we rely on two publicly available datasets on candidate characteristics, DIME data set (Bonica, 2015) and the Reflective Democracy (RefDem) dataset,[9] to improve the profile distributions of three demographic variables. In addition, we use our substantive knowledge to explore different theoretically relevant profile distributions on policy dimensions.

Data. For the DIME data (Bonica, 2015), we consider all major party candidates that ran for Congress in the 2014 general election, the last year of the dataset's coverage. This yields 1148 candidates. In the RefDem data, we consider all major party candidates who ran for the House or Senate in 2018. This yields 911 unique candidates. We use the DIME data to

[9]This dataset is available at `https://wholeads.us/resources/for-researchers/`

replace the marginal distribution for number of years in office (`Experience`), and the RefDem data to replace the marginal distributions of race and gender (`Race, Gender`). Figure A1 visualizes this new distribution. Comparing to Figure 2, the two most notable differences are the larger proportion of white candidates — particularly for Democrats — and the higher incidence of candidates with "no experience," a natural consequence of considering challengers. In a second profile distribution, we also augment these new demographic marginal distributions with changes to the marginal distributions of policy positions, making them more extreme to reflect the fact that winning candidates are systematically more moderate than losing candidates. Figure A2 visualizes this second new distribution. On all policy dimensions, we made the policy positions slightly more extreme, and it can be clearly seen in positions on `Deficit`. Our goal is to assess the robustness of the pAMCE based on the 115th Congress to these alternative target profile distributions.

Results. Figure A3 shows the pAMCE estimates of being female in Ono and Burden (2019) with three different profile distributions. The first row represents the pAMCE estimates reported in Section 5.1. The second and third rows show the results based on the first alternative profile distribution (with new marginals for three demographic factors) and the second alternative distribution (with improved demographic factors + more extreme policy positions), respectively. Although the difference between the uAMCE (black estimates) and pAMCE estimates for Republican (red) and Democrats (blue) are large, the change across different alternative target profile distributions is small. This suggests that even though the target profile distribution based on the 115th Congress is different from the ideal political candidate-level profile distribution, the pAMCE estimates based on the 115th Congress are robust to theoretically relevant changes in profile distributions that better reflect candidate-level data.

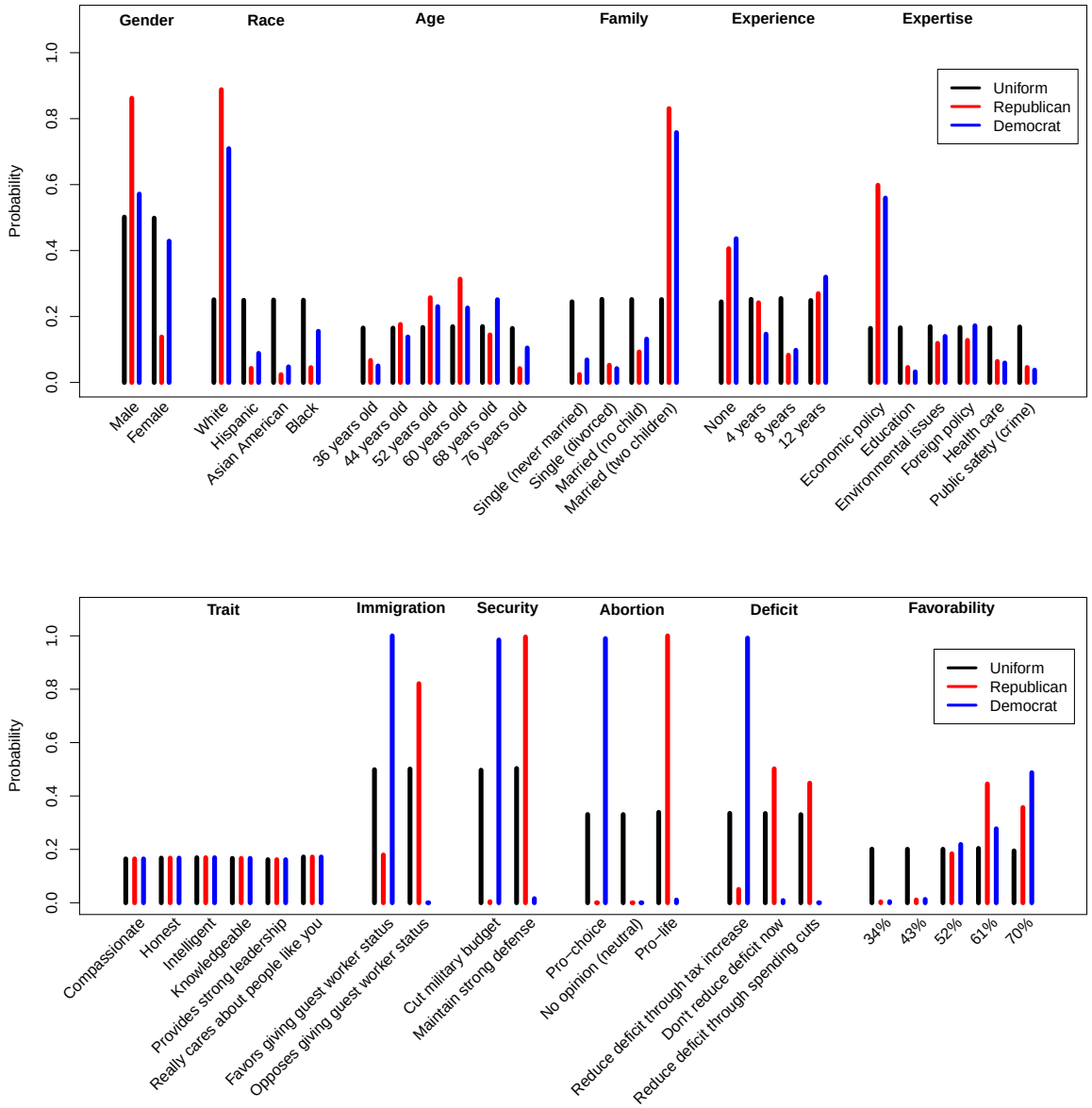

Figure A1: Experimental and Target Profile Distributions of Factors in Ono and Burden (2019) improved by candidate-level data sets.

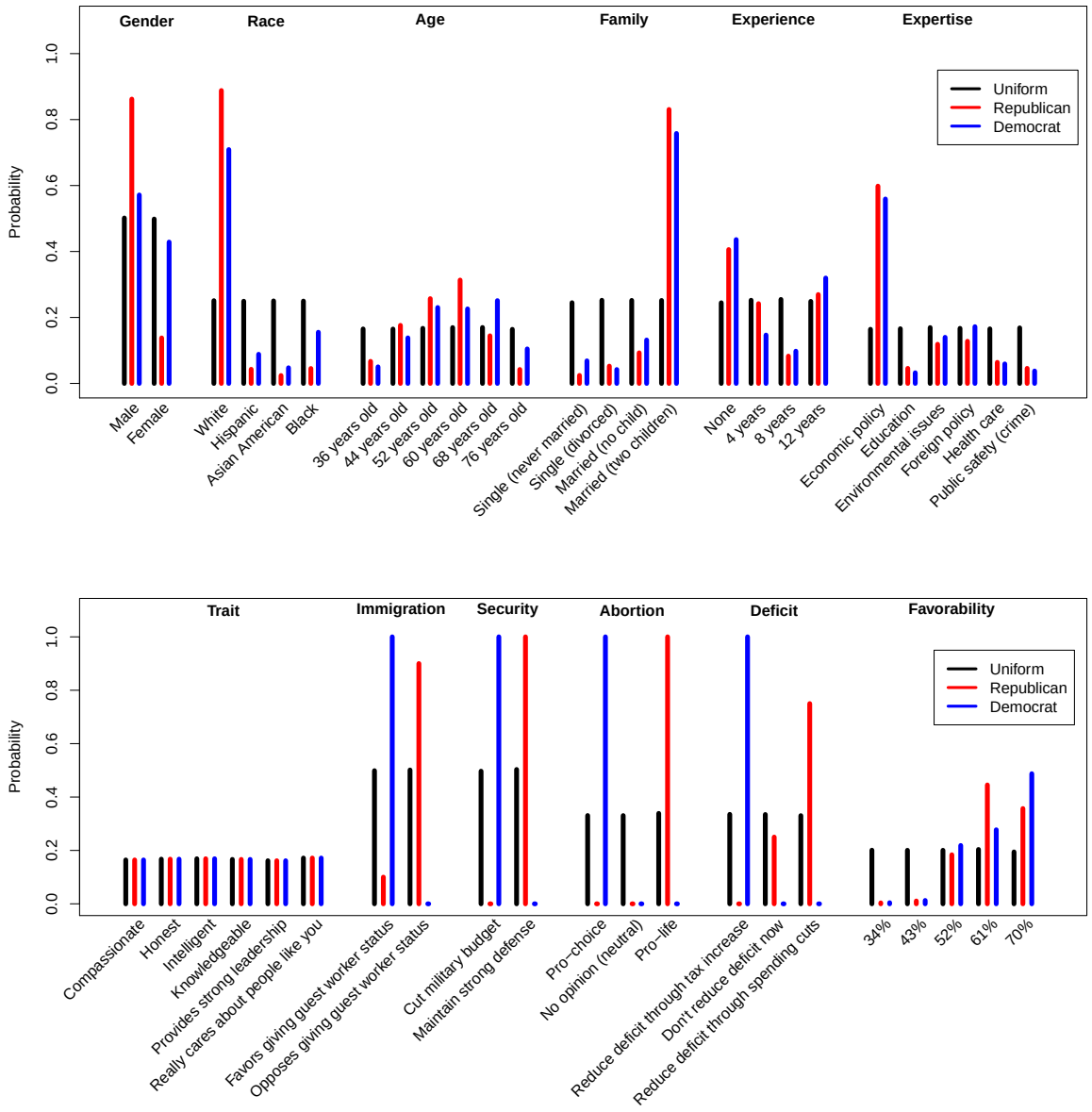

Figure A2: Experimental and Target Profile Distributions of Factors in Ono and Burden (2019) improved by candidate-level data sets and augmented with counterfactual more extreme policy positions.

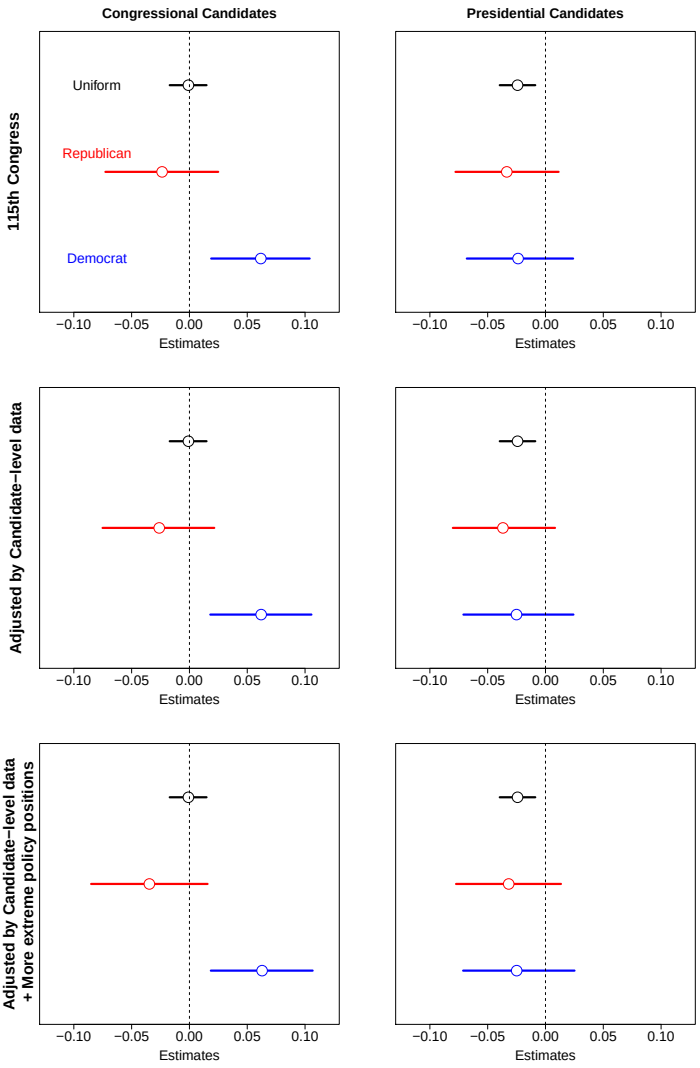

Figure A3: Estimates of the pAMCEs of Being Female in Ono and Burden (2019) with three different profile distributions. The first row represents the pAMCE estimates reported in Section 5.1. The second and third rows show results based on the first alternative profile distribution (with improved three demographic variables) and the second alternative profile distribution (with improved three demographic variables + more extreme policy positions).

D Proofs

D.1 Consistency of Weighted Difference-in-Means

Here, we formally prove that the proposed weighted difference-in-means estimator is consistent for the pAMCE under any randomization distribution that satisfies a set of positivity conditions.

THEOREM 1 (CONSISTENCY OF THE WEIGHTED DIFFERENCE-IN-MEANS ESTIMATOR) The weighted difference-in-means estimator defined in equation (5) is consistent for the pAMCE,

$$\widehat{\tau}^*_\ell(t_1, t_0) \xrightarrow{p} \tau^*_\ell(t_1, t_0), \tag{A1}$$

for any randomization distribution $\Pr^{\mathtt{R}}(\cdot)$ that satisfies the following positivity conditions,

$$\begin{aligned}
&\Pr^{\mathtt{R}}(T_{ijk\ell} = t_1 \mid (\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k}) = \mathbf{t}) > 0 \\
&\Pr^{\mathtt{R}}(T_{ijk\ell} = t_0 \mid (\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k}) = \mathbf{t}) > 0 \\
&\Pr^{\mathtt{R}}((\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k}) = \mathbf{t}) > 0
\end{aligned}$$

for all $\mathbf{t} \in \mathcal{T}^*$ where $\mathcal{T}^*$ is the support of $\Pr^*(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})$.

The positivity requirement guarantees that all possible profile combinations under the target population distribution have non-zero probabilities under the randomization distribution. The proposed three designs satisfy this requirement.

Proof. We want to prove that the following estimator is consistent for the pAMCE.

$$\widehat{\tau}^*_\ell(t_1, t_0) \;=\; \frac{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}}{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell}} - \frac{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K \mathbf{1}\{T_{ijk\ell} = t_0\} w_{ijk\ell} Y_{ijk}}{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K \mathbf{1}\{T_{ijk\ell} = t_0\} w_{ijk\ell}},$$

where the weights are defined as,

$$w_{ijk\ell} \;=\; \frac{1}{\Pr^{\mathtt{R}}(T_{ijk\ell} \mid \mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})} \times \frac{\Pr^*(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})}{\Pr^{\mathtt{R}}(\mathbf{T}_{ijk,-\ell}, \mathbf{T}_{i,-j,k})}.$$

We first focus on the numerator. By the law of large number, we can obtain

$$\frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk} \xrightarrow{p} \mathbb{E}[\mathbb{E}^{\mathtt{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}]],$$

where the first expectation is over a random sample of respondents i and task positions k, and the second expectation is over randomization of treatment assignment. We focus on the expression inside the first expectation.

$$\begin{aligned}
&\mathbb{E}^{\mathtt{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}] \\
&= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \mathbb{E}^{\mathtt{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk} \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}] \Pr^{\mathtt{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \\
&= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \big\{\mathbb{E}^{\mathtt{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}] \\
&\qquad \times \Pr^{\mathtt{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\} \\
&= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \{Y_{ijk}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) w_{ijk\ell}
\end{aligned}$$

$$\times \mathbb{E}^{\mathrm{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}]\Pr^{\mathrm{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\}$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \{Y_{ijk}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) w_{ijk\ell}$$

$$\times \Pr^{\mathrm{R}}(T_{ijk\ell} = t_1 \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^{\mathrm{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\}$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \big\{Y_{ijk}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^{*}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\},$$

where the first equality follows from the rule of conditional expectation, the second from the definition of potential outcomes, the third from the fact that potential outcomes and weights are fixed within the second expectation, the fourth from the definition of probability, and the final equality from the definition of weights we propose.

Due to the no profile-order assumption, we can average over j.

$$\frac{1}{NJK} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}$$

$$\xrightarrow{p} \mathbb{E}\left\{ \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} Y_{ik}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^{*}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \right\}.$$

For the denominator, we again use the law of large number.

$$\frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} \xrightarrow{p} \mathbb{E}[\mathbb{E}^{\mathrm{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell}]].$$

Focusing on the expression inside the second expectation.

$$\mathbb{E}^{\mathrm{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell}]$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \mathbb{E}^{\mathrm{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}] \Pr^{\mathrm{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \big\{w_{ijk\ell} \mathbb{E}^{\mathrm{R}}[\mathbf{1}\{T_{ijk\ell} = t_1\} \mid \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}] \Pr^{\mathrm{R}}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\}$$

$$= \sum_{(\mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})} \Pr^{*}(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})$$

$$= 1$$

where the first equality follows from the rule of conditional expectation, the second from the fact that weights are fixed within the second expectation, the third from the definition of probability and weights, and the final equality also from the definition of probability.

Therefore, we obtain,

$$\frac{1}{NJK} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} \xrightarrow{p} 1,$$

which completes the proof. □

D.2 Consistency of Simple Difference-in-Means Under Marginal Population Randomization

Under the assumption of no three-way or higher-order interactions, the following simple difference-in-means is consistent for the pAMCE after randomizing profiles according to the

marginal population randomization design (equation (3)).

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}} \xrightarrow{p} \tau^*_\ell(t_1,t_0)$$

Proof. Under the assumption of no three-way or higher-order interactions, the potential outcomes can be modeled as a function of all main terms and all two-way interactions between factors $(\mathbf{t}_{ijk}, \mathbf{t}_{i,-j,k})$ in the following fashion,

$$\begin{aligned} Y_{ik}(\mathbf{t}_{ijk},\mathbf{t}_{i,-j,k}) = \quad & \widetilde{\alpha}_{ik} + \sum_{j=1}^{J}\sum_{\ell=1}^{L}\mathbf{X}_{ijk\ell}^{\top}\widetilde{\beta}_{ik,j\ell} \\ & + \sum_{j=1}^{J}\sum_{\ell=1}^{L}\sum_{\ell'\neq\ell}(\mathbf{X}_{ijk\ell}\bigtimes\mathbf{X}_{ijk\ell'})^{\top}\widetilde{\gamma}_{ik,j\ell\ell'} + \sum_{j=1}^{J}\sum_{j'\neq j}\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L}(\mathbf{X}_{ijk\ell}\bigtimes\mathbf{X}_{ij'k\ell'})^{\top}\widetilde{\delta}_{ik,jj'\ell\ell'} + \tilde{\epsilon}_{ijk} \end{aligned}$$

where $\mathbf{X}_{ijk\ell}$ is a vector of $(D_\ell - 1)$ dummy variables for the levels of $t_{ijk\ell}$ excluding the baseline level, $\times$ represents the cartesian product operator, e.g., $(\mathbf{X}_{ijk\ell}\times\mathbf{X}_{ijk\ell'})^{\top}\widetilde{\gamma}_{ik,j\ell\ell'} = \sum_{d=1}^{D_\ell-1}\sum_{d'=1}^{D_{\ell'}-1}X_{ijk\ell d}X_{ijk\ell' d'}\widetilde{\gamma}_{ik,j\ell d\ell' d'}$, and $\tilde{\epsilon}_{ijk}$ is the error term. Then,

$$\begin{aligned} & \sum_{(\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})}\{Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k}) - Y_{ijk}(t_0,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})\}\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k}) \\ =& \sum_{(\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})}\{Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k}) - Y_{ijk}(t_0,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})\}\prod_{\ell'\neq\ell}\Pr^*(\mathbf{T}_{ijk\ell'}=\mathbf{t}_{ijk\ell'})\prod_{\ell''}\Pr^*(\mathbf{T}_{i,-j,k,\ell''}=\mathbf{t}_{i,-j,k,\ell''}) \end{aligned}$$

where the second expression only uses marginal distributions of each factor separately. Therefore, under the assumption of no three-way or higher-order interaction, the approximation of the joint distribution by the multiplication of each marginal distribution produces the same point estimate as the one based on the exact joint distribution.

Therefore, under the assumption of no three-way or higher-order interaction, weights are simplified as:

$$w_{ijk\ell} = \frac{1}{\Pr^{\mathtt{R}}(T_{ijk\ell}\mid\mathbf{T}_{ijk,-\ell},\mathbf{T}_{i,-j,k})}\times\frac{\prod_{\ell'\neq\ell}\Pr^*(\mathbf{T}_{ijk\ell'}=\mathbf{t}_{ijk\ell'})\prod_{\ell''}\Pr^*(\mathbf{T}_{i,-j,k,\ell''}=\mathbf{t}_{i,-j,k,\ell''})}{\Pr^{\mathtt{R}}(\mathbf{T}_{ijk,-\ell},\mathbf{T}_{i,-j,k})}.$$

When the marginal population randomization design is used,

$$\begin{aligned} w^{\mathtt{Mar}}_{ijk\ell} &= \frac{1}{\Pr^*(T_{ijk\ell})}\times\frac{\prod_{\ell'\neq\ell}\Pr^*(\mathbf{T}_{ijk\ell'}=\mathbf{t}_{ijk\ell'})\prod_{\ell''}\Pr^*(\mathbf{T}_{i,-j,k,\ell''}=\mathbf{t}_{i,-j,k,\ell''})}{\prod_{\ell'\neq\ell}\Pr^*(\mathbf{T}_{ijk\ell'}=\mathbf{t}_{ijk\ell'})\prod_{\ell''}\Pr^*(\mathbf{T}_{i,-j,k,\ell''}=\mathbf{t}_{i,-j,k,\ell''})} \\ &= \frac{1}{\Pr^*(T_{ijk\ell})}. \end{aligned}$$

Therefore, the weighted difference-in-means becomes the simple difference-in-means.

$$\begin{aligned} & \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}w^{\mathtt{Mar}}_{ijk\ell}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}w^{\mathtt{Mar}}_{ijk\ell}} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}w^{\mathtt{Mar}}_{ijk\ell}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}w^{\mathtt{Mar}}_{ijk\ell}} \\ =& \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}}. \end{aligned}$$

Based on Theorem 1,

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}}-\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}Y_{ijk}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}}\xrightarrow{p}\tau_\ell^*(t_1,t_0),$$

which completes the proof. □

D.3 Optimality of the Mixed Randomization Design

Here, we investigate the Neyman variance of the following inverse probability weighting estimator that corresponds to the weighted difference-in-means estimator (equation (5)).

$$\widehat{\tau}_\ell^{\texttt{IPW}}(t_1,t_0) \;=\; \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}-\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}\{T_{ijk\ell}=t_0\}w_{ijk\ell}Y_{ijk}, \tag{A2}$$

We show that the mixed randomization design minimizes the Neyman variance when there is a single main factor and the assumption of no cross-profile interactions holds.

Proof. We can write the variance of the estimator as

$$\begin{aligned}
&\mathrm{Var}(\tau_\ell^{\texttt{IPW}}(t_1;t_0))\\
&=\mathrm{Var}\bigg(\frac{1}{NK}\sum_{i,k}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}\bigg)+\mathrm{Var}\bigg(\frac{1}{NK}\sum_{i,k}\mathbf{1}\{T_{ijk\ell}=t_0\}w_{ijk\ell}Y_{ijk}\bigg)\\
&\quad-2\mathrm{Cov}\left(\frac{1}{NK}\sum_{i,k}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk},\frac{1}{NK}\sum_{i,k}\mathbf{1}\{T_{ijk\ell}=t_0\}w_{ijk\ell}Y_{ijk}\right).
\end{aligned}$$

First, we focus on the first term.

$$\mathrm{Var}\bigg(\frac{1}{NK}\sum_{i,k}\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}\bigg)=\frac{1}{N^2K^2}\sum_{i,k}\mathrm{Var}\bigg(\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}\bigg),$$

because treatments are independently randomized across individuals i and task positions k. Focusing on the expression inside the summation,

$$\begin{aligned}
&\mathrm{Var}\bigg(\mathbf{1}\{T_{ijk\ell}=t_1\}w_{ijk\ell}Y_{ijk}\bigg)\\
&=\mathrm{Var}\bigg(\sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}}\mathbf{1}\{T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k}\}Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})\\
&\qquad\qquad\times\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})}{\Pr^{\mathrm{R}}(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})}\bigg)\\
&=\sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}}Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})^2\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})^2}{\Pr^{\mathrm{R}}(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})^2}\\
&\qquad\times\mathrm{Var}\bigg(\mathbf{1}\{T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k}\}\bigg)\\
&+\sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}}\sum_{\substack{\mathbf{t}'_{ijk,-\ell},\\ \mathbf{t}'_{i,-j,k}}}Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})Y_{ijk}(t_1,\mathbf{t}'_{ijk,-\ell},\mathbf{t}'_{i,-j,k})\\
&\qquad\times\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})}{\Pr^{\mathrm{R}}(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})}\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}'_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}'_{i,-j,k})}{\Pr^{\mathrm{R}}(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}'_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}'_{i,-j,k})}\\
&\qquad\times\mathrm{Cov}\bigg(\mathbf{1}\{T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k}\},\mathbf{1}\{T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}'_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}'_{i,-j,k}\}\bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} Y_{ijk}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})^2 \frac{\Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})^2}{\Pr^{\mathtt{R}}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})^2} \\
&\qquad \times \Pr^{\mathtt{R}}\Big(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}\Big) \times \Big\{1 - \Pr^{\mathtt{R}}\Big(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}\Big)\Big\} \\
&- \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} \sum_{\substack{\mathbf{t}'_{ijk,-\ell},\\ \mathbf{t}'_{i,-j,k}}} Y_{ijk}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) Y_{ijk}(t_1, \mathbf{t}'_{ijk,-\ell}, \mathbf{t}'_{i,-j,k}) \\
&\qquad \times \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}'_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}'_{i,-j,k}),
\end{aligned}
$$

where the first equality follow from the definition of potential outcomes and weights we propose, the second from the definition of variance, the third from the definition of Bernoulli distribution. Therefore,

$$
\begin{aligned}
&\mathrm{Var}\Big(\frac{1}{NK}\sum_{i,k}^{n} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}\Big) \\
&= \frac{1}{N^2K^2} \sum_{i,k} \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} Y_{ijk}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k})^2 \frac{\Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})^2}{\Pr^{\mathtt{R}}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})} \\
&\qquad \times \Big\{1 - \Pr^{\mathtt{R}}\Big(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}\Big)\Big\} \\
&- \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} \sum_{\substack{\mathbf{t}'_{ijk,-\ell},\\ \mathbf{t}'_{i,-j,k}}} Y_{ijk}(t_1, \mathbf{t}_{ijk,-\ell}, \mathbf{t}_{i,-j,k}) Y_{ijk}(t_1, \mathbf{t}'_{ijk,-\ell}, \mathbf{t}'_{i,-j,k}) \\
&\qquad \times \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}'_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}'_{i,-j,k}),
\end{aligned}
$$

where the second term does not contain expressions related to $\Pr^{\mathtt{R}}()$ and hence it is the same for any randomized design.

Next, we focus on the third term of the variance.

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{NK}\sum_{i,k} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}, \frac{1}{NK}\sum_{i,k} \mathbf{1}\{T_{ijk\ell} = t_0\} w_{ijk\ell} Y_{ijk}\right) \\
&= \frac{1}{N^2K^2} \sum_{i,k} \mathrm{Cov}\Big(\mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}, \mathbf{1}\{T_{ijk\ell} = t_0\} w_{ijk\ell} Y_{ijk}\Big) \\
&= -\frac{1}{N^2K^2} \sum_{i,k} \Bigg\{ \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} \big\{Y_{ijk}(T_{ijk\ell} = t_1, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\} \\
&\qquad \times \sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} \big\{Y_{ijk}(T_{ijk\ell} = t_0, \mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k}) \Pr^*(\mathbf{T}_{ijk,-\ell} = \mathbf{t}_{ijk,-\ell}, \mathbf{T}_{i,-j,k} = \mathbf{t}_{i,-j,k})\big\} \Bigg\},
\end{aligned}
$$

where the first equality comes from the fact that treatments are independently randomized across individuals i and task positions k, and the second from the definition of covariance. Because all the expressions are not related to $\Pr^{\mathtt{R}}()$, this covariance is the same for any randomized designs.

We now solve the minimization problem of the Neyman variance with respect to $\Pr^{\mathtt{R}}()$. To compare alternative experimental designs, we average over the potential outcomes unknown to researchers.

$$
\mathbb{E}_{Y(\mathbf{t})}\left[\mathrm{Var}\Big(\frac{1}{NK}\sum_{i,k}^{n} \mathbf{1}\{T_{ijk\ell} = t_1\} w_{ijk\ell} Y_{ijk}\Big)\right]
$$

$$= \frac{1}{N^2K^2}\sum_{i,k}\sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}} \mathbb{E}_{Y(\mathbf{t})}[Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})^2]\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})^2}{\Pr^{\mathtt{R}}(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})}$$
$$\times\left\{1-\Pr^{\mathtt{R}}\left(T_{ijk\ell}=t_1,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k}\right)\right\}$$
$$-\sum_{\substack{\mathbf{t}_{ijk,-\ell},\\ \mathbf{t}_{i,-j,k}}}\sum_{\substack{\mathbf{t}'_{ijk,-\ell},\\ \mathbf{t}'_{i,-j,k}}} \mathbb{E}_{Y(\mathbf{t})}[Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})Y_{ijk}(t_1,\mathbf{t}'_{ijk,-\ell},\mathbf{t}'_{i,-j,k})]$$
$$\times\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}_{i,-j,k})\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}'_{ijk,-\ell},\mathbf{T}_{i,-j,k}=\mathbf{t}'_{i,-j,k}),$$

where $\mathbb{E}_{Y(\mathbf{t})}$ is the expectation over the uniform distribution of the potential outcomes table. Therefore, $\mathbb{E}_{Y(\mathbf{t})}[Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})^2]$ and $\mathbb{E}_{Y(\mathbf{t})}[Y_{ijk}(t_1,\mathbf{t}_{ijk,-\ell},\mathbf{t}_{i,-j,k})Y_{ijk}(t_1,\mathbf{t}'_{ijk,-\ell},\mathbf{t}'_{i,-j,k})]$ are both constants. In addition, to compare experimental designs, we can remove all the terms that don't have $\Pr^{\mathtt{R}}()$. Taken together, we can focus on the following minimization problem under the assumption of no cross-profile interactions.

$$\min_{\Pr^{\mathtt{R}}()}\sum_{d=0}^{D_\ell-1}\sum_{\mathbf{t}_{ijk,-\ell}}\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})^2}{\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})}\quad\text{s.t.}\quad\sum_{d=0}^{D_\ell-1}\sum_{\mathbf{t}_{ijk,-\ell}}\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})=1.$$

Then, by using the Lagrange multiplier, we can solve:

$$\min_{\Pr^{\mathtt{R}}()} L$$
$$\text{where}\quad L=\sum_{d=0}^{D_\ell-1}\sum_{\mathbf{t}_{ijk,-\ell}}\frac{\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})^2}{\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})}+\lambda\left(\sum_{d=0}^{D_\ell-1}\sum_{\mathbf{t}_{ijk,-\ell}}\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})-1\right)$$
$$\text{and}\quad\lambda>0.$$

Therefore,

$$\frac{\partial L}{\partial\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})}=0\iff\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})=\frac{1}{\sqrt{\lambda}}\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell}).$$

In addition,

$$\sum_{d=0}^{D_\ell-1}\sum_{\mathbf{t}_{ijk,-\ell}}\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})=1\iff\sqrt{\lambda}=D_\ell.$$

Hence, the optimal randomization distribution is the mixed randomization design.

$$\Pr^{\mathtt{R}}(T_{ijk\ell}=t_d,\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell})=\frac{1}{D_\ell}\Pr^*(\mathbf{T}_{ijk,-\ell}=\mathbf{t}_{ijk,-\ell}),$$

which completes the proof. □

E Diagnostic Tools for Model-based Analysis

Here, we introduce a set of diagnostic tools that are designed to help researchers assess the validity of modeling assumptions. These tools are essential for a successful implementation of the proposed model-based exploratory analysis.

Specification test. We first introduce a specification test that assesses the validity of all the modeling assumptions as a whole under the uniform randomization design. The idea is that if the modeling assumptions were violated, the estimated uAMCE using the model-based approach would differ from the simple difference-in-means estimator, which is unbiased under

the uniform randomization design. We test whether the difference in the two estimates are statistically distinguishable from zero using bootstrap.

If the model-based estimate of the uAMCE is significantly different from the difference-in-means estimate, at least one modeling assumption is likely to be violated. Because we rely on the same modeling assumptions when estimating the pAMCE, it is likely that a model-based estimate of the pAMCE is also biased. This diagnostic tool can be implemented with or without regularization. We caution that rejecting the null hypothesis of no difference does not tell us which modeling assumption is violated — the absence of higher-order interaction or the absence of strong regularization bias.

Regularization bias. The proposed regularization procedure given in Section 4.2.3 uses cross-fitting to minimize the possible bias due to incorrectly shrinking coefficients to zero. In practice, however, one should check how regularization affects the estimate of the pAMCE. We suggest examining the bootstrap distribution of the estimated pAMCE separately for each factor. When a regularization bias is substantial, the bootstrap distribution often differs significantly from the normal distribution.

F Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed methodology. Specifically, we examine the following three aspects: the relative efficiency of the mixed randomization design over the uniform and marginal population randomization designs, the bias-variance tradeoff of the regularization approach, and the advantage of the design-based confirmatory analysis over the model-based exploratory analysis.

F.1 The Setup

To make simulation settings realistic, we utilize the data from the conjoint analysis about attitudes toward immigrants (Hainmueller *et al.*, 2015). The original study has the following seven factors where the number of levels is shown in parentheses; `Age` (4), `Education` (3), `Gender` (2), `Integration` (4), `Language` (4), `Origin` (8), `Year since arrival` (4). To construct the population distribution of immigrant profiles, we follow Hainmueller *et al.* (2015) and use the information from the actual referendums conducted in Switzerland, giving us the marginal distribution of each factor (see Table A3). Finally, we use the following linear utility model as the true data generating process,

$$
\begin{aligned}
\widetilde{Y}_{ijk} = \;& 0.1 + \mathbf{X}_{ijk,\texttt{Gen}}^{\top}\widetilde{\beta}_{\texttt{Gen}} + \mathbf{X}_{ijk,\texttt{Ori}}^{\top}\widetilde{\beta}_{\texttt{Ori}} + \mathbf{X}_{ijk,\texttt{Age}}^{\top}\widetilde{\beta}_{\texttt{Age}} + \mathbf{X}_{ijk,\texttt{Year}}^{\top}\widetilde{\beta}_{\texttt{Year}} + \mathbf{X}_{ijk,\texttt{Edu}}^{\top}\widetilde{\beta}_{\texttt{Edu}} + \mathbf{X}_{ijk,\texttt{Int}}^{\top}\widetilde{\beta}_{\texttt{Int}} \\
& + \mathbf{X}_{ijk,\texttt{Lan}}^{\top}\widetilde{\beta}_{\texttt{Lan}} + (\mathbf{X}_{ijk,\texttt{Year}} \bigtimes \mathbf{X}_{ijk,\texttt{Edu}})^{\top}\widetilde{\gamma}_{\texttt{Year,Edu}} + (\mathbf{X}_{ijk,\texttt{Edu}} \bigtimes \mathbf{X}_{ijk,\texttt{Lan}})^{\top}\widetilde{\gamma}_{\texttt{Edu,Lan}} \\
& + (\mathbf{X}_{ijk,\texttt{Edu}} \bigtimes \mathbf{X}_{ij'k,\texttt{Edu}})^{\top}\widetilde{\delta}_{\texttt{Edu,Edu}}, \\
\Pr(Y_{ijk} = 1 \mid \mathbf{X}_{ijk}, \mathbf{X}_{ij'k}) \;=\; & \left(\widetilde{Y}_{ijk} - \widetilde{Y}_{ij'k}\right) + 0.5,
\end{aligned}
$$

Factors	Levels
`Gender`	Male (0.69), Female (0.31)
`Origin`	Netherlands (0.01), Germany (0.18), Austria (0.04), Italy (0.21), Turkey (0.21), Croatia (0.05), Former Yugoslavia (0.23), Bosnia-Herzegovina (0.07)
`Age`	21 years (0.30), 30 years (0.21), 41 years (0.33), 55 years (0.16)
`Years since arrival`	14 years (0.26), 20 years (0.30), 29 years (0.14), Born in CH (0.30)
`Education`	Primary school (0.31), High school (0.60), University (0.09)
`Language`	Adequate (0.03), Good (0.09), Perfect (0.88)
`Integration`	Assimilated (0.37), Integrated (0.33), Indistinguishable (0.08), Familiar with Swiss traditions (0.22)

Table A3: Factors, Levels, and Each Probability Used in Hainmueller *et al.* (2015). The factors were constructed in order to match the categories of the leaflets on which actual immigrants' characteristics were printed.

We choose the values of the coefficients such that there are substantial interaction effects, making the comparison of different methods clear.[10]

We compare four approaches, each of which corresponds to a different combination of an experimental design and its corresponding estimator; (1) Mixed Design: the mixed randomization design (equation (4)) and its corresponding weighted difference-in-means estimator (equation (5)), (2) Population Design: the marginal population randomization design (equation (2)) and its corresponding difference-in-means estimator (equation (7)), (3) Reg-regression: the uniform randomization and the regularized regression estimator (equation (14)), and (4) Regression: the uniform randomization and the non-regularized regression estimator (equation (11)). For Mixed Design, we specify one main factor of interest. For each simulation, the results reported here average over the results for each of the seven factors. Using a total of 1000 Monte Carlo simulations, we compute the bias, standard error, and root mean of squared error (RMSE) of each estimator as well as the coverage of 95% confidence intervals. We let the sample size vary from 1000 to 8000, i.e., $\{1000, 2000, 4000, 6000, 8000\}$.

F.2 The Results

Figure A4 presents the results. First, as we expect from Theorem 1, both Mixed Design and Population Design induce little bias (see the upper left plot). The correctly specified Regression also suffers from little bias, whereas Reg-Regression has also little bias due to its flexible two-way interaction model. Second, in terms of statistical efficiency, because Mixed

[10] Specifically, we set $\widetilde{\beta}_{\texttt{Gen}} = -0.01$, $\widetilde{\beta}_{\texttt{Ori}} = (0, 0, 0, 0, 0, -0.002, -0.002)$, $\widetilde{\beta}_{\texttt{Age}} = (-0.005, -0.01, -0.01)$, $\widetilde{\beta}_{\texttt{Year}} = (0, 0.01, 0.01)$, $\widetilde{\beta}_{\texttt{Edu}} = (0.005, 0.02)$, $\widetilde{\beta}_{\texttt{Int}} = (0, -0.01, 0.01)$, $\widetilde{\beta}_{\texttt{Lan}} = (0.005, 0.01)$, $\widetilde{\gamma}_{\texttt{Year,Edu}} = (0.005, 0.005, 0.005, 0.01, 0.01, 0.01)$, $\widetilde{\gamma}_{\texttt{Edu,Lan}} = (0.005, 0.005, 0.01, 0.01)$, and $\widetilde{\delta}_{\texttt{Edu,Edu}} = (-0.0024, -0.0048, -0.0024, -0.0048)$.

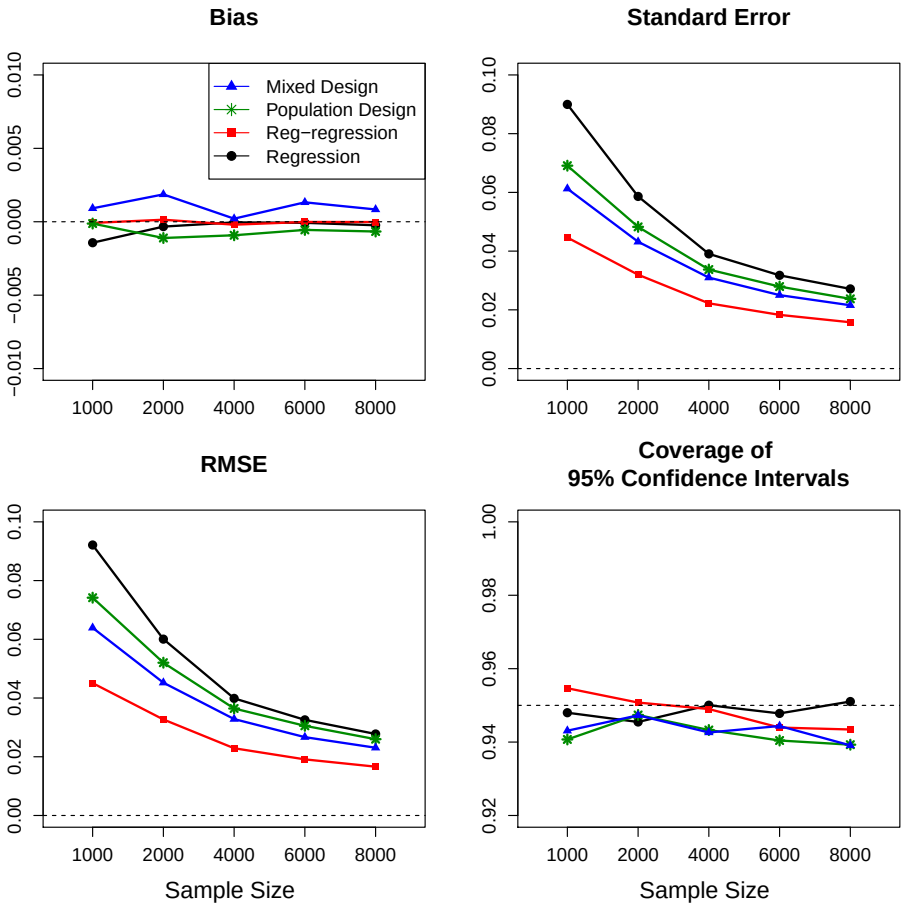

Figure A4: Comparison of Four Approaches in terms of Bias, Standard Error, RMSE, and the Coverage of 95% Confidence Intervals. We evaluate (1) the mixed randomization design and its corresponding weighted difference-in-means estimator (Mixed Design, blue square), (2) the joint population randomization design and its corresponding simple difference-in-means estimator (Population Design, green diamond), (3) the uniform randomization design and the regularized regression estimator (Reg-regression, red star), and (4) the uniform randomization design and the non-regularized regression estimator (Regression, black circle).

Design focuses on only one factor at a time, it has smaller standard errors than Population Design (see the upper right plot). Comparing the two model-based estimators, Reg-regression has smaller standard errors than Regression. The efficiency gain of Reg-regression is achieved by collapsing indistinguishable levels. In fact, this simulation shows that Reg-regression can achieve standard errors even smaller than the design-based confirmatory analysis when there are a lot of redundant levels. However, in some applications like Ono and Burden (2019), the design-based confirmatory analysis is more efficient. Whenever possible, we recommend the design-based confirmatory analysis because researchers can always implement the regularized approach after data collection if necessary. Finally, the coverage of the 95% confidence intervals is reasonable for all estimators.