Item Response Theory for Conjoint Survey Experiments

Devin Caughey* Hiroto Katsumata† Teppei Yamamoto‡

April 12, 2021

Abstract

In recent years, political scientists have increasingly used conjoint survey experiments to analyze preferences about objects that vary in multiple attributes. The dominant approach in these studies has been to apply the regression-based estimator for the average marginal component effect (AMCE) proposed by Hainmueller, Hopkins, and Yamamoto (2014). Although the standard approach enables model-free inference about preferences underlying conjoint survey data, it has important limitations for analyzing heterogeneity in respondents' preferences about attributes and investigating how attributes are related to each other in the formation of preference about profiles as a whole. In this paper, we propose an item response theory (IRT) model for conjoint survey data to analyze respondents' heterogeneous preferences about attributes. Our proposed approach builds upon a canonical spatial theory of voting to model preferences as a function of respondents' ideal points on a latent space capturing taste variation. The model also incorporates a set of valence parameters to identify the dimension of preference about attributes that is common to all respondents. We discuss identification conditions, inference via a Bayesian algorithm, and how to map model parameters to substantive quantities of interest. We illustrate the utility of the proposed approach through Monte Carlo simulations as well as a validation analysis of an original online conjoint experiment on presidential candidate choice.

*Associate Professor, Department of Political Science, Massachusetts Institute of Technology.

†Project Lecturer, Graduate School of Arts and Sciences, The University of Tokyo.

‡Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

1 Introduction

In recent years, political scientists have increasingly used conjoint survey experiments to analyze preferences about objects that vary in multiple attributes. The dominant approach in these studies has been to apply the regression-based estimator for the average marginal component effect (AMCE) proposed by Hainmueller, Hopkins, and Yamamoto (2014). This approach has the great advantage of permitting nonparametric (model-free) inferences regarding attributes' effects on subjects' choices under quite weak assumptions. A limitation, however, is that these inferences do not speak directly to the structure of the preferences that underlie subjects' choices, nor to variation in subjects' preferences or in attributes' relationship to (latent) preference structure.

Making the inferential leap from observed choices to unobserved preferences requires a model that represents the structure of those preferences and links them to the data. In this paper, we propose a framework for this task, which builds on the canonical spatial theory of choice. In this framework, a subject's choice between conjoint vignettes is determined by locations of the alternatives and of the subject in a latent Euclidean space, which captures "taste" variation across subjects, as well as of "valence" differences between the vignettes, which are common to all subjects. Spatial and valence differences between vignettes are in turn assumed to be a function of the attributes included in the vignette.

We operationalize this spatial model with a Bayesian item response theory (IRT) model tailored to conjoint data, which we estimate using Hamiltonian Monte Carlo in Stan (Stan Development Team 2017). The conjoint IRT model estimates subjects' ideal points in the latent space as well as attributes' mapping onto that space. From these parameters, we can calculate not only the AMCE, which represents the average effect of an attribute across subjects, but also quantities such as the conditional average marginal component effect (CAMCE), which characterizes heterogeneity in attributes' effects as a function of subjects' ideal points. The package **conjointIRT** (Caughey, Katsumata, and Yamamoto 2019) implements the conjoint IRT model in R.

We validate this model with simulations and an original conjoint experiment. The simulations confirm that the model accurately estimates true parameter values under conditions typical of most conjoint studies (e.g., 200 subjects, 10 tasks per subject). The original experiment, an analysis of presidential candidate choice, provides further validation of our approach and also illustrates its usefulness. First, the model successfully identifies a latent ideological dimension that coincides

with our substantive understanding of American politics. Specifically, it correctly estimates the effects of all of our six policy attributes to be highly heterogeneous along the dimension, while showing the other six valence attributes to have substantially less or no discernible variation. At the same time, the model uncovers subtle but significant variation in effects for two of these valence attributes—gender and past military experience—that matches our substantive understanding. Finally, we show that the model-based point estimates of the overall AMCEs given our IRT model almost exactly coincide with the nonparametric AMCE estimates with less statistical uncertainty. This implies that our simple model captures the underlying true structure of respondents' preferences very well without introducing bias in the estimation of nonparametrically identified causal quantities. Our approach reveals the richness of empirical information contained in conjoint survey data that is not fully exploited by the existing approaches.

2 Motivation

In many settings, especially economic ones, it is reasonable to treat decision makers' preferences as *nonsatiable* in the sense that for a given person, more (or less) of a given attribute or commodity is always better. In addition, it is frequently plausible to assume a substantial degree of consensus over whether more or less is better. Thus, for example, more money is generally preferred to less, lower prices to higher ones, greater durability to less, better health to worse, and so on. Political preferences differ from economic ones in two respects. First, it is frequently more plausible to regard preferences over political outcomes as *satiable* rather than nonsatiable (Brady 1990, 97–8; McCarty and Meirowitz 2007, 21–2). That is, there is an ideal amount of a given attribute, often represented as a point in Euclidean space, that a person prefers to either more or less of that attribute. Second, people often disagree over which outcome is best. A conservative, for example, might ideally prefer a $5 minimum wage, a $10,000 cap on employment discrimination damages, a legal limit of 100,000 immigrants per year, and a prohibition on abortions except to save the life of the mother, whereas a liberal might have very different policy preferences. In short, political preferences are both more likely to be satiable than economic preferences and more likely to conflict between different types of people.

Existing methods of conjoint analysis are not ideally suited to studying satiable, conflicting preferences. Although political conjoint experiments often focus on how subjects resolve trade-

offs across attributes, reactions to any given attribute are typically summarized with a single effect estimate, such as the AMCE. Nonparametric estimation of the AMCE has the advantage of providing model-free inference even in the face of individual-level heterogeneity. To the extent that individual heterogeneity is substantial, however, the AMCE's focus on average effects glosses over potentially interesting variation. In particular, if subjects have very different ideal points, then an attribute might have large but opposite individual-level effects that cancel to produce a negligible AMCE. Moreover, when individual-level effects are highly heterogenous, the AMCE in an unrepresentative sample may differ substantially in magnitude or even in sign from the average effect in the population of interest.

The main existing approach to subject-level effect variation in conjoint experiments is to examine how observed subject characteristics moderate the effect of a given attribute. Hainmueller and Hopkins (2015), for example, examine variation in the effects of immigrant attributes by subject race, ethnocentrism, partisan identification, and other factors. This approach, however, has several limitations. First, subjects' observed characteristics may not be strong predictors of variation in treatment effects. This is particularly problematic if characteristics are analyzed one by one, which may result in the dangerous combination of multiple comparisons and low statistical power (see, e.g., Gelman, Hill, and Yajima 2012, 194–5). Second, one downside of being fully nonparameteric is the lack of a connection to a generative theoretical model, such as the spatial framework often used to represent satiable preferences. Even if just an approximation, a theoretical model can aid interpretation of the estimates by uncovering common variation across subjects and revealing how different attributes relate to that underlying dimension.

For these reasons, we propose an alternative approach to analyzing conjoint experiments, one that explicitly models heterogeneous preferences in terms of variation in subjects' spatial ideal points. Using an IRT model derived from a spatial model of choice, our proposed approach jointly estimates subjects' ideal points and attributes' item parameters. These parameters provide an intuitive, low-dimensional characterization of variation in subjects' preferences and in attributes' mapping to the latent dimension(s) of variation.

3 The Proposed Methodology (multi-dimensional version)

We now present our proposed methodology for analyzing survey respondents' preferences expressed via conjoint tasks. Throughout this section, we assume a paired forced-choice conjoint design, in which respondents are presented pairs of conjoint profiles and required to choose the more preferred profile within each pairwise comparison. This design has been very commonly used in political science and shown to perform well when benchmarked against revealed preference data (Hainmueller, Hangartner, and Yamamoto 2015). Extending the proposed framework to other common conjoint designs should be straightforward and is left for future work.

3.1 The Model

We consider a conjoint experiment in which respondent $i \in \{1, \ldots, N\}$ completes K choice tasks. In each task $k \in \{1, \ldots, K\}$, the respondent is presented with a choice between two profiles. Each profile $j \in \{1, 2\}$ is characterized by a collection of L categorical attributes. For the sake of simplicity, we assume that the attributes are all binary, such that each of the $P \equiv 2^L$ unique profiles can be represented by a set of L binary indicators or dummies. Let $\mathbf{x}_p \equiv [1, x_1, \ldots, x_L]^\top$ be a (column) vector of L such dummies, plus an intercept term, where $x_l \in \{0, 1\}$ for any $l \in \{1, \ldots, L\}$. We use the subscript $p \in \{1, \ldots, P\}$ to index each unique set of attributes, so that $\mathbf{x}_1 = [1, 0, \ldots, 0]^\top$, $\mathbf{x}_2 = [1, 0, \ldots, 1]^\top$, etc., for example.

Our general approach is to model the choice between two profiles as a function of the relative appeal of the attribute vectors $\mathbf{x}_p$ and $\mathbf{x}_{p'}$ that characterize the two profiles. That is, using a random utility framework, we assume the following model of respondent i's potential utility from profile j in task k if the pth set of attributes were assigned to that profile:

$$U_{ikj}(\mathbf{x}_p) \;=\; -\|\theta_i - \zeta(\mathbf{x}_p)\|^2 + \nu(\mathbf{x}_p) + \varepsilon_{ikj}, \tag{1}$$

where θ_i is respondent i's *ideal point* in a D-dimensional latent space, representing taste variation, $\zeta(\mathbf{x}_p)$ is the location of profile p in the same taste space, $\nu(\mathbf{x}_p)$ is the *valence* of profile p, representing the desirability of $\mathbf{x}_p$ for all respondents, ε_{ikj} is a stochastic component of utility, and $\|\cdot\|$ is the Euclidean norm. This utility function thus contains two systematic components. The first, $-\|\theta_i - \zeta(\mathbf{x}_p)\|^2$, is what we call the spatial, taste, or location component, which we assume to take

on the form of a standard quadratic loss function. That is, the model posits that respondents evaluate how desirable a profile is based on the Euclidean distance between the position of the profile and her own position in the same latent space. She will dislike the position if the profile is located far away from the position, and vice versa. Second, the model also contains a valence component, $\nu(\mathbf{x}_p)$, which represents the desirability of profile p for all respondents. This component is assumed to not vary across respondents, implying that it captures the portion of utility that does not depend on the underlying taste dimension in our model.

Next, we specify the functional forms for the valance and location parameters in terms of the attributes. Here, we model each of these as a linear function of $\mathbf{x}_p$:

$$\nu(\mathbf{x}_p) = \mathbf{x}_p^\top \alpha \tag{2}$$

$$\zeta(\mathbf{x}_p) = \mathbf{x}_p^\top \beta, \tag{3}$$

where α is a $(L+1)$-vector of coefficients indicating the valence of the attributes, β is a $(L+1) \times D$ matrix of coefficients indicating how much each attribute shifts the D-dimensional spatial location of a profile in a certain direction. Note that the linear specification here amounts to assuming that there is no interaction between the attributes; this assumption can be easily relaxed by additionally including interaction terms. A larger value of α_l means that attribute l is more preferable to respondents regardless of their ideal points. Valence attributes, such as honesty and integrity of political candidates for example, should have large α. On the other hand, $\beta_{l,d}$ represents how heterogeneous respondents' preferences are with respect to attribute l in dimension d. This parameter is larger when the preference for the attribute differs greatly depending on respondents' ideal points. In other words, β indicates the extent of disagreement among respondents over the preferable levels of the attribute.

Now we map our behavioral model of conjoint preferences to the observed data arising from an actual conjoint experiment. As the jth profile for respondent i in kth task, we obtain a random draw of a combination of attributes, which represents one of the P possible profiles. We denote such a random vector of realized attributes by $\mathbf{X}_{ikj} \equiv [1, X_{ikj1}, \ldots, X_{ikjL}]^\top \in \{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$. Then, we follow the standard random utility framework for discrete choice (e.g., McFadden et al. 1973) to model respondent i's choice to be the profile that gives her the highest utility. Specifically, let $\varepsilon_{ik} \equiv \varepsilon_{ik2} - \varepsilon_{ik1}$ be distributed as $\varepsilon_{ik} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Then, we can model respondent i's observed

choice response in the kth task, denoted by $Y_{ik} \in \{1, 2\}$, as follows:

$$\begin{aligned}
&\Pr[Y_{ik} = 1 \mid \mathbf{X}_{ik1}, \mathbf{X}_{ik2}] && (4)\\
&= \Pr[U_{ik1}(\mathbf{X}_{ik1}) > U_{ik2}(\mathbf{X}_{ik2})] && (5)\\
&= \Pr[-\|\theta_i - \zeta(\mathbf{X}_{ik1})\|^2 + \nu(\mathbf{X}_{ik1}) + \varepsilon_{ik1} > (-\|\theta_i - \zeta(\mathbf{X}_{ik2})\|^2 + \nu(\mathbf{X}_{ik2})) + \varepsilon_{ik2}] && (6)\\
&= \Pr[-\|\theta_i - \zeta(\mathbf{X}_{ik1})\|^2 + \nu(\mathbf{X}_{ik1}) - (-\|\theta_i - \zeta(\mathbf{X}_{ik2})\|^2 + \nu(\mathbf{X}_{ik2})) > -\varepsilon_{ik1} + \varepsilon_{ik2}] && (7)\\
&= \Pr[2(\zeta(\mathbf{X}_{ik1}) - \zeta(\mathbf{X}_{ik2}))^\top \theta_i\\
&\quad - (\zeta(\mathbf{X}_{ik1})^\top \zeta(\mathbf{X}_{ik1}) - \zeta(\mathbf{X}_{ik2})^\top \zeta(\mathbf{X}_{ik2}) - (\nu(\mathbf{X}_{ik1}) - \nu(\mathbf{X}_{ik2}))) > \varepsilon_{ik}]. && (8)
\end{aligned}$$

If we let

$$b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2}) = 2(\zeta(\mathbf{X}_{ik1}) - \zeta(\mathbf{X}_{ik2}))/\sigma$$

denote the "discrimination" of i's kth profile contrast and

$$g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2}) = (\zeta(\mathbf{X}_{ik1})^\top \zeta(\mathbf{X}_{ik1}) - \zeta(\mathbf{X}_{ik2})^\top \zeta(\mathbf{X}_{ik2}) - (\nu(\mathbf{X}_{ik1}) - \nu(\mathbf{X}_{ik2})))/\sigma$$

its "difficulty," then the model takes on the same form as the two-parameter probit IRT model,

$$\Pr[Y_{ik} = 1 \mid \mathbf{X}_{ik1}, \mathbf{X}_{ik2}] = \Phi[b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})^\top \theta_i - g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})], \tag{9}$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution function.

Given the assumed linear specifications for the location and valence parameters, the discrimination and difficulty parameters can be written directly in terms of the model coefficients,

$$\begin{aligned}
b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2}) &= 2(\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^\top \beta/\sigma && (10)\\
g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2}) &= (\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^\top (\beta\beta^\top (\mathbf{X}_{ik1} + \mathbf{X}_{ik2}) - \alpha)/\sigma. && (11)
\end{aligned}$$

Note that the difficulty parameter $g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})$ depends on both the attribute valence coefficients α and the spatial coefficients β, whereas the discrimination parameter $b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})$ depends on β only. This means that the discrimination parameter determines a part of utility that is related to ideal points and the difficulty parameter is related to both ideal points and valence. The likelihood

of the model is

$$
\begin{aligned}
&L(\beta, \alpha, \theta, \sigma \mid \mathbf{Y}, \mathbf{X}) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \Phi\left(b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})^{\top} \theta_i - g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})\right)^{2-Y_{ik}} \\
&\quad \times \left\{1 - \Phi\left(b(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})^{\top} \theta_i - g(\mathbf{X}_{ik1}, \mathbf{X}_{ik2})\right)\right\}^{Y_{ik}-1} \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \Phi\left(\frac{2(\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^{\top} \beta \theta_i - (\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^{\top} (\beta\beta^{\top} (\mathbf{X}_{ik1} + \mathbf{X}_{ik2})^{\top} - \alpha)}{\sigma}\right)^{2-Y_{ik}} \\
&\quad \times \left\{1 - \Phi\left(\frac{2(\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^{\top} \beta \theta_i - (\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^{\top} (\beta\beta^{\top} (\mathbf{X}_{ik1} + \mathbf{X}_{ik2})^{\top} - \alpha)}{\sigma}\right)\right\}^{Y_{ik}-1} \qquad (12)
\end{aligned}
$$

where $\theta = [\theta_1, \ldots, \theta_N]^{\top}$ and $\mathbf{Y}$ and $\mathbf{X}$ respectively collect the observed responses and realized attributes of the NK task observations.

From an IRT perspective, conjoint survey data can be regarded as analogous to a particular kind of roll-call voting data, where each "legislator" (respondent) votes on a random subset of pairs of "bill proposals" (profiles) formed out of the P total proposals. That is, a paired forced-choice conjoint task can be interpreted as a respondent's vote for a particular bill against another. This view highlights an important challenge when fitting an IRT model to conjoint survey data. That is, unlike typical legislative roll-call data, researchers can only observe each respondent's conjoint vote choices for a small fraction of the bills that exist in the data. For example, consider a paired conjoint experiment with 10 binary attributes, each respondent completing 20 forced-choice tasks. In this scenario, the researcher can only observe less than 0.004% of the universe of possible pairwise profile comparisons per respondent (ignoring unlikely but possible duplicates), because the number of possible profile pairs is $\binom{2^{10}}{2} = 523{,}776$. This suggests that an IRT model fitted to conjoint data must be sufficiently simple to maintain required statistical power for meaningful inference.

In our proposed approach, this simplification is achieved by reducing the dimensionality of the parameter space substantially by modeling the bill parameters with $(1 + D)(L + 1)$ coefficients on the attributes. While this might seem too drastic, it is important to recall that conjoint "bills" are distinguishable from each other only based on the observed attribute values *by design*. That is, unlike observational legislative roll-calls where each bill can always be unique in ways not captured by observed variables in a dataset, conjoint profiles are virtually guaranteed to be conditionally exchangeable because of their experimental nature. Moreover, by virtue of the randomization of the

attributes, the set of profiles presented to each respondent is guaranteed to be an i.i.d. draw from the population of the possible profile sets: missingness is completely at random by design. This contrasts sharply with typical legislative roll-call data, in which abstentions are highly unlikely to occur at random. These advantages of conjoint data are likely to make our inference fairly robust to model misspecifications, even with a very concise model such as ours and with an extremely sparse observed data matrix by an IRT standard. Indeed, as we show in Section 5.5, we find our model-based AMCE estimates for our empirical example to be remarkably close to nonparametric estimates of AMCEs which do not rely on any modeling assumption.

3.2 Identification

As is typical of spatial voting and other latent-variable models, identification of our conjoint IRT model requires a number of parameter restrictions. We enumerate these explicitly because few published sources collect them in one place. We largely follow Rivers's (2003) oft-cited unpublished paper on the subject, but depart from it in a few places, notably in our approach to rotational invariance. It should be emphasized that although these restrictions are *theoretically* sufficient for identification, in practice effective sampling may require redundant additional constraints on the signs of the spatial coefficients (Restriction 5 below).

"Local" identification of the conjoint IRT model can be achieved by imposing $D(D+1)$ independent restrictions on the parameters (Rivers 2003, 2). The first D restrictions we impose are setting the mean of respondent ideal points to 0 in each dimension:

Restriction 1 (Zero-Mean Ideal Points)**.** $\sum_i^N \theta_{id} = 0,$ for all d.

The next D restrictions consist of setting the variance of the ideal points to 1:

Restriction 2 (Unit-Variance Ideal Points)**.** $\sum_i^N \theta_{id}^2 = N,$ for all d.

For multidimensional models ($D > 1$), we impose a further $D(D-1)$ restrictions. $D(D-1)/2$ of these consist of requiring ideal points to be uncorrelated across dimensions:

Restriction 3 (Orthogonal Dimensions)**.** $\sum_i^N \theta_{id}\theta_{id'} = 0,$ for all $d, d' : d \neq d'$.[1]

1. We implement Restriction 3 (along with Restriction 2) by multiplying the column-demeaned ideal-point matrix ($\mathring{\boldsymbol{\theta}}$) by the Cholesky decomposition of the precision matrix of the ideal points ($\boldsymbol{L} = \text{chol}(\text{cov}(\mathring{\boldsymbol{\theta}})^{-1})$). The columns of the "whitened" ideal-point matrix $\tilde{\boldsymbol{\theta}} = \mathring{\boldsymbol{\theta}}\boldsymbol{L}^\top$ have, in addition to means of 0, variances of 1 and covariances of 0 (Kessy, Lewin, and Strimmer 2018).

The other $D(D-1)/2$ restrictions, required for rotational invariance, consist of setting a subset of spatial coefficients to 0 (Quinn 2004, 340):

Restriction 4 (Rotational Invariance)**.** $\beta_{ld} = 0$, for some $D(D-1)/2$ combinations of l and d.[2]

As noted, the $D(D+1)$ restrictions in Restrictions 1–4 are sufficient for local identification of the model. To achieve "global" identification, however, we must impose an additional D sign restrictions on β, which fixes the polarity of the latent dimension(s):

Restriction 5 (Polarity)**.** $\beta_{ld} > 0$ or $\beta_{ld} < 0$, for some l in each d.

Finally, we impose the standard normalizing assumption that the differences in utility shocks, $\varepsilon_{ik2} - \varepsilon_{ik1} = \varepsilon_{ik}$, have unit variance:

Restriction 6 (Unit-Variance Utility Shocks)**.** $\sigma = 1$.

Even with these restrictions, some of the model's parameters remain unidentified. In particular, the attribute valence parameters α_l are not identified separately from the intercept of the spatial component β_0. We can, however, identify $\gamma_l = \alpha_l - 2\beta_0^\top \beta_l,\ \forall l \in \{1, 2, \ldots, L\}$.[3] This motivates the following reparameterization of the likelihood function:

$$
\begin{aligned}
&L(\beta, \alpha, \theta, \sigma \mid \mathbf{Y}, \mathbf{X}) = L(\tilde{\beta}, \gamma, \theta, \sigma \mid \mathbf{Y}, \mathbf{X}) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \Phi\left(2(\tilde{\mathbf{X}}_{ik1} - \tilde{\mathbf{X}}_{ik2})^\top \tilde{\beta}\theta_i - (\tilde{\mathbf{X}}_{ik1} - \tilde{\mathbf{X}}_{ik2})^\top (\tilde{\beta}\tilde{\beta}^\top (\tilde{\mathbf{X}}_{ik1} + \tilde{\mathbf{X}}_{ik2})^\top - \gamma)\right)^{2-Y_{ik}} \\
&\quad \times \left\{1 - \Phi\left(2(\tilde{\mathbf{X}}_{ik1} - \tilde{\mathbf{X}}_{ik2})^\top \tilde{\beta}\theta_i - (\tilde{\mathbf{X}}_{ik1} - \tilde{\mathbf{X}}_{ik2})^\top (\tilde{\beta}\tilde{\beta}^\top (\tilde{\mathbf{X}}_{ik1} + \tilde{\mathbf{X}}_{ik2})^\top - \gamma)\right)\right\}^{Y_{ik}-1}, \quad (13)
\end{aligned}
$$

where $\sigma = 1$ is suppressed and $\mathbf{X}_{ikj}^\top = [1, \tilde{\mathbf{X}}_{ikj}^\top]^\top$, $\beta = [\beta_0, \tilde{\beta}^\top]^\top$, and $\gamma = [\gamma_1, \ldots, \gamma_l, \ldots, \gamma_L]^\top$. Note that the spatial intercept β_0 does not appear in equation (13) and therefore is not identified.

Because γ depends on β as well as α, it does not directly measure attributes' valence. Nevertheless, interpretation of it is aided by the fact that its second term ($2\beta_0^\top \beta_l$) drops out if all elements of either β_0 or β_l equal zero. Thus, if the spatial component of attribute l is small ($\beta_l \approx \mathbf{0}_D$), γ_l

2. As Aguilar and West (2000, 340) note, an advantage of this approach to rotational invariance is that it helps imbue the latent dimensions with substantive meaning. For example, in a two-dimensional model, the $D(D-1)/2 = 1$ attribute l for which $\beta_{l2} = 0$ is equal to the first factor plus noise. For this reason, it may be advisable to select for restriction attributes whose spatial component is both strong and substantively interpretable.

3. The valence intercept, α_0, is entirely unidentifiable because it is differenced out of the likelihood in (12) when α is pre-multiplied by $(\mathbf{X}_{ik1} - \mathbf{X}_{ik2})^\top$.

approximates l's valence coefficient α_l. As we will see, several of the attributes in the empirical application in Section 5 satisfy this condition. Even if β_l is large, γ_l may still approximate α_l if $\beta_0 \approx \mathbf{0}_D$. Note that β_0, being the intercept, corresponds to the location of the respondent whose ideal profile is $\mathbf{x}_0 \equiv [0, \ldots, 0]^\top$ in the latent space. Thus, one practical possibility researchers might consider for interpretability is to set $\mathbf{x}_0$ to be a profile plausibly located at the center of the latent space (e.g., an ideologically moderate candidate). If successful, this will ensure that γ will approximate α even for attributes with a strong spatial component.

3.3 Inference

To estimate the model just described, we adopt a Bayesian framework and use the probabilistic programming language Stan, as implemented by the R package **rstan** (Carpenter et al. 2017; Stan Development Team 2017). Stan samples from the posterior of Bayesian models using Hamiltonian Monte Carlo (HMC) simulation, which for complex posteriors is often more efficient and robust than alternative techniques such as Gibbs sampling. We implement this approach in the R package **conjointIRT**.

We parameterize the model as it appears in (13), that is, in terms of the identifiable parameters $\tilde{\beta}$ and γ rather than β and α. To complete the Bayesian specification, we assign independent $\mathrm{N}(0, 1)$ priors to the elements of the raw (i.e., pre-transformed) θ and β matrices as well as to the elements of γ.[4] The identification constraints described in section 3.2 are implemented in the Stan code itself, so they occur within each iteration rather than in post-processing. For full Stan code, see Appendix A.1.

3.4 Causal Quantity of Interest

Although our model parameters are themselves interpretable in the context of the proposed spatial choice model, one might desire to translate these parameters to a quantity that can be interpreted as causal effects. In particular, one might ask: What is the average causal effect of an attribute on the choice probability for a profile, conditional on the location of a respondent's ideal point?

4. Note because the location and scale of the latent dimension is fixed by the identification restrictions on θ, the mean and variance of these priors could be any pair of arbitrary constants instead of 0 and 1 and still yield the same posterior distributions.

We propose the conditional average marginal component effect (CAMCE) as a quantity of interest that can be estimated with our methodology. Let $Y_{ikj}(\mathbf{x}_j, \mathbf{x}_{j'}) = Y_{ikj}(x_{1j}, \ldots, x_{Lj}, x_{1j'}, \ldots, x_{Lj'}) \in \{0, 1\}$ denote the binary potential outcome indicating whether respondent i would choose the jth profile in the kth task if the task was a comparison of the profile $\mathbf{x}_j$ against the other profile within the pair $\mathbf{x}_{j'}$, where $i \in \{1, \ldots, N\}$, $j \neq j' \in \{1, 2\}$ and $k \in \{1, \ldots, K\}$. (We suppress the intercept term in $\mathbf{x}$ to simplify the notation.) Note that $Y_{ik1}(\mathbf{x}_1, \mathbf{x}_2) = 1 - Y_{ik2}(\mathbf{x}_2, \mathbf{x}_1)$ by design. We make the consistency assumption such that the observed outcome can be written as $Y_{ik} = 2 - Y_{ik1}(\mathbf{X}_{ik1}, \mathbf{X}_{ik2}) = 1 + Y_{ik2}(\mathbf{X}_{ik2}, \mathbf{X}_{ik1})$. Using this notation, we define the CAMCE of the first attribute given ideal point $\theta_i = \theta^*$ as follows.

$$CAMCE_1(\theta^*) \;=\; \mathbb{E}[Y_{ikj}(1, X_{ikj2}, \ldots, X_{ikjL}, \mathbf{X}_{ikj'}) - Y_{ikj}(0, X_{ikj2}, \ldots, X_{ikjL}, \mathbf{X}_{ikj'}) \mid \theta_i = \theta^*], \quad (14)$$

where the expectation operator is defined with respect to both the joint distribution of the attributes other than the attribute of interest, i.e., $X_{ikj2}, \ldots, X_{ikjL}$ and $\mathbf{X}_{ikj'}$,[5] and the random sampling of the NK respondent-tasks from the population. The CAMCEs for the rest of the attributes in the design, $CAMCE_2(\theta^*), \ldots, CAMCE_L(\theta^*)$, are defined analogously.

Equation (14) is closely related to the AMCE as defined by Hainmueller, Hopkins, and Yamamoto (2014), which in our notation is given by,

$$AMCE_1 \;=\; \mathbb{E}[Y_{ikj}(1, X_{ikj2}, \ldots, X_{ikjL}, \mathbf{X}_{ikj'}) - Y_{ikj}(0, X_{ikj2}, \ldots, X_{ikjL}, \mathbf{X}_{ikj'})],$$

for the first attribute. That is, the CAMCE represents how the average marginal effect of an attribute on the probability of choosing a profile varies as a function of the respondent's ideal point. Hainmueller, Hopkins, and Yamamoto (2014, equation (7)) propose a similar conditional AMCE where the conditioning variable is an observed pre-treatment respondent characteristic. Their approach is likely to be preferable if the researcher is specifically interested in the heterogeneity of attribute effects with respect to a concrete respondent characteristic measured in their study (e.g.

5. This distribution, often referred to as the randomization distribution, can in theory be any distribution specified by the researcher, although the predominant practice is to use the distribution used in the random generation of the profiles in the actual experiment. Care must be taken when the actual randomization distribution includes restrictions that prohibit certain combinations of attributes. For details, see Hainmueller, Hopkins, and Yamamoto (2014). Incorporating such a non-standard randomization distribution in the current Bayesian framework is beyond the scope of this paper; in our empirical example, we define our CAMCEs in terms of the actual randomization used in the experiment, which contained no restriction.

ethnocentrism in Hainmueller, Hopkins, and Yamamoto's immigration example). Here, we consider an alternative scenario where a researcher is interested in the effect heterogeneity with respect to a latent characteristic of respondents (i.e. ideal points) that best explains the inter-respondent variation in the choice responses under the assumed model.

4 Simulations

To evaluate the finite sample performance of this model, we conducted a set of Monte Carlo simulations with varying numbers of respondents (N) and tasks (K). Our main goal is to provide numerical evidence that our (Bayesian) point estimates of the attribute parameters (β and γ) and respondent ideal points (θ) have desirable asymptotic properties with respect to N and K. We are also interested in determining the minimum values of N and K required for tolerably accurate parameter estimates. In brief, we find that for all parameters the accuracy of the estimates increases with the number of tasks K, and that for the attribute parameters accuracy increases with the number of respondents N as well. Reasonable estimates can be obtained at levels of N and K typical of conjoint studies in political science.

The simulations were designed as follows. $L = 8$ attributes were assigned the parameter values $\tilde{\alpha} = [0.5, 0, -0.5, 0.5, 0, 0, -0.5, 0.5]^\top$ where $\alpha = [\alpha_0, \tilde{\alpha}^\top]^\top$, $\tilde{\beta} = [-0.6, -0.3, -0.2, 0, 0, 0.2, 0.3, 0.6]^\top$, and $\beta_0 = 0.2$, which together imply $\gamma = [0.7, 0.1, -0.4, 0.5, 0.0, -0.1, -0.6, 0.3]^\top$. Respondent ideal points (θ) were generated from a mixture of three normal distributions corresponding to three ideological groups (e.g., Liberals, Moderates, and Conservatives) with equal sizes. The group distributions were respectively centered at -2, 0, and $+2$, all with standard deviation 1. For comparability, we rescaled the θ distibution in each simulation to have mean 0 and standard deviation 1. Using these parameter values, we generated hypothetical survey response data in accordance with the random utility model described in Section 3. Specifically, we randomly drew samples from

the following data generating process:

$$
\begin{aligned}
X_{ikjl} &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5), \\
\varepsilon_{ik} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\
\nu(\mathbf{X}_{ikj}) &= \mathbf{X}_{ikj}^{\top}\alpha, \\
\zeta(\mathbf{X}_{ikj}) &= \mathbf{X}_{ikj}^{\top}\beta, \\
Y_{ik} &= \begin{cases} 1 & \text{if } -(\theta_i - \zeta(\mathbf{X}_{ik1}))^2 + \nu(\mathbf{X}_{ik1}) - (-(\theta_i - \zeta(\mathbf{X}_{ik2}))^2 + \nu(\mathbf{X}_{ik2})) > \varepsilon_{ik}, \\ 2 & \text{otherwise.} \end{cases}
\end{aligned} \tag{15}
$$

For each of the 9 combinations of $N = \{100, 200, 400\}$ respondents and $K = \{5, 10, 20\}$ tasks, we conducted 1,000 simulations with respondents' ideal points fixed. For each set of simulations, we estimated the bias and root mean squared error (RMSE) of the posterior means of β and γ. We did the same for the ideal points of the respondents at the 10th, 25th, 50th, 75th, and 90th percentiles of the θ distribution. To evaluate our uncertainty estimates, we also calculated the proportion of simulations in which each parameter's 90% credible interval (CI) did not cover the true parameter value (its *noncoverage rate*).

The asymptotic properties of our estimators are well illustrated by Figure 1, which presents RMSE as a function of the number of subjects (x-axis) and tasks (columns). For all three sets of parameters, RMSE decreases monotonically as the number of tasks K increases. In addition, the RMSE of the attribute parameters (but not θ) is also decreasing in the number of subjects N. Within each parameter set, RMSE tends to be larger for relatively extreme values, and it is proportionally larger for θ than for γ and β. The bias of these estimators exhibits similar patterns, indicating that the RMSE results are not simply due to decreasing variance (see Figure A.1 in Appendix A.2). The only difference from RMSE is that the bias of γ estimates decreases in K but not N. For β, bias is positively correlated with the parameter value (i.e., positive values have positive bias), but for θ the correlation is negative (i.e., positive values have negative bias).

The coverage properties of the parameters' 90% CIs were insensitive to both K and N, but they differed systematically across parameter sets (see Figure A.2 in Appendix A.2). The CIs for θ tended to be conservative—that is, they failed to cover the true parameter value less frequently than their nominal level of 10%. In contrast, the CIs for β were anti-conservative, exhibiting

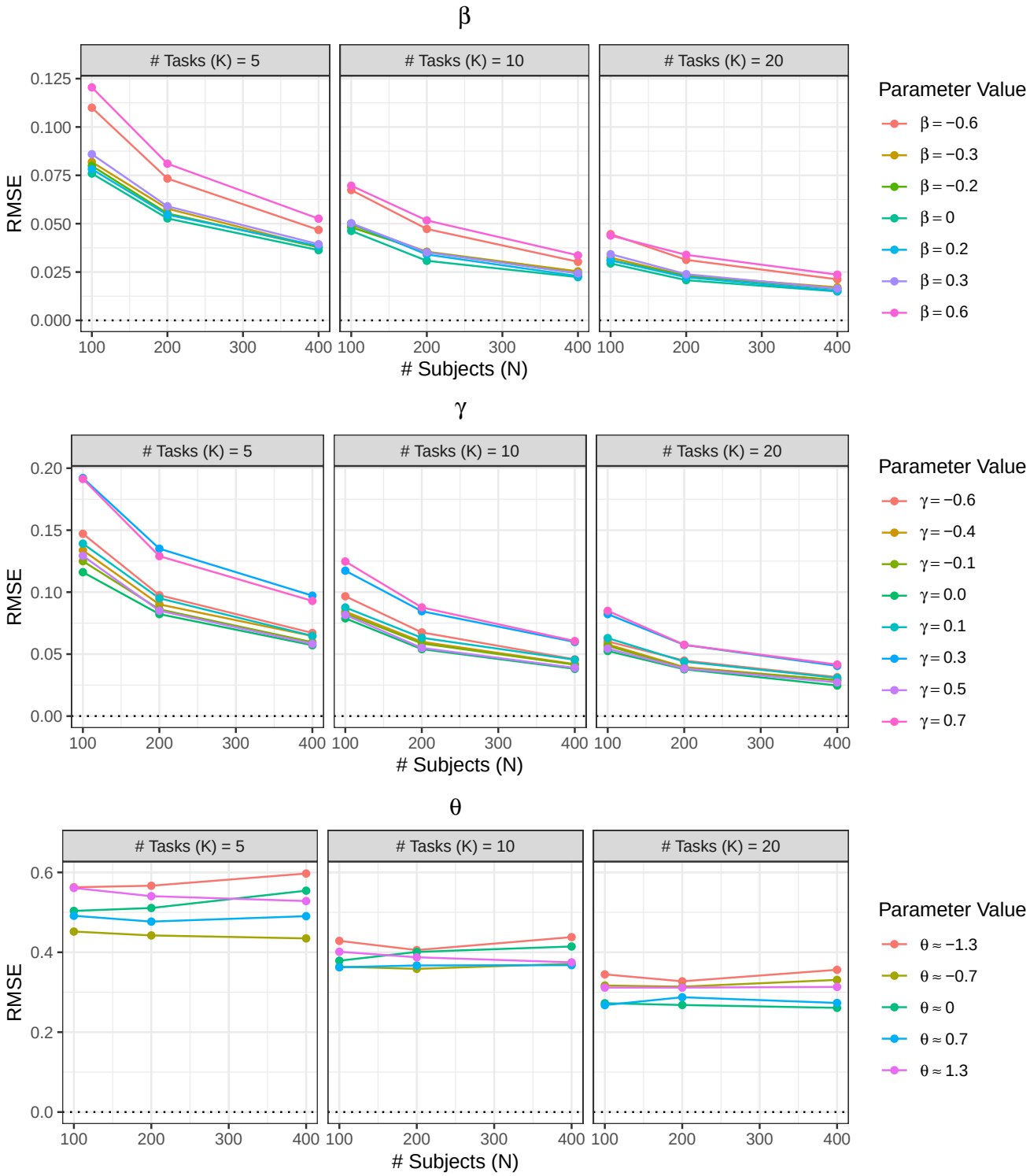

Figure 1: Results of the Monte Carlo Experiment in Terms of RMSE. The graphs show the simulated root mean squared error of posterior mean for β (top), γ (middle), and θ (bottom) for different combinations of N and K.

noncoverage rates as high as 25% for extreme values (e.g., $\beta = 0.6$). Finally, the noncoverage rates for γ clustered around their nominal level. Although we should not necessarily expect Bayesian CIs to perform optimally according to frequentist standards, these results do suggest that the CI for large values of $|\beta|$ may understate the uncertainty of the estimates.

On the whole, however, the simulation results validate our estimation approach. Although the estimates improve with larger N and especially K, they are quite accurate with even modest numbers of subjects and tasks (e.g., $N = 200, K = 10$). In fact, if the primary targets of inference are the attribute parameters rather than the ideal points, as few as 5 tasks per respondent may be sufficient. Since most conjoint experiments in political science have values of N and K as least as large as these thresholds, the simulations suggest that applying our method does not require major changes in how conjoint surveys are designed, and that the method can be applied retrospectively to surveys that have already been conducted.

5 Application: U.S. Presidential Candidates

We now present results from our original conjoint experiment conducted online via Amazon's Mechanical Turk platform. We designed the survey experiment so that it would serve as a validation dataset for our proposed methodology. Namely, we picked a substantive setting—choice between presidential candidates in the United States—for which we had a clear expectation about the latent structure of respondents' preferences on the basis of existing empirical evidence. The goal of our study, thus, is to investigate whether our proposed approach will return results that match our prior expectations based on substantive understanding of the context.

5.1 Data

For our survey experiment, we recruited a total of 1,003 respondents on Mechanical Turk, with the compensation of $1.50 for the expected completion time of 15 to 20 minutes. The respondents were residents of the United States who were 18 years old or older. The survey started with a battery of questions about respondents' socio-demographic characteristics, including gender, race, year of birth, state of residence, education, income, marital status, political ideology, and six-point party identification (omitting pure independent as an option). We then asked respondents about their

preferences for hypothetical candidates for President of the United States in a standard pairwise conjoint format with forced choice. Specifically, we told them: "This study is about voting and your views on potential candidates for President. We are going to present 30 pairs of hypothetical presidential candidates in the United States. For each pair, please indicate which of the two candidates you would prefer to see as President. Please make a choice even if you are not sure."[6]

Our candidate profiles consisted of 12 attributes, each randomly taking on one of the two possible levels.[7] Of those 12 attributes, six were designed to represent candidates' positions on policy issues (health care, abortion, gun control, immigration, same-sex marriage and tax), whereas the other six non-policy characteristics of the candidates (gender, education, profession, age, military service experience, and whether they have been investigated for any campaign violation). The exact wordings for these attributes and their levels are provided in Table A.3, Appendix A.3. We coded the policy attributes such that the baseline level would correspond to the more conservative position, so the positive values of β and θ would indicate the liberal direction.

The rationale behind our choice of these 12 attributes is threefold. First, we expect the preferences about the six policy attributes to vary across respondents along a single left-right ideological dimension. It has been shown that policy preferences in the United States can be largely characterized by a single latent dimension that closely aligns with partisanship both at the elite and mass level, especially in recent years (Jessee 2009; Tausanovitch and Warshaw 2013; Marble and Tyler 2018). In fact, we deliberately pick policy issues that we anticipate to correlate with respondents' latent ideology particularly closely for the purpose of this validation exercise. We thus expect the effects of these attributes to be highly heterogenous and vary systematically as functions of respondents' ideal points.

Second, we expect respondents to have more or less nonsatiable preferences about four of our six non-policy attributes. That is, we anticipate both left- and right-leaning respondents to have largely homogenous preferences about education, profession, campaign violations, and (probably) age. Our expectation is thus that the effects for these attributes are mostly picked up by the

6. Our choice of including as many as 30 tasks per respondent, a number that is larger than most existing conjoint studies in political science (though not necessarily in other fields such as marketing), is motivated by the recent empirical evidence reported by Bansak et al. (2018). They use a similar sample recruited from Mechanical Turk and ask a similar presidential candidate choice task 30 times to examine if there is any change in the way respondents complete those tasks. Remarkably, they find no noticeable degradation in response quality even after 30 tasks.

7. We made all of our attributes dichotomous for the sake of simplicity, although the proposed method can easily be applied to attributes with more than two discrete levels.

valence component of our model, such that the estimated CAMCEs will vary little as functions of the ideal points if at all.

Third, preferences about the two remaining non-policy attributes—gender and military service experience—are plausibly correlated with the ideological dimension that underlies the policy preferences. That is to say, we expect respondents who tend to prefer liberal policy positions to also prefer to see female candidates as President, and conversely conservative respondents to prefer male candidates. We also expect that respondents who tend to prefer conservative policy positions will value previous military service experience more positively than liberal respondents.

5.2 Estimation Results for Attribute Parameters

To make inference about the parameters in our proposed model, we applied the estimation procedure described in Section 3 to data from our presidential choice conjoint experiment. After sampling 1,000 draws from each of the four HMC chains, the sampler appeared to have achieved convergence, with Gelman and Rubin's potential scale reduction factor becoming very close to one for all of the model parameters. Visual inspection of traceplots for the parameters (see Figure A.3 in Appendix A.3 for illustrative examples) also indicated strong evidence of convergence.[8]

Figure 5.2 presents our estimates of the β parameters (except the intercept β_0, which is not separately identified from the valence parameters). To reiterate, the spatial (or taste) parameter β in our proposed model represents the "factor loadings" of the attributes on the common latent dimension that governs the preference heterogeneity across respondents. In other words, a large value of β_l indicates that the preference about attribute l is highly heterogenous among respondents with different ideal points. Since our validation experiment is designed to capture the common liberal-conservative ideological dimension in the U.S. public opinion, β can be interpreted as parameters representing how strongly respondents' preferences about each of the attributes are correlated with their underlying political ideology.

The estimation results not only corroborate our prior expectations but also produce additional empirical insights that are substantively interesting. First, the six policy attributes are all estimated to have large positive loadings on the latent scale, indicating that the model successfully captures the common ideological dimension underlying respondents' preferences. Interestingly, the estimates

8. The computation took several hours on a relatively modern workstation (Top-of-the-line mid-2010 Mac Pro) with the four chains parallelized to separate CPU cores.

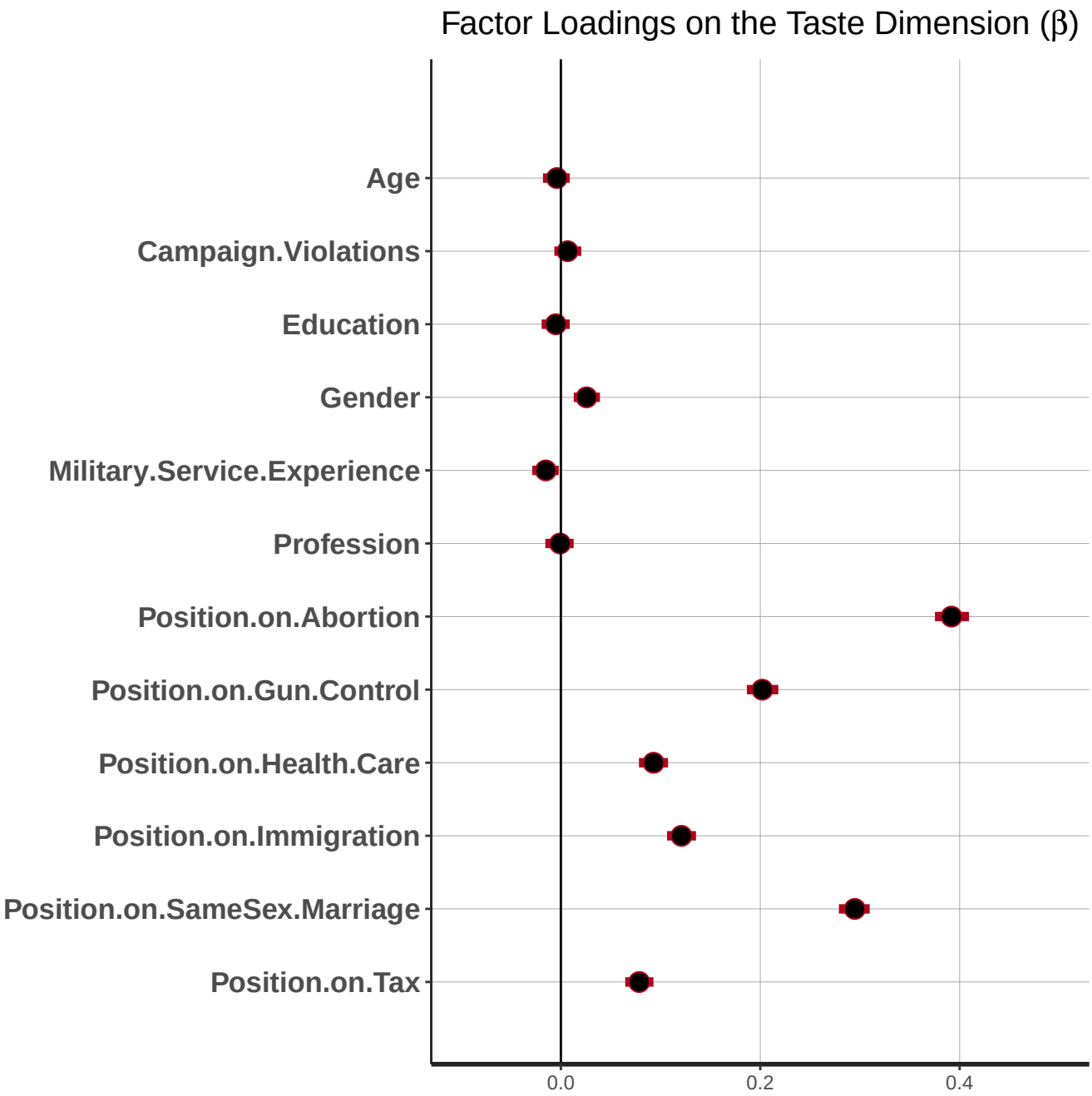

Figure 2: Estimates of the β Parameters for the Presidential Candidate Choice Experiment. The solid dots represent Bayesian point estimates (posterior means) for the parameters. The red and black horizontal bars represent 80% and 95% posterior central intervals, respectively.

indicate that position on abortion is by far the strongest correlate of the ideological dimension, followed by positions on same-sex marriage and gun control. Compared to these three policy positions, which can be thought of as primarily representing socio-cultural components of political ideology, the positions on the three economic and redistributive policies (immigration, health care, and taxation on the rich) are estimated to have weaker associations with the latent dimension.

Second, compared to the policy attributes, the six non-policy attributes are estimated to have much smaller β parameters, indicating that respondents' preferences about these attributes do not vary nearly as much in accordance with their ideological positions. In particular, the β values for four of these parameters (age, campaign violations, education, and military service experience) are estimated to be very close to zero. This implies that respondents' ideology does not correlate with their preferences about these four attributes.

Third, of the six non-policy attributes, gender and military service experience do appear to have significant loadings on the ideological dimension. That is, gender is estimated to have a positive β parameter that is not as large as the policy attributes in magnitude but still is statistically distinguishable from zero. This implies that liberal respondents tend to prefer female candidates more compared to conservative respondents, corroborating our prior expectation. Likewise, military service experience is estimated to have a smaller but statistically significant β parameter that is in the negative direction, indicating veteran status is valued more by conservative respondents than liberal respondents. Thus, overall, our model appears to capture respondents' taste variation about presidential candidates remarkably successfully in terms of the canonical liberal-conservative ideological dimension. It is also noteworthy that the parameters are highly precisely estimated even with our moderate-sized sample of 1,003 respondents.

We now briefly discuss the other set of attribute parameters γ. As discussed in Section 3.2, γ is difficult to interpret directly because it is a function of both the taste parameter β and the valence parameter α. An exception, however, is when the taste parameter happens to equal zero for an attribute, in which case γ exactly coincides with the valence parameter. Although we can never know whether this is indeed the case for any given parameter except by assumption, it is possible to empirically find attributes for which the condition approximately holds by examining the estimates for β. In our presidential candidate experiment, β for four non-policy attributes (age, campaign violations, education, and profession) are estimated to be negligibly small, which

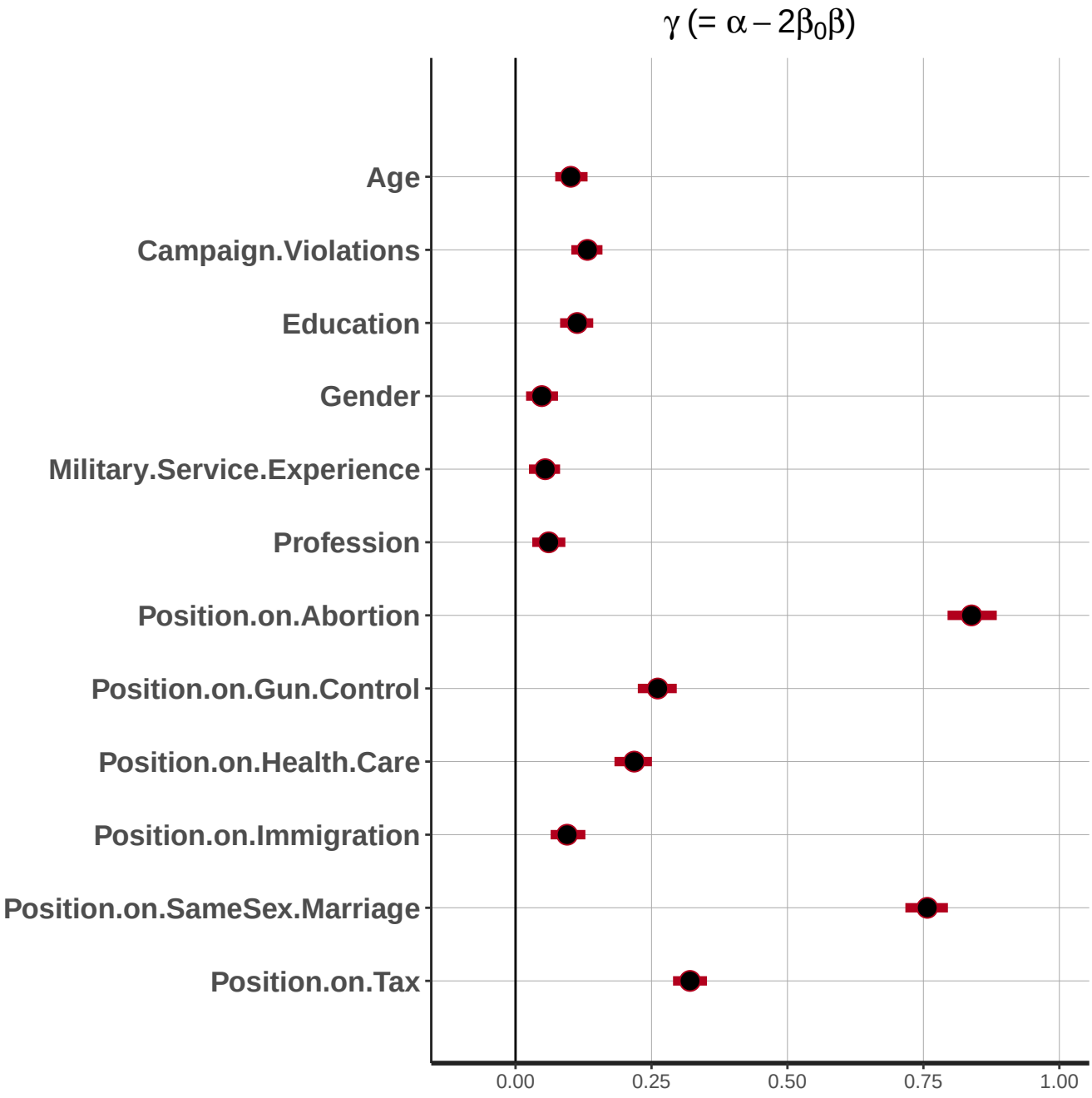

Figure 3: Estimates of the γ Parameters for the Presidential Candidate Choice Experiment. See caption for Figure 5.2 for the interpretations of the graph elements. Note that the γ parameters are difficult to interpret directly, except for attributes whose taste parameters (β) equal zero. In that case, the γ parameters coincide with the valence parameters (α) for the attributes. Four of the six non-policy attributes (age, campaign violations, education, and profession) approximate that scenario.

enables us to interpret γ for those attributes as representing their valence coefficients.[9]

Figure 5.2 presents our estimates for the γ parameters, including the four that can plausibly be interpreted as our α estimates. The estimates indicate that all of those attributes have sizable effects on the portion of the utility function that is common to all respondents regardless of their ideological leanings. This implies that respondents almost invariably prefer one of the two levels for each of these attributes (specifically, 45-year-old to 72-year-old candidates, candidates with a college degree to candidates with a high school degree, corporate executives to car dealers, and candidates who have never been investigated for campaign violations to those who have been), whether respondents are liberal or conservative.

5.3 Estimated Ideal Points

We now turn to the estimates of respondents' ideal points, represented by θ in our model. Figure 5.3 presents the empirical distribution of our point estimates of θ_i (posterior means). The distribution is standardized to mean 0 and variance 1 for the purpose of model identification, as discussed in Section 3.2. Figure A.4 in Appendix A.3 shows both posterior means and 95% central posterior intervals for each θ_i, indicating reasonable precision for our estimates even for individual ideal points.

The estimated distribution of ideal points clearly indicates the presence of at least three modes, resembling a mixture of three gaussians. Remarkably, the modes appear to correspond closely to conservatives, moderates, and liberals. The highest mode appears to represent ideological moderates and sits slightly to the left of the mean of the distribution (at around -0.4). This indicates that our Mechanical Turk respondent sample on average leans in the liberal direction, consistent with the existing evidence about the ideological tendency of survey respondents recruited on Mechanical Turk (e.g., Berinsky, Huber, and Lenz 2012). Moreover, the liberal mode appearing at around 0.8 is much higher than the conservative mode that sits at about -1.5, further confirming what has previously been known about the ideology of Mechanical Turk respondents. It is also interesting that the shape of the overall distribution is trimodal rather than uniform or unimodal:

9. Note that this approximation might be off by quite a bit even with $\beta_l \simeq 0$ for attribute l, if β_0 turns out to be very large. It is thus important to understand what the intercept parameter β_0 represents. It can be easily shown that β_0 corresponds to the ideal point of a respondent who prefers the "baseline profile" the most, i.e., the conjoint profile consisting of the baseline levels of all attributes. In our empirical example, this corresponds to a candidate who has all of the more undesirable non-policy attributes and all of the conservative policy positions.

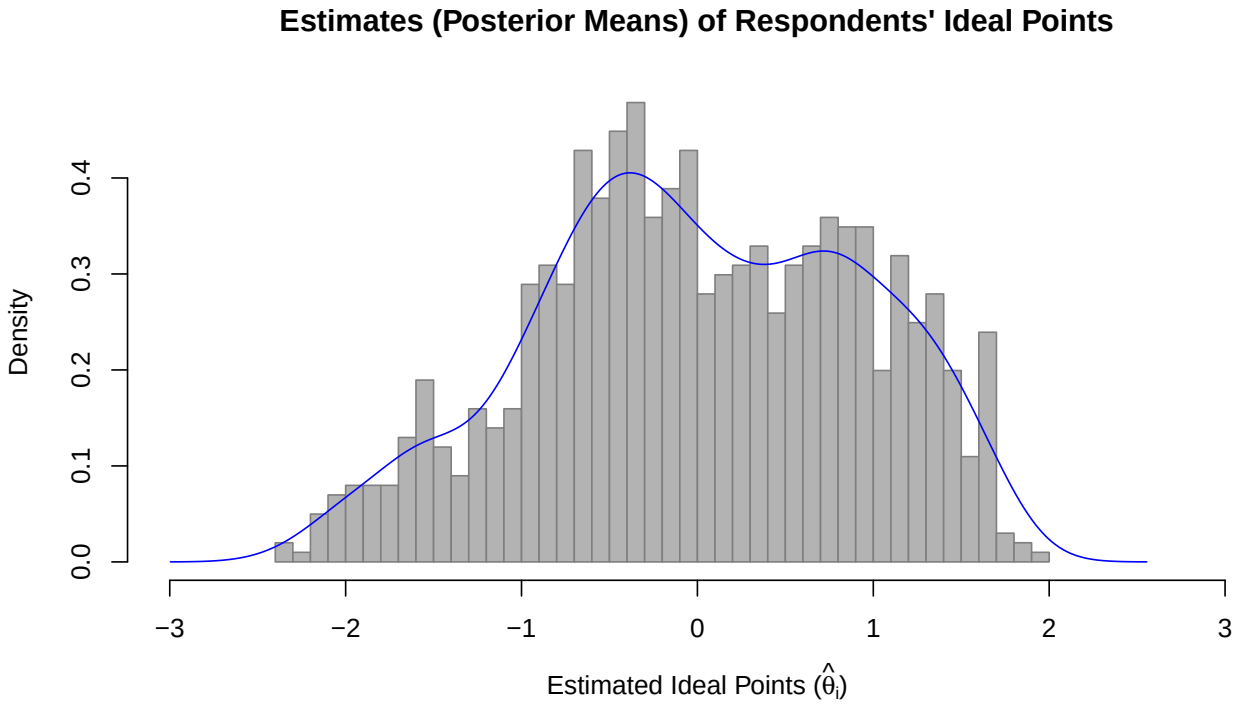

Figure 4: Estimates of Respondents' Ideal Points (θ) for the Presidential Candidate Choice Experiment. The histogram represents the distribution of the posterior means of the respondent ideal point parameters (θ_i for $i = 1, \ldots, 1003$), standardized to mean 0 and variance 1 for the purpose of identification. The blue curve represents a kernel density estimate for the same distribution.

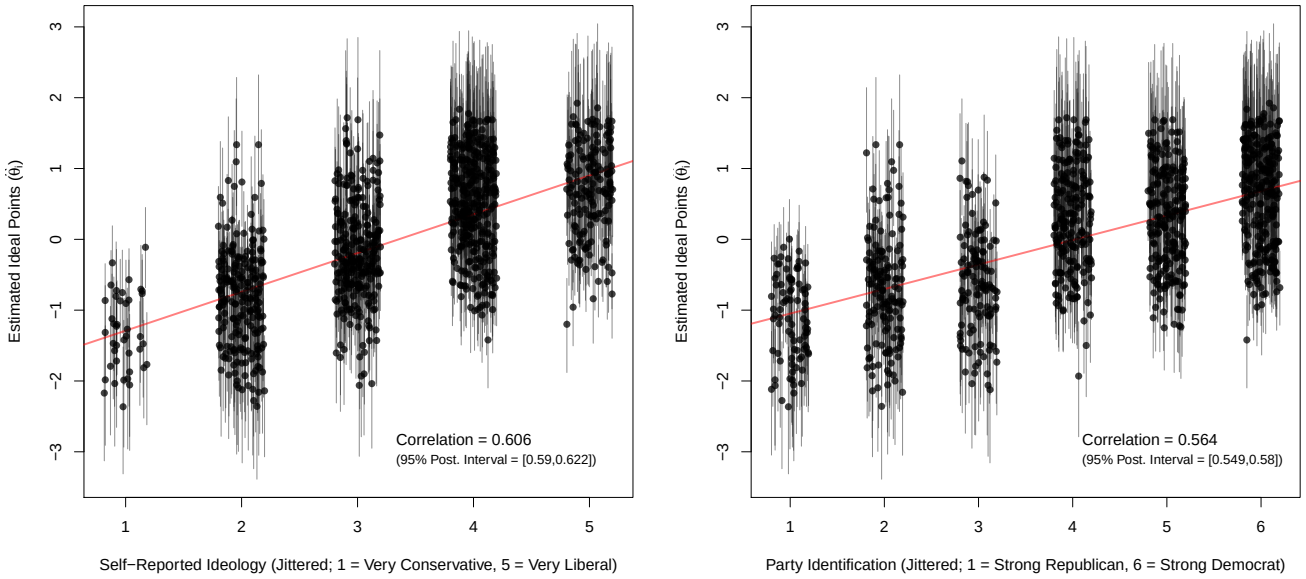

Figure 5: Comparison of the Estimated Ideal Points against Respondents' Self-Reported Ideology and Partisanship. The panels show estimated ideal points (θ_i) plotted against respondents' five-point ideological self-placement (left) and party identification (right). The solid dots represent point estimates (posterior means) and the vertical bars show 95% posterior central intervals. The red lines indicate ordinary least squares regression of the ideal point estimates on the self-reported ideology and partisanship measures.

there appears to exist distinct clusters of respondents into two "polarized" groups at both ends of the ideological spectrum and the remaining "moderate" group in the middle.

A question of interest is how our conjoint-based measure of respondents' ideology compares to more standard measures of ideology and partisanship based on direct survey questions. In Figure 5.3, we plot our estimated ideal points against respondents' self-reported five-point ideology scores (on the left) and six-point party identification (on the right). The result indicates that our ideal point estimates are strongly correlated with both of the self-reported measures. This is remarkable given that our estimates are solely based on responses to conjoint choice tasks on hypothetical presidential candidates which do not include any direct reference to either candidates' ideology or party affiliation. It is also noteworthy that our estimated ideal points correlate more strongly with ideological self-placement (the estimated correlation coefficient = 0.606, with the 95% central posterior interval of $[0.589, 0.623]$) than with party identification (0.565, $[0.547, 0.583]$). This suggests that subjects did not simply use candidates' attributes to guess their party and then vote

accordingly.

5.4 Inference for Quantities of Interest

A key component of our proposed approach is to translate the model parameters to quantities that can be substantively interpreted as causal effects. Specifically, we calculate the CAMCE as defined in Section 3.4 based on our estimates of the model parameters. To reiterate, the CAMCE represents the average (over the randomization distribution) effect of the "treatment" level of an attribute, relative to the baseline "control" level, on a respondent's probability of choosing a conjoint profile that contains the attribute, conditional on the respondent's ideal point. The CAMCE of 0.2 given $\theta_i = 0$, for example, indicates that the treatment level of the attribute on average increases the choice probability by 20 percentage points (compared to the baseline level) for a respondent whose ideal point sits right at the mean of the ideological distribution.

Figures 6 and 7 show our estimated CAMCEs for each of the 12 attributes included in our presidential candidate choice experiment. The results largely confirm what we found in our estimates for the model parameters but put them in terms of more interpretable quantities. First, the CAMCEs for the non-policy attributes show little to moderate variation across respondents as functions of their ideal points. In particular, age, education, profession, and campaign violations have almost equal effects for both liberal and conservative respondents, indicating that these attributes primarily signal candidates' valence as opposed to their ideological leanings. Interestingly, however, these four valence attributes appear to have slightly larger effects for ideological moderates, as indicated by the unimodal shapes of the estimated CAMCE curves. This can be interpreted as evidence that moderates' preferences tend to be driven relatively more strongly by non-policy attributes than policy attributes, wheras the choices of ideological extremists are almost completely based on candidates' policy positions, leaving less room for the effects of non-policy attributes.

Second, of the six non-policy CAMCE estimates, the effects for gender and military service experience do seem to vary depending on whether respondents are liberal or conservative. That is, female candidates are on average more likely to be chosen by an extremely liberal respondent (with θ_i at 2) by about two percentage points, but they are indeed about two percentage point *less* likely to be chosen by an extreme conservative whose θ_i is at -2. For military service experience, the CAMCE is estimated to be as large as about 3 to 4 percentage points for moderate to conservative

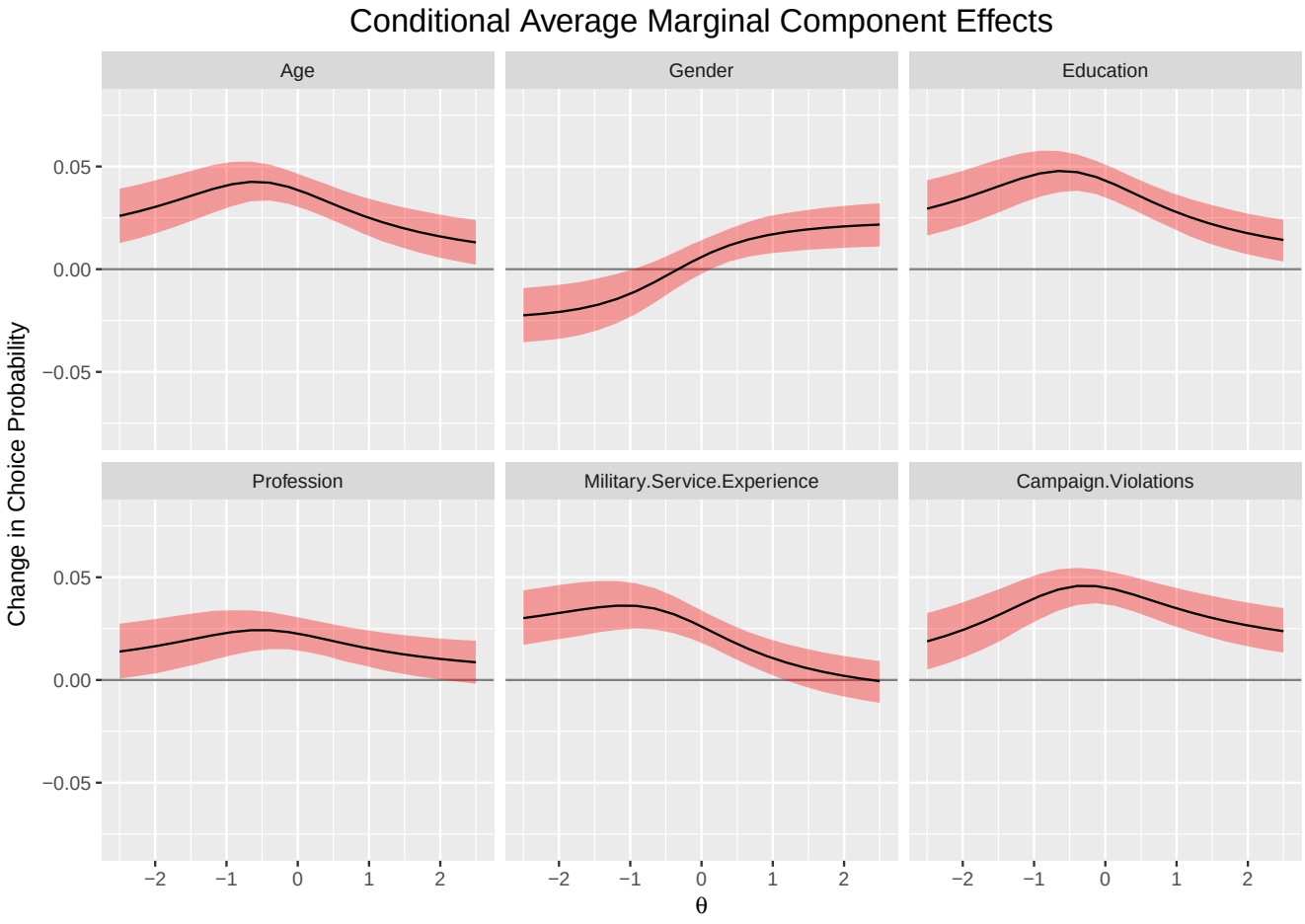

Figure 6: Estimates of the CAMCEs for the Non-Policy Attributes of Presidential Candidates. Each of the six plots shows the estimated marginal average component effect for an attribute as a function of respondents' ideal points on the taste dimension (θ_i). The solid lines represent Bayesian point estimates (posterior means) and the red ribbons show 95% posterior central intervals.

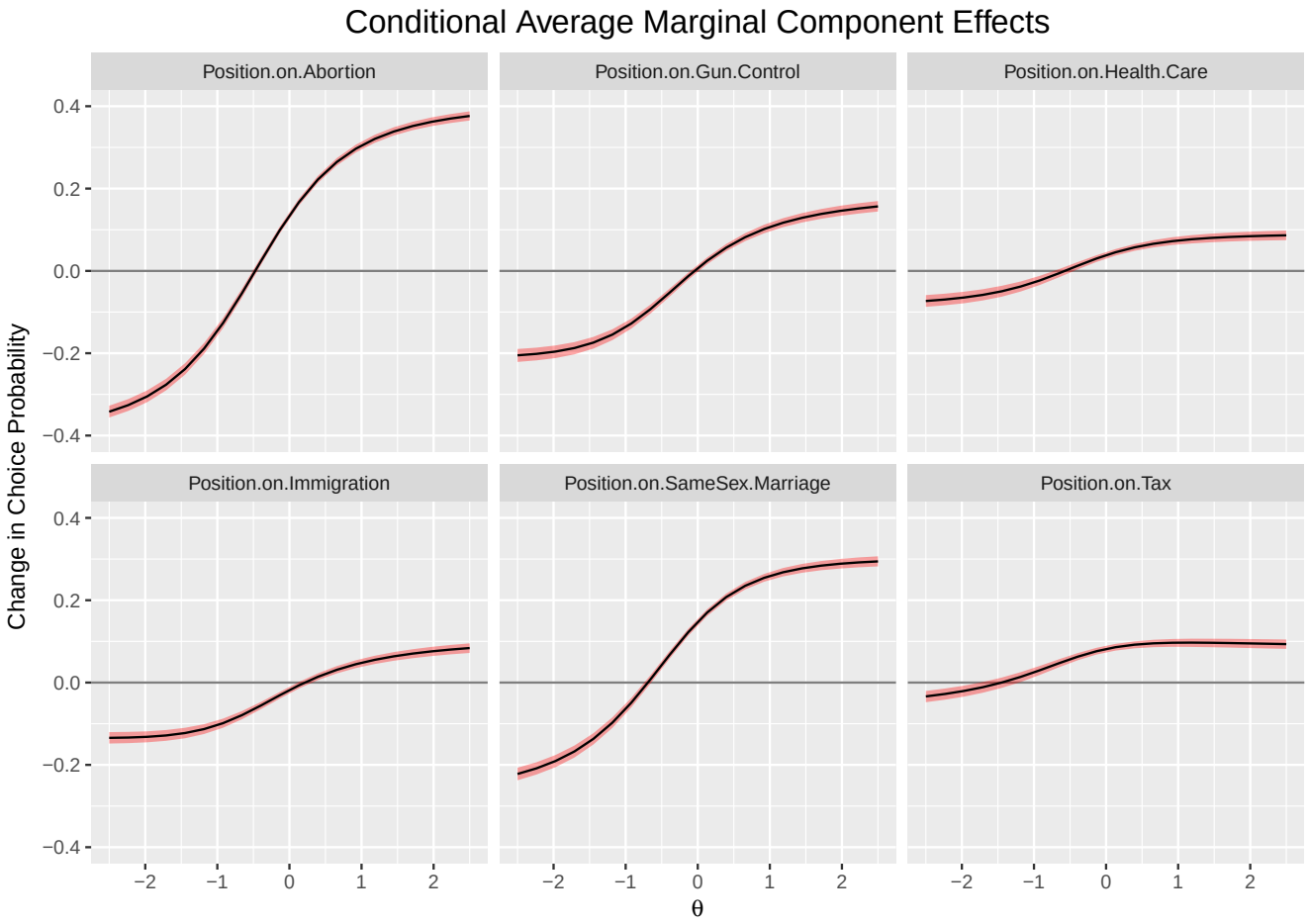

Figure 7: Estimates of the CAMCEs for the Policy Attributes of Presidential Candidates. See caption for Figure 6 for the interpretations of the graph elements. Note that the vertical axes for the plots are scaled differently from those in Figure 6.

respondents whose θ_i takes on negative values. However, the CAMCE estimate drops to nearly zero for respondents who are on the liberal extreme of the ideological spectrum.

Third, compared to the non-policy attributes, the effects of the six policy attributes (Figure 7) exhibit much greater variation as functions of respondents' ideal points. (Note that the vertical axes are differently scaled for Figures 6 and 7 for clearer presentation; Figure A.5 in Appendix A.3 shows the same estimates on a common scale for all attributes.) The heterogeneity is especially large for position on abortion: on average, the pro-choice position is found to increase the choice probability for an extremely liberal respondent ($\theta_i = 2$, which sits right at the end of the support of the empirical distribution for θ_i as shown in Figure 5.3) by as much as about 38 percentage points, whereas it *decreases* the choice probability by about 31 percentage points for an extremely conservative respondent ($\theta_i = -2$). The effect variation is also very large for position on same-sex marriage, where the pro-gay marriage position is found to increase the choice probability by about 29 percentage points for the extreme liberal respondent and decrease the probability by 20 percentage points for the extreme conservative respondent.

A natural question to ask is how our IRT-based CAMCE estimates compare to subgroup AMCEs conditional on standard survey measures of general policy preferences, such as self-reported ideology and party identification. That is, does our proposed approach add any substantive value to the more standard approach using these direct measures of ideology as respondent-specific moderators? To answer this question, we calculated nonparametric subgroup AMCE estimates as proposed by Hainmueller, Hopkins, and Yamamoto (2014) with respect to self-reported ideology and party identification. The results are presented in Figures A.7, A.8, A.9 and A.10 in Appendix A.3. Overall, we find the patterns of the variations in effect estimates to be largely similar between the model-based CAMCEs and the nonparametric subgroup AMCEs, especially for the self-reported ideology variable. This is unsurprising, given the strong correlation between these variables and our estimates of the ideal points, as we discussed in Section 5.3. It is however noteworthy that the IRT-based estimates tend to be more precise, particularly for the policy attributes. The model-based estimates also represents the covariation between ideology and effect sizes as a smoother, more interpretable functional relationship. Thus, our empirical result indicates a remarkable consistency between the inferences obtained from two starkly different kinds of ideology measurements—one based on direct questions about abstract concepts, and the other based on concrete choice tasks mimicking the real-world voting decisions—while suggesting important advantages of adopting a

more concise, model-assisted approach.

5.5 Robustness Checks

A potential downside of our proposed approach is that it imposes rather strong functional form assumptions for respondents' utility function, risking biased inferences if those assumptions are far from the true data generating process. Indeed, our assumed behavioral model is highly simplistic in that it allows systematic heterogeneity of preferences only in terms of the assumed latent spatial dimension. This is in stark contrast with the standard approach to conjoint experimental data proposed by Hainmueller, Hopkins, and Yamamoto (2014) which requires no modeling assumption for valid inference other than randomization of the attributes.

A natural question, then, is whether our simple model-based approach does any harm to the data for making inferences about causal quantities that could be estimated without making modeling assumptions, such as the overall AMCEs. In other words, if we translated our model parameter estimates into the AMCEs, would our estimates noticeably different from the nonparametric estimates of AMCEs à la Hainmueller, Hopkins, and Yamamoto (2014)?

Figure 5.5 shows the result of comparison between the unconditional AMCE estimates based on our model (blue triangles and bars) and the estimates using the nonparametric procedure proposed by Hainmueller, Hopkins, and Yamamoto (2014) (red circles and bars). The result shows remarkable agreement between the two sets of estimates. That is, the point estimates for the AMCEs are virtually identical for both estimation approaches, the gaps between the two estimates never exceeding one percentage point. This implies that the arguably strong assumptions entailed in our model-based Bayesian approach do not appear to cause any noticeable harm to inference in terms of point estimates for the AMCEs. Interestingly, the uncertainty estimates indicate markedly more precision in our approach than the nonparametric approach, especially for the attributes that have large taste (β) parameters such as positions on abortion, same-sex marriage and gun control. This is unsurprising because our estimates incorporate a lot more information given by the distributional and functional form assumptions embedded in the model as well as priors. Hainmueller, Hopkins, and Yamamoto's approach, in contrast, is specifically designed to be valid without any such assumptions, producing generally conservative uncertainty estimates at the potential cost of statistical power. Overall, however, these results suggest that our proposed approach is remark-

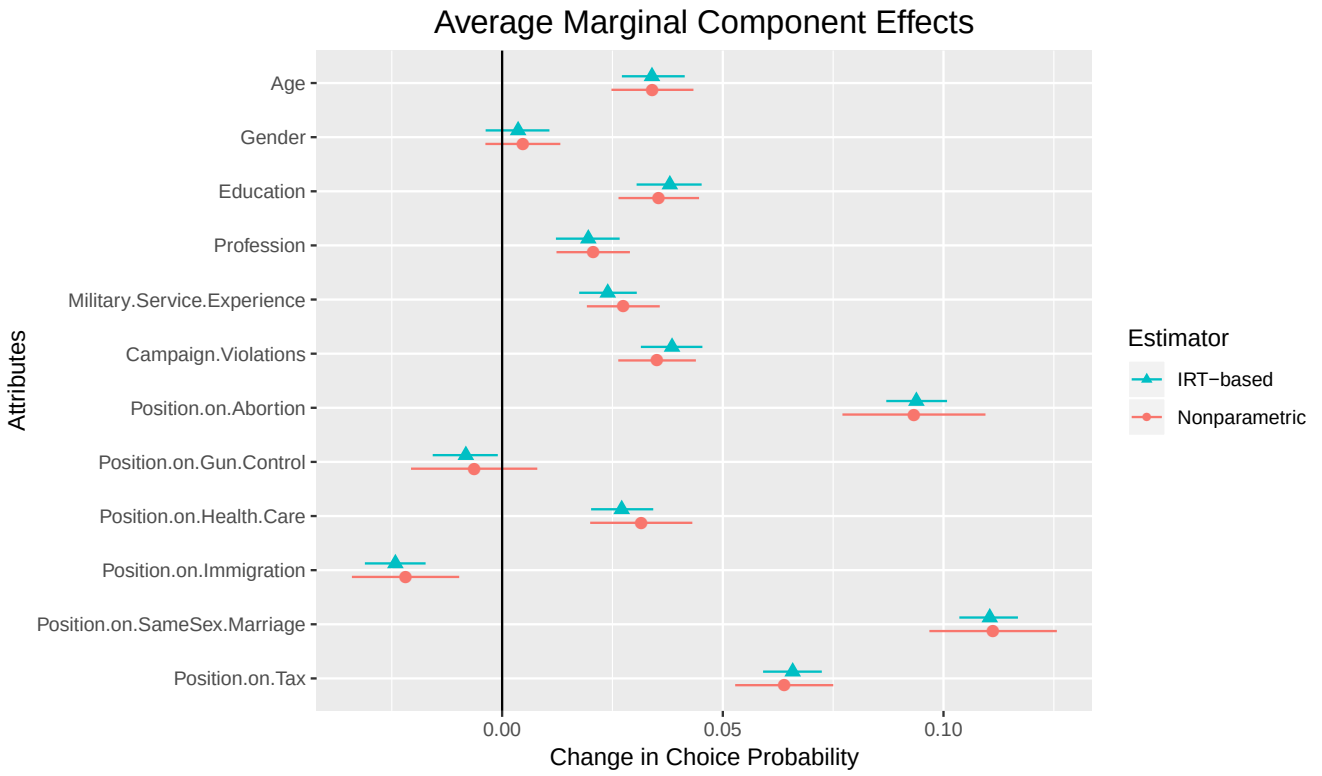

Figure 8: Comparison of the Model-Based and Nonparametric Estimates of the AMCEs for the Presidential Candidate Choice Experiment. The blue triangles and horizontal bars represent the estimated AMCEs for the twelve attributes based on the proposed Bayesian IRT methodology. The red circles and bars represent the nonparametric estimates of the AMCEs based on the regression-based procedure proposed by Hainmueller, Hopkins, and Yamamoto (2014). The result indicates that the point estimates are virtually identical between the two procedures and the uncertainty estimates are reduced for our proposed approach, particularly for the policy attributes.

ably successful in capturing the key latent structure in respondents' preferences that govern the patterns of effect heterogeneity in our conjoint data, while not distorting unconditional average effect estimates.

6 Conclusion

In this paper, we proposed an IRT model for conjoint survey data as a method for analyzing respondents' heterogenous preferences about attributes. The proposed method takes advantage of recent developments in two research traditions for the analysis of survey data — causal inference and Bayesian scaling — to enable inferences about latent heterogeneity in causal effects. As

demonstrated by our empirical example, conjoint survey data contain rich information about the structure of respondents' preferences that is largely untapped by the current standard methodology.

References

Aguilar, Omar, and Mike West. 2000. "Bayesian Dynamic Factor Models and Portfolio Allocation." *Journal of Business & Economic Statistics* 18 (3): 338–357.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2018. "The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments." *Political Analysis* 26 (1): 112–119.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (1): 351–368.

Brady, Henry E. 1990. "Traits versus Issues: Factor versus Ideal-Point Analysis of Candidate Thermometer Ratings." *Political Analysis* 2:97–129.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Caughey, Devin, Hiroto Katsumata, and Teppei Yamamoto. 2019. *conjointIRT: Item response theory for conjoint experiments.*

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5 (2): 189–211.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating vignette and conjoint survey experiments against real-world behavior." *Proceedings of the National Academy of Sciences* 112 (8): 2395–2400.

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

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

Horiuchi, Yusaku, Daniel M Smith, and Teppei Yamamoto. 2018. "Measuring Voters' Multidimensional Policy Preferences with Conjoint Analysis: Application to Japan's 2014 Election." *Political Analysis* 26 (2): 190–209.

Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103 (1): 59–81.

Kessy, Agnan, Alex Lewin, and Korbinian Strimmer. 2018. "Optimal Whitening and Decorrelation." *The American Statistician* 72 (4): 309–314.

Marble, William, and Matthew Tyler. 2018. *The Structure of Political Choices: Distinguishing Between Constraint and Multidimensionality.* Unpublished working paper.

McCarty, Nolan, and Adam Meirowitz. 2007. *Political Game Theory: An Introduction.* New York: Cambridge University Press.

McFadden, Daniel, et al. 1973. "Conditional logit analysis of qualitative choice behavior."

Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12 (4): 338–353.

Rivers, Douglas. 2003. "Identification of Multidimensional Spatial Voting Models." Unpublished manuscript.

Stan Development Team. 2017. *RStan: the R interface to Stan.* Version 2.17.0. `http://mc-stan.org`.

Tausanovitch, Chris, and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* 75 (2): 330–342.

Appendix

A.1 Stan Code for IRT Model

```
functions {
  matrix whiten(matrix XX) {
    /* De-means and "whitens" (cov = I) XX */
    matrix[rows(XX), cols(XX)] DM;
    matrix[cols(XX), cols(XX)] SS;
    matrix[cols(XX), cols(XX)] PP;
    matrix[cols(XX), cols(XX)] WW;
    for (d in 1:cols(XX)) {
      DM[, d] = XX[, d] - mean(XX[, d]); /* de-mean each column */
    }
    SS = crossprod(DM) ./ (rows(XX) - 1.0); /* covariance of XX */
    PP = inverse_spd(SS); /* precision of XX */
    WW = cholesky_decompose(PP); /* Cholesky decomposition of precision */
    return DM * WW; /* de-meaned and whitened XX */
  }
  real probit_to_logit_scale (real x) {
    return 0.07056 * pow(x, 3) + 1.5976 * x;
  }
}
data {
  int<lower=1> D;                    // num. latent dimensions
  int<lower=1> L;                    // num. attributes
  int<lower=1> I;                    // num. respondents
  int<lower=1> N;                    // num. observations
  int<lower=1, upper=I> ii[N];       // respondent for obs n
  int<lower=0, upper=1> choice[N];   // response (0/1)
  matrix<lower=0,upper=1>[N, L] XX1;  // attributes of profile 1
  matrix<lower=0,upper=1>[N, L] XX2;  // attributes of profile 2
  int<lower=-1, upper=1> beta_nonzero[L, D]; // betas set to 0
  int<lower=-1, upper=1> beta_sign[L, D]; // betas set to +/-
}
transformed data {
  matrix<lower=-1,upper=1>[N, L] XX1_minus_XX2; // attribute differences
  matrix<lower=0,upper=2>[N, L] XX1_plus_XX2;   // attribute sums
  XX1_minus_XX2 = XX1 - XX2;
  XX1_plus_XX2 = XX1 + XX2;
}
parameters {
  vector[L] gamma;            // function of valence and spatial coefficients
  matrix[L, D] beta_free; // unconstrained spatial coefficients
  matrix[I, D] theta_raw;             // ideal point (pre-normalized)
}
transformed parameters {
```

```
  matrix[N, D] disc;            // discriminations
  vector[N] diff;               // difficulty
  matrix[I, D] theta;           // ideal points (normalized)
  matrix[L, D] beta;            // spatial coefficients
  /* IDENTIFICATION */
  /* Normalize thetas [D + D + D(D - 1)/2 restrictions] */
  theta = whiten(theta_raw);
  /* Restrict betas [D(D - 1)/2 + D restictions] */
  for (l in 1:L) {
    for (d in 1:D) {
      /* Set D(D - 1) betas to 0 */
      if (beta_nonzero[l, d] == 0) {
        beta[l, d] = 0;
      } else {
        if (beta_sign[l, d] == 0) {
          beta[l, d] = beta_free[l, d];
        } else {
          /* Restrict sign of D betas */
          beta[l, d] = beta_sign[l, d] * fabs(beta_free[l, d]);
        }
      }
    }
  }
  /* Convert spatial and valence terms into IRT item parameters */
  for (n in 1:N) {
    disc[n, 1:D] = 2 * XX1_minus_XX2[n, 1:L] * beta[1:L, 1:D];
    diff[n] = XX1_minus_XX2[n, 1:L] *
      (beta[1:L, 1:D] * beta[1:L, 1:D]' * XX1_plus_XX2[n, 1:L]' - gamma[1:L]);
  }
}
model {
  real pi[N];
  // Priors
  to_vector(theta_raw) ~ normal(0, 1);
  to_vector(beta_free) ~ normal(0, 1);
  gamma ~ normal(0, 1);
  // Likelihood
  for (n in 1:N) {
    pi[n] = probit_to_logit_scale(disc[n, 1:D] * theta[ii[n], 1:D]' - diff[n]);
  }
  choice[1:N] ~ bernoulli_logit(pi[1:N]);
}
```

A.2 Additional Simulation Results

A.3 Details of the Empirical Application

Attribute	Levels
Position on health care	government should do more, government should do less
Position on abortion	abortion should be legal in most cases, abortion should be illegal in most cases
Position on gun control	protect the right to own guns, control gun ownership
Position on immigration	increase immigration, decrease immigration
Position on same-sex marriage	favor same-sex marriage, oppose same-sex marriage
Position on tax	support raising tax on incomes over $300,000 per year, oppose raising tax on incomes over $300,000 per year
Age	45 years old, 72 years old
Gender	female, male
Education	college degree, high school degree
Profession	corporate executive, car dealer
Military service experience	served in the military, never served in the military
Campaign violations	has never been investigated for campaign violation, investigated for campaign violation but acquitted

Table A.1: Candidate Attributes for the Example Conjoint Experiment. The first six attributes represent candidates' positions on policy issues, whereas the other six attributes are candidates' personal characteristics. We expect the preferences about the policy attributes to differ across respondents due to their taste variation along the latent ideological dimension, and the preferences about the personal attributes to have less variation along the dimension.

A.4 Additional Application: 2014 Japanese General Election

We apply our proposed methodology to the conjoint survey data on Japanese voters originally collected by Horiuchi, Smith, and Yamamoto (2018) (hereafter HSY) as an additional illustration of the proposed methodology. In this example, we have far less established prior understanding of how individual attributes are related to the underlying latent ideological dimension than in our first application. Thus, our goal here is to demonstrate how the proposed method can be used for a new, meaningful empirical discovery.

A.4.1 Background

In their study, HSY examined Japanese voters' preferences about political parties in the 2014 House of Representatives election via a conjoint experiment implemented during the actual campaign period. Their conjoint profiles represented party manifestos, or policy bundles, which consisted of positions on nine policy issues identified as particularly salient in the election based on the authors' analysis of the actual party manifestos. The authors presented randomized combinations of the nine issue positions as hypothetical party manifestos in pairs and asked respondents to choose the one they would vote for in the upcoming election.

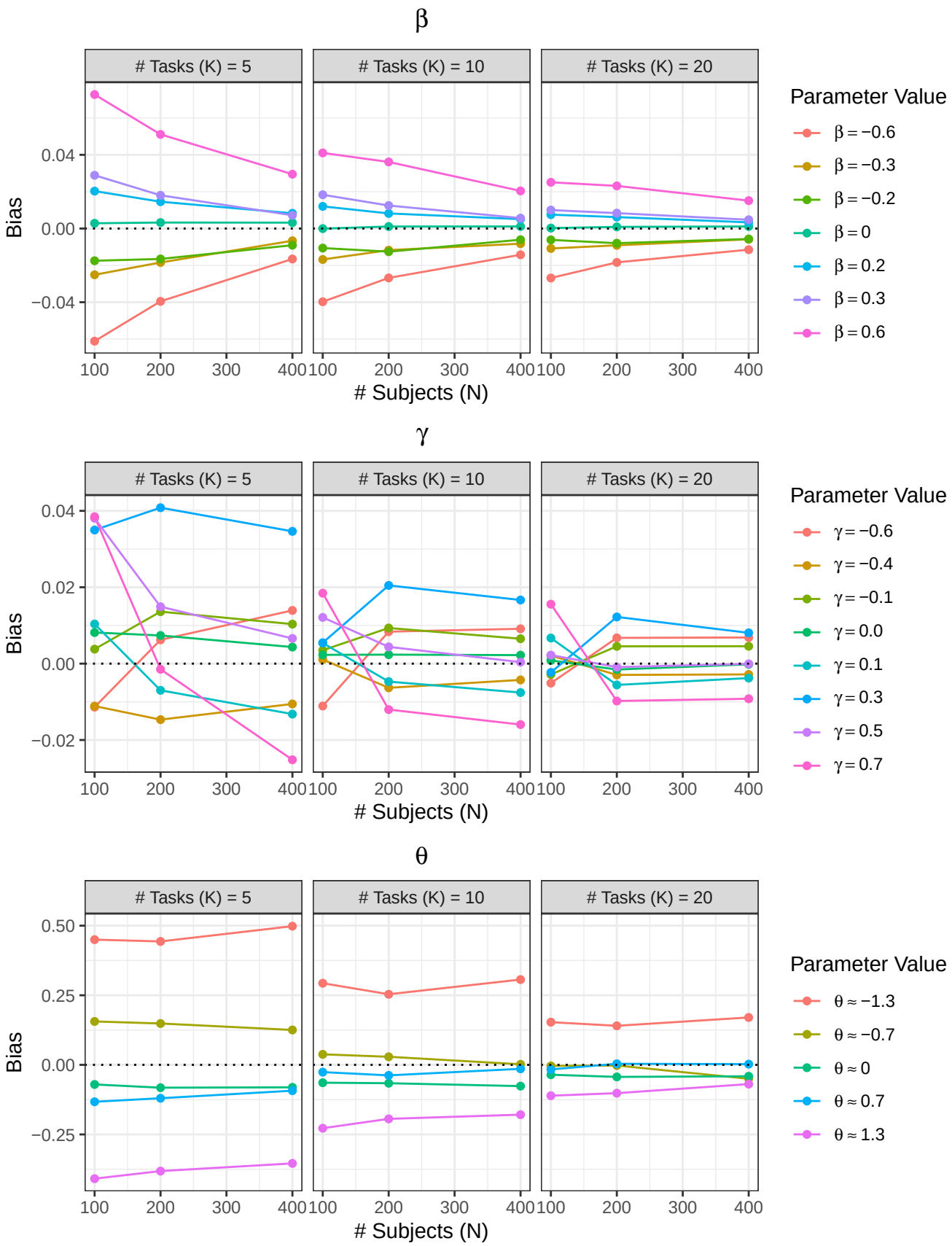

Figure A.1: Estimated bias of posterior mean for β (top), γ (middle), and θ (bottom) for different combinations of N and K.

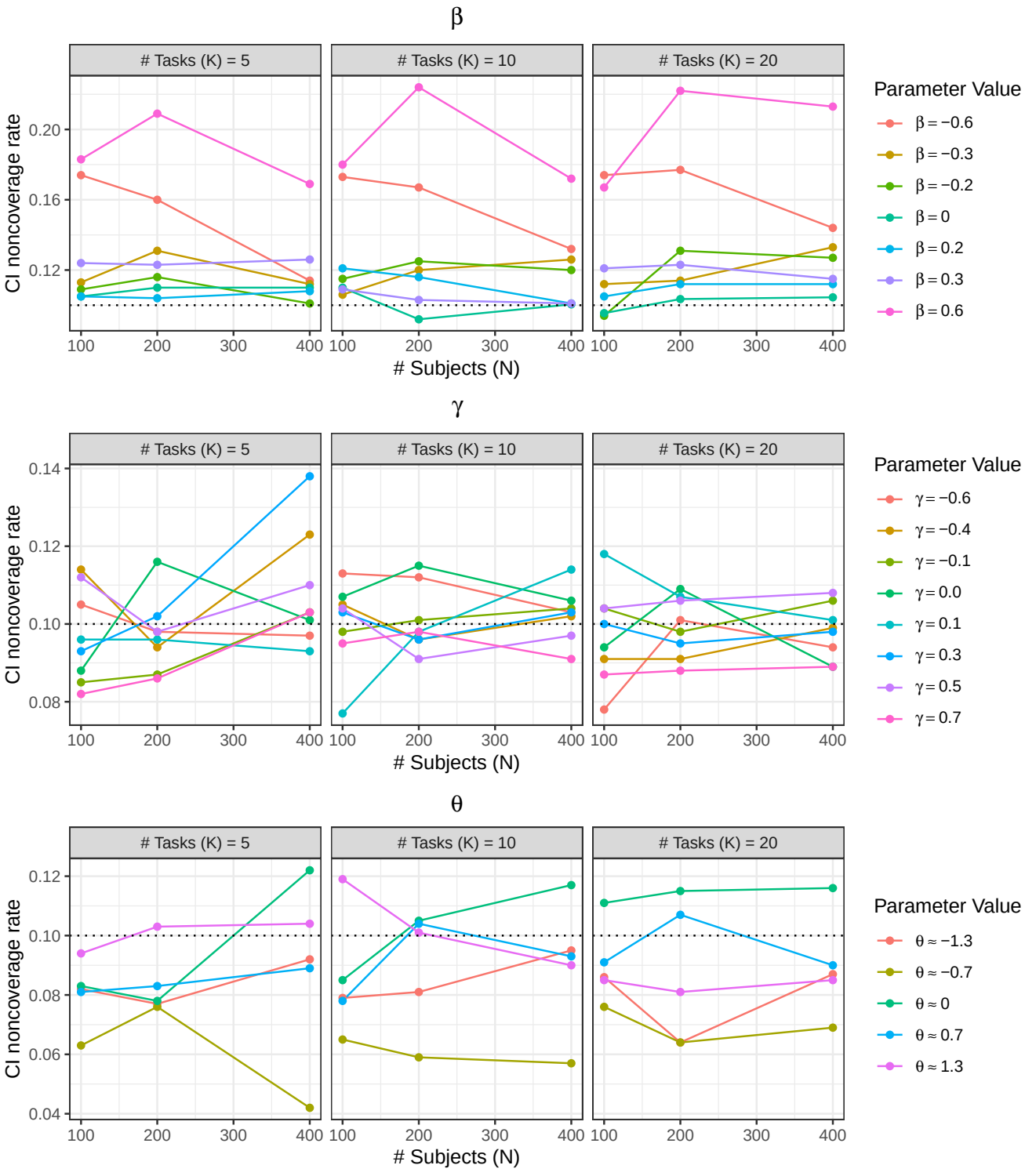

Figure A.2: Estimated noncoverage rate of 90% posterior CI for β (top), γ (middle), and θ (bottom) for different combinations of N and K.

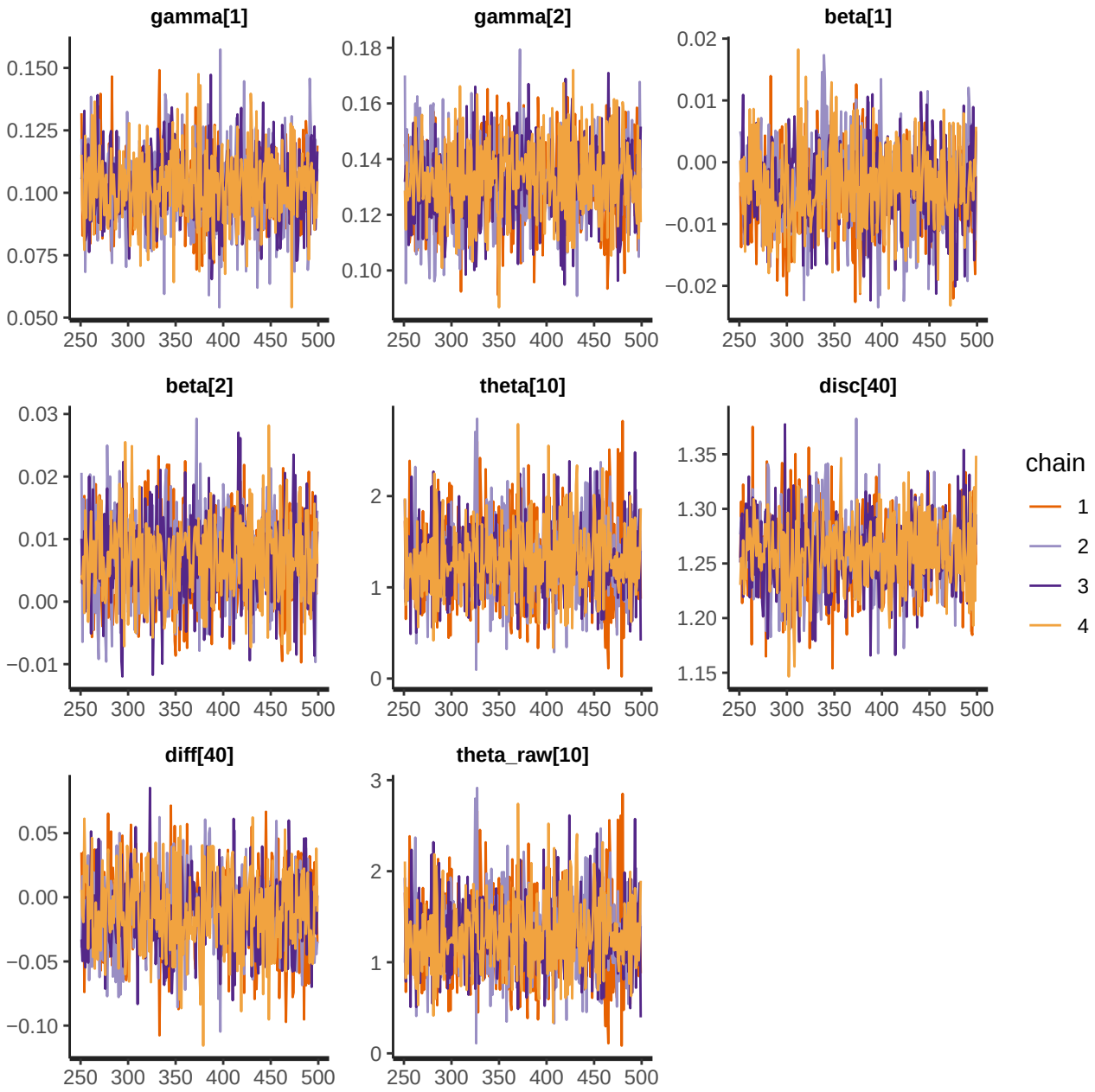

Figure A.3: Sample Parameter Traceplots for the Presidential Candidate Choice Experiment.

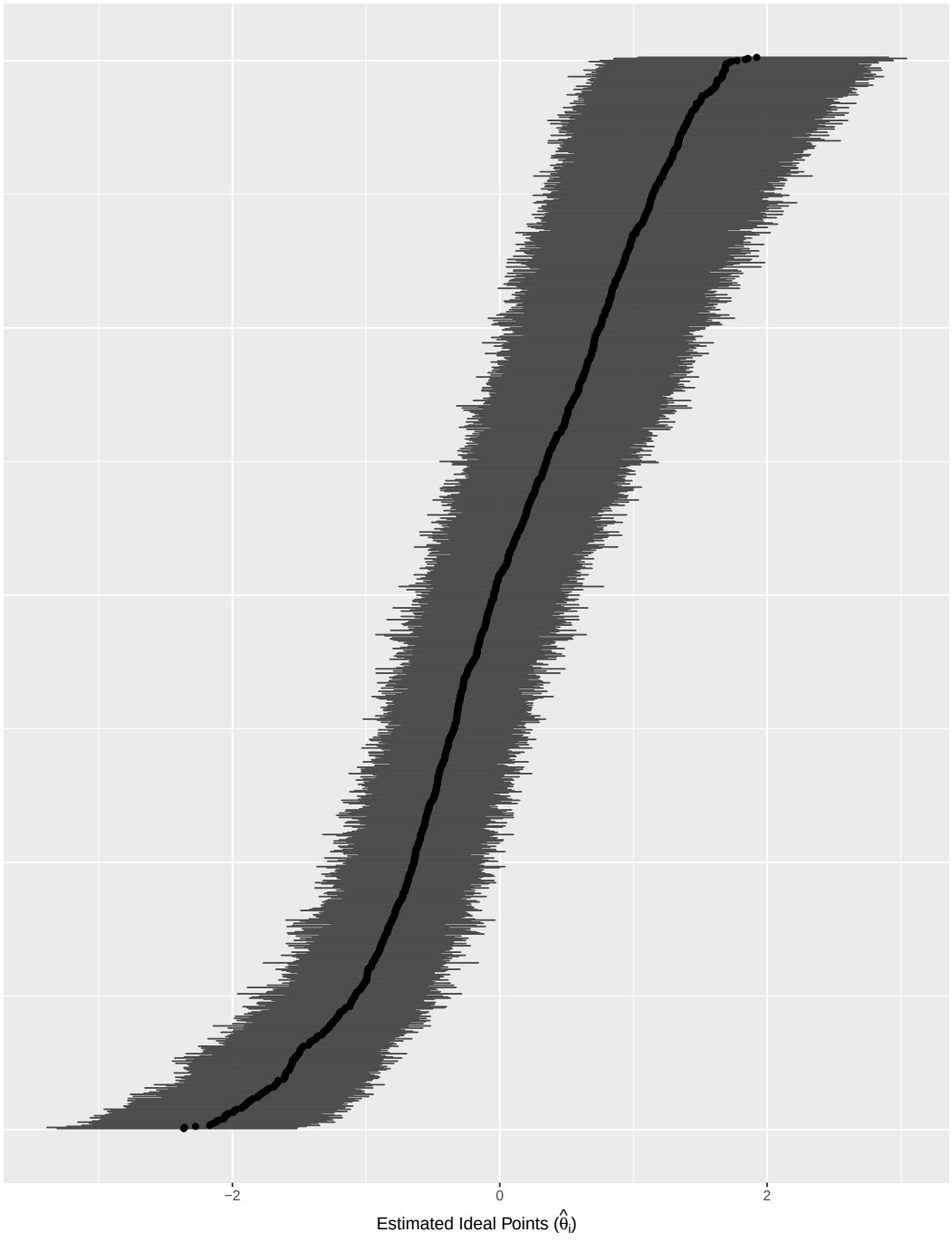

Figure A.4: Estimates of Respondents' Ideal Points (θ).

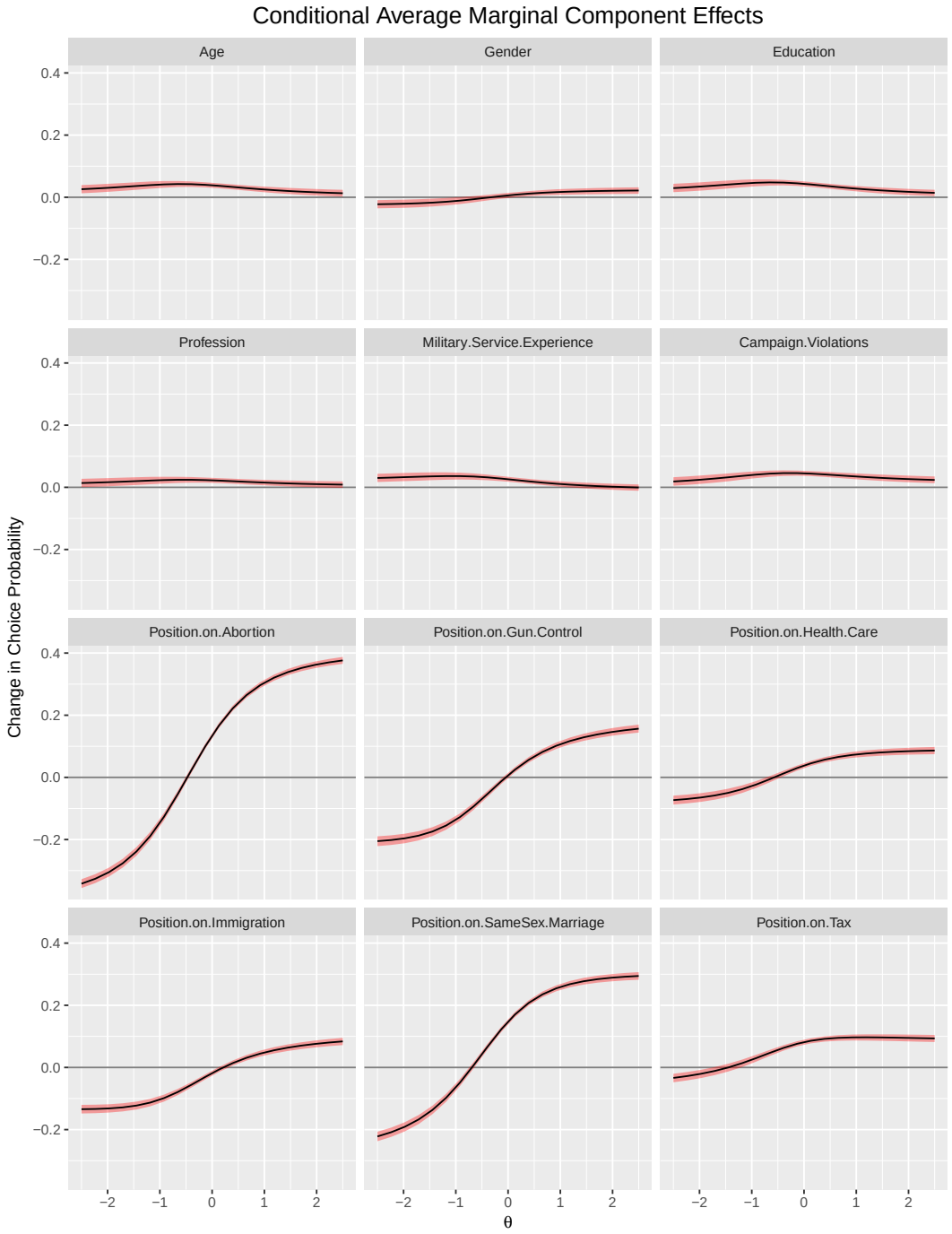

Figure A.5: Estimates of CAMCEs for all Attributes Presented on the Same Scale.

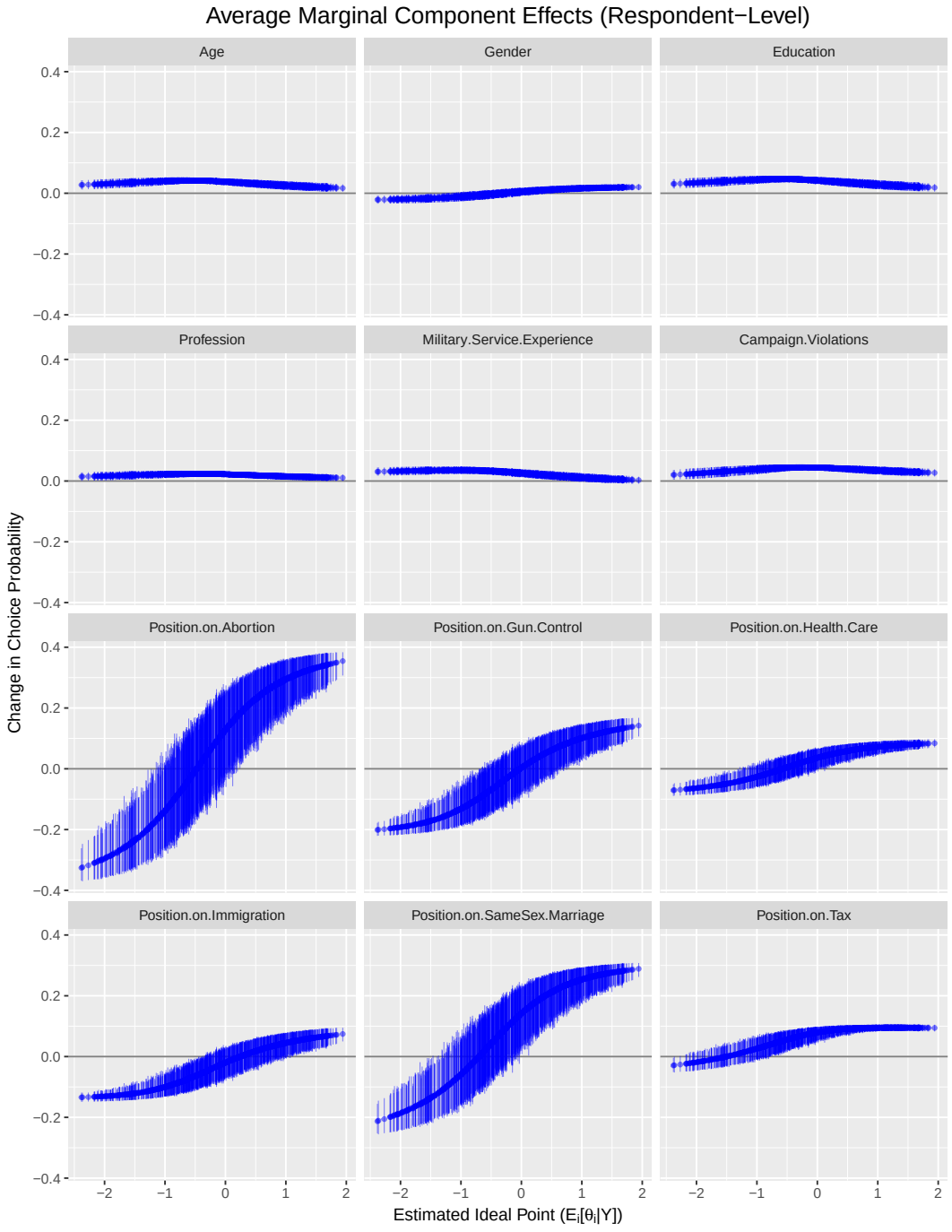

Figure A.6: Estimates of Respondent-Level AMCEs.

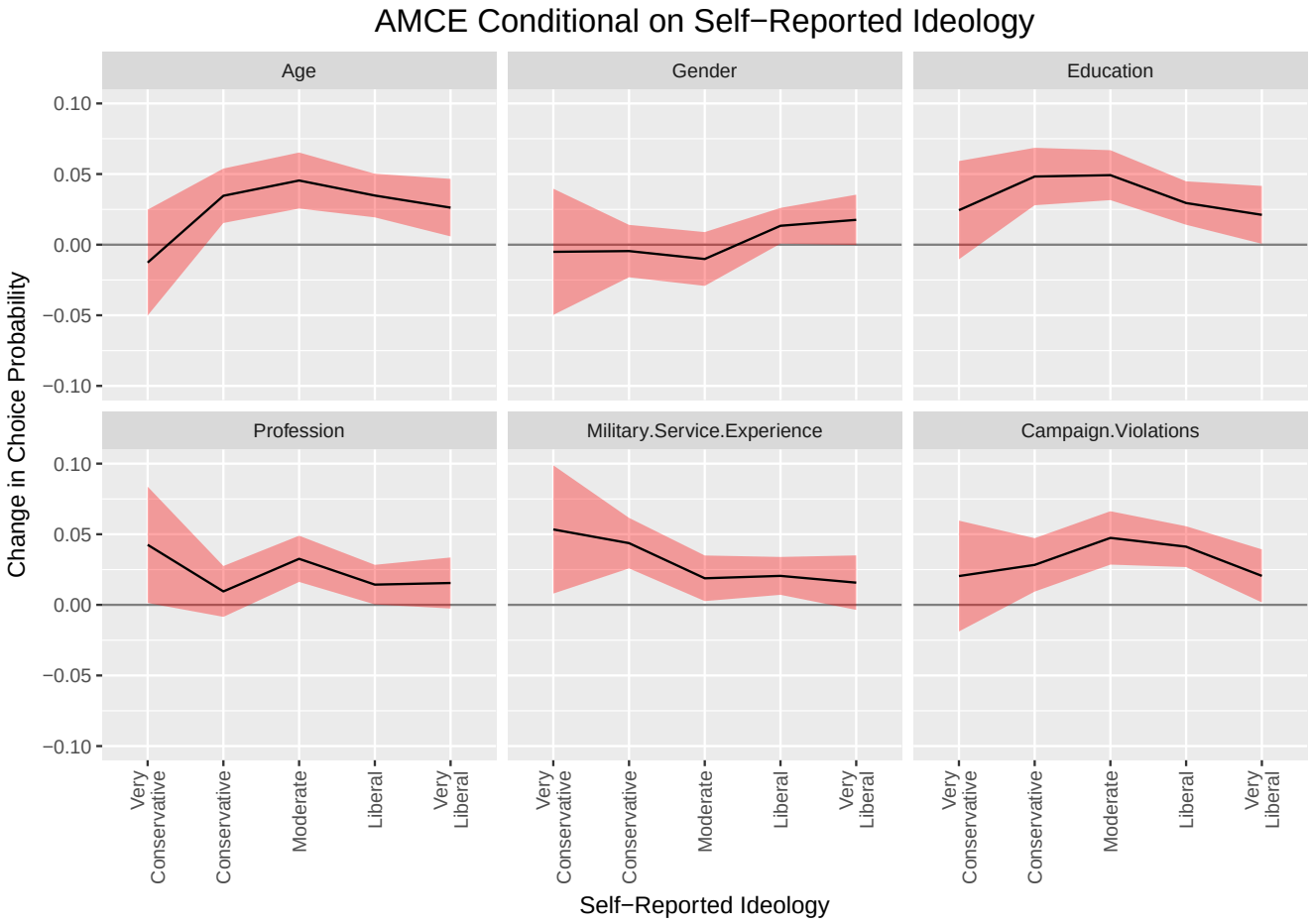

Figure A.7: Estimates of Subgroup AMCEs Conditional on Self-Reported Ideology, Non-Policy Attributes.

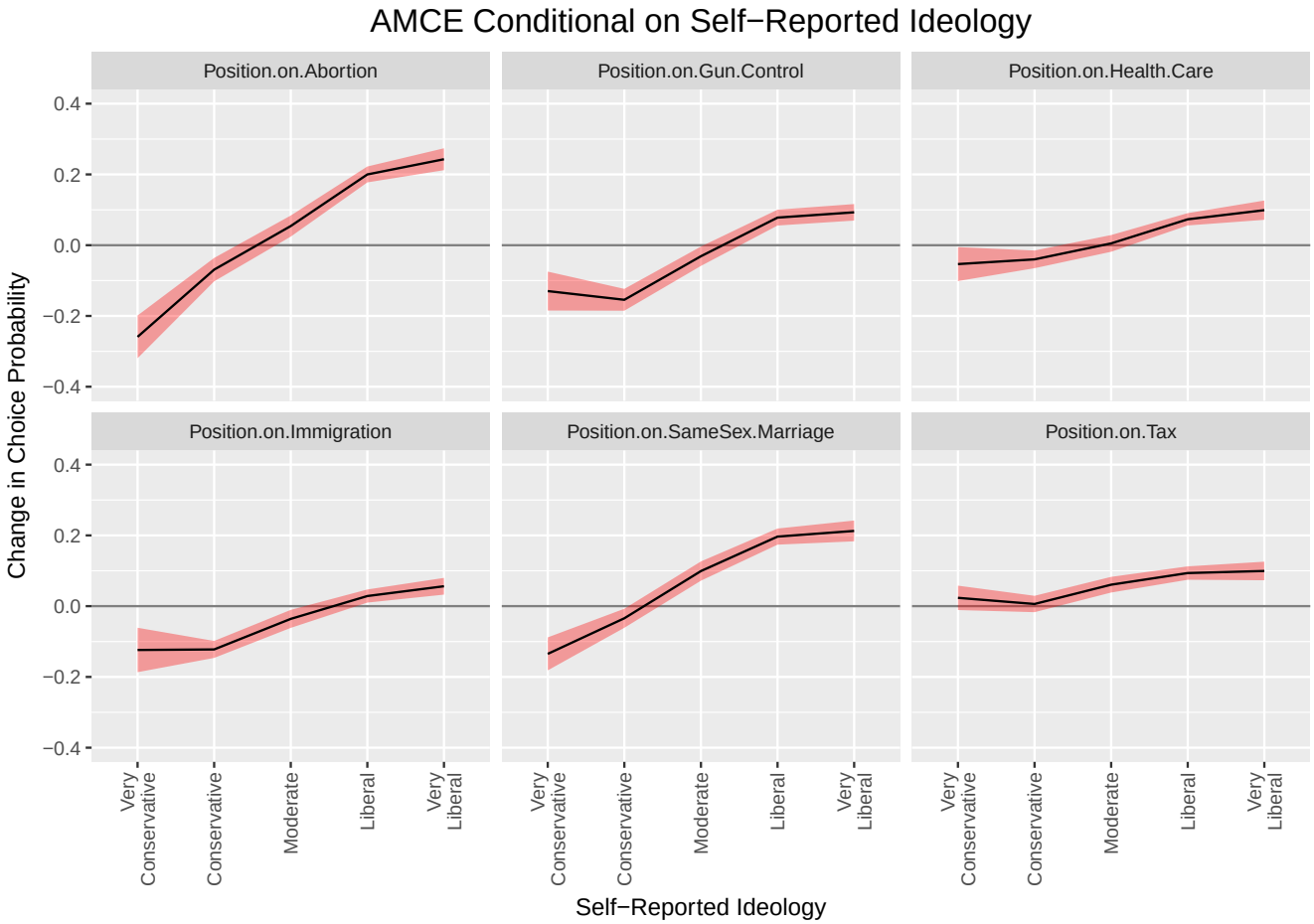

Figure A.8: Estimates of Subgroup AMCEs Conditional on Self-Reported Ideology, Policy Attributes.

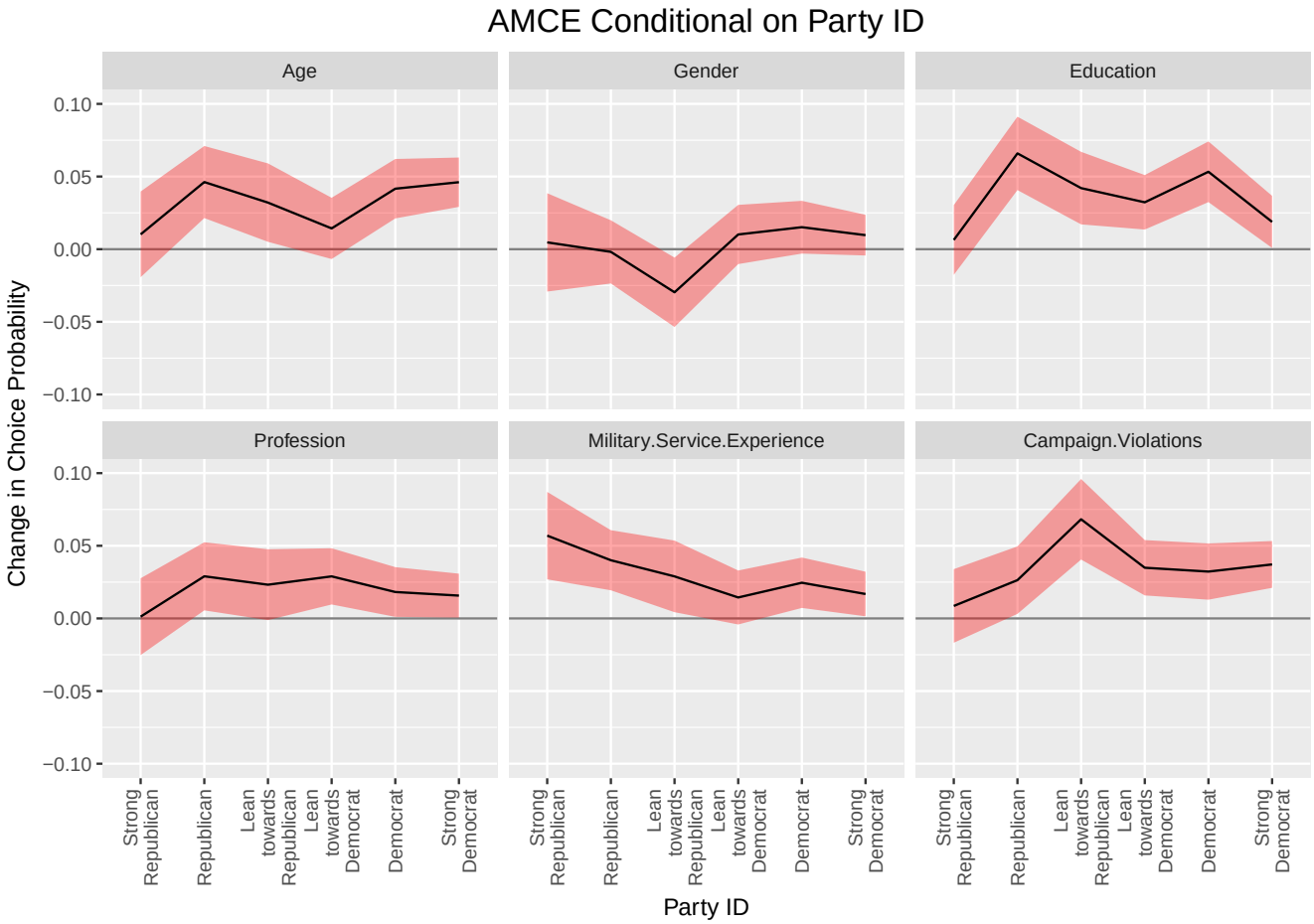

Figure A.9: Estimates of Subgroup AMCEs Conditional on Party Identification, Non-Policy Attributes.

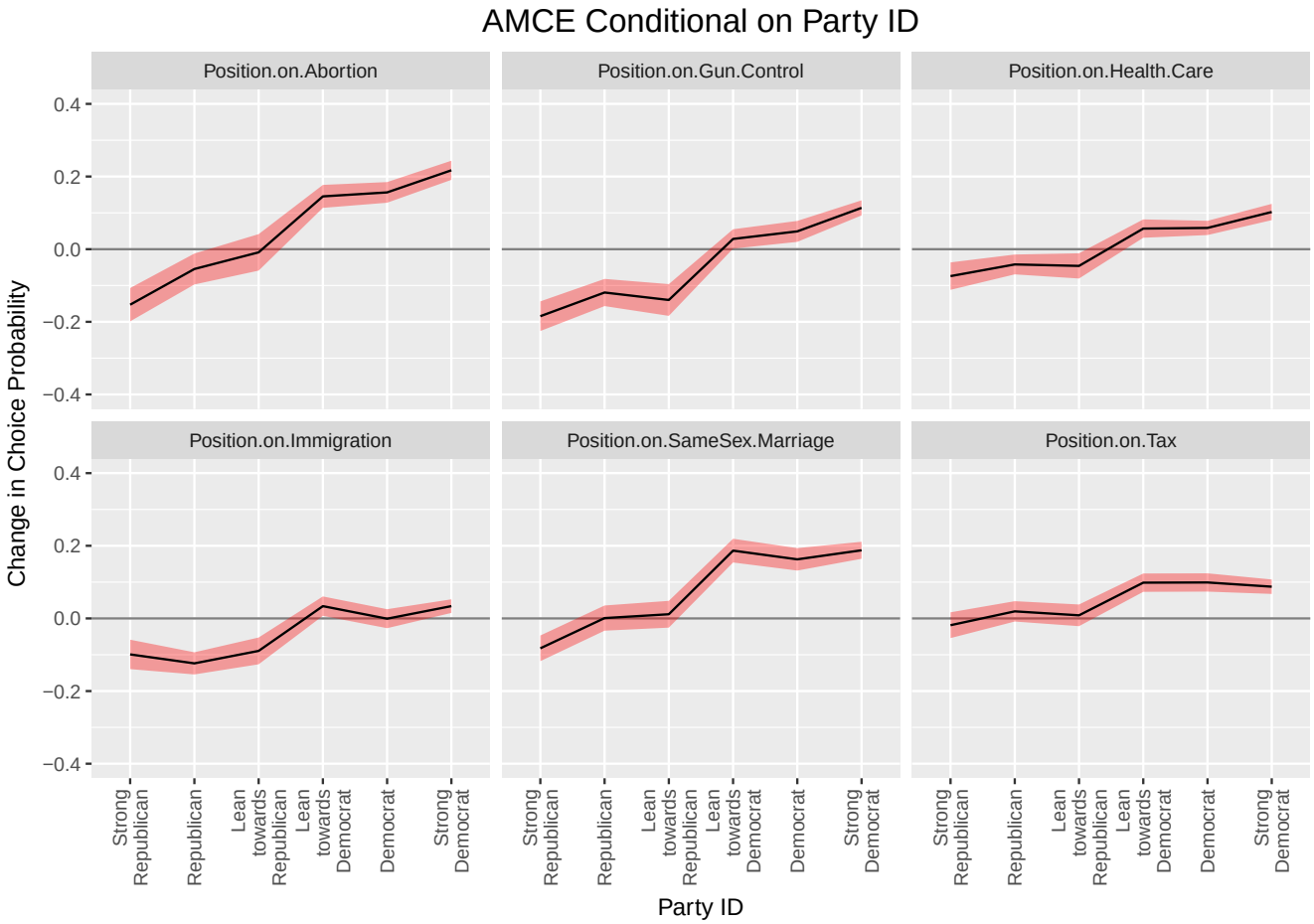

Figure A.10: Estimates of Subgroup AMCEs Conditional on Party Identification, Policy Attributes.

In their original analysis, HSY estimated the AMCEs for each of the nine policy attributes and interpreted them as estimates of the Japanese voters' "multidimensional policy preferences." While those AMCE estimates are interesting in and of themselves, an important question HSY left unanswered is whether there is any latent lower-dimensional structure behind voters' preferences about individual policy issues. More specifically, previous research has shown that the Japanese ideological space can be largely described by a single left-right dimension predominantly determined by international security issues (CITE), although more recent authors argue that the importance of this dimension has been gradually eroding since the 2000s (CITE). Of the nine policy issues in HSY's experiment, two are widely believed to constitute the core of the national discourse on Japan's security policy (Collective Self-defense and Constitutional Revision), whereas the remaining seven issues represent a mixture of economics (Consumption Tax, Employment, Fiscal and Monetary Policy, and Economic Growth Strategy), diplomacy (Trans-Pacific Partnership), energy (Nuclear Power), and political reform (National Assembly Seat Reduction). Is the latent space behind Japanese voters' policy preferences still primarily determined by the classical security policies in 2014? If so, how are the other, more current policy issues associated with the latent left-right ideological dimension? How do voters on both ends of the ideological spectrum coalesce in terms of these new issues?

A.4.2 Results