

# Learning Preferences from Conjoint Data: A Structural Deep Learning Approach

Avidit Acharya\*      Jens Hainmueller†      Yiqing Xu‡

April 12, 2026

## Abstract

Conjoint experiments randomize multidimensional profiles, offering a powerful design for recovering structural preference parameters—including marginal rates of substitution, willingness to pay, and the distribution of preferences across a population. Yet the dominant approach in political science has focused on nonparametric causal estimands that do not leverage this potential. We propose a structural approach that embeds a deep neural network within a random utility logit model, allowing preference parameters to vary as a fully flexible function of respondent characteristics. The neural network addresses the concern that a parametric specification may not capture the true data generating process, while double/debiased machine learning provides valid inference on average preference parameters. We apply our method to three prominent conjoint studies and find rich preference heterogeneity masked by reduced-form averages: a near-zero gender effect coexists with 83% preferring female candidates, opposition to undemocratic behavior is near-universal but varies sharply in intensity, and progressive tax preferences cut across every partisan subgroup.

**Keywords:** conjoint analysis, preference heterogeneity, deep neural networks, double machine learning, random utility model, structural estimation

---

\*Department of Political Science, Stanford University. Email: [avidit@stanford.edu](mailto:avidit@stanford.edu).

†Department of Political Science, Stanford University. Email: [jhain@stanford.edu](mailto:jhain@stanford.edu).

‡Department of Political Science, Stanford University. Email: [yiqingxu@stanford.edu](mailto:yiqingxu@stanford.edu).

# 1 Introduction

Conjoint experiments hold great promise for helping researchers understand how individuals trade off different dimensions of preference. By presenting respondents with multidimensional profiles and asking them to choose, they directly elicit the multi-attribute tradeoffs inherent in real-world decisions. This insight was recognized early: [Greenhalgh and Neslin \(1981\)](#) used conjoint analysis to study negotiator preferences over contract terms, and [Shamir and Shamir \(1995\)](#) explicitly framed conjoint designs as a way to measure value tradeoffs and interactions in mass opinion. Since [Hainmueller et al. \(2014\)](#) introduced conjoint experiments to political science, applications have proliferated across a number of areas from immigration policy ([Hainmueller and Hopkins, 2015](#); [Bansak et al., 2016](#)) to candidate evaluation ([Saha and Weeks, 2022](#)), democratic accountability ([Graham and Svulik, 2020](#)), and tax policy preferences ([Ballard-Rosa et al., 2017](#)).

The ability of conjoints to elicit multi-attribute tradeoffs connects tightly to theoretical quantities at the core of our understanding of political choice—quantities that arise from specifying an underlying utility function that governs choice behavior. These quantities, such as the marginal rate of substitution across preference dimensions, willingness to pay for a policy attribute, or the fraction of a population that prefers one alternative over another, are central to theories of voting, representation, and political economy, and conjoint experiments are, in principle, ideally suited to estimate them. But because estimating these quantities requires imposing parametric structural assumptions on preferences, and because the recent turn toward design-based causal inference in political methodology has been skeptical of such structural assumptions, the development of conjoint analysis in political science focused primarily on nonparametric causal estimands. The average marginal component effect (AMCE) introduced by [Hainmueller et al. \(2014\)](#) and subsequent extensions using BART ([Robinson and Duch, 2024](#)), causal forests ([Abramson et al., 2022](#)), Bayesian mixtures ([Goplerud et al., 2025](#)), and design-based hypothesis tests ([Ham et al., 2024](#)) have dominated applied work in the field. This has meant that the underlying ability of conjoint designs to speak to structural quantities of interest is underdeveloped and under-explored.

Skepticism of structural assumptions is well-motivated—structural models can be wrong and the consequences of misspecification can be severe. In this paper, we address this concern by introducing a method that combines the conjoint design with deep neural networks (DNNs) ([Farrell et al., 2021, 2025](#)) and double/debiased machine learning (DML) ([Chernozhukov et al., 2018](#)). We embed a DNN within a random utility logit model: the network takes each respondent’s observed characteristics as input and outputs that respondent’s complete preference vector, the full set of marginal utilities over all attribute levels in the conjoint

design. This preference vector enters a structural logit choice model, so the predicted choice probability is determined by the randomly assigned profile contrast and the respondent’s estimated preferences jointly. Because the logit structure is preserved, all structural quantities remain well-defined and computable from the estimated preference vectors. The DNN replaces the distributional assumptions of standard mixed logit with a fully flexible mapping from respondent characteristics to preferences, addressing the concern that any particular parametric specification may not capture the data generating process. For inference on population-average preference parameters, we use cross-fitting and the influence-function correction of DML, which provides valid confidence intervals (CIs) despite the flexibility of the first-stage estimator. We think of this approach as *lean structure*—the upside of a structural model without its usual parametric downside.

The payoff is substantial, and it speaks directly to an active debate about what conjoint experiments can and cannot deliver. [Bansak et al. \(2023\)](#) provide a systematic treatment of what conjoint experiments can—and cannot—aggregate, showing that different quantities of interest (the AMCE, the effect of an attribute on the probability of winning, the fraction of voters who prefer a given attribute level) carry different substantive meanings and that the choice of quantity should be driven by the underlying question; a central implication of their analysis is that many theoretically natural quantities—in particular, the fraction of voters who prefer an attribute level—are essentially infeasible to recover using standard reduced-form methods, because those methods identify only marginal population averages and therefore cannot point-identify individual-level distributional features of preferences without additional structure. Our paper develops the structural machinery to recover individual-level preferences directly, and therefore to deliver precisely those quantities of interest that [Bansak et al.](#) and others have identified as substantively desirable but out of reach under existing reduced-form approaches. The power of our approach comes from combining two sources of leverage: the *flexible structure* of a deep neural network, which lets each respondent’s preference vector depend on covariates in an essentially unrestricted way, and the *identification power of randomization*, since the conjoint design randomly assigns profile contrasts to each respondent, pinning down the structural preferences without the strong functional-form assumptions that traditional structural models typically invoke.

We apply our method to three prominent conjoint studies, showing how the structural approach reveals deep preference heterogeneity. In the [Saha and Weeks \(2022\)](#) candidate-choice experiment, we recover the full distribution of preferences across respondents and find that the vast majority prefer female candidates—near-consensus in one direction, masked by a near-zero average effect. Candidate empathy shows the opposite pattern: its AMCE is near zero, but this masks a sharp partisan cleavage—Democrats reward empathy while Republicans

penalize it—so the positive and negative individual-level preferences cancel almost exactly in the average. In the [Graham and Svulik \(2020\)](#) democracy conjoint, the structural model reveals that virtually all voters *oppose* undemocratic behavior, but many weight party and policy more heavily—the disagreement is about intensity, not direction. In the [Ballard-Rosa et al. \(2017\)](#) tax-plan experiment, we recover an individual-level preferred rate schedule for each respondent and find that nearly all have progressive revealed preferences—a near-universal feature across every partisan, income, and ideological subgroup—while Democrats and Republicans differ primarily in *which* brackets drive their choices. The same application provides an individual-level external validation: the DNN-recovered progressivity slope correlates well with each respondent’s self-reported ideal tax rates—a measure collected independently of the conjoint task and never seen by the model during training. Our results show that by focusing on nonparametric causal estimands, prior conjoint studies have been unable to uncover this rich preference heterogeneity.

In the marketing literature, it has long been known that conjoint studies can identify the structural parameters of an underlying utility model. The conjoint measurement foundations laid by [Luce and Tukey \(1964\)](#) and [Green and Rao \(1971\)](#) were linked to random utility models by [McFadden \(1974\)](#), and the connection between conjoint experiments and structural discrete choice has been a staple of marketing research ([Green and Srinivasan, 1978, 1990](#); [Louviere et al., 2000](#); [Train, 2009](#)).<sup>1</sup> Random utility models have also been used extensively in political science—notably the probabilistic spatial-utility frameworks of [Poole and Rosenthal \(1985\)](#) and [Palfrey and Poole \(1987\)](#), the heterogeneous voter-choice models of [Rivers \(1988\)](#) and [Alvarez and Nagler \(1998\)](#), and the Bayesian ideal-point models of [Martin and Quinn \(2002\)](#) and [Clinton et al. \(2004\)](#)—and the formal literature on spatial voting and preference aggregation ([Enelow and Hinich, 1984](#); [Hinich and Munger, 1997](#); [Austen-Smith and Banks, 1999](#)) provides the theoretical foundations for the structural quantities we recover. But these traditions have not been linked to conjoint experiments. Our paper makes this connection, and by using deep neural networks to flexibly estimate preference heterogeneity, addresses the main worry that political scientists have about parametric identification: that the specified model may not be capturing the underlying data generating process.

---

<sup>1</sup>The introduction of conjoint analysis to political science by [Hainmueller et al. \(2014\)](#) does discuss the random utility foundation, citing [Luce and Tukey \(1964\)](#), [Green and Rao \(1971\)](#), and [McFadden \(1974\)](#); see [Bansak et al. \(2021\)](#) for a comprehensive handbook treatment. Our contribution is to develop this structural approach and the quantities it enables.

## 2 Theoretical Framework

This section develops the structural model that underlies our approach. We begin with the conjoint experimental setup and the random utility model that connects profile attributes to respondent choices. We then define the structural quantities of interest that the model enables and show that these quantities are identified by the conjoint randomization design.

### 2.1 Setup

Consider a conjoint experiment in which each of  $M$  respondents evaluates  $T$  choice tasks.<sup>2</sup> In each task  $t$ , respondent  $i$  views two profiles (alternatives) and selects one. Each profile  $j \in \{1, 2\}$  is characterized by a vector of randomly assigned attribute levels, encoded as dummy variables relative to a reference category:  $\mathbf{X}_{ijt} \in \mathbb{R}^p$ , where  $p$  is the total number of non-reference attribute levels across all attributes. To fix ideas, consider the candidate-choice experiment of [Saha and Weeks \(2022\)](#), where respondents evaluate hypothetical candidates for public office described by five attributes—policy agenda, talent, number of children, gender, and prior candidacy—yielding  $p = 13$  dummy-coded attribute levels.

We define the *profile-pair difference*

$$\Delta \mathbf{X}_{it} = \mathbf{X}_{i1t} - \mathbf{X}_{i2t} \in \mathbb{R}^p, \tag{1}$$

which captures the contrast between the two profiles on all attribute dimensions simultaneously. Let  $Y_{it} \in \{0, 1\}$  denote whether respondent  $i$  chose profile 1 in task  $t$ . The total number of choice observations is  $N = MT$ .

Respondent characteristics are collected in a vector  $\mathbf{Z}_i \in \mathbb{R}^{pz}$ , which may include demographics (age, gender, education, income), attitudinal measures (political ideology, party identification), and contextual variables (region, urban/rural residence). For example, in the candidate-choice experiment,  $\mathbf{Z}_i$  includes the respondent’s party identification, ideology, gender, age, education, income, race, and region—15 variables in total. Crucially,  $\mathbf{Z}_i$  is constant across all tasks for respondent  $i$ .

---

<sup>2</sup>We use constant- $T$  notation because the number of tasks per respondent is fixed by design in most conjoint studies; all formulas extend immediately to varying  $T_i$ .

## 2.2 The Random Utility Model

We ground our framework in the canonical random utility model of [McFadden \(1974\)](#). Respondent  $i$  derives utility from profile  $j$  in task  $t$  according to

$$U_{ijt} = \mathbf{X}_{ijt}^\top \boldsymbol{\beta}(\mathbf{Z}_i) + \varepsilon_{ijt}, \quad (2)$$

where  $\boldsymbol{\beta}(\mathbf{Z}_i) \in \mathbb{R}^p$  is a vector of *marginal utilities* that varies as an unknown function of respondent characteristics, and  $\varepsilon_{ijt}$  are independent and identically distributed Type I Extreme Value (Gumbel) taste shocks.

Each coefficient  $\beta_k(\mathbf{Z}_i)$  has a direct interpretation: it is the marginal utility that respondent  $i$  assigns to attribute level  $k$  relative to the reference level of the corresponding attribute. For instance, in the candidate-choice experiment,  $\beta_{\text{Female}}(\mathbf{Z}_i)$  measures how much respondent  $i$  values a female candidate relative to a male candidate—and this valuation may differ across respondents with different party identification, education, or ideology. The key modeling choice is that  $\boldsymbol{\beta}(\cdot)$  is treated as a *fully nonparametric* function: we impose no distributional assumptions on how preferences relate to respondent characteristics. This is what distinguishes our approach from standard mixed logit, which assumes  $\boldsymbol{\beta}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  or restricts the dependence on covariates to be linear.

The respondent chooses profile 1 if  $U_{i1t} > U_{i2t}$ . Since the difference of two independent and identically distributed Gumbel random variables follows a logistic distribution, the forced-choice probability is:

$$\Pr(Y_{it} = 1 \mid \Delta \mathbf{X}_{it}, \mathbf{Z}_i) = G(\Delta \mathbf{X}_{it}^\top \boldsymbol{\beta}(\mathbf{Z}_i)), \quad (3)$$

where  $G(v) = (1 + e^{-v})^{-1}$  is the logistic cumulative distribution function. The profile-pair differencing (1) eliminates any alternative-specific constant, so we need only estimate the single parameter function  $\boldsymbol{\beta}(\cdot)$ .

The key assumption of this model is an additive utility that is linear in attribute levels, which rules out attribute interactions unless explicitly included. The logistic link from the Gumbel error distribution is a secondary modeling choice—common alternatives such as probit are nearly indistinguishable in practice. While  $\boldsymbol{\beta}(\mathbf{Z})$  is left fully nonparametric, the additive utility structure is what distinguishes the approach from purely design-based methods and is the source of its additional identifying power. We discuss its limitations in [Section 5](#), where we also discuss some possible extensions to address these limitations.

## 2.3 Quantities of Interest

The structural model defined by equations (2) and (3) enables a set of quantities that are inaccessible to reduced-form analysis. We focus here on the quantities that we estimate in our applications below; all require the individual-level preference vector  $\beta(\mathbf{Z}_i)$  and cannot be recovered from non-parametric causal estimands or from heterogeneous treatment effect estimators that operate one attribute at a time.

1. *Average preference parameters.* The primary estimand is the population-average marginal utility of attribute level  $k$ :

$$\theta_k = \mathbb{E}[\beta_k(\mathbf{Z}_i)], \quad k = 1, \dots, p.$$

Substantively,  $\theta_k$  summarizes how much the average voter in the population rewards or penalizes attribute  $k$ —the baseline quantity for describing aggregate preferences. On the logit scale,  $\theta_k$  is directly comparable to the homogeneous logit coefficient.

2. *Average marginal effects.* The average preference parameter  $\theta_k$  lives on the logit scale. Its probability-scale counterpart is the average marginal effect (AME):

$$\text{AME}_k = \mathbb{E}_{\mathbf{Z}_i, \mathbf{X}_{-k}} \left[ G\left(\beta_k(\mathbf{Z}_i) + \mathbf{X}_{-k}^\top \beta_{-k}(\mathbf{Z}_i)\right) - G\left(\mathbf{X}_{-k}^\top \beta_{-k}(\mathbf{Z}_i)\right) \right],$$

which measures the average change in the probability of choosing a profile when attribute level  $k$  is switched on, averaging over respondent characteristics and the randomization distribution of all other attributes. Under correct specification of the random utility model, the structural AME is exactly the AMCE of [Hainmueller et al. \(2014\)](#): the two estimands coincide because the expectation over the experimental randomization distribution is built into both. This equivalence is a testable implication of the model: if the random utility specification is correct, the structural AME and the nonparametric AMCE should agree; a discrepancy between the two could signal misspecification, though in finite samples it may also reflect estimation noise, particularly when the sample is small. We use this comparison as a diagnostic in each application below, where the structural AME and the linear probability model AMCE are nearly indistinguishable.

3. *Individual preference vectors.* The estimator recovers  $\hat{\beta}(\mathbf{Z}_i) \in \mathbb{R}^p$  for each respondent—the complete vector of marginal utilities across all attribute levels simultaneously. Precisely,  $\hat{\beta}(\mathbf{Z}_i)$  is a function of observed respondent characteristics  $\mathbf{Z}_i$ , capturing heterogeneity across respondents to the extent that their preferences vary with observables; [Section 5](#) discusses the distinction from respondent-specific latent preferences. It is the central

structural object, because every other quantity below is a functional of it, and it is the quantity that reduced-form methods cannot deliver.

4. *Preference polarization.* For each attribute  $k$ , the structural model reveals the fraction of respondents with  $\beta_k(\mathbf{Z}_i) > 0$  versus  $< 0$ , decomposing the *direction* of a preference (does the respondent favor or oppose the attribute?) from its *intensity* (how strongly?). This matters because the AMCE cannot distinguish consensus indifference from hidden polarization: a near-zero average can mask an electorate split into strong supporters and strong opponents whose effects cancel out, and only the direction-vs.-intensity decomposition can tell the two cases apart.
5. *Attribute importance.* Under conjoint randomization, the variance of utility decomposes additively across attributes:

$$\text{Var}(\mathbf{X}^\top \boldsymbol{\beta}(\mathbf{Z}_i)) = \sum_{a=1}^A \text{Var}(\mathbf{X}_a^\top \boldsymbol{\beta}_a(\mathbf{Z}_i)),$$

where  $a$  indexes attributes. Normalizing to shares yields a respondent-specific importance ranking that answers, for each voter, which attributes dominate the choice and which are essentially ignored. This matters because voters who appear similar on average can weight attributes very differently—some are near-singly focused on a single dimension while others spread attention across many—and that heterogeneity in *what voters care about* is invisible to an average-effect analysis.

6. *Marginal rates of substitution (MRS).* The trade-off between attributes  $j$  and  $k$  for respondent  $i$  is:

$$\text{MRS}_{jk}(\mathbf{Z}_i) = \frac{\beta_j(\mathbf{Z}_i)}{\beta_k(\mathbf{Z}_i)},$$

which measures how many units of attribute  $k$  must change to compensate for a one-unit change in attribute  $j$ , holding utility constant. The MRS is the natural measure of how voters *trade off* competing considerations—party against policy, democratic norms against partisan loyalty, environmental protection against cost—and is expressed in the units of the denominator attribute, which makes it directly interpretable. When  $k$  indexes a monetary attribute, the MRS reduces to willingness to pay (WTP), expressed in dollars. The MRS requires the ratio of two preference parameters *for the same respondent*, which reduced-form methods cannot deliver; moreover, the ratio of population averages  $\mathbb{E}[\beta_j]/\mathbb{E}[\beta_k]$  does not equal the average of the individual-level ratio  $\mathbb{E}[\beta_j/\beta_k]$  under heterogeneity (Jensen’s inequality), so the MRS cannot be recovered from aggregate estimates alone.

7. *Compensating differentials.* For a penalty attribute with  $\beta_j(\mathbf{Z}_i) < 0$ , one can identify which attribute levels  $k$  satisfy  $\beta_j(\mathbf{Z}_i) + \beta_k(\mathbf{Z}_i) \geq 0$ , and compute the fraction of respondents for whom compensation holds. This is the binary “would you take the deal?” question—the discrete counterpart to the MRS—and it matters because many substantive questions in politics take exactly this form: how many voters will tolerate one thing they dislike in exchange for something they prefer, and how does that tolerance vary across the electorate? It requires the joint distribution of preferences across attributes within each respondent.
8. *Counterfactual choice probabilities.* For any pair of hypothetical profiles  $A$  and  $B$  with attribute vectors  $\mathbf{X}_A$  and  $\mathbf{X}_B$ :

$$\Pr(\text{choose } A \text{ over } B \mid \mathbf{Z}_i) = G\left((\mathbf{X}_A - \mathbf{X}_B)^\top \hat{\boldsymbol{\beta}}(\mathbf{Z}_i)\right).$$

This predicts the outcome of any head-to-head contest between fully specified profiles for any targeted respondent or subgroup—for example, the probability that voters with given demographics would choose one candidate configuration over another. This requires the respondent’s joint preference vector, not separate attribute-by-attribute effects, because the nonlinear logistic link means the combined contrast is not the sum of marginal effects.

The preference vectors also support a number of additional structural quantities that may be of interest depending on the application, including: consumer surplus via the closed-form welfare measure of [McFadden \(1981\)](#), indifference sets that map individual-specific substitution patterns geometrically, preference clustering via  $k$ -means or other algorithms applied to  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i) \in \mathbb{R}^p$ , preference inequality measures such as  $\text{Var}(\hat{\beta}_k(\mathbf{Z}_i))$  or Gini coefficients, decisiveness measures  $|(\mathbf{X}_A - \mathbf{X}_B)^\top \hat{\boldsymbol{\beta}}(\mathbf{Z}_i)|$  capturing the strength of preference for a given profile comparison, and the *majority preference function*—for any profile pair  $(A, B)$ , the fraction of respondents whose deterministic utility index  $(\mathbf{X}_A - \mathbf{X}_B)^\top \hat{\boldsymbol{\beta}}(\mathbf{Z}_i) > 0$ .<sup>3</sup> We refer the reader to [Train \(2009\)](#), [Louviere et al. \(2000\)](#), and [Green and Srinivasan \(1990\)](#) for detailed treatments of these quantities.

**Relations to the AMCE.** As noted above, the structural AME recovers the AMCE under correct specification, so the structural model complements rather than replaces what [Bansak et al. \(2023\)](#) show is a valuable estimand that maps directly to aggregate vote shares. The key limitation of the AMCE is that it estimates the effect of one attribute marginalized over all others and therefore cannot recover the joint preference vector for a given respondent. This

---

<sup>3</sup>Unlike logit choice probabilities, the majority preference function abstracts from idiosyncratic taste shocks and directly addresses the concern raised by [Abramson et al. \(2022\)](#) that the AMCE can indicate the opposite of the true majority preference. See also our discussion of this criticism below.

means it cannot deliver marginal rates of substitution, counterfactual choice probabilities for complete profiles, compensating differentials, attribute importance decompositions, or preference polarization—each of which is a functional of  $\beta(\mathbf{Z}_i)$ . Conditional AMCEs interact treatment indicators with covariates or estimate the AMCE within subgroups (Hainmueller et al., 2014; Leeper et al., 2020), but they still average within subgroups rather than recovering individual-level preferences. Relatedly, de la Cuesta et al. (2022) note that under preference heterogeneity the AMCE depends on the profile-randomization distribution and is not a structural parameter of the utility model, and Abramson et al. (2022) show that the AMCE can indicate the opposite of the true majority preference—a quantity the structural approach point-identifies for each respondent. Section 4 illustrates these distinctions in practice.

## 2.4 Identification

All of the quantities of interest defined above—average preference parameters, average marginal effects, individual preference vectors, preference polarization, attribute importance, marginal rates of substitution, compensating differentials, and counterfactual choice probabilities—are derived from the individual-level preference function  $\beta(\mathbf{Z})$ . It is therefore sufficient to show that  $\beta(\mathbf{Z})$  itself is identified, since every quantity of interest listed above is a known functional of  $\beta(\mathbf{Z})$  and inherits identification from it. We now show that  $\beta(\mathbf{Z})$  is identified whenever (i) the profile contrast  $\Delta\mathbf{X}_{it}$  is randomly assigned independently of respondent characteristics  $\mathbf{Z}_i$ , and (ii) the random utility model (2)–(3) is correctly specified. Both conditions are built into the conjoint design.

By construction,  $\Delta\mathbf{X}_{it}$  is randomly assigned and therefore independent of  $\mathbf{Z}_i$ :

$$\Delta\mathbf{X}_{it} \perp\!\!\!\perp \mathbf{Z}_i.$$

This exogeneity condition, guaranteed by the experimental design, means that the conditional choice probability (3) identifies the preference function  $\beta(\mathbf{Z})$  without concerns about omitted variable bias or endogenous sorting that plague observational discrete choice.

To see what this buys us, consider a simplified version of the candidate-choice experiment with two binary attributes (candidate gender and party) and a binary respondent characteristic (college degree). Randomization means that within each education group, differences in choice probabilities across profile contrasts identify the preference coefficients directly: a comparison of female versus male candidates (holding party fixed) isolates  $\beta_{\text{Female}}(z)$ , and a comparison of Democrat versus Republican candidates (holding gender fixed) isolates  $\beta_{\text{Party}}(z)$ . Once these two coordinates are identified, the structural logit model implies the probability of choosing a female Democrat over a male Republican as  $G(\beta_{\text{Female}}(z) + \beta_{\text{Party}}(z))$ —the logit

model links the isolated effects into a joint preference vector. Randomization identifies each attribute’s marginal utility, and the structural model combines them.

We emphasize that identification here has two components: experimental randomization nonparametrically identifies the conditional choice probability  $\Pr(Y_{it} = 1 \mid \Delta \mathbf{X}_{it}, \mathbf{Z}_i)$ , eliminating the endogeneity concerns that dominate observational discrete choice, and the logit functional form inverts these probabilities to recover the preference function  $\beta(\mathbf{Z})$ . The logit link is therefore the remaining structural assumption.

Two features, however, limit its bite. The DNN is fully flexible in how it maps respondent characteristics to preferences, so the parametric restriction applies only to how preferences map to choice probabilities, not to the heterogeneity itself. In practice, common link functions (logit, probit) are nearly indistinguishable in the interior of the probability space, and the DNN can partially compensate for link misspecification by rescaling the preference vector; the binding structural assumption is additive utility, not the choice of error distribution.

### 3 Estimation

We parameterize  $\beta(\mathbf{Z})$  as a deep neural network (DNN) with a structural output layer, following Farrell et al. (2025). The estimation procedure has three stages: (i) train the DNN to learn the preference function, (ii) cross-fit to avoid overfitting bias, and (iii) apply the DML influence-function correction for valid inference.

The architecture consists of a *feature network* that maps respondent characteristics to preference parameters, and a *model layer* that embeds these parameters in the structural logit model. The feature network takes  $\mathbf{Z} \in \mathbb{R}^{p_z}$  as input and applies  $L$  hidden layers with ReLU activations:

$$\mathbf{h}_\ell = \text{ReLU}(\mathbf{W}_\ell \mathbf{h}_{\ell-1} + \mathbf{b}_\ell), \quad \ell = 1, \dots, L,$$

with  $\mathbf{h}_0 = \mathbf{Z}$ , where  $\text{ReLU}(x) = \max(0, x)$  is a standard nonlinear activation function applied elementwise, and  $\mathbf{W}_\ell, \mathbf{b}_\ell$  are weight matrices and bias vectors learned during training. The final hidden layer maps to the  $p$ -dimensional preference vector:

$$\beta(\mathbf{Z}) = \mathbf{W}_{L+1} \mathbf{h}_L + \mathbf{b}_{L+1} \in \mathbb{R}^p,$$

with no activation function on the output, since preference parameters are unrestricted in sign and magnitude. The model layer then computes the logit index  $\Delta \mathbf{X}^\top \beta(\mathbf{Z})$ , which enters the logistic choice probability. This is the key architectural choice: the model layer enforces the linear-in-parameters utility structure of the logit model, so the DNN learns the preference function  $\beta(\mathbf{Z})$  rather than a free-form mapping from  $(\Delta \mathbf{X}, \mathbf{Z})$  to choice probabilities. The

network is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^M \sum_{t=1}^T \left[ Y_{it} \log G(\Delta \mathbf{X}_{it}^\top \boldsymbol{\beta}(\mathbf{Z}_i)) + (1 - Y_{it}) \log(1 - G(\Delta \mathbf{X}_{it}^\top \boldsymbol{\beta}(\mathbf{Z}_i))) \right],$$

which is exactly the negative log-likelihood of the heterogeneous logit model—the same objective function a researcher would maximize when estimating a standard logit, except that the coefficients now depend on respondent characteristics through the network. This structural loss ensures that the network learns parameters with the economic interpretation of marginal utilities, not merely prediction-optimal features. The estimation pipeline can be summarized schematically as:

$$\mathbf{Z} \xrightarrow[\text{(nonparametric)}]{\text{hidden layers}} \hat{\boldsymbol{\beta}}(\mathbf{Z}) \xrightarrow[\text{(parametric)}]{\text{model layer}} \Delta \mathbf{X}^\top \hat{\boldsymbol{\beta}}(\mathbf{Z}) \xrightarrow[\text{(link function)}]{G(\cdot)} \hat{Y}.$$

Naively plugging in  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i)$  to estimate  $\theta_k = \mathbb{E}[\beta_k(\mathbf{Z})]$  yields biased estimates due to overfitting. We address this via  $K$ -fold cross-fitting (Chernozhukov et al., 2018): each of  $M$  respondents is assigned to one of  $K$  folds (all tasks from the same respondent go to the same fold), the DNN is trained on  $K - 1$  folds, and  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i)$  is computed for held-out respondents. After cycling through all folds, every observation has a predicted preference vector estimated without using that observation’s data.

Even with cross-fitting, a first-order bias remains from the estimation error of  $\hat{\boldsymbol{\beta}}$ . The DML framework corrects this via an influence function. For attribute  $k$ , respondent  $i$ , and task  $t$ , define:

$$\psi_{ikt} = \hat{\beta}_k(\mathbf{Z}_i) + \left[ \hat{\Lambda}^{-1}(\mathbf{Z}_i) \cdot \Delta \mathbf{X}_{it} \cdot (Y_{it} - \hat{G}_{it}) \right]_k,$$

where  $\hat{G}_{it} = G(\Delta \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}}(\mathbf{Z}_i))$  and  $[\cdot]_k$  denotes the  $k$ -th element. The first term is the plug-in estimate; the second is the debiasing correction, which projects the logit residual  $(Y_{it} - \hat{G}_{it})$  through the treatment  $\Delta \mathbf{X}_{it}$  and the inverse local information matrix:

$$\Lambda_{jk}(\mathbf{Z}) = \mathbb{E} \left[ G'(\Delta \mathbf{X}^\top \boldsymbol{\beta}(\mathbf{Z})) \cdot \Delta X_j \cdot \Delta X_k \mid \mathbf{Z} \right],$$

where  $G'(v) = G(v)(1 - G(v))$  is the logistic density. The matrix  $\Lambda(\mathbf{Z})$  measures how informative the typical profile contrast is about the preference parameters for a respondent of type  $\mathbf{Z}$ : it is large when profile contrasts vary widely and when choice probabilities are near one-half (where the logit model is most informative about the underlying preferences). Intuitively, if the DNN underpredicts the choice probability for certain profile contrasts, the correction adjusts the preference estimate upward for the corresponding attributes. The

debiased estimator of the average preference parameter  $\theta_k = \mathbb{E}[\beta_k(\mathbf{Z})]$  is then:

$$\hat{\theta}_k = \frac{1}{N} \sum_{i=1}^M \sum_{t=1}^T \psi_{ikt}.$$

The influence function  $\psi_{ikt}(\theta, \boldsymbol{\beta})$  satisfies the Neyman orthogonality condition:

$$\left. \frac{\partial}{\partial r} \mathbb{E}[\psi_{ikt}(\theta_k^0, \boldsymbol{\beta}_0 + r \mathbf{h})] \right|_{r=0} = 0 \quad \text{for all } \mathbf{h},$$

where  $\mathbf{h}$  is an arbitrary perturbation direction. This means that first-order errors in  $\hat{\boldsymbol{\beta}}$  do not propagate into  $\hat{\theta}_k$ , permitting  $\sqrt{N}$ -consistent inference despite the slower convergence rate of the nonparametric first stage.

The variance is estimated with clustering at the respondent level:

$$\widehat{\text{Var}}(\hat{\theta}_k) = \frac{M}{M-1} \cdot \frac{1}{N^2} \sum_{m=1}^M \left( \sum_{t:i(t)=m} \psi_{mkt} \right)^2.$$

Clustering at the respondent level is essential: a single respondent contributes multiple observations (one per task), and these observations are correlated through the shared preference vector  $\boldsymbol{\beta}(\mathbf{Z}_i)$ ; ignoring this correlation would understate the standard errors (SEs), as we confirm empirically in each application below. Under regularity conditions given in [Farrell et al. \(2025\)](#) and [Chernozhukov et al. \(2018\)](#)—which require the DNN to converge to the true  $\boldsymbol{\beta}(\mathbf{Z})$  at a rate faster than  $N^{-1/4}$ —the DML estimator is  $\sqrt{N}$ -consistent and asymptotically normal:

$$\frac{\hat{\theta}_k - \theta_k}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_k)}} \xrightarrow{d} N(0, 1).$$

As a basic consistency check, [Supplementary Materials A](#) shows that the DNN’s average  $\hat{\theta}_k$  reproduces the coefficient of a pooled homogeneous logit on the same data almost exactly ( $r = 0.998$  across 28 attribute levels), and that its within-subgroup conditional means reproduce subgroup-specific logit coefficients with comparable accuracy ( $r = 0.955$  across 15 country subgroups  $\times$  28 attributes), so any finding a reduced-form AMCE analysis would deliver is recoverable as a special case of the structural estimator. We further validate the finite-sample performance of the estimator—including coverage, individual-level preference recovery, and comparison to alternative methods—via Monte Carlo simulations in [Supplementary Materials B](#). [Table 1](#) summarizes the methodological landscape.

|                                  | <i>Structural</i>          |                              |                             |                            | <i>Reduced-form</i> |             |
|----------------------------------|----------------------------|------------------------------|-----------------------------|----------------------------|---------------------|-------------|
|                                  | DNN<br>(ours) <sup>†</sup> | Homog.<br>Logit <sup>†</sup> | Mixed<br>Logit <sup>†</sup> | Bayes.<br>HLM <sup>†</sup> | AMCE                | BART/<br>CF |
| Utility model                    | Yes                        | Yes                          | Yes                         | Yes                        | Implicit            | No          |
| Preference heterogeneity         | Nonparam.                  | None                         | Parametric                  | Parametric                 | None                | Nonparam.   |
| Systematic by $\mathbf{Z}$       | Flexible                   | —                            | Linear                      | Linear                     | —                   | Flexible    |
| Individual $\beta(\mathbf{Z}_i)$ | Direct                     | —                            | Posterior*                  | Posterior*                 | —                   | —           |
| Structural quantities            | Yes                        | Yes                          | Yes                         | Yes                        | No                  | No          |
| Distributional assumption        | None                       | —                            | Normal                      | Normal                     | None                | None        |
| Inference on $\theta_k$          | DML                        | MLE                          | MLE                         | MCMC                       | OLS                 | —           |

*Notes.* Structural quantities include MRS, WTP, counterfactual choice probabilities, and compensating differentials—all of which require the utility model and individual preference vectors. \*Individual-level posteriors from Mixed Logit and Bayesian HLM are heavily shrunk toward the population mean when the number of tasks per respondent is small relative to the number of parameters (see Supplementary Materials B). BART/CF denotes BART (Robinson and Duch, 2024) and causal forests (Abramson et al., 2022). “—” indicates the feature is not applicable or not provided by the method.

**Table 1:** Comparison of approaches to conjoint analysis. Methods marked with † are included as benchmarks in the Monte Carlo simulations (Supplementary Materials B).

## 4 Applications

We illustrate the structural DNN estimator with three published conjoint experiments that highlight distinct strengths of the approach. The first application recovers the distribution of preferences across respondents—including preference polarization and compensating differentials—in a candidate-choice experiment. The second reframes a prominent finding about democratic accountability by recovering the direction–intensity decomposition and the marginal rate of substitution between democracy and partisanship. The third demonstrates willingness to pay as a structural quantity, enabled by a conjoint design with a monetary attribute.

### 4.1 Political Candidate Preferences

Saha and Weeks (2022) study how voters evaluate candidates for public office using a conjoint experiment fielded to 1,191 respondents. Each respondent evaluated three pairs of hypothetical candidates described by five attributes: policy agenda (3 levels), talent (7 levels), number of children (4 levels), gender (2 levels), and prior candidacy (2 levels), yielding  $p = 13$  dummy-coded attribute levels. We specify the structural DNN with architecture  $\mathbf{Z} \rightarrow 32 \rightarrow 32 \rightarrow 16 \rightarrow \beta \in \mathbb{R}^{13}$ , using  $K = 10$  cross-fitting folds and 2,000 training epochs. The covariate vector  $\mathbf{Z}_i$  includes 15 respondent characteristics (party identification, ideology, gender, age, education, income, race, and region).

Panel A of Figure 1 reports the DML-corrected average preference estimates  $\hat{\theta}_k$  with 95% confidence intervals on the logit scale. The dominant attributes are policy agenda: both

Moderate Changes ( $\hat{\theta} = 0.81$ ,  $p < 0.001$ ) and Complete Overhaul ( $\hat{\theta} = 0.81$ ,  $p < 0.001$ ) are strongly preferred over the reference category (Very Few Changes). Among talent attributes, only Hard-Working is statistically significant ( $\hat{\theta} = 0.45$ ,  $p < 0.001$ ). The gender coefficient—Male relative to Female—is negative but not significant ( $\hat{\theta} = -0.11$ , 95% CI  $[-0.24, 0.01]$ ), consistent with the original AMCE finding. Clustered standard errors exceed unclustered standard errors by a factor of 1.78, confirming the importance of respondent-level clustering.

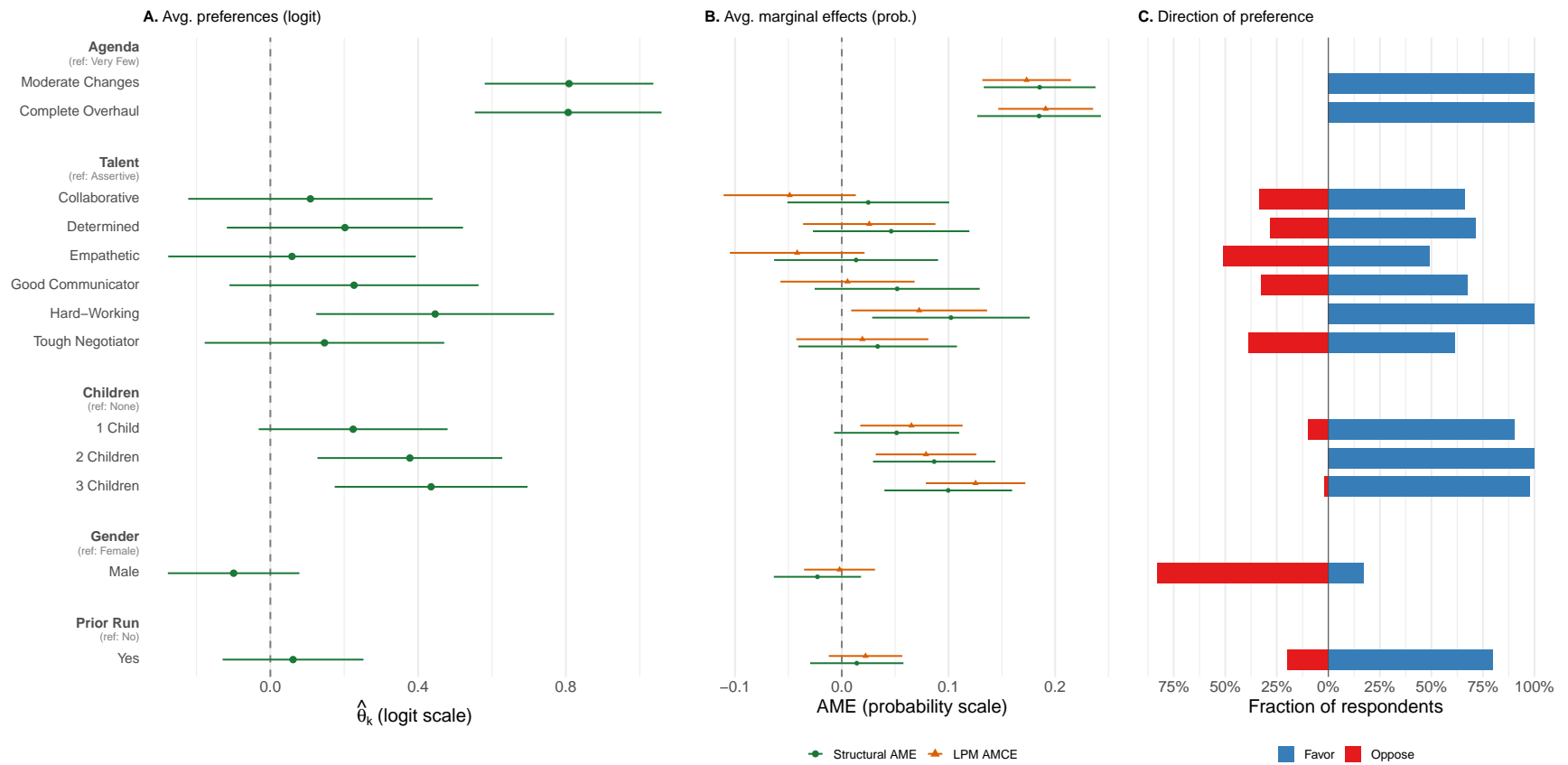
Panel B converts these estimates to the probability scale—average marginal effects—and overlays the AMCE from a standard linear probability model. The two sets of estimates are nearly indistinguishable, confirming that the structural model’s population average  $\hat{\theta}_k = n^{-1} \sum_i \hat{\beta}_k(\mathbf{Z}_i)$  reproduces the standard reduced-form AMCE as a special case. This equivalence is a feature of the method: any finding that a conventional AMCE analysis would deliver remains fully recoverable, while the structural model additionally provides the individual-level preference vectors  $\hat{\beta}_k(\mathbf{Z}_i)$  that underlie these averages.

Panel C reveals what lies behind the averages. Because the structural model recovers  $\hat{\beta}_k(\mathbf{Z}_i)$  for each respondent, we can compute the fraction with  $\hat{\beta}_k(\mathbf{Z}_i) > 0$  (favoring the attribute level) versus  $< 0$  (opposing it). Three patterns are striking.

First, despite the near-zero average effect for gender, 83% of respondents have  $\hat{\beta}_{\text{Male}} < 0$ , meaning they prefer female candidates all else equal. This is not consensus indifference—it is near-consensus in one direction.

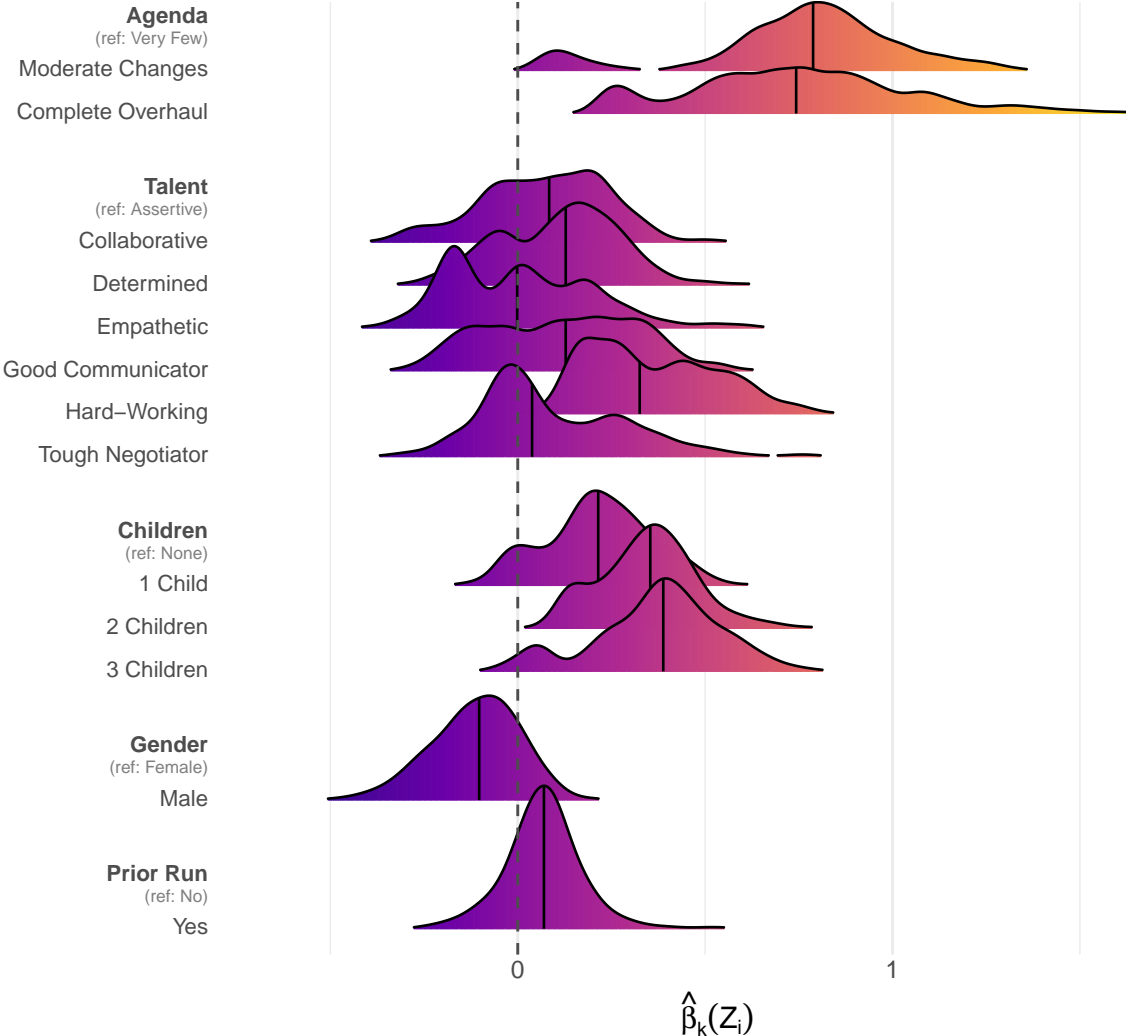
Second, Empathetic has a near-zero average preference parameter ( $\hat{\theta} = -0.02$ ), but the structural model reveals an almost perfectly polarized electorate: 49% favor empathetic candidates while 51% oppose them. This polarization is sharply partisan: Democratic respondents have a mean  $\hat{\beta}_{\text{Empathetic}}$  of +0.42, while Republicans average  $-0.28$ . The near-zero average is not indifference—it is a cancellation of opposing preferences across party lines.

Third, some attributes show genuine consensus. Hard-Working is unanimously favored (100% positive), and both Moderate Changes and Complete Overhaul are unanimously favored. For these consensus attributes, the average preference parameter and the polarization measure tell the same story; the structural model’s additional power becomes evident precisely when preferences are heterogeneous.



**Figure 1:** Saha & Weeks (2022) candidate choice conjoint. **A:** Average preference parameters  $\hat{\theta}_k$  on the logit scale with 95% DML confidence intervals. **B:** Average marginal effects on the probability scale; the structural estimate (green circles) and the linear probability model AMCE (orange triangles) nearly perfectly coincide, confirming that the structural model's average nests the standard reduced-form estimand. **C:** Fraction of respondents who favor (blue) vs. oppose (red) each attribute level, based on the sign of  $\hat{\beta}_k(\mathbf{Z}_i)$ . Gender has a near-zero average but 83% prefer female; Empathetic has a near-zero average but a 49/51 split reflecting partisan polarization.

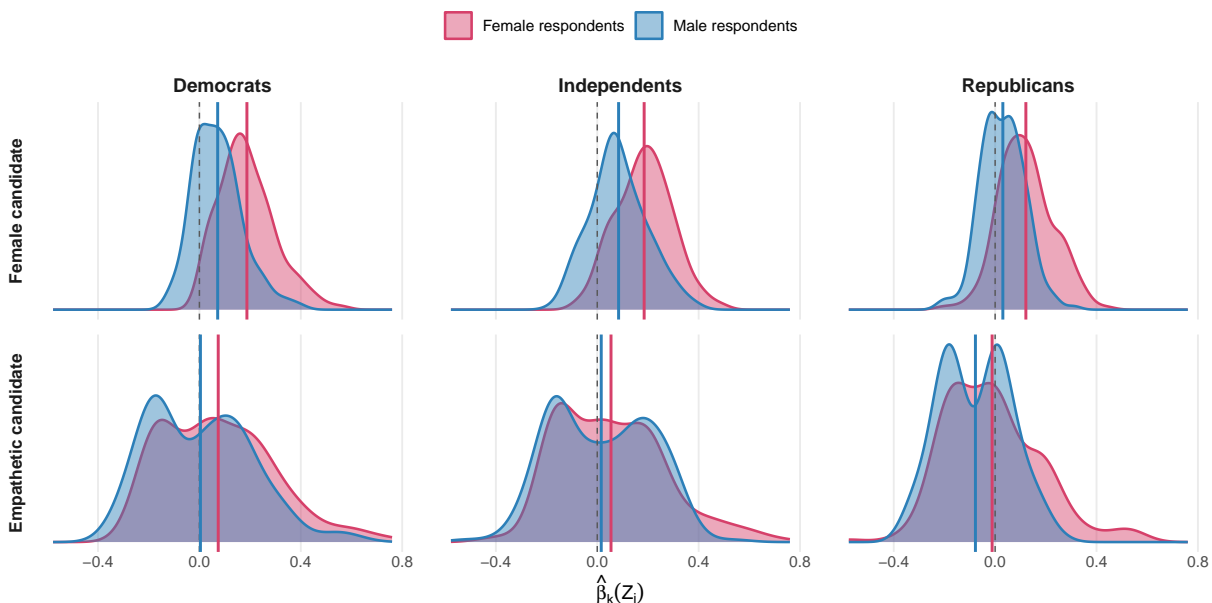
Figure 2 shows the full density of  $\hat{\beta}_k(\mathbf{Z}_i)$  across respondents for each attribute level, ordered by variance. Complete Overhaul is the most heterogeneous: while nearly all respondents favor it, the intensity ranges from near-zero to over 1.5 in logit units. The Empathetic distribution straddles zero, while Hard-Working is tightly concentrated in positive territory. These densities illustrate the direction-versus-intensity distinction: attributes can agree in direction yet differ sharply in intensity (Hard-Working), or can split in both dimensions simultaneously (Empathetic).



**Figure 2:** Ridgeline densities of  $\hat{\beta}_k(\mathbf{Z}_i)$  across respondents for each attribute level in the Saha & Weeks (2022) conjoint, ordered by variance. Vertical dashed line at zero; solid line at median.

Figure 3 drills further into the two most striking attributes by cutting the individual-level preference parameters along two respondent dimensions simultaneously—gender and party. The top row shows the preference for a female over a male candidate. The positive skew is nearly universal across all six subgroups, consistent with the 83% figure in the right panel

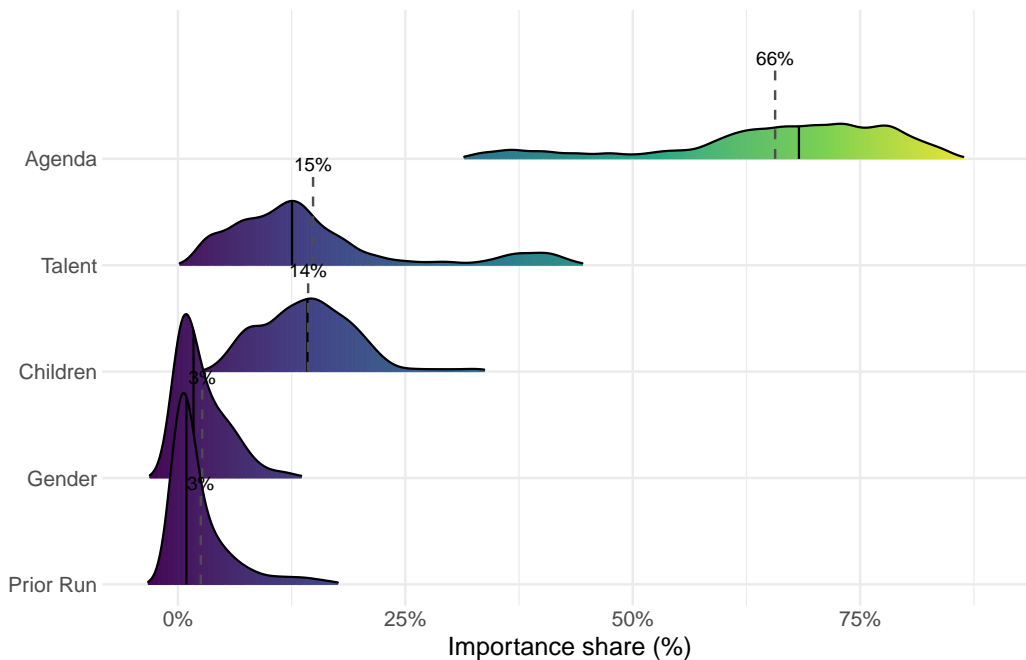
of Figure 1, but the gender gap is visible throughout: female respondents reward female candidates roughly twice as heavily as male respondents within each party (+0.19 vs. +0.07 among Democrats, +0.18 vs. +0.08 among Independents, +0.12 vs. +0.03 among Republicans). Party differences are small by comparison. The bottom row shows the preference for empathetic candidates, and the picture is qualitatively different. Here the party cleavage is the dominant axis: Democrats and Independents are centered in positive territory while Republicans are centered in negative territory, with female Republicans essentially indifferent (−0.01) and male Republicans actively penalizing empathy (−0.08). The fraction of respondents with  $\hat{\beta}_{\text{Empathetic}} > 0$  falls from 56% among Democrats to 38% among Republicans.



**Figure 3:** Individual-level preference parameters  $\hat{\beta}_k(\mathbf{Z}_i)$  for preferring a female candidate (top) and an empathetic candidate (bottom), by respondent party (columns) and respondent gender (fill). Solid vertical lines mark subgroup means; dashed vertical line at zero. Gender preference is positive nearly everywhere with a pronounced respondent-gender gap; empathy preference is structured primarily by party, with Republicans centered in negative territory.

The sign and spread of  $\hat{\beta}_k(\mathbf{Z}_i)$  tell us which way voters lean on each attribute, but not how much each attribute actually drives their vote. Two attributes can generate equally large coefficients yet contribute very differently to the decision once we account for how far the attribute levels are spread apart in the design. We quantify this with an individual-level importance share, defined for respondent  $i$  and attribute group  $g$  as  $\text{Imp}_{i,g} = \sum_{k \in g} \hat{\beta}_{i,k}^2 \cdot \text{Var}(X_k) / \sum_{g'} \sum_{k \in g'} \hat{\beta}_{i,k}^2 \cdot \text{Var}(X_k)$ , the share of the variance in the respondent’s utility that each attribute accounts for. Aggregated at the respondent level, these shares map directly onto the question a campaign strategist cares about: given the range of candidate profiles a voter might actually see, which attributes are moving their decision? Figure 4 shows the

distribution of these shares across respondents. Policy agenda dominates decisively (mean 66%), consuming two-thirds of the decision variance for the median voter—substantially more than its  $\hat{\theta}$  alone would suggest, because the agenda attribute has three highly differentiated levels whose effects compound. Talent and children each average around 15%, and gender and prior candidacy contribute less than 3% on average. The spread of the agenda share is wide: for some respondents it accounts for over 80% of utility variance, while for others it drops below 40% as talent or children attributes carry more weight. The substantive point is that Saha and Weeks’ near-zero average for gender is not just a canceled preference—it is an attribute that the electorate, with only a handful of exceptions, treats as a minor tiebreaker behind policy and competence. The structural model lets us see that the 83% female-preference finding of Figure 1 and the low importance share in Figure 4 are both true simultaneously: voters do prefer female candidates, but for most of them this preference is worth only a small fraction of the weight they place on the candidate’s policy agenda.



**Figure 4:** Distribution of individual-level attribute importance shares in the Saha & Weeks (2022) conjoint. Dashed lines and labels mark the mean. Policy agenda dominates on average (66%) but with considerable individual heterogeneity.

## 4.2 The Democracy Tradeoff

Graham and Svulik (2020) study how American voters trade off democratic principles against policy and partisan considerations. Their conjoint experiment presents 1,605 respondents with

pairs of hypothetical state legislative candidates described by eight attribute groups: party (co-partisan vs. not), two policy positions (economic and social), six good-governance practices, seven undemocratic actions, two valence violations, sex, race, and profession—yielding  $p = 30$  attribute levels. Each respondent evaluated approximately 13 choice tasks (20,657 total). We specify the DNN with architecture  $\mathbf{Z} \rightarrow 64 \rightarrow 64 \rightarrow 32 \rightarrow \boldsymbol{\beta} \in \mathbb{R}^{30}$ ,  $K = 10$  folds, and 2,000 epochs. The covariate vector  $\mathbf{Z}_i$  includes 12 variables: ideology (7-point), party ID (7-point), Trump approval, age, education, household income, authoritarianism, political knowledge, gender, and race indicators.

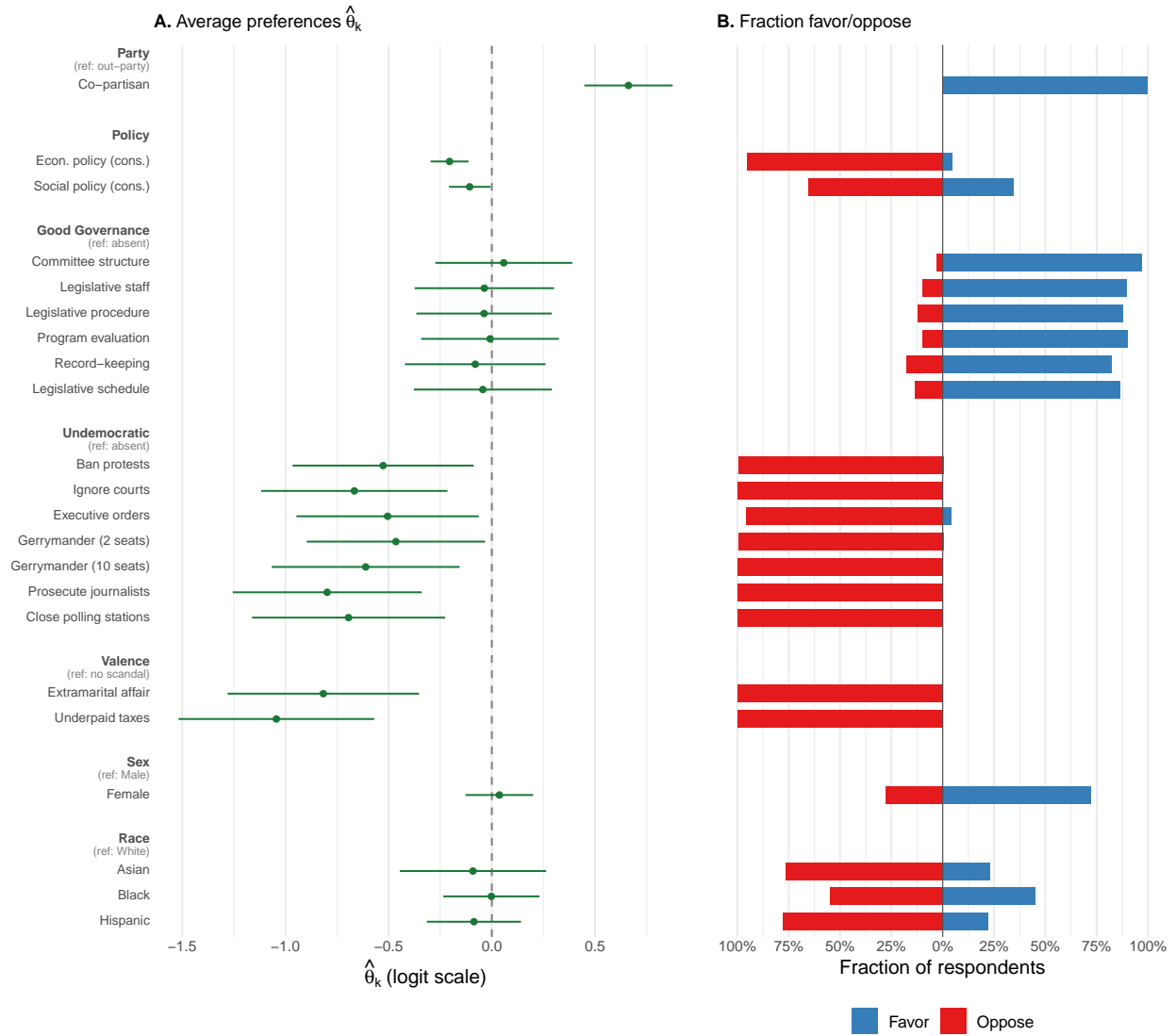
Co-partisanship has the largest positive effect ( $\hat{\theta} = 0.66$ ,  $p < 0.001$ ). All seven undemocratic actions carry significant negative coefficients, ranging from  $-0.46$  (executive orders) to  $-0.80$  (prosecuting journalists). DML-clustered standard errors exceed unclustered standard errors by a factor of 3.72, reflecting the 13 tasks per respondent.

The central claim of [Graham and Svulik](#)’s paper is that only 11.7% of Americans prioritize democratic principles in their vote choice. Our structural model reframes this finding. Figure 5 shows that for every undemocratic action, fewer than 5% of respondents in this sample have  $\hat{\beta}_k(\mathbf{Z}_i) > 0$ . For four of seven actions—ignoring courts, gerrymandering (10 seats), prosecuting journalists, and closing polling stations—the fraction is exactly 0%. There is near-universal rejection of undemocratic behavior in this sample.

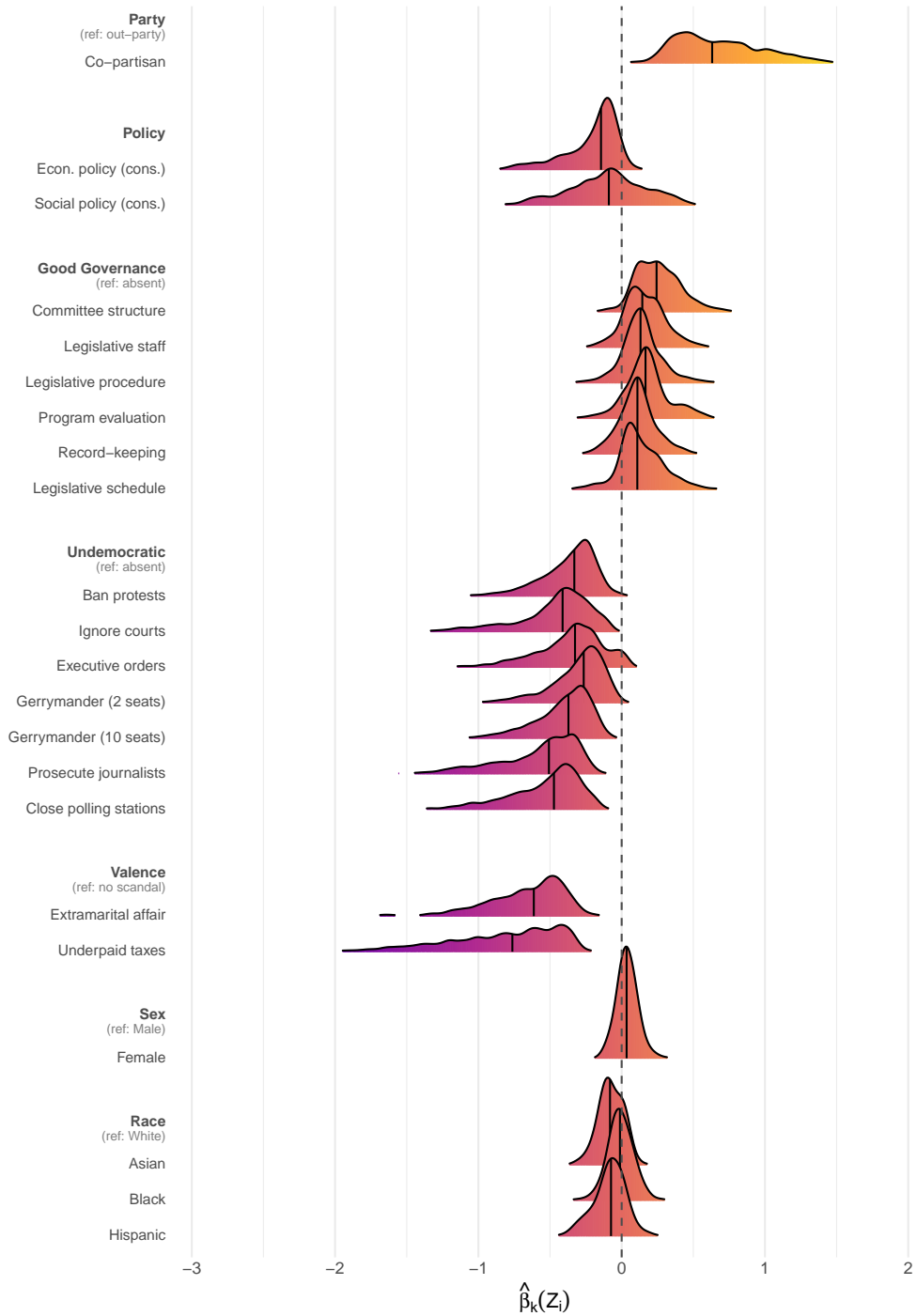
The apparent contradiction dissolves once we distinguish between direction and magnitude. [Graham and Svulik](#)’s 11.7% counts respondents who chose the more democratic candidate regardless of all other attributes—a measure that conflates the strength of democratic preferences with the strength of competing considerations. Our structural model shows that, in this sample, virtually all voters *dislike* undemocratic behavior (direction), but many voters weight party and policy considerations *more heavily* (magnitude). The disagreement across the electorate is not about whether democracy matters—it is about how much.

Figure 6 displays the full density of  $\hat{\beta}_k(\mathbf{Z}_i)$  for each attribute level. The undemocratic actions are entirely concentrated in negative territory, but the spread varies considerably, indicating that while virtually everyone opposes these actions, the intensity of opposition varies substantially.

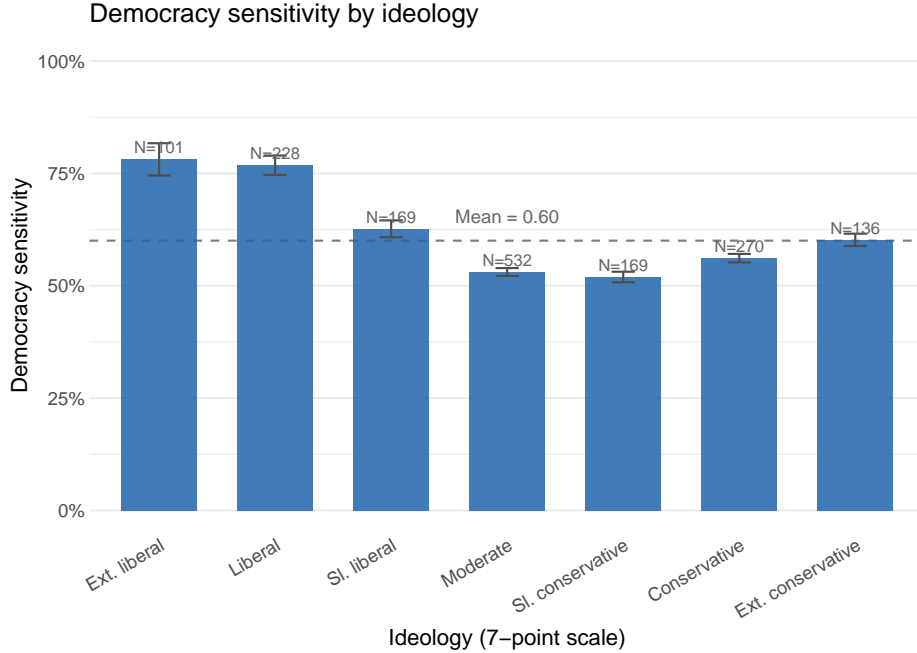
We define *democracy sensitivity* for respondent  $i$  as the mean absolute value of their undemocratic-action coefficients,  $\text{DemSens}_i = (1/7) \sum_{k \in \mathcal{U}} |\hat{\beta}_k(\mathbf{Z}_i)|$ . Figure 7 plots average democracy sensitivity across the 7-point ideology scale, revealing a striking U-shape. Liberals are most sensitive (0.78), sensitivity drops to its minimum for moderates (0.53), and rises again for extreme conservatives (0.60). This contradicts a simple partisan narrative: the “democracy deficit” is concentrated in the disengaged middle, not among strong partisans of either stripe.



**Figure 5:** Graham & Svulik (2020) candidate choice conjoint (profession dummies omitted). **A:** Average preference parameters  $\hat{\theta}_k$  with 95% DML CIs. **B:** Fraction favoring vs. opposing each attribute level. All undemocratic actions are opposed by >95% of respondents.



**Figure 6:** Ridgeline densities of  $\hat{\beta}_k(\mathbf{Z}_i)$  in Graham & Svulik (2020), ordered by variance (profession dummies omitted). Undemocratic actions cluster entirely below zero; the spread captures heterogeneity in intensity.



**Figure 7:** Democracy sensitivity by 7-point ideology. The U-shape shows that both ideological poles penalize undemocratic behavior more heavily than the moderate middle.

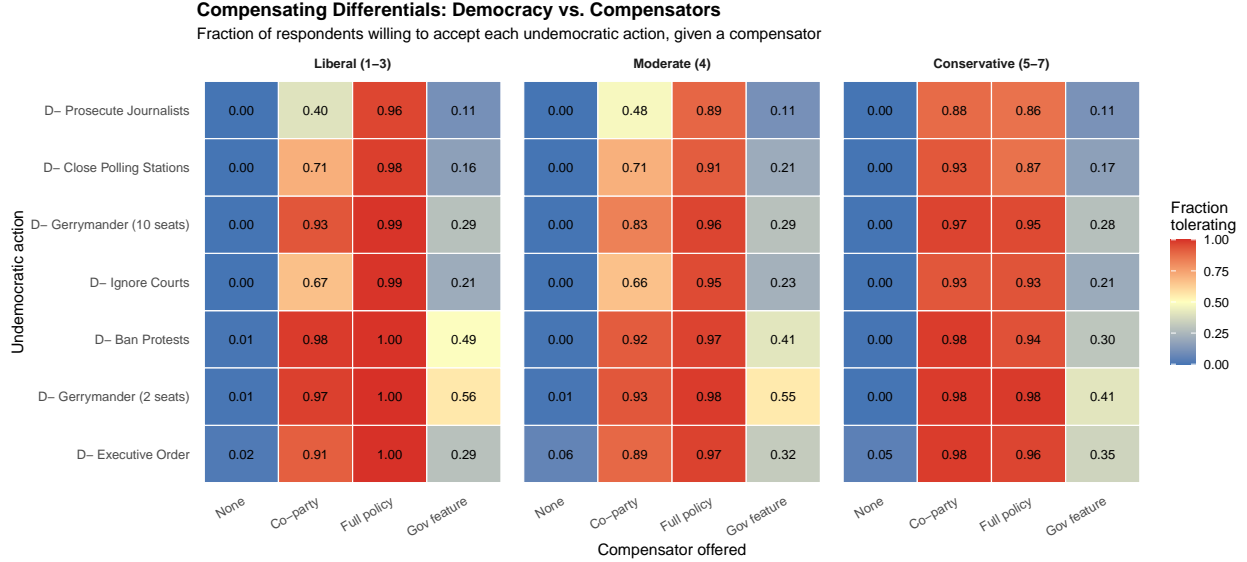
The marginal rate of substitution between undemocratic actions and co-partisanship quantifies how much partisan benefit a voter must sacrifice to avoid a democratic violation. Table 2 reports the MRS for three undemocratic actions. On average, voters would sacrifice 93% of the partisan benefit to avoid journalist prosecution, 76% to avoid ignoring courts, and 66% to avoid a 10-seat gerrymander. These ratios vary sharply by ideology: liberals sacrifice more than one full unit of partisan benefit for journalist prosecution (MRS = -1.05), while conservatives sacrifice only 70%.

|                        | Liberal | Moderate | Conservative | Overall |
|------------------------|---------|----------|--------------|---------|
| Prosecute journalists  | 1.05    | 0.93     | 0.70         | 0.93    |
| Ignore courts          | 0.86    | 0.76     | 0.55         | 0.76    |
| Gerrymander (10 seats) | 0.78    | 0.66     | 0.55         | 0.66    |

**Table 2:** Marginal rate of substitution: undemocratic actions relative to co-partisanship, by ideology tercile. Each entry reports  $|\text{MRS}| = |\hat{\beta}_{\text{undem}}/\hat{\beta}_{\text{party}}|$ .

The MRS measures how much partisan benefit a voter would have to sacrifice *on the margin* to avoid a democratic violation. A complementary and more discrete question—closer in spirit to the headline defection rates in [Graham and Svulik’s](#) original analysis—asks whether a specific compensator is *sufficient* to induce a voter to accept a given undemocratic action. For each voter  $i$  and undemocratic action  $j$ , we compute the compensating-differentials indi-

cator  $\mathbf{1}\{\hat{\beta}_{u_j}(\mathbf{Z}_i) + \text{benefit}_k(\mathbf{Z}_i) \geq 0\}$  under three alternative compensators: (i) co-partisanship (benefit =  $+\hat{\beta}_{\text{party}}(\mathbf{Z}_i)$ ), (ii) a full-range policy swing on both economic and social policy ( $3(|\hat{\beta}_{p_1}(\mathbf{Z}_i)| + |\hat{\beta}_{p_2}(\mathbf{Z}_i)|)$ ), and (iii) the voter’s single favorite good-governance feature ( $\max_k \hat{\beta}_{g_k}(\mathbf{Z}_i)$ ). The quantity of interest is the fraction of respondents for whom compensation holds, reported in Figure 8 by ideology tercile.



**Figure 8:** Compensating differentials in Graham & Svulik (2020). Each cell reports the fraction of respondents (in a given ideology tercile) whose individual preferences satisfy  $\hat{\beta}_{u_j}(\mathbf{Z}_i) + \text{benefit}_k(\mathbf{Z}_i) \geq 0$ —i.e., the share for whom the compensator is sufficient to induce acceptance of the undemocratic action. Rows are sorted by severity (fewer tolerators with any compensator at top). The “None” column reports the baseline fraction with  $\hat{\beta}_{u_j}(\mathbf{Z}_i) \geq 0$ .

Three features stand out. First, the “None” column confirms near-universal resistance to democratic erosion in this sample: across all 7 actions and all ideology groups, fewer than 5% of respondents have a positive coefficient on any undemocratic action. Second, co-partisanship alone is a remarkably powerful compensator: for 5 of 7 actions, a majority of respondents in every ideology tercile would accept the violation in exchange for voting for their own party. Third, and most strikingly, the co-partisanship column reveals a sharp ideological asymmetry on the hardest cases. For prosecuting journalists, only 40% of liberals would accept the violation in exchange for a co-partisan, compared to 88% of conservatives—a 48 percentage-point gap. For closing polling stations the gap is 22 points (71% vs. 93%); for ignoring courts it is 26 points (67% vs. 93%). These three actions—the ones most closely tied to free elections, a free press, and judicial independence—are also where liberals are most willing to cross party lines to defend democracy. The full-range policy-swing column shows that a sufficiently large policy concession can compensate essentially everyone for essentially

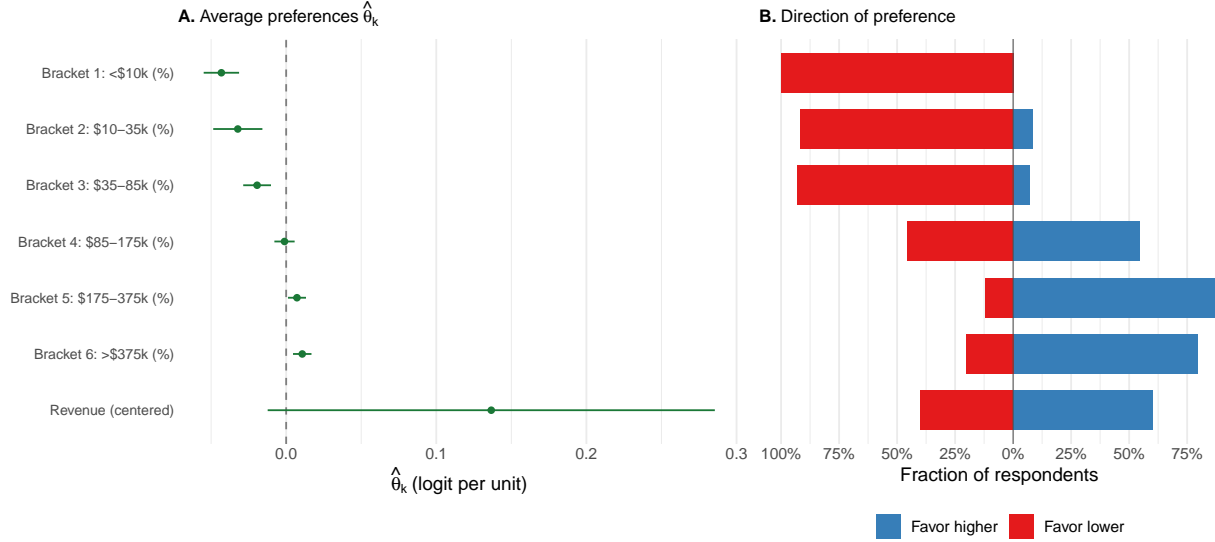
any violation, which re-emphasizes why policy alignment dominates democratic principles in observational voting patterns: the policy stakes are simply very large.

### 4.3 The Structure of Tax Policy Preferences

Ballard-Rosa et al. (2017) study American mass preferences over federal income-tax policy, asking 2,000 US adults to choose between pairs of hypothetical tax plans in which each plan specifies the marginal rate on six income brackets together with a revenue indicator describing how much revenue the plan raises relative to current policy. Their central reduced-form finding is that Americans have generally progressive preferences—opposing higher rates on the poor and supporting higher rates on the rich—but that support for a plan responds much more elastically to rates on the poor than to rates on the rich. The analysis sample contains 2,000 US respondents, each evaluating 8 paired comparisons (16,000 tasks). The seven continuous attributes produce  $p = 7$  treatment variables: marginal rates on six income brackets (entering in percentage points) and a five-level revenue indicator rescaled to  $\{-2, -1, 0, 1, 2\}$ . Every attribute is continuous, so the DNN recovers an individual-level *slope*—utility per percentage point of rate—for every bracket and every respondent, producing a full preferred tax schedule  $\hat{\beta}(\mathbf{Z}_i) \in \mathbb{R}^7$  per person rather than a handful of average level effects. We specify the DNN with architecture  $\mathbf{Z} \rightarrow 32 \rightarrow 32 \rightarrow 16 \rightarrow \beta \in \mathbb{R}^7$ ,  $K = 10$  folds, and 2,000 epochs. The covariate vector  $\mathbf{Z}_i$  includes 12 respondent characteristics: age, gender, party ID, education, race, income, inequality aversion, and beliefs about the tax-economy relationship. Because each respondent provides 8 tasks, within-respondent correlation is substantial: the DML/independent and identically distributed SE ratio is 3.00—the highest of any application we run—so naive independent and identically distributed inference would overstate precision by a factor of three.

Figure 9 reports the DML-corrected average effects. A 1 percentage-point increase in the bottom-bracket rate lowers log-odds of plan support by 0.043; the same increase on the top bracket raises log-odds by only 0.011. This fourfold absolute asymmetry is the quantitative form of the central Ballard-Rosa et al. finding: plan support is highly elastic with respect to rates on low incomes but nearly inelastic with respect to rates on high incomes. The \$85–175k bracket is indistinguishable from zero on average—the “dead zone”—and revenue has a small positive but non-significant average effect.

Because each respondent has a full 7-element vector  $\hat{\beta}(\mathbf{Z}_i)$ , we can compute an *individual-level progressivity slope*—the within-respondent regression of bracket-specific marginal utilities



**Figure 9:** Ballard-Rosa et al. (2017) tax-plan conjoint. **A:** Average preference parameters  $\hat{\theta}_k$  on the logit scale (per 1 percentage point of rate, or per unit of revenue). **B:** Fraction of respondents with  $\hat{\beta}_{i,k} > 0$  (favor raising that rate) vs.  $\hat{\beta}_{i,k} < 0$  (oppose). The \$85–175k bracket is the most polarized dimension—55% favor, 45% oppose—despite its near-zero average effect.

on the log of the bracket midpoint:

$$s_i = \frac{\sum_{k=1}^6 (\log m_k - \overline{\log m}) \cdot \hat{\beta}_{i,k}}{\sum_{k=1}^6 (\log m_k - \overline{\log m})^2}.$$

A positive  $s_i$  means respondent  $i$  prefers higher rates on higher incomes; zero means a flat tax; negative means regressive. Remarkably, in this sample 99.9% of respondents have a positive slope, and  $\hat{\beta}_{i,\text{top}} > \hat{\beta}_{i,\text{bottom}}$  for 99.1%. The progressive direction is essentially universal here: in every partisan, income, ideological, and inequality-aversion subgroup of this sample, fewer than 0.4% of respondents have regressive revealed preferences. Democrats’ mean slope is +0.0131, Republicans’ is +0.0100—a real but modest 31% gap—and 99.7% of Republican respondents still reveal progressive preferences. This is the sharpest possible individual-level form of the Ballard-Rosa et al. finding that progressive preferences cut across partisan lines.

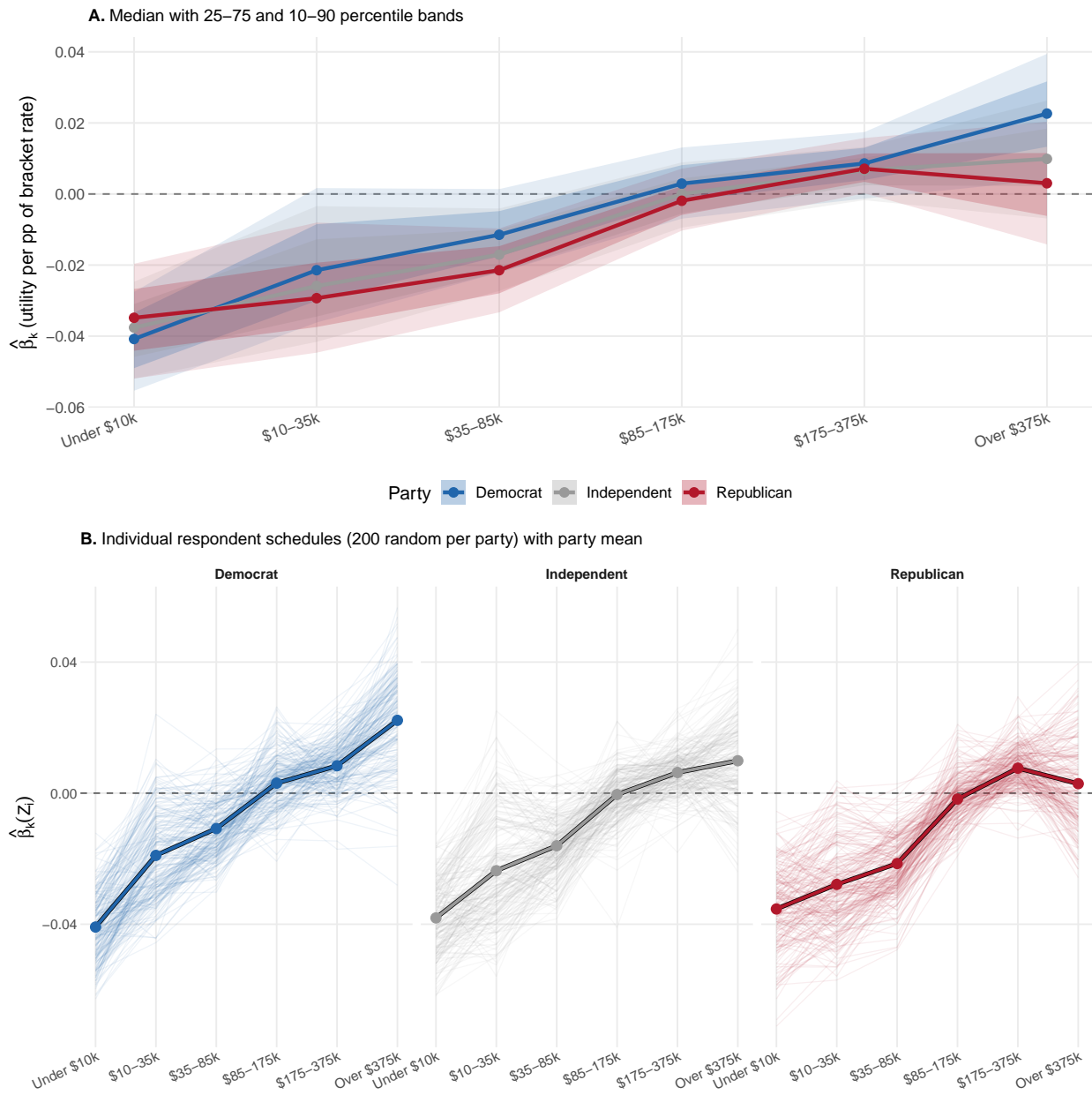
Figure 10 exposes the within-party heterogeneity that the structural model recovers. The upper panel overlays each party’s median schedule with its inter-quartile (25–75) and 10–90 percentile bands. Three facts are immediately visible: the party-level medians are essentially indistinguishable at the bottom brackets but fan out at the top, with Democrats above Republicans on the top two brackets and below on the bottom-middle brackets; the within-party 25–75 bands overlap substantially across parties at every bracket, so within-party heterogeneity dwarfs the between-party gap; and the 10–90 band is wide and straddles zero

in every middle bracket, meaning each party contains both bracket-level “hawks” and “doves.” The lower panel shows a random sample of 200 individual respondent schedules per party as semi-transparent lines, with the party mean overlaid. Almost every individual line in this sample slopes upward, so the progressive direction is near-universal here, but the levels vary enormously—some Republicans want a schedule steeper than the Democratic median, and some Democrats want schedules nearly as flat as the Republican median. The between-party difference is a shift in central tendency, not a separation of populations.

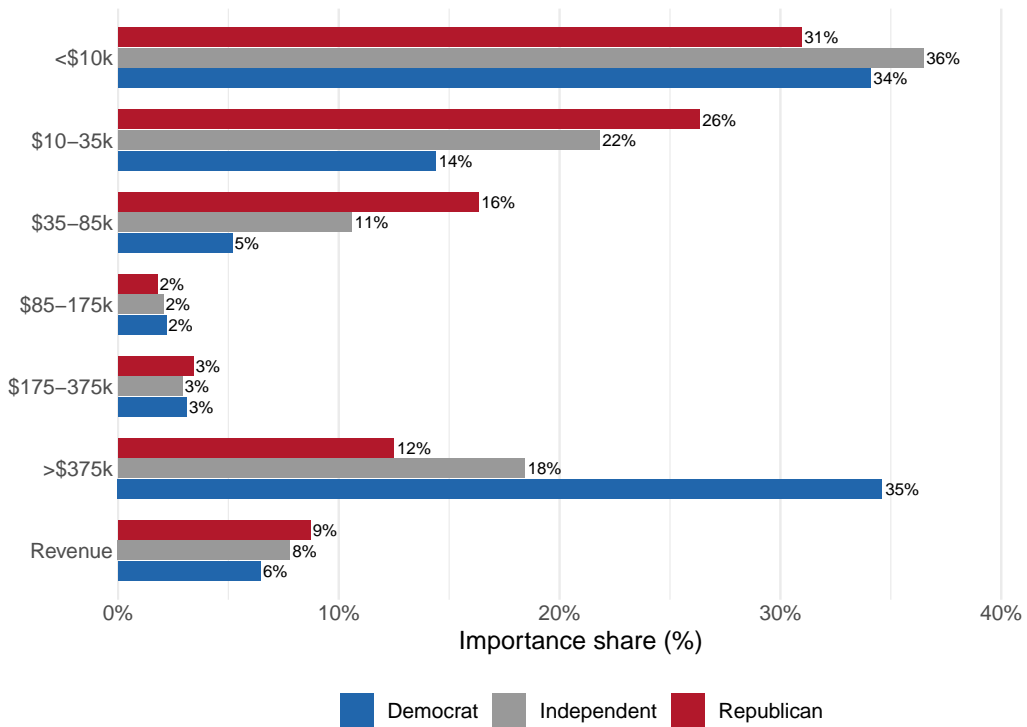
The structural model also enables a variance decomposition by subgroup that is invisible to reduced-form subgroup AMCEs. Figure 11 reports the share of plan-choice variance attributable to each attribute, by party. Democrats allocate 34.6% of variance to the top bracket; Republicans only 12.5%—a ratio of nearly three-to-one. Republicans, in turn, weight the \$10–35k and \$35–85k brackets substantially more than Democrats do (26.3% and 16.3% vs. 14.4% and 5.2%). Democrats’ plan-choice variance is driven 35% by what happens to the rich; Republicans’ variance is driven 43% by what happens to working-class and middle-class incomes. This is a substantive reframing of the standard partisan narrative: Republican opposition to progressive taxation is primarily opposition to taxing the working class, not primarily protection of the rich.

Alongside the conjoint, [Ballard-Rosa et al. \(2017\)](#) collected each respondent’s self-reported ideal marginal tax rate for each of the six brackets. These ideal rates were elicited outside the conjoint task and are not used by the DNN at any stage of training, so they provide an independent benchmark against which the structural model’s recovered individual preferences can be validated. No reduced-form AMCE estimator can be validated at the individual level, because it does not produce an individual-level preference parameter. Using the 401 respondents who answered both the bottom- and top-bracket ideal questions, the correlation between the revealed progressivity slope  $s_i$  and its self-reported analog is  $r = 0.37$ , and the correlation between the revealed and self-reported top-minus-bottom rate gap is  $r = 0.43$ . Using all  $n = 2,000$  respondents, the DNN’s individual-level top-bracket coefficient correlates  $r = 0.37$  with the self-reported ideal top rate. These correlations survive disaggregation by party ( $r \approx 0.26$ – $0.30$  within each partisan subgroup), confirming that the DNN recovers the correct individual ordering of tax progressivity within, not just across, parties. A correlation of  $r = 0.43$  between a DNN-recovered preference parameter and an independent self-report is substantively strong—comparable to typical cross-measure correlations in political psychology and well above what one would expect from noise.

The individual-level preference vector  $\hat{\beta}(\mathbf{Z}_i)$  also enables evaluation of any hypothetical tax schedule. We compare four stylized plans: a revenue-neutral flat 15% tax, a steeply progressive plan with rates rising 0/5/15/25/35/45 and “much more” revenue, a regressive plan with the



**Figure 10:** Distribution of individual-level  $\hat{\beta}_{i,k}$  schedules, by party. **A:** Party medians overlaid with inter-quartile (25–75) and 10–90 percentile bands. **B:** A random sample of 200 individual respondent schedules per party (semi-transparent), with the party mean in bold. Almost every individual line slopes upward, but the level varies enormously: within-party heterogeneity dwarfs the between-party gap.



**Figure 11:** Attribute importance shares (variance decomposition) by party. Democrats allocate nearly three times as much variance to the top bracket as Republicans; Republicans allocate roughly twice as much to the low and middle brackets as Democrats.

reverse rate pattern and “much less” revenue, and a revenue-neutral status-quo analog with rates 5/15/25/25/35/35. The progressive plan yields the only positive mean systematic utility in the set. Head-to-head probabilistic comparisons are striking: the progressive plan beats the flat plan for 78.7% of respondents overall and for 73.1% of Republicans, while it beats the status-quo plan for roughly 70% of respondents. Americans prefer the status quo to a flat or regressive plan, but they prefer an explicitly more progressive alternative to the status quo even more strongly. The [Ballard-Rosa et al.](#) conclusion that Americans’ preferences lie “close to current policy” is consistent with the status-quo plan ranking second-best, but it understates how much the public would prefer a more aggressively progressive schedule if offered one. A purely partisan model of tax preferences—“Republicans want flat, Democrats want progressive”—is not consistent with the data: nearly three-quarters of Republicans prefer a steeply progressive schedule with more revenue to a flat, revenue-neutral tax.

## 5 Limitations and Extensions

Our method rests on several assumptions, which we now discuss alongside a set of possible extensions that addresses these limitations.

The first consideration is what the method estimates when preference heterogeneity is not fully captured by observed respondent characteristics. If the true process is  $\beta_i = \boldsymbol{\beta}(\mathbf{Z}_i) + \boldsymbol{\eta}_i$  with a respondent-specific latent component, the method targets  $\mathbb{E}[\beta_i \mid \mathbf{Z}_i]$ —the conditional mean preference given observables. This is a well-defined estimand, not a source of bias: omitting covariates from  $\mathbf{Z}_i$  does not bias the average preference parameters  $\theta_k = \mathbb{E}[\beta_k(\mathbf{Z}_i)]$ , and the individual-level  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i)$  consistently estimates the conditional expectation  $\mathbb{E}[\boldsymbol{\beta}_i \mid \mathbf{Z}_i]$ , which is the best prediction of a respondent’s preferences given what is observed about them. What changes is the interpretation:  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i)$  recovers the average preference of respondents who share  $i$ ’s observable profile, rather than  $i$ ’s exact personal preference. Richer covariates narrow this gap, which is why we recommend collecting a rich set of  $\mathbf{Z}_i$  variables.

The practical consequence is that nonlinear functionals of the conditional mean can differ from the conditional mean of those functionals: the distribution of  $\hat{\boldsymbol{\beta}}(\mathbf{Z}_i)$  across respondents understates the true dispersion of  $\beta_i$  (because it misses the within- $\mathbf{Z}$  component  $\boldsymbol{\eta}_i$ ), and the MRS computed from conditional means need not equal the conditional mean of the MRS, since  $\mathbb{E}[\beta_1 \mid \mathbf{Z}]/\mathbb{E}[\beta_2 \mid \mathbf{Z}] \neq \mathbb{E}[\beta_1/\beta_2 \mid \mathbf{Z}]$  by Jensen’s inequality. These smoothed quantities remain valid and interpretable as properties of  $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{Z}]$ , but researchers should bear in mind that they provide a lower bound on the true degree of preference heterogeneity. A natural extension is a hybrid model that combines the DNN mean function with latent random

coefficients  $\boldsymbol{\eta}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\mathbf{Z}_i))$ , using within-respondent dependence across repeated tasks to identify the latent covariance structure.

A second key assumption is an additive random utility model with a linear-in-parameters utility index. If respondents care about attribute interactions not included in  $\mathbf{X}_{ijt}$ , use noncompensatory decision rules, or evaluate profiles relative to task-specific reference points, the model is misspecified. Interactions can be addressed by including interaction terms in  $\mathbf{X}_{ijt}$ , but noncompensatory rules are fundamentally inconsistent with the linear utility framework and would require a different modeling strategy. Relatedly, the Gumbel distribution of taste shocks fixes the error scale, so preferences are identified only up to this normalization: if the true utility includes respondent-specific scale  $\sigma_i$ , the data identify  $\boldsymbol{\beta}_i/\sigma_i$  rather than  $\boldsymbol{\beta}_i$ , and what appears as taste heterogeneity may partly reflect scale heterogeneity. Modeling scale heterogeneity jointly with preference heterogeneity is a natural and straightforward extension of the current framework.

As the utility index is linear in  $\mathbf{X}$ , it rules out complementarities between attributes unless interaction terms are explicitly included. Manual inclusion of pairwise interactions is straightforward but quickly becomes high-dimensional ( $p(p+1)/2$  terms for  $p$  attributes). A more systematic extension is a low-rank interaction layer in the DNN architecture, in which the network outputs both the main-effect vector  $\boldsymbol{\beta}(\mathbf{Z}) \in \mathbb{R}^p$  and a low-rank factor  $\mathbf{V}(\mathbf{Z}) \in \mathbb{R}^{p \times r}$  with  $r \ll p$ , so that the utility index becomes  $\boldsymbol{\Delta} \mathbf{X}^\top \boldsymbol{\beta}(\mathbf{Z}) + \|\mathbf{V}(\mathbf{Z})^\top \boldsymbol{\Delta} \mathbf{X}\|^2$ . The rank constraint regularizes the interaction space while preserving structural interpretability;  $r$  can be chosen by cross-validation. More broadly, the structural AME provides a built-in diagnostic for these functional-form assumptions: as discussed in Section 2.3, a discrepancy between the structural AME and the nonparametric AMCE would signal misspecification of either the logit link or the additive utility structure.

A third set of caveats concerns inference. Formal DML guarantees apply to the average parameters  $\theta_k$ , but many of the structural quantities reported in the applications—including the MRS, WTP, and the AME—are smooth functionals of these parameters, for which valid inference extends naturally via the delta method applied to the DML influence functions, or equivalently, via Neyman orthogonal moment conditions for the functionals themselves. What remains genuinely open is inference on distributional quantities that depend on the full shape of  $\boldsymbol{\beta}(\mathbf{Z})$  rather than its mean—polarization fractions, importance-share distributions, and individual-level preference rankings, for instance. Our simulations in Supplementary Materials B show that these quantities are recovered with high accuracy when the sample is large enough, and Supplementary Materials C provides design guidance on the sample sizes required. Developing formal inferential tools for these distributional quantities is an important direction for future work.

## 6 Conclusion

This paper develops a structural estimator for conjoint experiments that combines a random utility logit model with a deep neural network mean function, yielding an individual-level preference vector  $\hat{\beta}(\mathbf{Z}_i)$  for every respondent together with valid double/debiased machine learning inference on population averages. The approach nests the reduced-form AMCE as a special case—our average  $\hat{\theta}_k$  reproduces a pooled homogeneous-logit coefficient almost exactly both overall and within subgroups (Supplementary Materials A)—so nothing a standard AMCE analysis would deliver is lost. What is gained is access to the full distribution of preferences and to structural quantities—polarization, marginal rates of substitution, willingness to pay, compensating differentials, counterfactual welfare—that an average effect cannot express.

The three applications show what this buys. In the candidate-choice experiment, a near-zero average preference for gender coexists with the vast majority of respondents preferring female candidates, and a near-zero average preference for empathetic candidates dissolves into a sharp partisan split; the structural model recovers the direction–intensity decomposition that are obscured by reduced-form averages. In the democracy-versus-partisanship tradeoff, virtually every respondent has a negative individual-level coefficient on every undemocratic action, resolving the apparent paradox of the paper’s headline finding: opposition to democratic erosion is near-universal in direction, but many voters weight partisan and policy considerations more heavily in magnitude, and liberals are substantially more willing than conservatives to cross party lines on the hardest cases. In the tax-plan experiment, we recover a full preferred tax schedule for each respondent and find that nearly all respondents have progressive revealed preferences, including nearly all Republicans; the recovered progressivity slope correlates meaningfully with independently elicited self-reported ideal rates, providing external validation at the individual level.

Structural heterogeneity is a complement to reduced-form AMCE analysis rather than a replacement. When the research question is about the average effect of an attribute, the AMCE is appropriate and remains the natural quantity to report. But when the research question is about polarization, cross-attribute tradeoffs, individual welfare, or counterfactual choice—questions about whether preferences are shared, about how much one attribute is worth relative to another, or about what a population would choose under a policy that was never part of the experimental design—an individual-level preference vector is required, and the structural DNN approach developed here makes this practical at the sample sizes that are typical of applied conjoint research.

For researchers planning to use this approach, our simulation results (Supplementary Materials B and C) suggest several practical guidelines. First, collect a rich set of respondent covariates  $\mathbf{Z}_i$ . The DNN learns heterogeneity entirely through observed characteristics; richer covariates mean finer-grained preference recovery. Demographics, ideology, party identification, attitudes, and behavioral measures all contribute. Second, ensure a sufficient total sample size. Population-level quantities (average preferences, polarization, importance shares) are reliably recovered at  $NT \approx 3,000\text{--}5,000$  total observations, while accurate individual-level preference vectors require  $NT \approx 8,000\text{--}10,000$ . The total number of observations  $NT$  is the key design parameter: our factorial simulations show that neither the number of respondents  $N$  nor the number of tasks  $T$  alone is a sufficient statistic, though  $N$  contributes more than  $T$  (53% vs. 30% of variance in recovery accuracy). Third, include enough tasks per respondent. The marginal gain from additional tasks is largest at the bottom: going from one to two tasks per respondent yields the single largest improvement in individual-level recovery. We recommend at least three to five tasks per respondent, consistent with standard conjoint practice. Fourth, the method scales straightforwardly with the number of attributes  $p$ : all three applications in this paper have different attribute counts ( $p = 7, 13, \text{ and } 30$ ), and the estimator performs well across this range without manual tuning beyond the network architecture.

The method is implemented in the R package, `sconjoint`, which provides functions for estimation, inference, and visualization.<sup>4</sup>

## References

- Abramson, S. F., K. Koçak, and A. Magazinnik (2022). What do we learn about voter preferences from conjoint experiments? *American Journal of Political Science* 66(4), 1008–1020.
- Alvarez, R. M. and J. Nagler (1998). When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science* 42(1), 55–96.
- Austen-Smith, D. and J. S. Banks (1999). *Positive Political Theory I: Collective Preference*. University of Michigan Press.
- Ballard-Rosa, C., L. Martin, and K. Scheve (2017). The structure of American income tax policy preferences. *Journal of Politics* 79(1), 1–16.

---

<sup>4</sup>A user tutorial is available at <https://yiqingxu.org/packages/sconjoint/>. The package was developed using `StasClaw`, an AI-collaborative workflow described in Qin and Xu (2026).

- Bansak, K., J. Hainmueller, and D. Hangartner (2016). How economic, humanitarian, and religious concerns shape european attitudes toward asylum seekers. *Science* 354(6309), 217–222.
- Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto (2021). Conjoint survey experiments. In J. N. Druckman and D. P. Green (Eds.), *Advances in Experimental Political Science*, pp. 19–41. Cambridge University Press.
- Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto (2023). Using conjoint experiments to analyze election outcomes: The essential role of the average marginal component effect. *Political Analysis* 31(4), 500–518.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.
- Clinton, J., S. Jackman, and D. Rivers (2004). The statistical analysis of roll call data. *American Political Science Review* 98(2), 355–370.
- de la Cuesta, B., N. Egami, and K. Imai (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis* 30(1), 19–45.
- Enelow, J. M. and M. J. Hinich (1984). *The Spatial Theory of Voting: An Introduction*. Cambridge University Press.
- Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica* 89(1), 181–213.
- Farrell, M. H., T. Liang, and S. Misra (2025). Deep learning for individual heterogeneity: An automatic inference framework. Working paper, arXiv:2010.14694.
- Goplerud, M., K. Imai, and N. E. Pashley (2025). Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis. *Annals of Applied Statistics* 19(2), 866–888.
- Graham, M. H. and M. W. Svobik (2020). Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *American Political Science Review* 114(2), 392–409.
- Green, P. E. and V. R. Rao (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research* 8(3), 355–363.

- Green, P. E. and V. Srinivasan (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research* 5(2), 103–123.
- Green, P. E. and V. Srinivasan (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing* 54(4), 3–19.
- Greenhalgh, L. and S. A. Neslin (1981). Conjoint analysis of negotiator preferences. *Journal of Conflict Resolution* 25(2), 301–327.
- Hainmueller, J. and D. 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, J., D. J. Hopkins, and T. Yamamoto (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis* 22(1), 1–30.
- Ham, D. W., K. Imai, and L. Janson (2024). Using machine learning to test causal hypotheses in conjoint analysis. *Political Analysis* 32(3), 329–344.
- Hinich, M. J. and M. C. Munger (1997). *Analytical Politics*. Cambridge University Press.
- Leeper, T. J., S. B. Hobolt, and J. Tilley (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis* 28(2), 207–221.
- Louviere, J. J., D. A. Hensher, and J. D. Swait (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University Press.
- Luce, R. D. and J. W. Tukey (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* 1(1), 1–27.
- Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10(2), 134–153.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press.
- McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 198–272. MIT Press.
- Palfrey, T. R. and K. T. Poole (1987). The relationship between information, ideology, and voting behavior. *American Journal of Political Science* 31(3), 511–530.

- Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science* 29(2), 357–384.
- Qin, T. and Y. Xu (2026). Statsclaw: An ai-collaborative workflow for statistical software development. arXiv:2604.04871.
- Rivers, D. (1988). Heterogeneity in models of electoral choice. *American Journal of Political Science* 32(3), 737–757.
- Robinson, T. S. and R. Duch (2024). How to detect heterogeneity in conjoint experiments. *Journal of Politics* 86(2), 412–427.
- Saha, S. and J. L. P. Weeks (2022). Welfare over democracy? State capacity signals and self-interested voting. *Journal of Conflict Resolution* 66(4–5), 629–658.
- Shamir, M. and J. Shamir (1995). Competing values in public opinion: A conjoint analysis. *Political Behavior* 17(1), 107–133.
- Train, K. E. (2009). *Discrete Choice Methods with Simulation* (2nd ed.). Cambridge University Press.

# Supplementary Materials

|   |            |
|---|------------|
| <b>A Validation Against Reduced-Form Benchmarks</b> | <b>A-1</b> |
| <b>B Benchmark Monte Carlo Simulation</b>           | <b>A-2</b> |
| B.1 Data Generating Process . . . . .               | A-3        |
| B.2 Results . . . . .                               | A-3        |
| <b>C Factorial Simulation: Design Guidance</b>      | <b>A-4</b> |

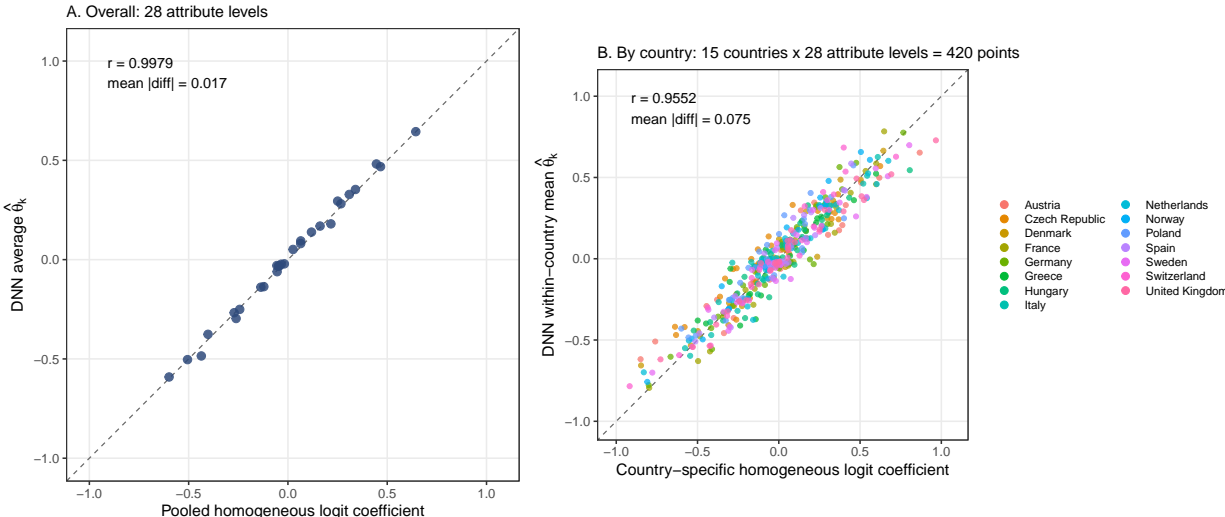
## A Validation Against Reduced-Form Benchmarks

A useful sanity check for the structural DNN is that, under correct specification of the logit model, the average of its individual-level coefficients should reproduce the standard pooled homogeneous-logit estimate. Specifically, our estimand  $\theta_k = \mathbb{E}[\beta_k(\mathbf{Z})]$  coincides with the coefficient of a pooled logit of  $Y$  on  $\Delta\mathbf{X}$ , and within any subgroup  $S$  the conditional mean  $\mathbb{E}[\beta_k(\mathbf{Z}) \mid S]$  coincides with the logit coefficient fit on  $S$ -only rows. A reduced-form subgroup AMCE therefore provides a target that the DNN’s within-subgroup average should hit. This validation covers only the *mean* of  $\beta(\mathbf{Z})$ —the individual-level heterogeneity that the DNN recovers is precisely what reduced-form estimators cannot access, and therefore cannot be validated against them.

We use the [Bansak et al. \(2023\)](#) European immigration conjoint for this check because of its large sample size ( $NT = 74,090$  tasks from 14,818 respondents across 15 countries), which makes the within-country logit benchmarks themselves precisely estimated. For the overall comparison, we fit a pooled homogeneous logit of  $Y$  on the 28 attribute-level dummies and compare each coefficient to the DNN’s  $\hat{\theta}_k = n^{-1} \sum_i \hat{\beta}_k(\mathbf{Z}_i)$ . For the subgroup comparison, we refit the logit separately within each of the 15 country subsamples ( $\approx 5,000$  tasks each) and compare each country-specific coefficient to the DNN’s within-country conditional mean  $\mathbb{E}_n[\hat{\beta}_k(\mathbf{Z}_i) \mid \text{country}_i = c]$ .

Figure [A.1](#) displays the two scatters. The overall comparison (Panel A, 28 points) lies almost perfectly on the 45-degree line, with a correlation of  $r = 0.998$  and a mean absolute difference of 0.017 log-odds units—well within the pooled logit’s own sampling variability. The subgroup comparison (Panel B, 15 countries  $\times$  28 attributes = 420 points) also tracks the 45-degree line closely, with  $r = 0.955$  and a mean absolute difference of 0.075. The

slight attenuation of the subgroup correlation relative to the overall correlation is expected and has two sources: the country-specific logits are noisier because each uses only  $\approx 5,000$  observations, and the DNN’s within-country averages smooth slightly across observationally similar respondents from different countries. In both panels, the DNN is reproducing the reduced-form target without any special calibration—the individual-level  $\hat{\beta}_k(\mathbf{Z}_i)$ ’s average up to what a pooled logit would say. We view this as a necessary consistency check rather than a substantive contribution: it confirms that the structural DNN’s average theta is identified and estimated correctly, which is the foundation on which the individual-level heterogeneity results rest.



**Figure A.1:** Validation of DNN-recovered averages against reduced-form homogeneous logit estimates on the [Bansak et al. \(2023\)](#) immigration conjoint ( $NT = 74,090$ ). **A:** 28 attribute-level coefficients; each point compares the DNN’s population mean  $\hat{\theta}_k$  to the pooled homogeneous-logit coefficient. **B:**  $15 \times 28 = 420$  (country  $\times$  attribute) pairs; each point compares the DNN’s within-country conditional mean to the logit coefficient from a country-specific regression. Dashed lines show  $y = x$ . The DNN reproduces both the overall and within-country reduced-form coefficients to very high accuracy, confirming that  $\hat{\theta}_k$  is a well-identified estimate of  $\mathbb{E}[\beta_k(\mathbf{Z})]$  and that its subgroup analogs match what a researcher would obtain by running a reduced-form subgroup AMCE regression.

## B Benchmark Monte Carlo Simulation

We validate the structural DNN estimator with two simulation studies. The first is a benchmark Monte Carlo that uses the estimated  $\hat{\beta}(\mathbf{Z})$  from the [Bansak et al. \(2016\)](#) asylum-seeker conjoint ( $M = 14,818$  respondents,  $T = 5$  tasks,  $p = 28$  attribute levels) as the ground truth, generates  $R = 50$  synthetic datasets, and compares the DNN to homogeneous logit, random-coefficients

(RC) logit, and Bayesian hierarchical linear model (HLM). The second is a factorial design that provides practical guidance on when the estimator works well.

## B.1 Data Generating Process

The baseline DNN is first estimated on the full dataset ( $p_Z = 19$  respondent covariates), and the fitted  $\hat{\beta}(\mathbf{Z})$  is taken as the true mapping  $\beta^*(\mathbf{Z})$ . In each replication, choices are generated as  $Y_{it}^{(r)} \sim \text{Bernoulli}(G(\Delta \mathbf{X}_{it}^\top \beta^*(\mathbf{Z}_i)))$ .

## B.2 Results

Across 50 replications, the DNN achieves a mean absolute bias of  $|\hat{\theta} - \theta| = 0.006$  for the 28 average preference parameters—comparable to homogeneous logit (0.008) and substantially better than the Bayesian HLM (0.116). Figure B.1 plots estimated versus true  $\hat{\theta}_k$ : all 28 points lie on the 45-degree line, and 95% DML confidence intervals achieve 100% coverage. By contrast, plug-in standard errors (without the DML correction) yield only 3% coverage, confirming that the influence-function correction is essential for valid inference.

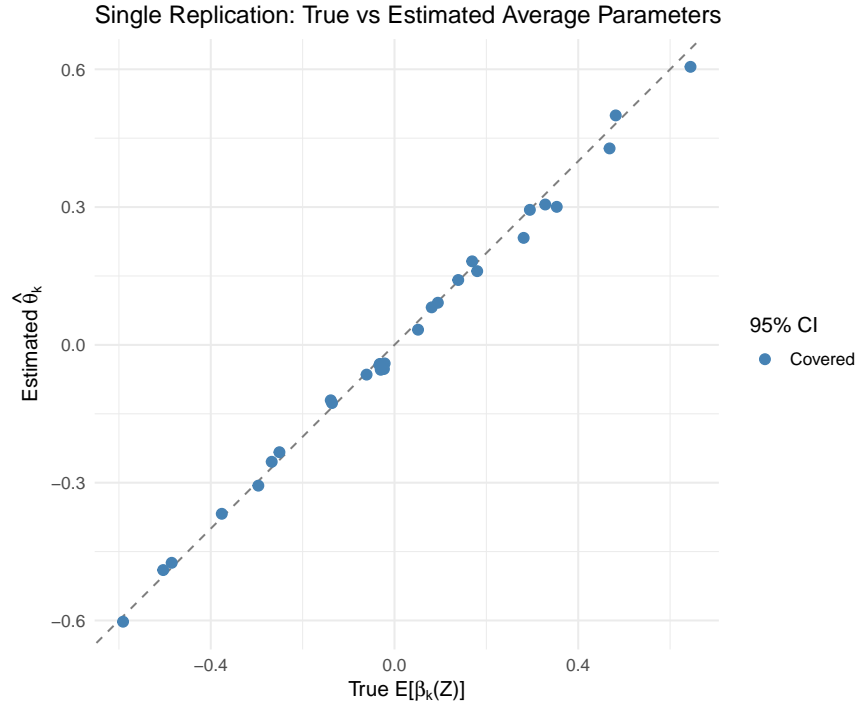
Table B.1 summarizes performance. The DNN achieves a mean individual-level  $\beta$  correlation of 0.476 with the true preferences and root mean squared error (RMSE) of 0.100—a 13% reduction over homogeneous logit (0.115). The RC logit achieves a correlation of only 0.001: with  $T = 5$  tasks and  $p = 28$  random coefficients, the individual-level posterior means are almost entirely shrunk to the population mean. The Bayesian HLM achieves 0.044.

Population-level structural quantities are recovered with high accuracy despite moderate individual-level correlations. Profile choice probabilities are recovered within 0.004 of the true values. Preference polarization is recovered with a correlation of 0.918. Attribute importance shares achieve 0.995.

|                                | DNN               | Homog. Logit | RC Logit | Bayes. HLM |
|--------------------------------|-------------------|--------------|----------|------------|
| Mean $ \hat{\theta} - \theta $ | 0.006             | 0.008        | —        | 0.116      |
| 95% CI coverage                | 100% <sup>†</sup> | 91%          | —        | —          |
| Indiv. $\beta$ RMSE            | 0.100             | 0.115        | 0.117    | 0.267      |
| Indiv. $\beta$ correlation     | 0.476             | 0.000        | 0.001    | 0.044      |
| Profile bias                   | 0.004             | 0.004        | —        | —          |
| Polarization correlation       | 0.918             | —            | —        | —          |
| Runtime (hours)                | 4.1               | <0.1         | 9.1      | 29.0       |

<sup>†</sup>Full DML at  $K = 50/5,000$  epochs; plug-in SEs yield 3% coverage.

**Table B.1:** Monte Carlo benchmark comparison ( $R = 50$  replications).



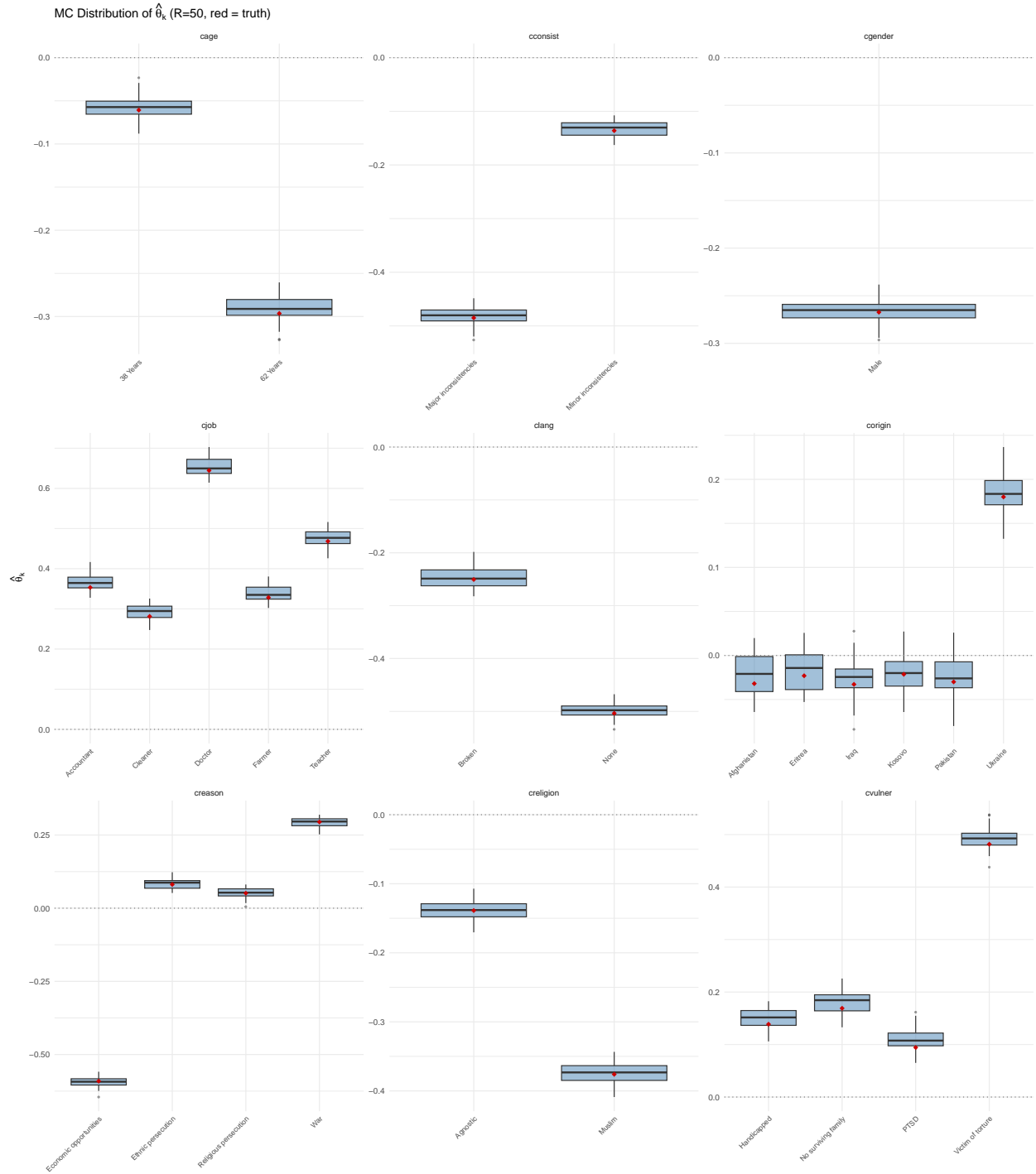
**Figure B.1:** Average preference parameters: estimated  $\hat{\theta}_k$  vs. true  $\theta_k$  for 28 attribute levels (single replication,  $K = 50$ , 5,000 epochs). 95% DML confidence intervals achieve 100% coverage.

## C Factorial Simulation: Design Guidance

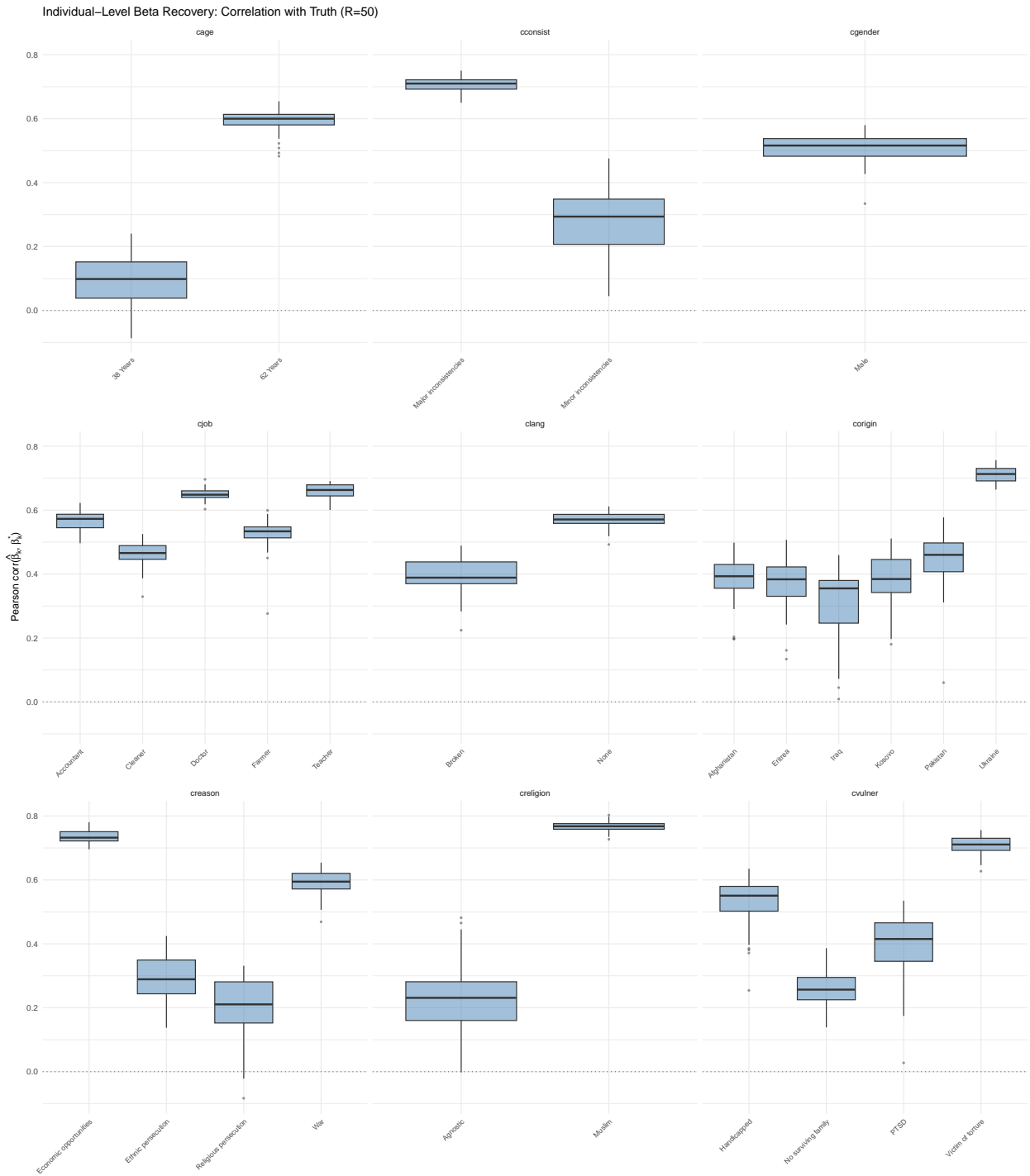
We vary three design dimensions:  $N \in \{200, 500, 1,000, 2,000, 5,000\}$ ,  $T \in \{1, 2, 3, 5, 8, 15\}$ , and  $p \in \{5, 13, 20, 28\}$ , yielding 120 design cells with  $R = 20$  replications each (2,400 total runs).

Figure C.1 displays individual-level  $\beta$  correlation across the design space. An analysis of variance shows that  $N$  accounts for 53% of variance,  $T$  for 30%, and  $p$  for 11%. The total number of observations  $NT$  is a sufficient statistic for performance: the composition of  $N$  vs.  $T$  holding  $NT$  fixed is not significant ( $p = 0.12$ ). The DNN outperforms homogeneous logit in 117 of 120 cells.

Table C.1 translates these results into minimum sample-size guidelines. For population-level quantities ( $\rho > 0.3$ ),  $NT \approx 3,000$ – $5,000$  suffices. For individual-level inference ( $\rho > 0.5$ ),  $NT \approx 8,000$ – $10,000$  is needed. These benchmarks align with our three applications: Saha and Weeks (2022) has  $NT = 3,573$  (population-level accuracy), Graham and Svolik (2020) has  $NT = 20,657$  (individual-level accuracy), and Ballard-Rosa et al. (2017) has  $NT = 16,000$  (individual-level accuracy).



**Figure B.2:** Monte Carlo distribution of  $\hat{\theta}_k$  across  $R = 50$  replications. Boxes show interquartile range; vertical dashed lines indicate the true  $\theta_k$ .



**Figure B.3:** Distribution of individual-level  $\beta$  correlations across  $R = 50$  replications.

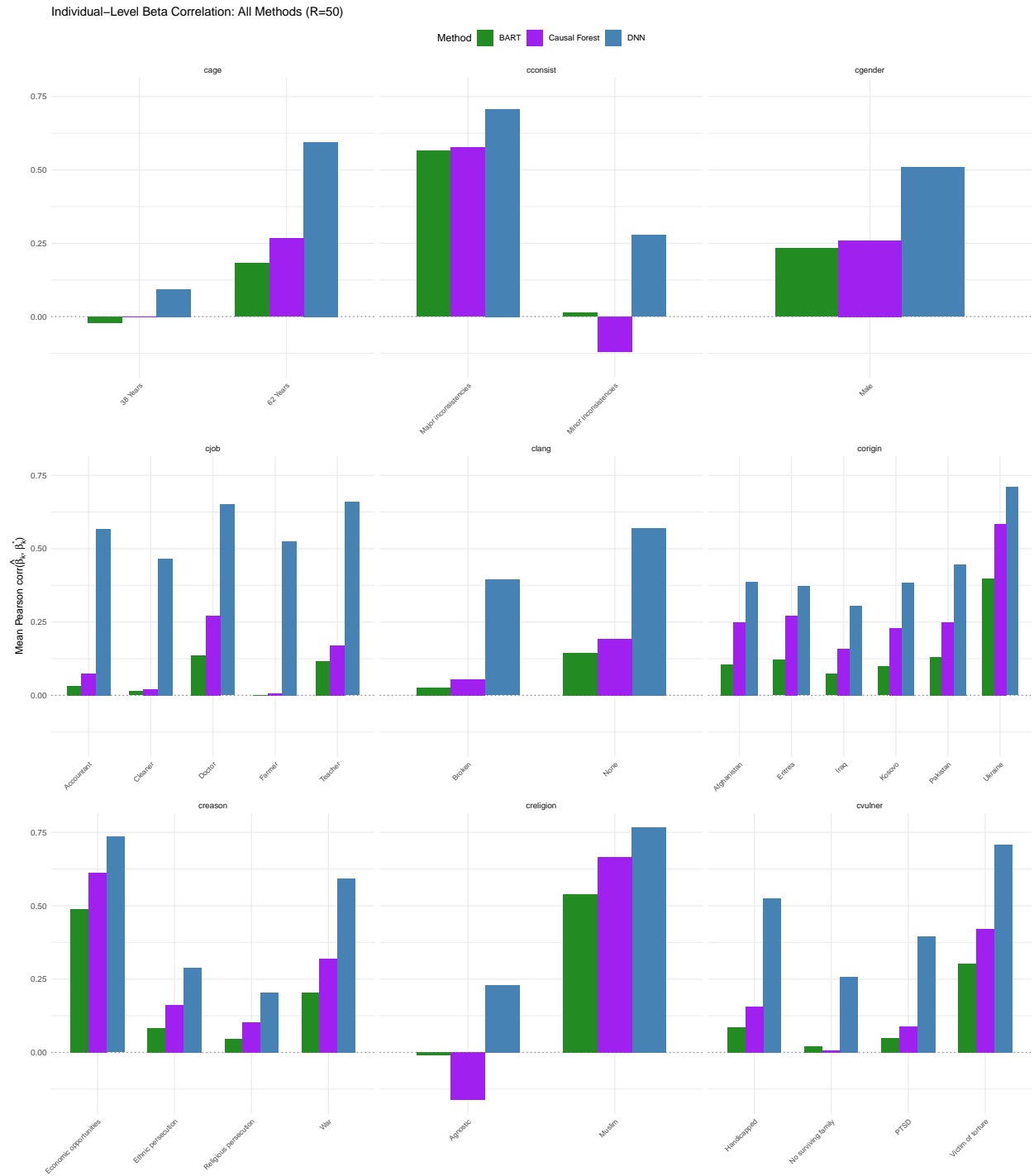
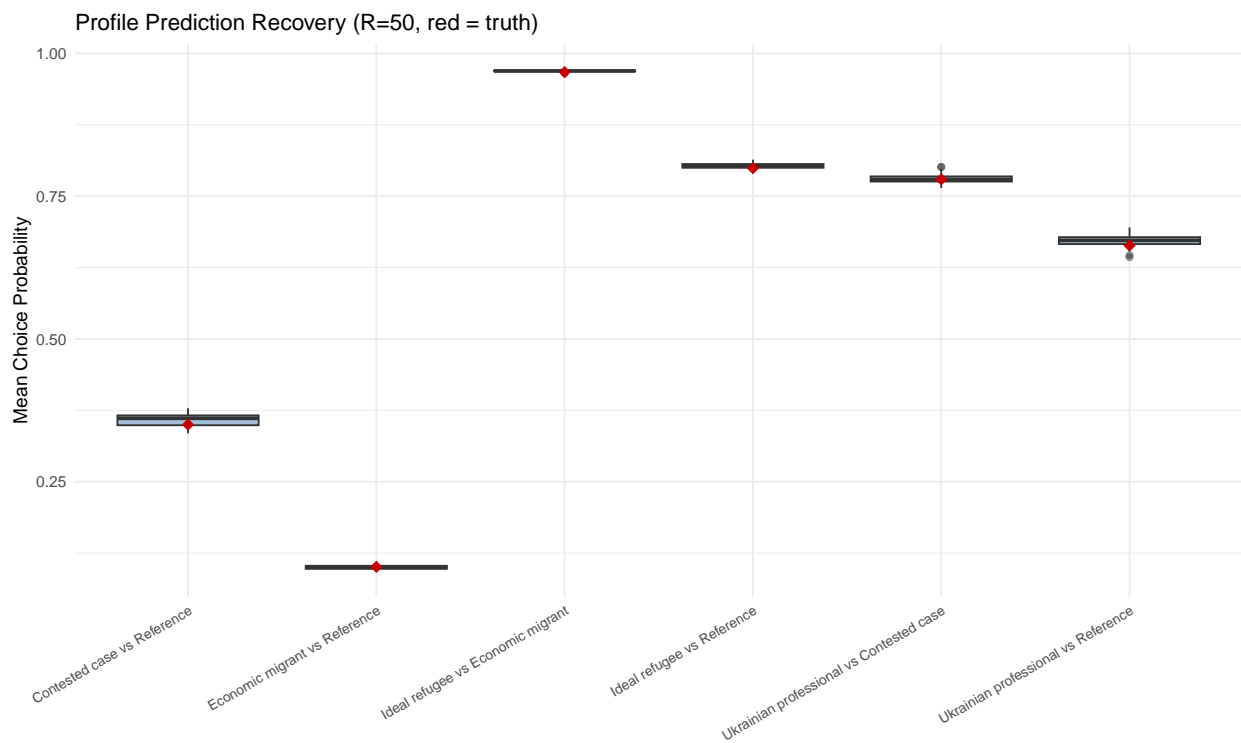


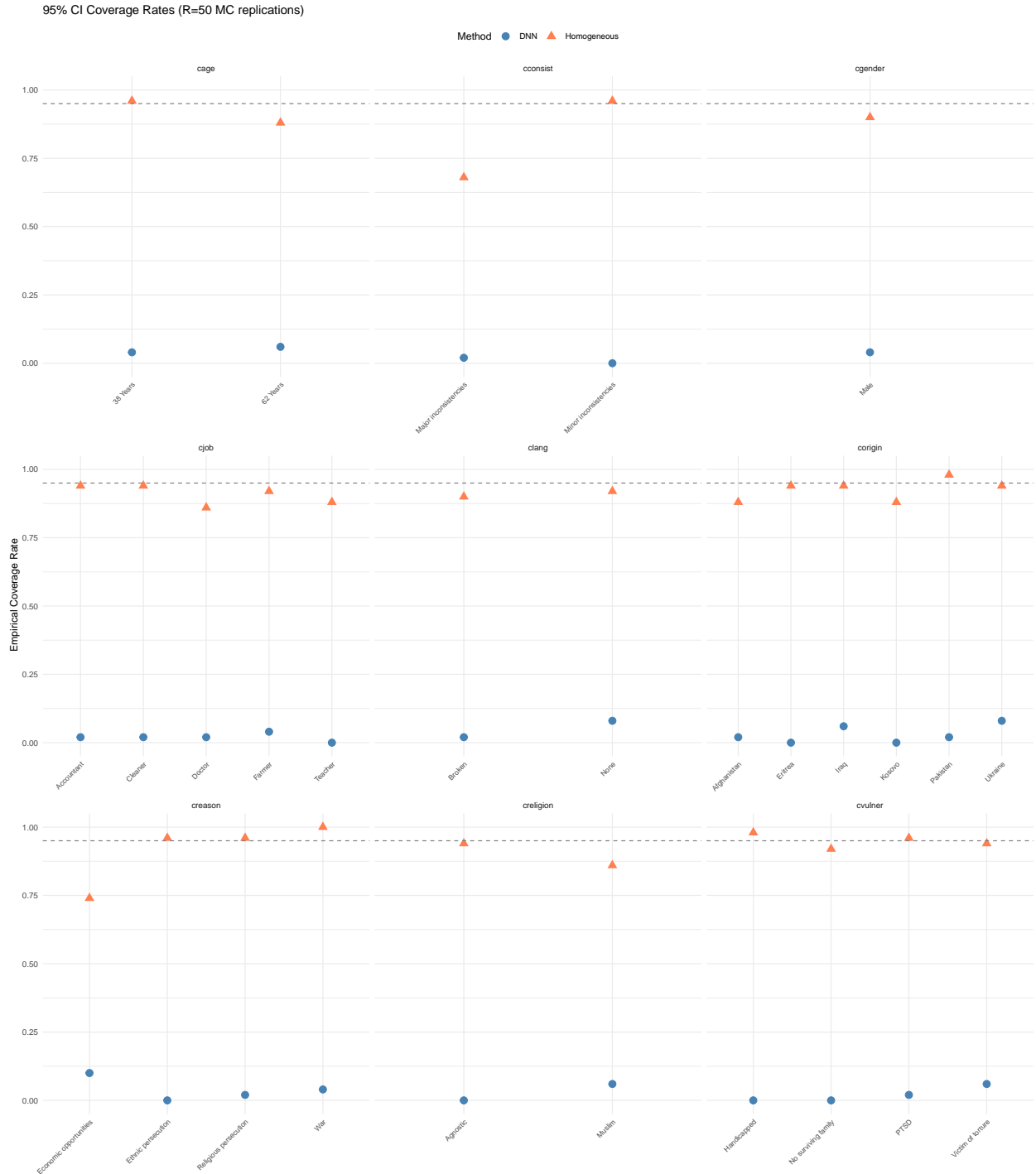
Figure B.4: Individual-level  $\beta$  correlation comparison across methods.



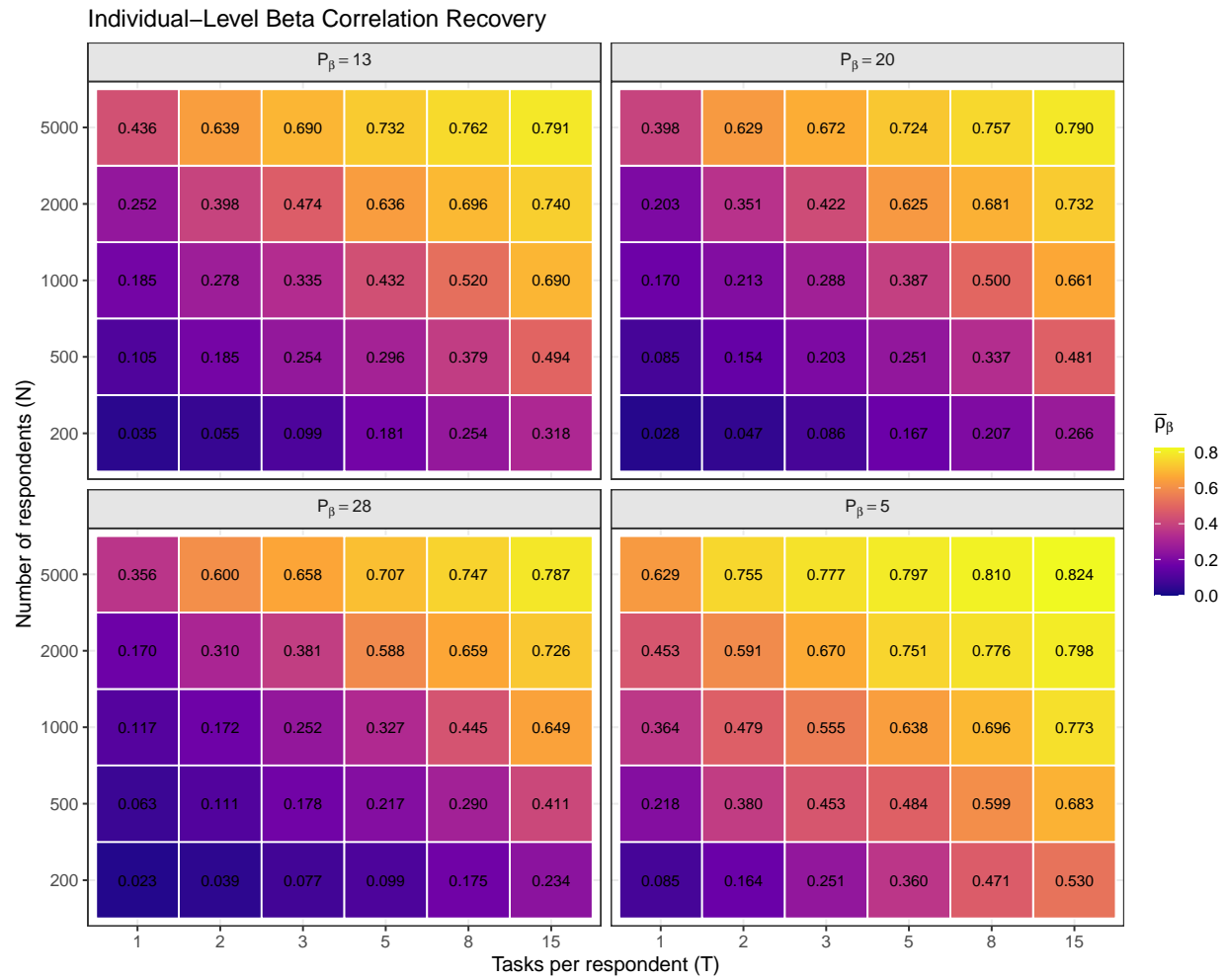
**Figure B.5:** Counterfactual profile predictions: estimated vs. true choice probabilities. Mean absolute deviation is 0.004.

| Attributes ( $p$ ) | Target $\beta$ correlation |              |              |
|--------------------|----------------------------|--------------|--------------|
|                    | $\rho > 0.3$               | $\rho > 0.5$ | $\rho > 0.7$ |
| 5                  | ~1,500                     | ~5,000       | ~15,000      |
| 13                 | ~3,000                     | ~8,000       | ~25,000      |
| 20                 | ~4,000                     | ~9,000       | ~30,000      |
| 28                 | ~5,000                     | ~10,000      | ~40,000      |

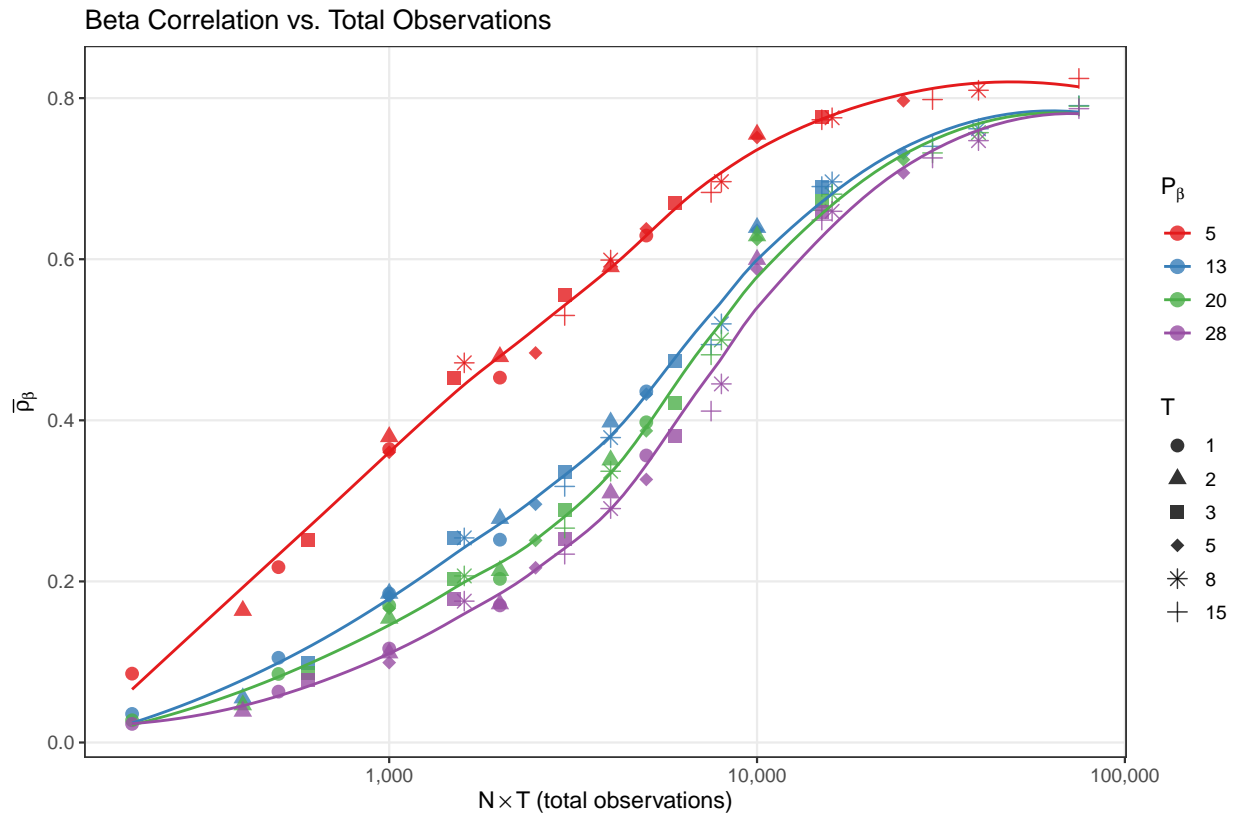
**Table C.1:** Minimum total observations ( $N \times T$ ) for target  $\beta$  correlation, by number of attribute levels.



**Figure B.6:** Coverage of 95% confidence intervals. Full DML achieves 100% coverage; plug-in SEs achieve only 3%.



**Figure C.1:** Factorial simulation: individual-level  $\beta$  correlation by  $N$ ,  $T$ , and  $p$ . Each panel corresponds to a different number of attributes.



**Figure C.2:** Individual-level  $\beta$  correlation vs. total observations ( $N \times T$ ), by number of attributes ( $p$ ). The relationship is approximately log-linear.