How does partisan type influence affective polarization?*

A comparative study of 25 European democracies

Tristan Muno†

December 8, 2025

Affective polarization (AP) is increasingly viewed as multidimensional, yet many studies restrict inference to explicit partisans who report subjective attachement. In multiparty systems, many politically engaged citizens do not identify with any party, raising the question of whether explicit attachement is necessary for partisan bias. I introduce partisan type — explicit versus implicit partisans (non-identifiers with clear vote anchors) — and test whether it conditions AP. Using a conjoint experiment fielded in 25 European democracies (N~29,800), I combine randomized partisan cues with Bayesian models to estimate AP and its components in two behavioral games. Across both, explicit and implicit partisans display nearly identical ingroup favoritism and outgroup derogation, with conditional effects centered near zero and consistent across countries. These findings imply that self-reported attachement adds little behavioral discriminatory content beyond partisan anchoring, supporting recent arguments that identity and substantive pathways yield similar partisan behaviros and clarifying AP measurement in multiparty contexts.

1 Introduction

1. Problem statement

- AP is widely studies, but who counts as a partisan varies dramatically across studies
- Many AP measures implicitly restrict inferene to explicit partisans (PID identifiers), while others include the entire electorate

*You can add acknowledgements here. Wordcount: 7617.
†University of Mannheim; Mail: tristan.muno@uni-mannheim.de

- In multiparty Europe, the prevelance of identifiers varies sharply -> population of inference becomes unstable

2. Conceptual motivation

- The literature increasingly views AP as multidimensional
- Theories of social identity expect explicit identifiers to polarize more, but identity alone may not drive AP (Orr, Fowler, and Huber 2023)
- Hence: Is explicit subjective attachement actually the driver of behaviroal AP?

3. Key innovation:

- Introduce partisan type = explicit vs implicit partisans (vote anchor without felt attachement)
- Future-proof, generalizable concept beyond U.S. "leaners"
- Clarifies measurement and inference in multiparty contexts

4. Research questions

- Do explicit and implicit partisans differ in affective polarization, ingroup favoritism, and outgroup derogation?
- Does partisan type moderate responses to randomized partisan cues?
- Does this vary cross-nationally?

5. Summary of approach

- Reanalysis of 25 country conjoint experiment (29,000 respondents)
- Profile randomization gives causal leverage on AP (within-type)
- Bayesian multilevel model estimate the CATE(TxR) contrast

6. Central empirical takeaway

- Explicit attachement adds little to no additional AP in either dictator or trust games
- Implicit partisns behave equally polarized despite denying subjective attachement
- Striking cross-national consistency

7. Contribution statement

- Measurement: demonstrates that restricting AP analyses to identifiers is unnecessary and potentially distorting
- Theory: shows the limits of identity-only accounts; supports multidimensional and cue-driven models of AP
- Comparative inference: partisan anchors, not explicit identity, capture the behaviorally relevant component of partisanship across Europe

2 Theory

1. Background: Affective polarization as multidimensional

- Rapid expansion since Iyengar, Sood, and Lelkes (2012) -> many measures, many conceptualizations
- Campos and Federico (2025): AP is not a single latent construct; different measures capture different processes (emotional, social, behavioral)
- In behavioral AP (allocations), identity may be only one pathway

2. Partisanship as social identity

- Classic social identity theory expectations:
  - Explicit identifiers -> stronger ingroup favoritism and outgroup derogation
- Expressive vs instrumental models: identity vs policy alignment

- BUT: identity strength, closeness, leaners, latent identity all complicacte the picture

3. Measurement problem: Who counts as a partisan?

- In multiparty settings, % identifiers varies dramatically (40-80% in our data)
- Restricting samples to identifiers implicitly changes the population of inference
- Eurobarometer and national surveys often capture attachement differently -> conceptual instability

4. Introducing "partisan type"

- Grabs the core idea from U.S. leaners literature, but generalizes it for multiparty contexts
- Explicit partisans: report attachement
- Implicit partisans: deny attachement but express a partisan anchor via intended/reported vote
- Rationale:
  - Vote anchor = behavioral signal of group alignment
  - Subjective attachement = psychological signal
  - These need not coincide

5. Why partisan type should matter (theoretical expectation)

- Strong identity pathway: explicit identifiers shoullds exhibit more IF and OD
- Mechanisms: categorization -> group affect -> threat -> exclusivity -> discrimination

6. Why partisan type may not matter (alternative pathways)

- Identity loyalty vs substantive alignment produce observationally equivalent behaviors (Orr, Fowler, and Huber 2023)

- Vote choice carries information about:
  - Ideology
  - Issue positions
  - Social composition of supporters
  - Expected norms of reciprocity
- Thus implicit partisans may activate the same heuristics as explicit identifiers even without subjective attachement

7. Hypotheses

- H1: Explicit > implicit in AP
- H2: Explicit > implicit in IF
- H3: Explicit > implicit in OD
- Note: These hypotheses represent the identity-centered expectation, but competing accounts (multidimensional AP, cue-based reasoning, normative expectations) predict null differences

3 Research Design

In this section, we detail our empirical strategy for identifying the effect of partisan type on affective polarization. We begin by introducing the dataset and experimental design, then describe the construction of partisan types and anchors, including a cross-national overview of their distribution. This distributional evidence clarifies the scope of the analytical sample and motivates our comparative research design. We then outline the key variables of interest and present a directed acyclical graph (DAG) that summarizes our causal model of the data-generating process. Next, we specify the statistical model

and explain how we obtain the quantities of interest, followed by a justification of our covariate set. Finally, we enumerate the assumptions required for causal interpretation and situate our design within the broader literature, highlighting both strengths and remaining limitations.

3.1 Data and Experimental Design

To evaluate how different partisan types influence AP, we reanalyze the cross-national conjoint experiment conducted by Hahm, Hilpert, and König (2024). The study was fielded by Dynata (formerly SSI) in 25 EU member states between late May and mid-August 2019.[1] National samples consist of approximately 1,100 respondents per country (29,827 in total), drawn to be broadly representative on key sociodemographic dimensions. Summary statistics appear in table XX in the appendix.

The experiment comprised two behavioral games: a dictator game capturing unilateral prosociality and a trust game introducing reciprocal incentives. As the games capture conceptually distinct behavioral settings, they will be analyzed separately. In each round, respondents interacted with a hypothetical partner with attributes ranging from age, gender, class, religion, to nationality and partisanship.[2] All conjoint attributes, including the partisan cue, were randomized independently with one restriction: partisan labels were displayed only when the profile's nationality matched the respondent's own country. This restriction was implemented to avoid implausible cross-national partisan combinations and maintain realism of the profile space. As a consequence, our analysis is restricted to the subset of profiles in which the partner is a conational. Within this restricted set, the control condition corresponds to conational profiles without a displayed partisan label.

[1] The countries covered are all EU member states (which included the UK at the time of data collection) except Luxembourg, Malta and Cyprus.

[2] Detailed information, instruction wording and an exemplary profile are presented in the appendix.

These restrictions do not affect randomization of the remaining attributes and preserve the "as-if random" variation in partisan cues among conational partners. Each respondent completed three rounds of each game, yielding six observations per participant. Our unit of analysis is thus a single game round. The resulting data are inherently nested: multiple rounds within respondents, respondents nested within party affiliations, and respondents and parties clustered within countries. We therefore rely on hierarchical models, described in detail in Section 3.3.

These behavioral measures offer several advantages for studying affective polarization. First, the design is explicitly suited to assessing *horizontal* AP among mass partisans (Areal and Harteveld 2024). Common observational indicators — such as like/dislike scales or feeling thermometers — conflate horizontal evaluations of fellow citizens with *vertical* evaluations of party elites, thereby limiting their interpretability (Druckman and Levendusky 2019). Second, the inclusion of a nonpartisan baseline condition enables us to causally decompose AP into its constituent components: ingroup favoritism (IF) toward co-partisans and outgroup derogation (OD) toward rival partisans. We estimate these components using standard conjoint estimands that capture the average marginal effect of each attribute on token allocations (Hainmueller, Hopkins, and Yamamoto 2014).

Third, respondents hold multiple, potentially cross-cutting social identities (Roccas and Brewer 2002). The randomized conjoint design mitigates bias arising from such heterogeneity by independently varying several attributes that might otherwise be confounded with partisan cues. This feature allows us to isolate the specific effect of partisanship and reduces the risks of "aliasing" (Hainmueller, Hopkins, and Yamamoto 2014) and "masking" (Bansak et al. 2021), which occur when omitted attributes correlate with treatment dimensions (Dafoe, Zhang, and Caughey 2018).[3] Finally, by embedding partisan

[3]Naturally, a conjoint design can only include a limited set of attributes and therefore cannot preclude all forms of aliasing; we make no claim to exhaustiveness. Nonetheless, by randomizing several

information among multiple randomized cues, conjoint tasks are less susceptible to social desirability pressures than direct survey questions (Horiuchi, Markovich, and Yamamoto 2022).

3.2 Measurement and Variable Construction

We construct our main variables of interest following a tripartite framework to ensure conceptual coherence and terminological consistency in all party-related measures. First, we define T_i as the *partisan type* of respondent i. If respondent i reported a party to which she feels attached to and is thus considered an *explicit* partisan, we define $T_i = 1$. Conversely, we set $T_i = 0$ for *implicit* partisans, i.e., respondent i who indicated no partisan attachement but reported a vote preference. We term the party reported by respondents *partisan anchor* A_i.[4]

To evaluate the scope conditions of our design and to clarify the population of inference, we begin by documenting the distribution of partisan types in each of the 25 country samples. Figure 1 presents the share of respondents classified as explicit partisans, implicit partisans, and nonpartisans (type undefined). Two features of these descriptive patterns are substantively important. First, the proportion of nonpartisans is remarkably stable across countries (ranging mostly between 10 and 15%, least in Estonia (6%), most in Croatia (23%)), suggesting that the exclusion of this group — required for treatment consistency in our design — does not introduce systematic cross-national biases in

plausibly correlated characteristics (e.g., age, gender), the design reduces the risk that inferences about partisanship merely reflect respondents' assumptions about the social composition of party supporters. In this respect, it is substantially less susceptible to the confounding that affects like/dislike scales or feeling thermometers, which cannot disentangle the effect of partisanship from correlated beliefs about associated social groups.

[4]We use the term *partisan anchor* to denote the party reference point reported by respondents, regardless of whether they express subjective attachment to that party. For explicit partisans, this anchor corresponds to the conventional *partisan ingroup* in the literature (as in Hahm, Hilpert, and König (2024)). For implicit partisans, however, respondents explicitly deny a felt attachment, making social-identity terminology inappropriate. The term *anchor* therefore provides a neutral label that applies consistently across both groups.

the analytical sample. Second, and more consequential for our theoretical claims, the balance between explicit and implicit partisans varies dramatically across national contexts. In some electorates, explicit identifiers constitute the clear majority of partisans (e.g. strongest case being Denmark with 85% explicit partisans), in others, implicit partisans are equally prevalent (e.g. in Latvia both explicit and implicit partisans account for 45% of respondents). This heterogeneity underscores the need for a comparative framework capable of assessing whether implicit and explicit partisans differ systematically in their affective responses. In short, the figure motivates both the internal validity of our sample restriction and the broader comparative question that animates the study: whether distinct partisan types, whose prevalence differs markedly across countries, exhibit different patterns of affective polarization.

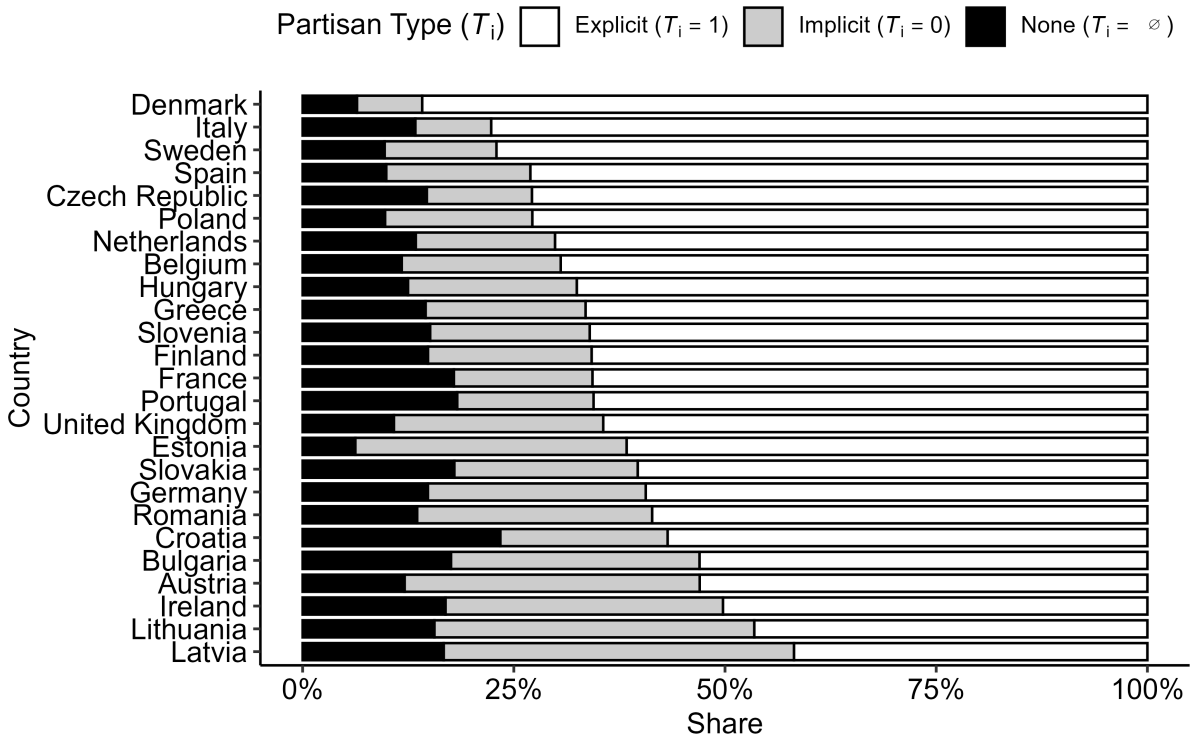

Figure 1: Distribution of partisan types, by country. Stacked horizontal bars show the within-country share (%) of three partisan types: explicit partisans (respondents who reported a subjective attachement to a party, $T_i = 1$), implicit partisans (respondents who reported no attachement but did report a vote preference or intention, $T_i = 0$), and respondents who reported neither (none, $T_i = \emptyset$). Percentages sum to 100% within each country, with country samples containing about $1,100$ respondents each (detailed numbers are reported in appendix section X).

In the experiment, each conjoint profile in round r contains a randomized party cue (Z_r^{party}). We compare this cue to the respondent's anchor to classify the *partisan relationship* R_{ri} as co-partisan, out-partisan, or neutral ($R_{ri} \in \{Co, Out, None\}$). This coding corresponds to the "average across out-groups" operationalization and offers a tractable strategy for multi-party contexts, although it constitutes a substantial reduction of the inherent complexity of AP in multi-party systems (Röllicke 2023). We adopt this measure as it provides a uniform and comparable classification scheme across countries.

We conceptually frame explicit partisanship as the treatment and implicit partisanship as a control. Respondents reporting a subjective attachement to a political party ($T_i = 1$) thus form the treated group, whereas respondents expressing vote choice without attachement ($T_i = 0$) constitute the comparison group.[5] The underlying research question is: Does explicit partisan attachement amplify affective polarization relative to implicit partisan attachement? In causal terms, we ask how a hypothetical intervention shifting T from 0 to 1 would change affective polarization, holding other factors constant.

Figure 2 depicts the underlying causal data-generating process as a directed acyclical graph (DAG). As noted above, respondents report a partisan anchor A. Partisan type T is defined only for the subset of respondents who either explicitly or implicitly report a party and therefore appears as a child node of A. In each round r of the conjoint games, respondents receive a set of conjoint attributes, denoted by Z_r. The partisan cue Z_r^{party} and respondents' anchors A_i jointly determine the partisan relationship category via the mapping $f : (A_i, Z_r^{\text{party}} \mapsto R_{ij} \in \{\text{Co}, \text{Out}, \text{None}\})$.

The dependent variable Y_{riac} is the token allocation made in round r by respondent i with anchor a from country c. Token allocations Y are affected by both R and Z_r^{other},

[5] For readability, we omit the prefix "quasi-" when referring to treatment and control, without implying random assignement. Our design remains observational with respect to our treatment of interest, partisan type T.

with the effect of R theorized to be conditional on T. Because T is observational, we cannot assume conditional independence. We therefore include C to capture individual-level confounders affecting both T and Y. Graphically, our substantive interest lies in the causal chain $T \rightarrow R \rightarrow Y$, namely the effect of R on Y conditional on T, represented as the central vertical path in Figure 2.

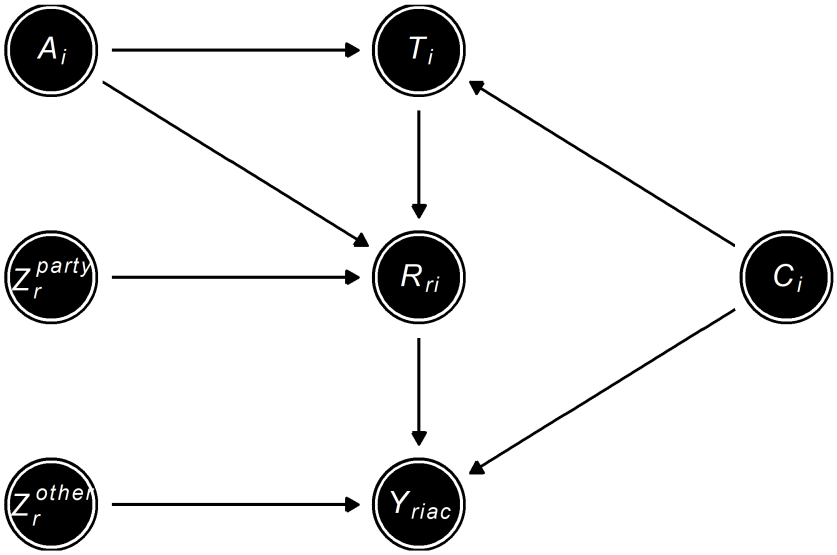

Figure 2: Directed acyclic graph of the causal data-generating process. Respondents have a partisan anchor A_i, representing the party they feel attached to (explicit partisans) or intend to vote for (implicit partisans). Partisan type T_i (explicit vs. implicit) is only defined for respondents with a partisan anchor and is therefore a child node of A_i. Each conjoint profile shown in round r presents randomized attributes: a partisan cue Z_r^{party} and other attributes Z_r^{other}. The partisan-relationship variable $R_{ri} = f(A_i, Z_r^{party})$ determines whether the profile is interpreted as a co-partisan, out-partisan, or neutral for respondent i. Token allocations Y_{riac} are affected by both R_{ri} and Z_r^{other}, with the effect of R_{ri} theorized to depend on T_i. Because T_i is observational, C_i denotes potential confounders of both T_i and Y_{ri}, highlighting the assumptions required for causal interpretation.

3.3 Model Specification and Estimation

We model the data in their natural hierarchical structure. The unit of analysis is a single conjoint profile evaluation in round r of a game. Profile evaluations (r) are nested within respondents (i), respondents are grouped by their partisan anchors (a), i.e., the party to which they report attachement or vote intention, and anchors are nested within countries (c). This yields a four-level structure: Level 1 (round r) $\rightarrow$ Level 2 (respondent i) $\rightarrow$ Level 3 (anchor a) $\rightarrow$ Level 4 (country c).

Let Y_{riac} denote the number of tokens allocated (0-10) in round r by respondent i with partisan anchor a in country c. Respondent partisan type is $T_i \in \{0, 1\}$ (implicit vs explicit). The respondent-profile partisan relationship $R_{ri} \in \{\text{Co, Out, None}\}$ is a deterministic function of the respondent's anchor and the profiles partisan cue. Z_r^{other} denotes the matrix of all other conjoint attributes, and C_i constitutes a matrix of respondent-level covariates discussed in Section 3.5.

To estimate how explicit partisan type conditions responsiveness to partisan cues, we specify a multilevel model with a cross-level interaction between T_i and R_{ri} and random effects at the respondent, anchor, and country levels, as shown in Equation 1:

$$\begin{aligned} Y_{riac} = {} & \beta_0 + \beta_1 T_i + \beta_2 R_{ri} + \beta_3 (T_i \times R_{ri}) + \beta_4 Z_{ri}^{\text{other}} + \beta_5 C_i \\ & + u_{i0} \\ & + v_{a0} + v_{aT} T_i + v_{aR} R_{ri} \\ & + w_{c0} + w_{cT} T_i + w_{cR} R_{ri} \\ & + \varepsilon_{riac} \end{aligned} \tag{1}$$

We assume a Gaussian likelihood for the bounded (0-10, see Figure 10 in the appendix) token allocation. While the discrete nature and strict bounds of the outcome violate the technical assumptions of the Gaussian family (as the distribution is constrained and not continuous), the approach is justified by its focus on the marginal means effects (AP, IF, OD). The estimator for the regression coefficients derived from this likelihood family is

mathematically equivalent to the Ordinary Least Squares (OLS) estimator under fixed effects and Normal errors (Wooldridge 2010). Crucially, the OLS estimator provides unbiased and consistent estimates of the conditional mean $E(Y \mid X)$ under large samples, even when the residuals are non-Normal or heteroskedastic, relying on asymptotic properties (White 1980). This pragmatic choice allows for directly interpretable coefficients on the original token allocation scale, avoiding the complexity of nonlinear link functions required by models such. Furthermore, the inclusion of a rich covariate set and the respondent-level random intercept (u_{i0}) helps to absorv substantial structured variance, mitigating the severity of residual non-Normality that often arises from bounding effects. Respondent-level random intercepts are assumed to follow a normal distribution (Equation 2):

$$u_{i0} \sim \mathcal{N}(0, \sigma^2_{u0}) \tag{2}$$

To capture heterogeneous treatment and cue effects, we include a random intercept and random slopes for T and R at the partisan anchor level. These random effects are assumed to be jointly normal (Equation 3), with intercepts and slopes allowed to be correlated:

$$\begin{pmatrix} v_{a0} \\ v_{aT} \\ v_{aR} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_a \right), \tag{3}$$

Similarly, we include random intercepts and random slopes for T and R at the country level, also assumed to be jointly normal (Equation 4):

$$\begin{pmatrix} w_{c0} \\ w_{cT} \\ w_{cR} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_c \right), \tag{4}$$

Both Σ_a and Σ_c are unstructured 3×3 covariance matrices, allowing intercepts and slopes to be estimated as correlated at the anchor and country levels. Residuals are assumed to be normally distributed around 0 (Equation 5):

$$\varepsilon_{riac} \sim \mathcal{N}(0, \sigma^2). \tag{5}$$

The interaction $\beta_3(T_i \times R_{ri})$ is the key parameter: it captures how explicit partisan type ($T_i = 1$) conditions sensitivity to co-, out- and nonpartisan cues (R_{ri}). Because all profile attributes – including the partisan cue – are randomized within respondents, R_{ri} varies as-if at random conditional on A_i, supporting causal interpretation of cue effects within each partisan type. We allow the effect of R_{ri} to vary across partisan anchors (v_{aR}) and across countries (w_{cR}) to capture heterogeneity plausibly structured by party and national context. Put differently, the magnitude of affective polarization and its subcomponent likely varies across parties and across countries, and the random slopes are designed to account for this structure. Similarly, the effect of T_i is allowed to vary across partisan anchors (v_{aT}) and countries (w_{cT}), reflecting that the meaning and salience of explicit attachement differ across parties and settings. Respondent-level heterogeneity in baseline generosity is captured by the random intercept u_{i0}. We do not include respondent-level random slopes for R_{ri} as recommended by Heisig and Schaeffer (2019), because within-respondent variation in partisan-relationship categories is insufficient to reliably estimate individual-specific cue sensitivities. Excluding this slope shifts heterogeneity in cue sensitivity to the anchor and country levels, remaining individual-level variation is absorbed by higher-level and residual variances. The correlated random-effects structure enables partial pooling, stabilizing heterogeneous treatment and cue-effect estimates while preserving systematic contextual variation.

We estimate the model in a Bayesian framework via Markov Chain Monte Carlo (MCMC) simulation implemented in Stan (Stan Development Team 2025) using the `brms` package in R (Bürkner 2017, 2018, 2021). This approach propagates uncertainty through the multilevel interaction structure and regularizes the maximal random-effects specification through priors. Our priors are weakly informative and tailored to the 0-10 outcome scale. Fixed-effect parameters (β_k) receive $\mathcal{N}(0, 2)$ priors, which center effects at zero while constraining 95% of prior mass to approximately ± 4 tokens — a substantively large shift (40% of the scale) given the initial findings of Hahm, Hilpert, and König (2024). The model intercept receives a $\mathcal{N}(5, 2)$ prior centered at the midpoint of the token scale. Standard deviations of the random effects σ_{RE} are assigned Half-student-$t(3, 0, 1)$ priors. This distribution is defined exclusively on the positive domain ($[0, \infty)$) and has heav-

ier tails compared to the Half-Normal distribution, allowing for the possibility of larger group-level variability while still shrinking effects towards zero. The residual standard deviation σ is assigned a Half-student-$t(3, 0, 2)$ prior. These priors improve convergance, facilitate shrinkage in a high-dimensional multilevel structure, and reduce the probability of implausible prior-predictive draws (e.g., $Y > 10$).

The randomized conjoint design secures internal validity for profile-level effects (R_{ri} and other Z attributes) within partisan types. The contrast across partisan types (i.e, the effect of R conditional on T) remains observational and is identified under adjustmend for C_i and the multilevel structure. Our hierarchical specification – allowing anchor- and country-varying slopes – and weakly informative priors stabilize these comparisons while transparently reflecting remaining uncertainty in the posterior.

3.4 Quantities of Interest: AP, IF and OD

In this section we define our three substantive quantities of interest — AP and its two constituent components, IF and OD. Our empirical focus lies in comparing these measures across implicit and explicit partisan types. Framed causally, we aim to recover the effect of a hypothetical intervention: holding all covariates constant, how do these quantities of interest change when T is set from 0 (implicit type) to 1 (explicit type)?

We conceptualize affective polarization as the difference in the expected number of tokens allocated to co-partisans versus out-partisans in the conjoint games. This follows the approach used in prior work (CHECK Wagner Reiljan), which averages across all political outgroups — a simplification in multi-party systems but one that yields a conservative measure and facilitates cross-national comparison, as both the number of viable parties and the size of the outgroup set vary across countries. Formally, we define affective polarization AP as:

$$AP = E(Y \mid R = Co) - E(Y \mid R = Out) \tag{6}$$

With respect to partisan type, we define our causal estimand as the conditional average treatment effect:

$$\begin{aligned} CATE_{AP} &= AP_{T=1} - AP_{T=0} \\ &= [E(Y \mid R = Co, T = 1) - E(Y \mid R = Out, T = 1)] \\ &\quad - [E(Y \mid R = Co, T = 0) - E(Y \mid R = Out, T = 0)] \end{aligned} \tag{7}$$

To disentangle AP into its components, we exploit the neutral control condition in the conjoint experiment, in which no partisan cue was displayed ($R = \text{None}$). This allows us to estimate the extent to which co-partisanship increases token allocations relative to a neutral baseline, conditional on T. We define ingroup favoritism IF as:

$$IF = E(Y \mid R = Co) - E(Y \mid R = None) \tag{8}$$

Refining this with respect to partisan type yields the causal estimand:

$$\begin{aligned} CATE_{IF} &= IF_{T=1} - IF_{T=0} \\ &= [E(Y \mid R = Co, T = 1) - E(Y \mid R = None, T = 1)] \\ &\quad - [E(Y \mid R = Co, T = 0) - E(Y \mid R = None, T = 0)] \end{aligned} \tag{9}$$

The same neutral condition ($R = \text{None}$) also enables estimation of outgroup derogation as the reduction in tokens allocated to out-partisans compared to the baseline. Outgroup derogation OD is defined as:

$$OD = E(Y \mid R = None) - E(Y \mid R = Out) \tag{10}$$

The corresponding causal estimand is:

$$\begin{aligned} CATE_{OD} &= OD_{T=1} - OD_{T=0} \\ &= [E(Y \mid R = None, T = 1) - E(Y \mid R = Out, T = 1)] \\ &\quad - [E(Y \mid R = None, T = 0) - E(Y \mid R = Out, T = 0)] \end{aligned} \tag{11}$$

By focusing on expected values (marginal means), we ensure that our substantive results do not depend on arbitrary choices of reference categories (Leeper, Hobolt, and Tilley 2020). Consistent with our hypotheses stated in **?@sec-ht**, we expect all

$CATE_{AP,IF,OD} > 0$: respondents with explicit partisan attachements ($T = 1$) should display higher affective polarization than implicit types ($T = 0$), caused simoultaneously by higher degrees of IF as well as stronger OD.

3.5 Covariates and Confounder Adjustment

To block potential backdoor paths between partisan type and allocation behavior (C_i in Figure 2), we adjust for a comprehensive set of individual-level covariates capturing the principal dimensions that may jointly predict explicit partisan attachement and prosocial behavior in experimental settings.

First, we include standard sociodemographic characteristics: age, gender, education, social class, religious affiliation, and urban-rural residence. Second, we control for respondents' political orientations and ideological predispositions, including left-right self-placement, cultural and economic nativism, attitudes toward the EU, and satisfaction with democracy. Third, we account for political interest, engagement, and knowledge. Fourth, we incorporate attitudes toward parties, elites, and democratic processes, capturing broader orientations toward the political system that plausibly shape both group attachement and intergroup conduct. Finally, we include economic evaluations and behavioral despositions (risk-taking and temporal discounting) to adjust for contextual and psychological factors associated with baseline generosity.

Together, these covariates represent substantively plausible confounders and help approximate conditional independence between partisan type and the outcome of interest. Summary statistics for all covariates by partisan type appear in Table 4, and English-language questionnaire items are provided in the appendix.

3.6 Identification Strategy and Causal Assumptions

Our empirical strategy aims to estimate the causal effect of explicit partisan type ($T = 1$) relative to implicit partisan type ($T = 0$) on affective polarization (AP, IF, OD). The

identification approach combines strong internal validity provided by the conjoint experiment for the profile attributes (Z) with an extensive model-based adjustmend strategy for the observed contrast across partisan types (T).

The conjoint design provides strong leverage: all profile attributes (Z) are randomized within each round. Since the respondent-profile relationship (R_{ri}) is a deterministic function of a randomized cue (Z_r^{party}) and the respondent's partisan anchor (A_i), the profile-level variation that drives AP, IF, and OF is as-if random. Consequently, the effect of R (and other randomized attributes) on the token allocation Y_{riac} is causally identified by design, conditional on the respondent's partisan type T.

The key identification challenge is the conditioning of cue effects on partisan type, estimated via the interaction $T \times R$. Identification of this contrast relies on the observational assumption that T is conditionally independent of potential outcomes, given the comprehensive set of covariates C and the full hierarchical structure. Our modeling strategy is explicitly designed to manage this residual confounding.

We formally state the key assumptions and detail our strategy for addressing the principal threat of unobserved confounding:

SUVTA

We assume no interference across respondents and a well-defined treatment. The survey was administered individually and anonymously, rendering cross-respondent spillovers unlikely. Dependence across rounds within respondents is mitigated by randomized profiles and explicitly modeled via the respondent-level random intercepts. Treatment consistency is ensured through a precise coding of T, exclusion of nonpartisan respondents from the analysis, and by allowing its effect to vary across parties and countries.

Conditional Independence (Ignorability) of T

Conditional on the set of observed covariates C and the grouping units (partisan anchor, country), assignment to explicit versus implicit partisan type is assumed intependent of

potential outcomes. While randomization of Z secures identification of the effects of R, the causal interpretation of the Conditional Average Treatment Effects ($CATE$s) across T rests on this ignorability assumption.

Covariate Balance

Observed covariates should be comparably distributed across $T = 1$ and $T = 0$. Descriptive assessments in Table 4 in the appendix indicate a high degree of distributional symmetry across the two groups. The medians are identical for 21 out of 23 scaled political and economic variables, suggesting that the regression adjustment (via $\beta_5 C_i$) starts from a strong, balanced position, substantially mitigating confounding by observed characteristics. Categorical sociodemographic variables also show similar proportions, with the main observed difference being that explicit partisans ($T = 1$) are more likely to be male (53% vs 43%).

Common Support (Overlap)

Both partisan types must occur with positive probability for observed covariate profiles across all grouping units.

The principal threat to the causal interpretation of the $CATE$s is residual confounding: unmeasured factors may jointly influence selection into T and allocation behavior Y. We deploy a multilevel strategy that leverages the repeated-measures design to control for unobserved condounding at all three hierarchical levels:

1. Respondent-level adjustment (u_{i0}): The inclusion of a respondent-level random intercept acts as a powerfull control for all stable, time-invariant, unobserved individual characteristics (e.g., latent personality traits, baseline generosity, or stable intensity of social identity) that may affect both selection into T and the outcome Y. This leverage is fundamental to our observational identification strategy.

2. Partisan anchor-level adjustment (v_{a0}, v_{aT}, v_{aR}): The random effects at the partisan anchor level account for unobserved confounding systematic across party lines.

They absorb average unobserved differences (e.g., party organizational culture or ideological uniformity) that influence both the propensity for explicit attachement (T) and baseline prosocioality (Y). The random slopes for T and R further allow the confounding or effect size to vary by party.

3. Country-level adjustment (w_{c0}, w_{cT}, w_{cR}): The country-level random effects address unobserved confounding operating at the national context level. This accounts for country-specific factors (e.g., general political climate, institutional trust, or overall polarization intensity) that influences both the prevalence of explicit attachement (T) and the general levels of prosocial behavior (Y).

Our hierachical Bayesian model encodes this strategy. All reported estimands are obtained via g-computaion (observed-values approach), averaging posterior expectations over the empirical covariate distribution and relevant grouping units, which propagates uncertainty through the entire multilevel structure.

We are confident in the internal validity of profile-level effects (R and other Z attributes) and of the derived quantities (AP, IF, OD) conditional on T. as these rely on random assignement of Z. The contrast between explicit and implicit partisans — the $CATE$s — is necessarily more assumption-intensive. Our modeling strategy, combining rich covariate adjustment and hierarchical partial pooling, substanially reduces, though cannot eliminate, the possibility of unobserved confounding.

Our goal is to approximate the effect of a hypothetical intervention shifting a respondent's partisan type from implicit ($T = 0$) to explicit ($T = 1$), holding other factors constant. Because T is observational, this interpretation is valid only under the assumptions outlined above. We thus present the resulting estimates as the best model-based causal approximation afforded by the available data and the experimental design.

4 Empirical analysis

We organize the presentation of results in two steps. We begin with pooled analyses based on posterior expected values generated from the fitted hierarchical model. These expected

values are computed for prespecified combinations of respondent's partisan type (T=0 implicit, T=1 explicit) and partisan relationship category (R = Co, Out, None) in a given game round, averaging over the empirical distribution of covariates and integrating over respondent random effects. From these expected values, we derive the three substantive quantities of interest — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — for each value of T. The difference between the quantities under explicit and implicit partisanship yields the CATE for each QoI.

The pooled results sections display, for each QoI, the posterior distributions for explicit and implicit partisans (left panels) alongside the posterior distributions of the corresponding CATE (right panels). All panels visualise full posterior densities along with respective medians, 95% and 99% credible intervals.

We then turn to country-specific analyses, where we recompute expected values separately within each national subsample (again averaging over that country's empirical covariate distribution, but this time incorporating random effects). Because the resulting CATEs show little cross-national variation, we additionally report the underlying country-level expected values for each relationship category. These reveal substantial heterogeneity in the levels of AP, even as the implicit-explicit comparison remains consistently close to zero across countries.

All results are presented separately for the dictator and trust games. For scale reference, a one-token difference corresponds to 42.6% of a standard deviation in the dictator game and 40.2% in the trust game.

4.1 The effect of partisan type on affective polarization in Europe

Dictator game

Figure 3 displays the pooled dictator-game estimates. Interpreted within the causal framework introduced above — where the estimand approximates the effect of a hypothetical intervention shifting individuals from implicit to explicit partisanship — the findings indi-

cate that such an intervention would have, at most, minimal consequences for allocative discrimination.

The overall AP contrast is highly similar across partisan types. Explicit identifiers allocate on average 0.99 tokens more to co-partisans than to out-partisans (median AP = 0.99; 95% CrI: 0.91, 1.08), virtually matching the behavior of implicit partisans (0.93; 95% CrI: 0.75, 1.10). The estimated causal effect is therefore small (CATE = 0.07; 95% CrI: -0.11, 0.24), with roughly three-quarters of its posterior mass concentrated below 0.1 tokens. Although the mode of the posterior aligns with H1, the uncertainty dominates, and the key empirical signal is the strength of AP among implicit partisans: even in the absence of self-reported attachement, they display clear and precisely estimated discriminatory allocations.

A similar pattern appears for ingroup favoritism. Explicit partisans show a strong tendency to reward co-partisans relative to neutral alters (median IF = 0.54; 95% CrI: 0.46, 0.63), and implicit partisans exhibit a nearly equivalent tendency (median = 0.45; 95% CrI: 0.26, 0.64). The corresponding causal contrast again remains modest (CATE = 0.09; 95% CrI: -0.10, 0.25). The posterior probability that explicit identifiers would show stronger IF under the hypothetical intervention is 83%, but the implied effect sizes are substantively negligible. As with AP, implicit partisanship alone suffices to generate meaningful levels of IF, consistent with accounts of partisanship as a cue-based or behavioral orientation rather than purely a self-ascribed identity alone.

Outgroup derogation yields the clearest indication of equivalence. Explicit partisans withhold 0.45 tokens from out-partisans (95% CrI: 0.36, 0.54), nearly identical to implicit partisans (0.48; 95% CrI: 0.35, 0.60). The causal effect is essentially null (CATE = -0.03; 95% CrI: -0.16, 0.10). The posterior even places slightly more mass on implicit partisans derogating somewhat more ($\Pr(\text{CATE} < 0) = 0.67$), though again the magnitudes are trivial. These results do not support H3.

Viewed jointly, the dictator-game evidence suggests that explicitly endorsing a partisan identity adds little to the behavioral AP dynamics already present among implicit partisans. Posterior medians close to zero, symmetric uncertainty, and concentration around

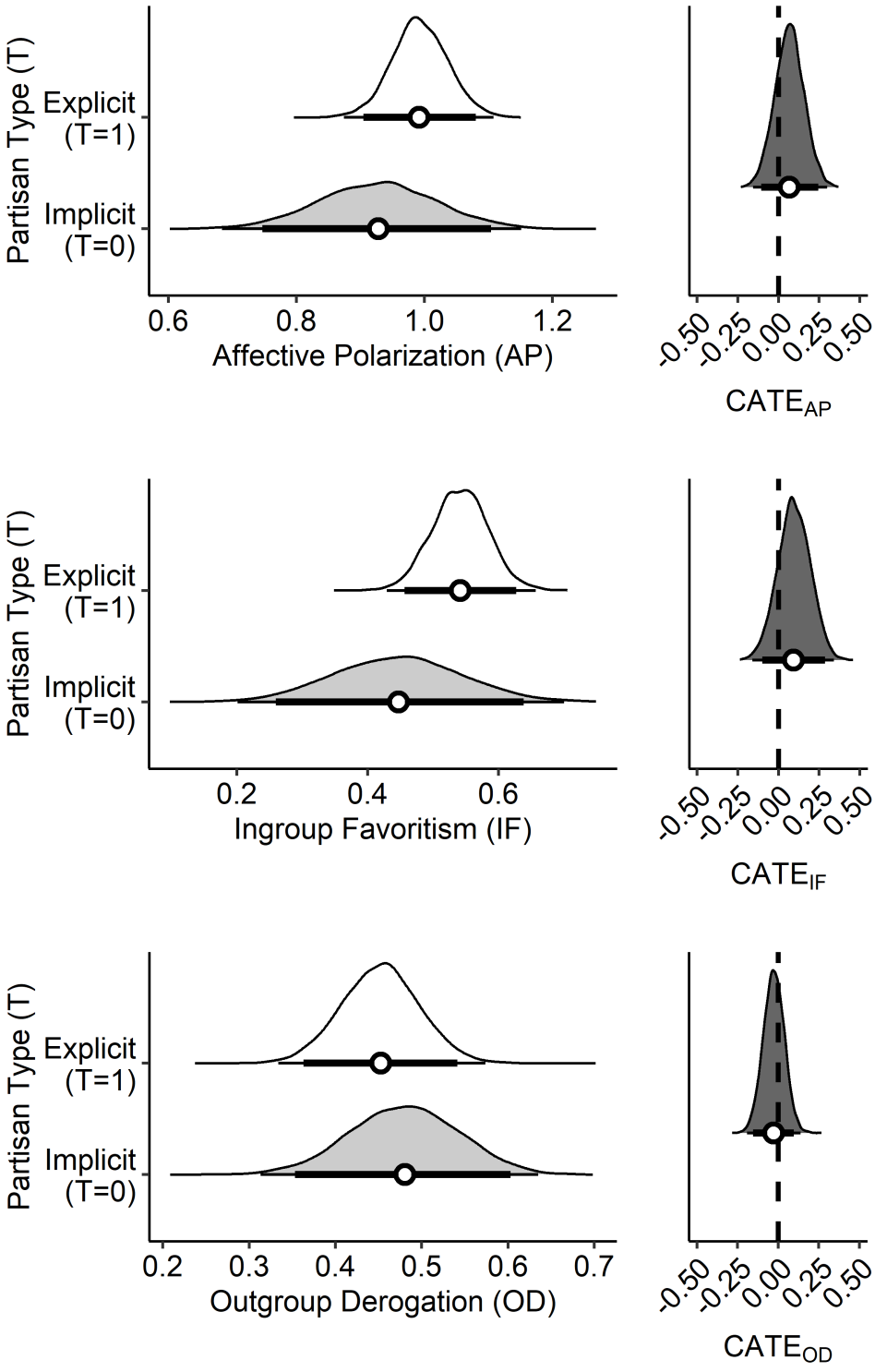

Figure 3: Pooled posterior distributions for AP, IF, and OD and the respective conditional average treatment effects in the dictator game. The figure displays pooled posterior estimates of three derived quantities — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — on the token outcome $Y_r iac$ in the dictator game. Rows correspond to AP, IF, and OD. In each row the left panel shows the marginal posterior distirbutions by partisan type (explicit $T_i = 1$, implicit $T_i = 0$), and the right panel shows the corresponding Conditional Average Treatment Effect (CATE), i.e., the difference between explicit and implicit partisans for that quantity. All posterior quantities are obtained from the covariate-adjusted hierarchical model described in section XX using the observed-values approach (g-computation). The panels depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

substantively small effects all imply that — under the intervention logic motivating our design — explicit identification does not meaningfully amplify partisan discrimination. Instead, the behavioral expression of AP appears grounded in categorization and anchoring, aligning with theoretical accounts emphasizing informational or norm-based pathways rather than subjective self-placement. We next examine whether this pattern extends to the trust game, where reciprocal structure may alter the relative contribution of ingroup and outgroup components.

Trust game

Figure 4 summarizes the pooled trust-game results. Although the broader conclusion mirrors the dictator game — limited causal influence of explicit over implicit partisanship — the component patterns differ in informative ways. Relative to the dictator game, the trust game produces (i) no directional signal fo AP, (ii) a reversed pattern for IF, and (iii) the clearest support for the hypothesized effect for OD.

For AP, explicit and implicit partisans behave indistinguishably: both allocate 1.08 tokens more to co-partisans (explicit 95% CrI: 1.00, 1.17; implicit 95% CrI: 0.89, 1.26). The estimated causal contrast is centered exactly at zero (CATE = 0.00; 95% CrI: -0.17, 0.19), producing a posterior split nearly evenly around the null. This stands in contrast to the slight directional tendency observed in the dictator game and offers no evidence for H1 in the reciprocal context.

Ingroup favoritism, by contrast, exhibits a modest reversal. Explicit identifiers reward co-partisans by 0.63 tokens (95% CrI: 0.53, 0.72), whereas implicit identifiers display a slightly larger premium (0.75; 95% CrI: 0.53, 0.95). The resulting causal estimate is negative (CATE = -0.12; 95% CrI: -0.32, 0.09), with the posterior assigning a probability of 0.86 to implicit partisans favoring co-partisans more strongly. Although the effect remains small, its direction differs from the dictator-game pattern, suggesting that reciprocal incentives condition the distribution of group-based benefits differently across partisan types.

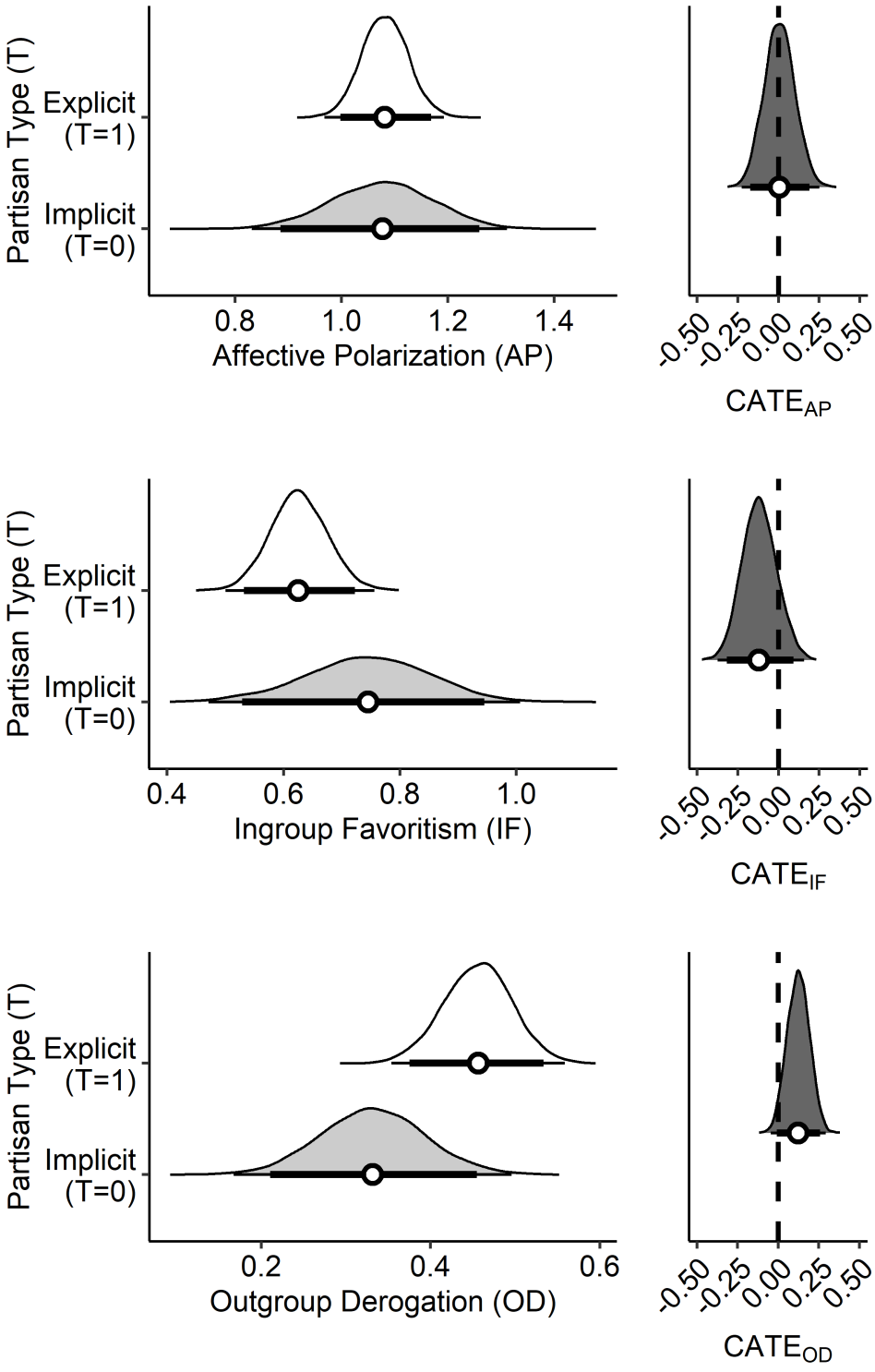

Figure 4: Pooled posterior distributions for AP, IF, and OD and the respective conditional average treatment effects in the trust game. The figure displays pooled posterior estimates of three derived quantities — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — on the token outcome $Y_r iac$ in the trust game. Rows correspond to AP, IF, and OD. In each row the left panel shows the marginal posterior distirbutions by partisan type (explicit $T_i = 1$, implicit $T_i = 0$), and the right panel shows the corresponding Conditional Average Treatment Effects (CATE), i.e., the difference between explicit and implicit partisans for that quantity. All posterior quantities are obtained from the covariate-adjusted hierarchical model described in section XX using the observed-values approach (g-computation). The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

Outgroup derogation provides the most supportive trust-game evidence for the hypothesized influence of explicit identification. Explicit partisans withhold 0.46 tokens (95% CrI: 0.38, 0.53) from out-partisans, compared with 0.33 among implicit partisans (95% CrI: 0.21, 0.46). The estimated causal contrast is positive (CATE = 0.12; 95% CrI: -0.01, 0.26), and the posterior probability that explicit identifiers derogate more is 0.97. Even here, however, the substantive magnitude remains limited, and implicit partisans still exhibit meaningful OD.

Synthesizing across both games, the evidence consistently indicates that explicit self-reported partisanship exerts only small, uncertain, and context-dependent causal effects on discriminatory AP behaviors. In several cases the directional tendency aligns with theoretical expectations; in others it reverses; and in no instance does the estimated effect approach a substantially large magnitude within conventional ranges of certainty. The lack of cross-game convergence is particularly informative: behaviors attributed to explicit partisan identity do not replicate across allocative and reciprocal environments, undermining the expectation that subjective attachement reliably intensifies partisan discrimination.

Instead, the pooled findings point to a robust conclusion: individuals who decline to label themselves as partisans nonetheless exhibit levels of AP — including both favoritism and derogation — that closely match those of explicit partisans. Implicit identification alone suffices to produce substantial partisan bias in both behavioral contexts under study.

Finally, the trust-game evidence underscores a key measurement implication. Aggregate AP can mask compensating differences in IF and OD. As observed here, similar levels of AP can arise from distinct combinations of its components, potentially obscuring theoretically relevant heterogeneity. Analyses relying solely on co-partisan versus out-partisan contrasts may therefore miss meaningful variation in the underlying mechanisms of affective polarization.

4.2 Do explicit and implicit partisans differ across countries?

We next examine whether the effect of partisan type varies across national contexts. If the causal influence of explicit identity depends on characteristics such as party-system structure, political culture, or the salience of partisan conflict, we would expect meaningfull cross-country variation in the Conditional Average Treatment Effect (CATE).

Figure 5 presents the country-specific CATE estimates for the dictator game. Strikingly, each country's posterior distribution is nearly identical to the pooled estimate. No country exhibits a directional deviation, and the uncertainty intervals overlap almost perfectly. The trust-game results in Figure 6 replicate this pattern: the country-level CATEs again mirror the pooled contrasts, with no systematic cross-national dispersion.

Such uniformity initially raises the concern that partial pooling or model specification might have obscured meaningful heterogeneity. To address this, we turn to the conditional expected values (EVs) that underpin the QoIs and their respective CATEs and examine the underlying behavioral patterns directly.

In the dictator game (Figure 7), clear cross-national differences emerge when comparing across rows. Countries vary in baseline giving (as indicated by contributions to "None") and in the magnitude of affective polarization and its components (visible as the relative distances between the shaded posteriors associated with co-partisan, neutral, and out-partisan recipients). These differences demonstrate that the model is not surpressing genuine cross-country variation: respondents in some contexts exhibit more AP, IF, or OD than those in others.

However, when comparing across partisan types within each country (the two columns in each panel), the expected valzes are virtually indistinguishable. In every case, implicit and explicit partisans display the same pattern of allocations, differing mainy in posterior precision (implicit types tend to have wider distributions) rather than in their locations. For example, Ireland exhibits lower overall AP than Spain, but this holds for both implicit and explicit partisans. The same cross-national ordering — and the same within-country equality — appears throughout the sample. This symmetry explains why the CATEs are identical across countries: while countries differ in how much partisans

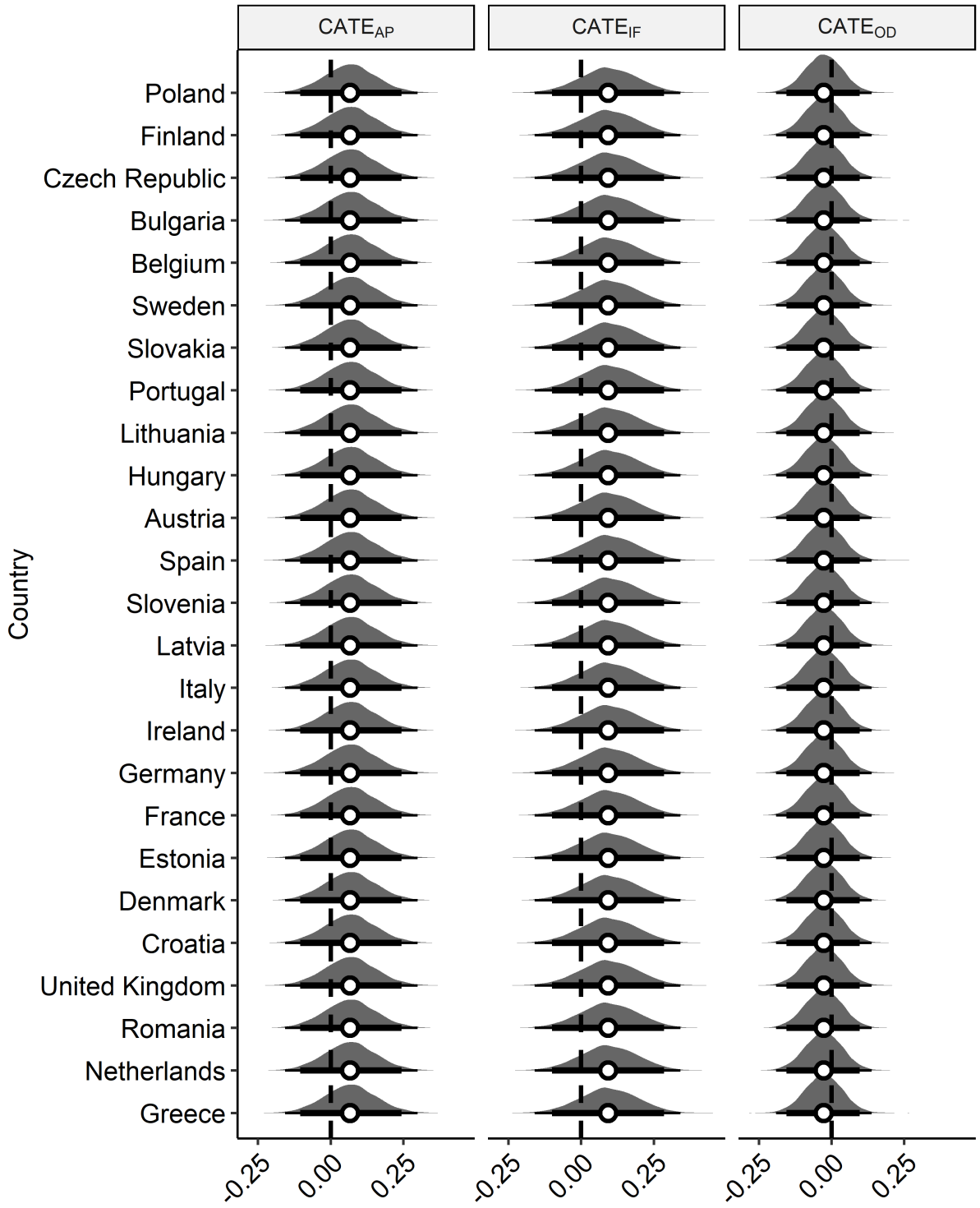

Figure 5: Country-specific CATEs for AP, IF, and OD in the dictator game. The figure displays country-specific posterior distributions of the Conditional Average Treatment Effects (CATEs) of explicit ($T_i = 1$) versus implicit ($T_i = 0$) partisanship on three derived quantities – affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) – measured on the token outcome $Y_r iac$. Columns correspond to $CATE_{AP}$, $CATE_{IF}$, and $CATE_{OD}$, respectively. The y-axis lists countries and the x-axis reports effect sizes in tokens. Positive values indicate larger AP/IF/OD among explicit relative to implicit partisans within a country. All estimates are sampled from the covariate-adjusted hierarchical model (seciton XX) using the observed-values (g-computation) approach. For each country, we counstruct country-specific datasets that fix T and R to the focal values, generate posterior predictions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

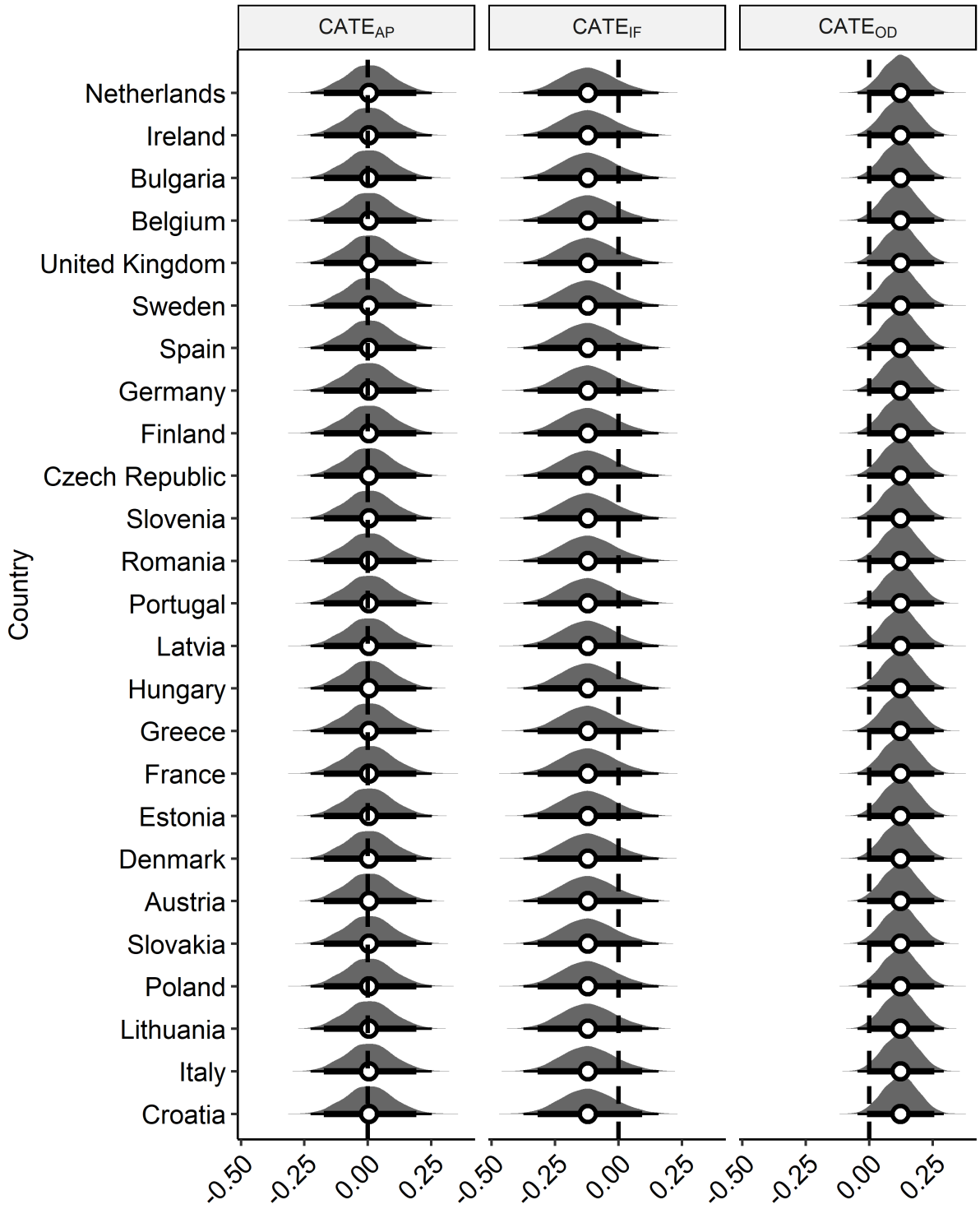

Figure 6: Country-specific CATEs for AP, IF, and OD in the trust game. The figure displays country-specific posterior distributions of the Conditional Average Treatment Effects (CATEs) of explicit ($T_i = 1$) versus implicit ($T_i = 0$) partisanship on three derived quantities – affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) – measured on the token outcome Y_{riac}. Columns correspond to $CATE_{AP}$, $CATE_{IF}$, and $CATE_{OD}$, respectively. The y-axis lists countries and the x-axis reports effect sizes in tokens. Positive values indicate larger AP/IF/OD among explicit relative to implicit partisans within a country. All estimates are sampled from the covariate-adjusted hierarchical model (seciton XX) using the observed-values (g-computation) approach. For each country, we counstruct country-specific datasets that fix T and R to the focal values, generate posterior predictions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

discriminate, they do not differ in how much *additional* discrimination is associated with explicit identification.

The trust-game EVs in Figure 8 reinforce this conclusion. Again substantial between-country variation is visible across rows, in both aggregate AP and the composition of IF and OD. Yet within each country, implicit and explicit identifiers behave nearly identically, with differences emerging only in the variance of the posterior distributions. The absence of divergence across partisan types is consistent and systematic.

Taken together, the country-level analyses provide strong evidence that the causal effect of shifting from implicit to explicit partisan type is negligible in every national context examined. Although countries differ in their overall levels of allocative and reciprocal discrimination, the *difference between partisan types* is remarkably stable and effectively zero across all 25 settings. This conclusion is further supported by the hierachical model: the estimated variance of the country-level random slopes for the $T \times R$ interaction is extremely small (SD < 0.11), indicating minimal cross-national dispersion in the causal effect.

In sum, while affective polarization itself varies meaningfully across Europe, the behavioral consequences of explicit versus implicit partisan identification do not. Implicit identification appears sufficient to generate the levels of AP observed in each country, and making partisanship explicit adds little — if anything — to the discriminatory behaviors captured in either experimental task.

5 Robustness

6 Conclusion

1. Recap of main empirical finding

- Across 25 European democracies, and two behavioral games, -> explicit partisan attachement does not meaningfully amplify affective polarization
- CATEs extremely small, posterior mass near zero, often symmetric

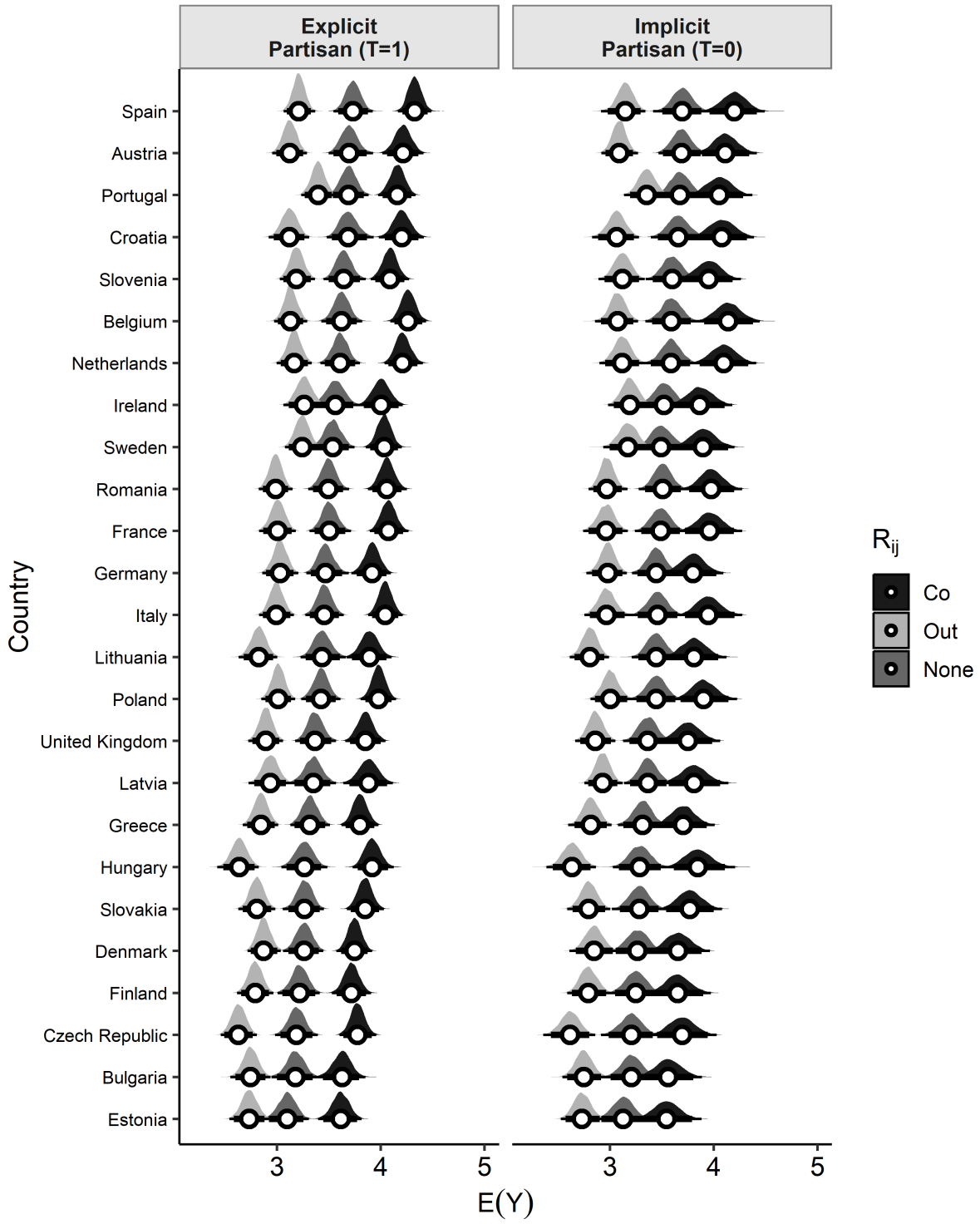

Figure 7: Country-specific posterior expected values by partisan type T and relationship R in the dictator game. The figure plots country-specific posterior expected token allocations E_{riac} on the 0-10 scale, separatly by partisan type T and partisan relationship R. The x-axis shows the expected tokens, the y-axis lists countries. Columns split partisan type (left: explicit, $T_i = 1$; right: implicit, $T_i = 0$). Within each column, densities display the posterior distributions for the three relationship conditions $R \in Co, Out, None$. Larger separation between Co and Out within a country indicates greater affective polarization (AP). Comparisons across columns reveal how levels vary by partisan type within country. Estimates are obtained from the covariate-adjusted hierarchical model (section XX) using the observed-value (g-computation) approach. For each country, we construct country-specific datasets fixing T and R to the focal values, generate posterior preditions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. Distributions show full posterior densities, points mark posterior medians, thick and thin horizontal bars denote 95% and 99% cresible intervals, respectively.

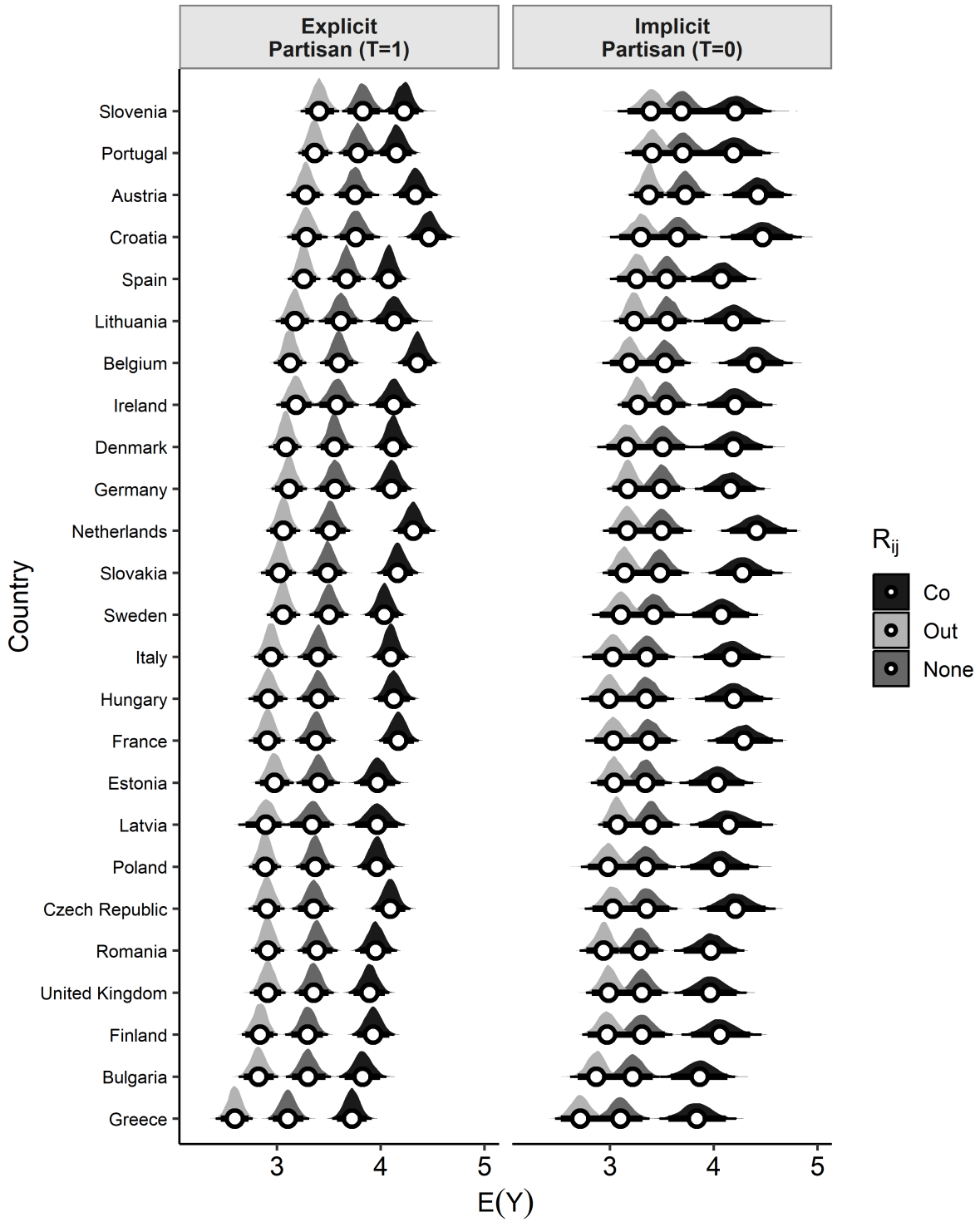

Figure 8: Country-specific posterior expected values by partisan type T and relationship R in the trust game. The figure plots country-specific posterior expected token allocations E_{riac} on the 0-10 scale, separatly by partisan type T and partisan relationship R. The x-axis shows the expected tokens, the y-axis lists countries. Columns split partisan type (left: explicit, $T_i = 1$; right: implicit, $T_i = 0$). Within each column, densities display the posterior distributions for the three relationship conditions $R \in Co, Out, None$. Larger separation between Co and Out within a country indicates greater affective polarization (AP). Comparisons across columns reveal how levels vary by partisan type within country. Estimates are obtained from the covariate-adjusted hierarchical model (section XX) using the observed-value (g-computation) approach. For each country, we construct country-specific datasets fixing T and R to the focal values, generate posterior preditions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. Distributions show full posterior densities, points mark posterior medians, thick and thin horizontal bars denote 95% and 99% cresible intervals, respectively.

- Implicit partisans exhibit robust AP (IF and OD) despite rejecting attachement

2. Implication 1: Measurement and inference

- Widespread practice of including leaners / vote anchor respondents is justified for behavioral AP
- Measuring relying *only* on identifiers risk selecting a shrinking subgroup in many countries
- Partisan type clarifies when populations of inference shift

3. Implication 2: Theory of partisan identity

- Explicit identity (PID attachement item) captures only one dimension of partisan identity
- Behavioral AP seems driven by latent identity, behavioral anchoring, and cue-based expectations more than by self-reported identity intensity
- Supports Orr, Fowler, and Huber (2023) view:
  - identity-like behaviors need not come from identity strensght
- Supports Campos and Federico (2025) view:
  - AP is multidimensional; behavioral AP does not cleanly map onto subjective identification

4. Implication 3: Comparative politics

- Cross-national consistency is notable: the meaning of explicit identification differs across countries, yet its *conditional effect on AP* is remarkably stable
- Suggests a general mechanism of partisan cue processing operating across democracies
- Highlights how multi-party systems structure partisan identity differently from U.S. two-party contexts

5. Methodological implications

- Conjoint experiments allow decomposing AP into IF and OD

- Design shows how randomization of cues + hierarchical modeling yields strong causal leverage
- Provides a model for how to estimate AP with heterogeneous partisanship structures

6. Limitations

- Behavioral AP not equal to all AP (thermometers, social distance, elite affect may differ)
- Single-item PID measure may be noisy; multiple-item batteries could refine type definitions
- Cross-sectional study; identity formation dynamics remain open

7. Future research avenues

- Test whether partisan type matters more for elite evaluations or symbolic identity
- Explore interactions with ideological extremety or policy sophistication
- Examine whether partisan type moderates responses to elite cues (vs mass cues)

8. Concluding statement

- Partisan identity is real, but not reducible to reported subjective attachement alone
- Implicit partisanship — the behavioral anchor — carries enough informational and social content to generate the same AP behaviors as explicit identity
- This challenges identity-only explanations and strengthens the case for multidimensional- cue-driven models of AP in comparative politics.

7 References

Areal, J., and E. Harteveld. 2024. "Vertical Vs Horizontal Affective Polarization: Disentangling Feelings Towards Elites and Voters." *Electoral Studies* 90. https://doi.org/10.1016/j.electstud.2024.102814.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, Teppei Yamamoto, James N. Druckman, and Donald P. Green. 2021. "Conjoint Survey Experiments." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, 19:19–41. Cambridge University Press Cambridge. https://doi.org/10.1017/9781108777919.004.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

———. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

———. 2021. "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software* 100 (5): 1–54. https://doi.org/10.18637/jss.v100.i05.

Campos, Nicolas, and Christopher Federico. 2025. "A New Measure of Affective Polarization." *American Political Science Review*, 1–19. https://doi.org/10.1017/S0003055425000255.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26 (4): 399–416. https://doi.org/10.1017/pan.2018.9.

Druckman, James, and Matthew Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83 (1): 114–22. https://doi.org/10.1093/poq/nfz003.

Hahm, Hyeonho, David Hilpert, and Thomas König. 2024. "Divided We Unite: The Nature of Partyism and the Role of Coalition Partnership in Europe." *American Political Science Review* 118 (1): 69–87. https://doi.org/10.1017/S0003055423000266.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30. https://doi.org/10.1093/pan/mpt024.

Heisig, Jan Paul, and Merlin Schaeffer. 2019. "Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction." *European Sociological Review* 35 (2): 258–79. https://doi.org/10.1093/esr/jcy053.

Horiuchi, Yusaku, Zachary Markovich, and Teppei Yamamoto. 2022. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 30 (4): 535–49. https://doi.org/10.1017/pan.2021.30.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–31. https://doi.org/10.1093/poq/nfs038.

Leeper, Thomas J, Sara B Hobolt, and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28 (2): 207–21. https://doi.org/10.1017/pan.2019.30.

Orr, Lilla V, Anthony Fowler, and Gregory A Huber. 2023. "Is Affective Polarization Driven by Identity, Loyalty, or Substance?" *American Journal of Political Science* 67 (4): 948–62. https://doi.org/10.1111/ajps.12796.

Roccas, Sonia, and Marilynn B. Brewer. 2002. "Social Identity Complexity." *Personality and Social Psychology Review* 6 (2): 88–106. https://doi.org/10.1207/S15327957PSPR0602_01.

Röllicke, L. 2023. "Polarisation, Identity and Affect - Conceptualising Affective Polarisation in Multi-Party Systems." *Electoral Studies* 85. https://doi.org/10.1016/j.electstud.2023.102655.

Stan Development Team. 2025. "RStan: The R Interface to Stan." https://mc-stan.org/.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 817–38. https://doi.org/10.2307/1912934.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

8 Appendix

8.1 Sample descriptives

Table 1

Country	N	Percent
Austria	1277	0.04
Belgium	1305	0.04
Bulgaria	982	0.03
Croatia	1240	0.04
Czech Republic	1135	0.04
Denmark	1200	0.04
Estonia	944	0.03
Finland	1160	0.04
France	1156	0.04
Germany	1188	0.04
Greece	1161	0.04
Hungary	986	0.03
Ireland	1061	0.04
Italy	1172	0.04
Latvia	1148	0.04
Lithuania	1265	0.04
Netherlands	1221	0.04
Poland	1198	0.04
Portugal	1187	0.04
Romania	1480	0.05
Slovakia	1297	0.04
Slovenia	1135	0.04
Spain	1396	0.05
Sweden	1254	0.04
United Kingdom	1279	0.04
Total	29827	0.99

Table 2

meta_country	1	2	3
Austria	645	630	2
Belgium	729	574	2
Bulgaria	463	518	1
Croatia	545	694	1
Czech Republic	515	618	2
Denmark	690	508	2
Estonia	331	611	2
Finland	586	567	7
France	562	594	0
Germany	587	597	4
Greece	588	572	1

Table 2

meta_country	1	2	3
Hungary	493	492	1
Ireland	481	577	3
Italy	603	569	0
Latvia	415	733	0
Lithuania	462	803	0
Netherlands	642	577	2
Poland	540	658	0
Portugal	593	593	1
Romania	837	641	2
Slovakia	550	746	1
Slovenia	569	566	0
Spain	677	718	1
Sweden	648	602	4
United Kingdom	635	641	3
Total	14386	15399	42

8.2 Experimental setup

Before the behavioral games, Hahm, Hilpert, and König (2024) presented respondents a short background information overview and instructions. For the dictator game, these were: *This game is played by pairs of individuals. Each pair is made up of a Player 1 and a Player 2. Each player will have some information about the other player, but you will not be told who the other players are during or after the experiment. The game is conducted as follows: A sum of 10 tokens will be provisionally allocated to Player 1 at the start of each round. Player 1 will then decide how much of the 10 tokens to offer to Player 2. Player 1 could give some, all, or none of the 10 tokens. Player 1 keeps all tokens not given to Player 2. Player 2 gets to keep all the tokens Player 1 offers. You will play this game three times with three different people.* In the trust game, the provided information and instruction were: *This game is played by pairs of individuals. Each pair is made up of a Player 1 and a Player 2. Each player will have some information about the other player, but you will not be told who the other players are during or after the experiment. Each player will receive 10 tokens. Player 1 then has the opportunity to give a portion of his or her 10 tokens to Player 2. Player 1 could give some, all, or none of the 10 tokens. Whatever amount Player 1 decides to give to Player 2 will be tripled before it is passed on to Player 2. Player 2 then has the option of returning any portion of this tripled amount to Player 1. Then, the game is over. Player 1 receives whatever he or she keeps from the original 10 tokens, plus anything returned to him or her by Player 2. Player 2 receives their original 10 tokens, plus whatever he or she keeps after returning any portion of the tripled amount to Player 1. You will play this game three times, with three different people. The more tokens you obtain, the more successful you will be.*

In both games respondents were shown a tabular overview of Player 2 after the instructions. Figure 9 shows an example of such a profile along with the interface respondents were provided to assign the 10 tokens. Each round, a new profile was displayed to respondents.

	Player 2
Nationality	United Kingdom
Age	18
Party Affiliation	Labour Party (Labour)
Gender	Female
Religion	Muslim
Social Class	Middle Class

So put the number of tokens you wish to keep in the box labeled "Player 1." Put the tokens you wish to go to Player 2 in the box labeled "Player 2."

Player 1 (You)	0 Token(s)
Player 2	0 Token(s)
Total	0 Token(s)

Figure 9: Example of potential co-player profile.

8.3 Distribution of Y

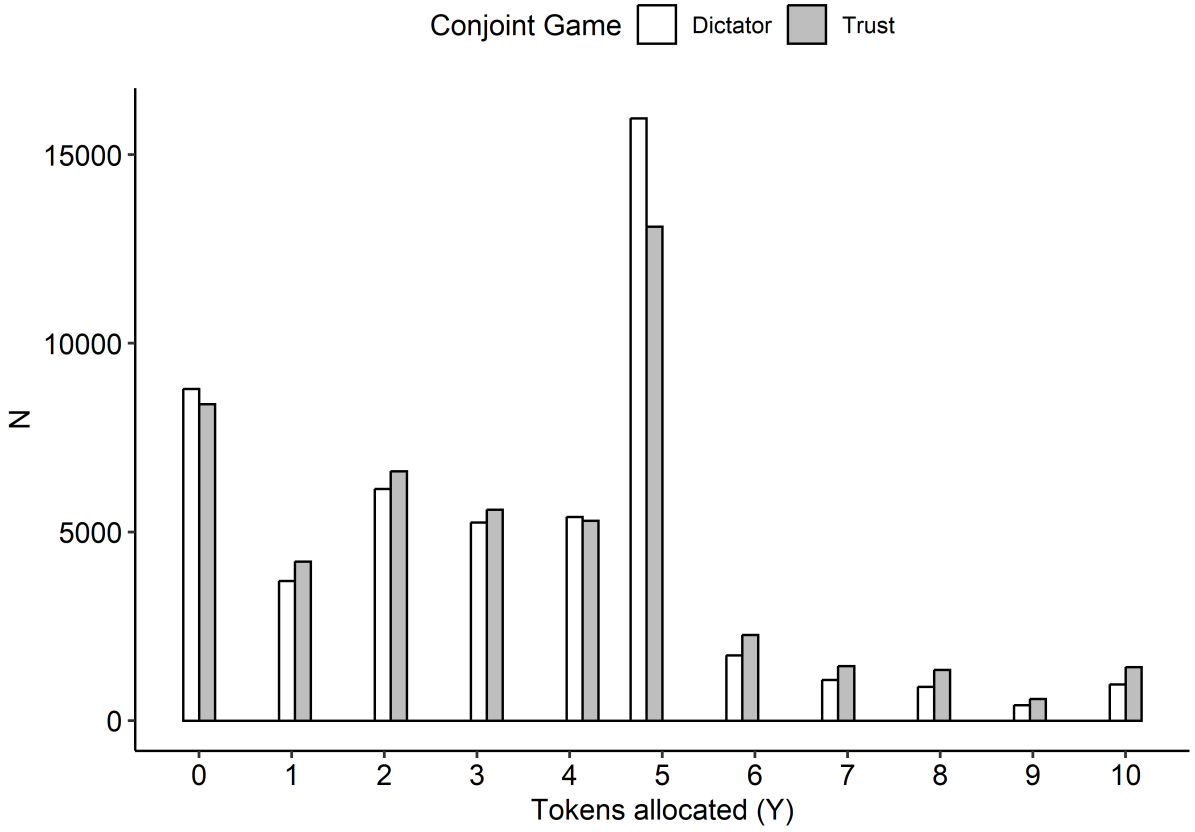

Figure 10: Distribution of token allocation (Y) by game. Dictator game: $Mean = 3.41$, $median = 4$, $SD = 2.35$. Trust game: $Mean = 3.48$, $median = 4$, $SD = 2.49$

8.4 Distribution of T and R

Table 3

der_partisan_type	None	Co	Out
0	5123	1948	18465
1	14898	30820	29366

8.5 Distribution of Covariates by T

Table 4

Variable	**0** N = 25,536[1]	**1** N = 75,084[1]
q_lrpos2_z	-0.08 (-0.45, 0.29)	-0.08 (-0.82, 0.66)
q_eupos2_z	0.08 (-0.68, 0.46)	0.08 (-0.68, 0.84)
q_econ_nativism_z	0.15 (-1.03, 0.74)	0.15 (-1.03, 0.74)
q_cult_nativism_z	0.07 (-1.03, 0.62)	0.07 (-1.03, 0.62)
q_satis_demo_country_z	0.30 (-0.81, 1.40)	0.30 (-0.81, 0.30)

Table 4

Variable	**0** N = 25,536[1]	**1** N = 75,084[1]
q_understand_nat_pol_z	0.12 (-0.65, 0.12)	0.12 (-0.65, 0.90)
q_understand_eu_pol_z	0.22 (-0.50, 0.22)	0.22 (-0.50, 0.94)
q_parties_harm_z	0.25 (-0.38, 0.88)	0.25 (-0.38, 0.88)
q_officials_talk_action_z	0.41 (-0.33, 1.15)	0.41 (-0.33, 1.15)
q_politics_good_evil_z	-0.14 (-0.76, 0.48)	-0.14 (-0.76, 0.48)
q_people_unaware_z	0.40 (-0.81, 1.00)	-0.20 (-0.81, 1.00)
q_leaders_educated_z	0.39 (-0.32, 1.10)	0.39 (-1.03, 1.10)
q_expert_decisions_z	0.16 (-0.49, 0.80)	0.16 (-0.49, 0.80)
q_listen_other_groups_z	0.18 (-0.74, 1.10)	0.18 (-0.74, 1.10)
q_democracy_compromise_z	-0.29 (-0.29, 0.57)	-0.29 (-0.29, 0.57)
q_interest_pol_country_z	0.06 (-0.60, 0.72)	0.06 (-0.60, 0.72)
q_interest_pol_eu_z	-0.28 (-0.96, 0.39)	0.39 (-0.28, 1.06)
q_eval_finance_household_z	0.01 (-0.99, 1.00)	0.01 (-0.99, 1.00)
q_eval_job_z	0.11 (-0.80, 1.03)	0.11 (-0.80, 0.11)
q_eval_econ_country_z	-0.14 (-1.03, 0.76)	-0.14 (-1.03, 0.76)
q_eval_econ_eur_z	0.07 (-0.94, 1.09)	0.07 (-0.94, 1.09)
q_risk_taking_z	0.11 (-0.56, 0.78)	0.11 (-0.56, 0.78)
q_future_discount_z	-0.17 (-0.84, 0.50)	-0.17 (-0.84, 0.50)
q_edu_z	-0.08 (-0.73, 0.57)	-0.08 (-0.73, 0.57)
q_age_z	-0.11 (-0.92, 0.64)	0.09 (-0.79, 0.91)
q_religion_en		
catholic	8,606 (34%)	27,379 (36%)
no religion	9,462 (37%)	25,678 (34%)
protstnt	1,881 (7.4%)	7,570 (10%)
other religion	5,470 (21%)	13,791 (18%)
muslim	114 (0.4%)	664 (0.9%)
q_perc_class		
Working class	5,307 (22%)	15,198 (21%)
Lower middle class	4,866 (20%)	13,924 (19%)
Middle class	12,060 (49%)	35,019 (48%)
Upper middle class	2,055 (8.4%)	7,622 (10%)
Upper class	191 (0.8%)	1,310 (1.8%)
q_rural_urban		
Rural area or village	5,819 (23%)	17,093 (23%)
Small or middle sized town	9,069 (36%)	28,418 (38%)
Large town	10,573 (42%)	29,298 (39%)
q_gender		
Male	10,939 (43%)	39,456 (53%)
Female	14,565 (57%)	35,514 (47%)
Other	32 (0.1%)	114 (0.2%)

[1] Median (Q1, Q3); n (%)

8.6 Robustness