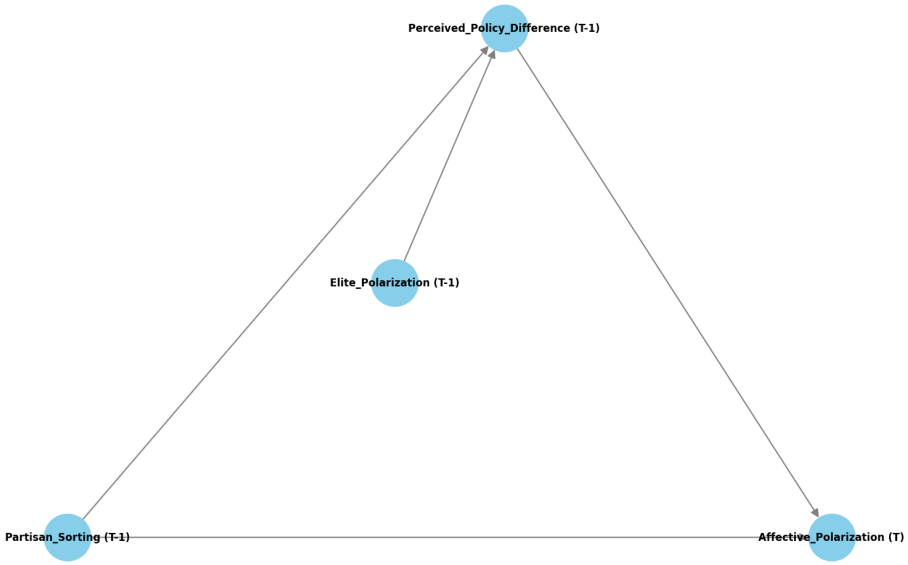
Appendix

A1 Excluded Variables

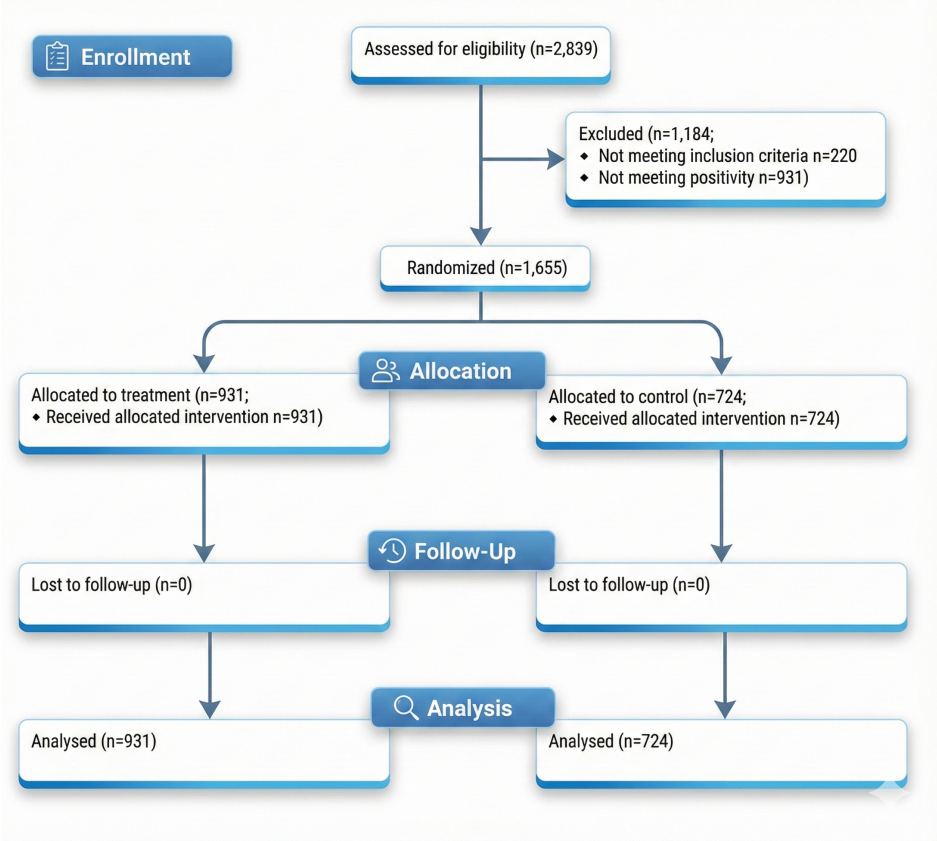
Figure 1: DAG of Excluded Variables



In this case, perceived policy difference works as a mediator, which does not require adjustment when focusing on the direct causal relationship of partisan sorting on affective polarization. In extension, elite polarization does not carry any direct causal relationship with both the treatment and outcome variable. Thus both variables are excluded from the analysis.

B1 CONSORT Chart

Figure 2: CONSORT-like Chart for the Target Trial



B2 SMD

Table 1: Standardized Mean Differences Before and After Adjustment

Variable	Diff. Unadjusted	Diff. Adjusted
pol_interest	0.2323	-0.0629
participation	0.2355	0.0081
social_media_usage	0.0610	0.0377
Campaign_Exposure	0.0877	-0.0118
age.group.4.18–29	-0.0098	0.0081
age.group.4.30–44	0.0122	0.0124
age.group.4.45–59	-0.0109	-0.0060
age.group.4.60+	0.0086	-0.0189
race.Black	0.0213	0.0212
race.Hispanic	0.0049	0.0122
race.Other	-0.0138	-0.0021
race.White	-0.0124	0.0005
region.Midwest	-0.0010	0.0001
region.Northeast	-0.0049	-0.0059
region.South	0.0110	0.0116
region.West	-0.0051	-0.0057

B3 Augmented Inverse Probability Weighting (AIPW)

To complement the methodological description in the main text, this appendix provides the formal definition of the stabilized Augmented Inverse Probability Weighting (AIPW) estimator used to compute the Average Treatment Effect (ATE). The estimator combines stabilized inverse probability weights with an outcome regression model, yielding a doubly robust estimatorconsistent if either the propensity score model or the outcome model is correctly specified.

Stabilized Inverse Probability Weights. Let $A_i \in \{0, 1\}$ denote treatment status, X_i the confounders, and $\hat{e}(X_i) = \Pr(A_i = 1 \mid X_i)$ the estimated propensity score. The stabilized inverse probability weights are defined as:

$$SW_i = \begin{cases} \dfrac{\hat{\pi}}{\hat{e}(X_i)} & \text{if } A_i = 1, \\ \dfrac{1-\hat{\pi}}{1-\hat{e}(X_i)} & \text{if } A_i = 0, \end{cases} \qquad \hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

Stabilization anchors individual weights to the marginal probability of treatment, thereby reducing the influence of extreme propensity score values and improving finite-sample behavior.

Stabilized AIPW Estimator. Let $\hat{\mu}(A_i, X_i) = \hat{E}(Y_i \mid A_i, X_i)$ denote the outcome regression model. The stabilized AIPW estimator of the ATE is:

$$\widehat{ATE}^{stab}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \left[\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i) + SW_i \left(A_i \left(Y_i - \hat{\mu}(1, X_i)\right) - (1 - A_i)\left(Y_i - \hat{\mu}(0, X_i)\right)\right)\right].$$

This estimator retains the double robustness property: consistency is achieved if either the propensity score model or the outcome regression model is correctly specified (Chernozhukov et

al., 2018). It is worth noting that while standard AIPW estimators typically utilize unstabilized inverse probability weights to ensure asymptotic double robustness (Chernozhukov et al., 2018), I employ stabilized weights (SW_i) in the correction term. This modification is an intentional choice to trade off a slight potential increase in bias for improved finite-sample stability, specifically by mitigating the influence of extreme propensity scores.

B4 Bootstrap Distribution for Short-term and Long-term effect (AIPW)

Figure 3: Short-term

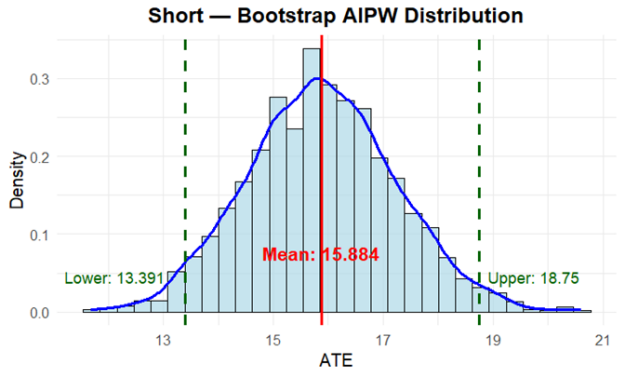
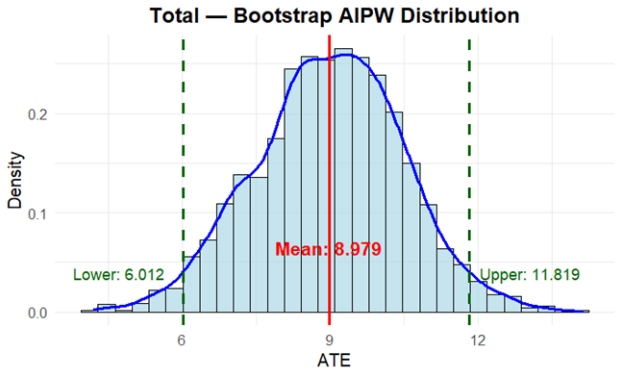


Figure 4: Long-term



B5 Bootstrap Distribution for Short-term and Long-term effect (DML)

Figure 5: Short-term

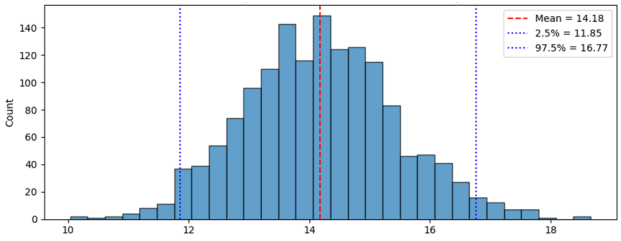
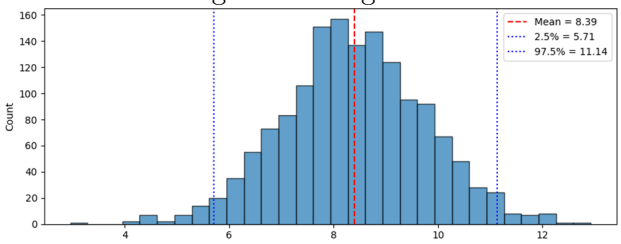


Figure 6: Long-term



B6 DML refuter test results

To assess the robustness of the causal estimates, I implemented five refutation strategies following the procedures in `DoWhy`.

Random Common Cause Refuter. This refuter injects a synthetic confounder $U \sim \text{Uniform}(0, 1)$ into the adjustment set. If including this irrelevant variable substantially changes the estimated ATE, the identification strategy may be overly sensitive to random noise, indicating potential residual confounding.

Random Outcome (Negative Outcome) Refuter. This refuter replaces the true outcome with an independent random variable and re-estimates the model. A well-specified causal model should return an ATE close to zero. Any non-negligible effect would suggest that the estimator is capturing spurious relationships rather than genuine causal effects.

Placebo Treatment Refuter. This procedure replaces the true treatment with a randomly generated variable that is independent of both confounders and outcomes. A correctly specified model should again produce an ATE near zero; a sizeable estimated effect would indicate structural misspecification or unaddressed confounding.

Placebo Permutation Refuter. Here, the original treatment assignment is randomly permuted, destroying any true association while preserving the empirical distribution. If the permuted treatment still leads to a non-zero ATE, the model may be fitting artifacts of the data rather than meaningful causal structure.

Subset Refuter. This refuter repeatedly recomputes the ATE using random subsamples (e.g., 50% of the observations). Large fluctuations across subsamples would imply that the results depend heavily on a small subset of influential units rather than reflecting a stable, population-level effect.

All refutation procedures were repeated 50 times for reliability, except the random-outcome refuter, which was repeated 500 times to obtain stable empirical null intervals.

Table 2: Refuter Results for Short-Term Effect Model (Original Estimate: 8.85)

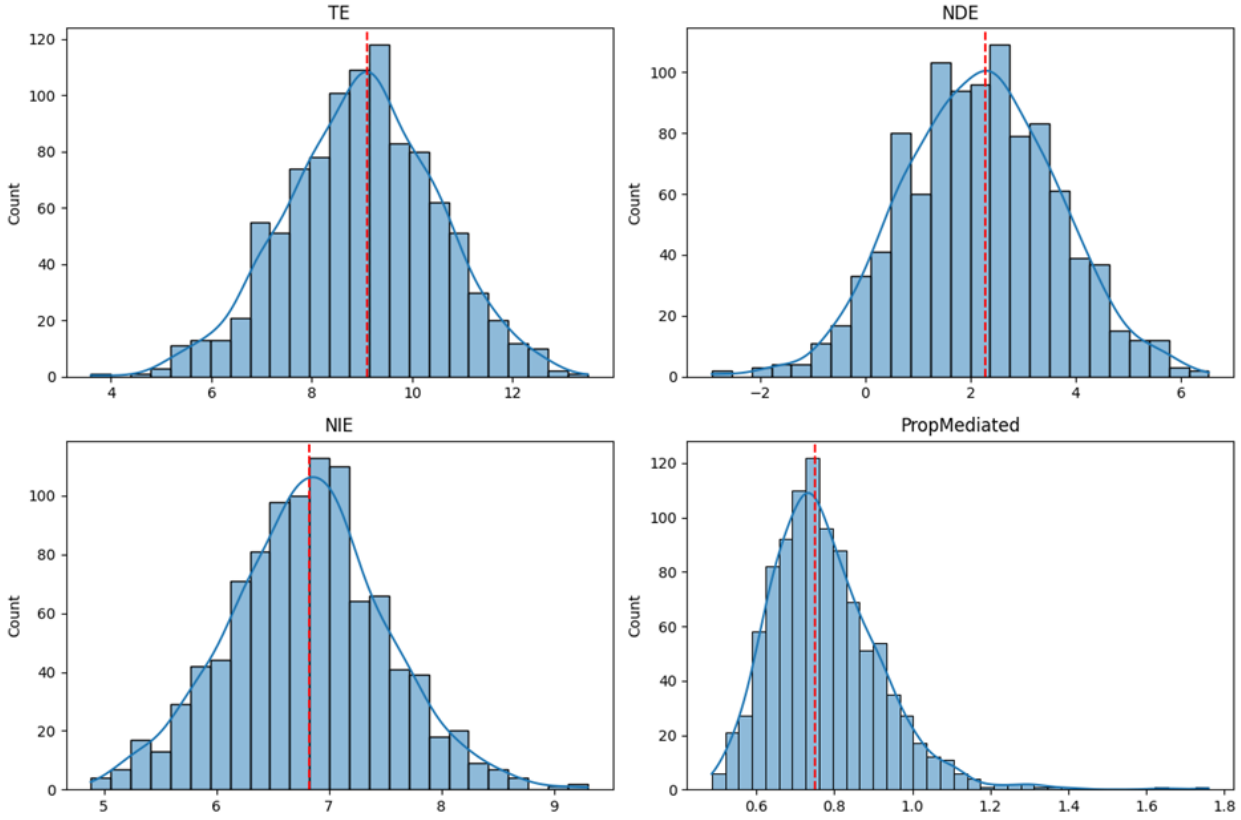
Refuter	**New Effect**	**p-value / CI**
Random common cause	8.77	0.82
Placebo treatment	-0.03	0.99
Placebo permute	0.07	0.99
Subset (0.5)	8.73	0.84
Random outcome	-0.00	[-0.08, 0.08]

Table 3: Refuter Results for Long-Term Effect Model (Original Estimate: 15.06)

Refuter	New Effect	p-value / CI
Random common cause	15.38	0.16
Placebo treatment	0.09	0.90
Placebo permute	-0.13	0.93
Subset (0.5)	15.25	0.68
Random outcome	-0.00	[-0.04, 0.07]

B7 Sequential Mediation Analysis Bootstrap Distribution

Figure 7: Bootstrap Distribution



B8 Controlled Direct Effect of Partisan sorting in 2016 on Affective Polarization 2020

To verify the statistical insignificance of the direct effect, I additionally estimated the Controlled Direct Effect (CDE) using a cross-fitted Gradient Boosting Regressor (nonparametric regression). The mediator (affective polarization in 2016) was fixed at its sample mean. The estimated CDE was 3.26, with a 95% bootstrap confidence interval of [0.03, 6.55] based on 1,000 replications. Because the interval includes zero, the direct effect is not statistically significant, confirming that most of the total effect of partisan sorting on later affective polarization operates through the mediator rather than directly.