

Leveraging Target Trial Emulation in Political Science: Assessing the Causal Effect of Partisan Sorting on Affective Polarization

Kang, Ji Hun

Graduate Student (Master's Program) at Dongguk University, Seoul

Abstract

This paper investigates the causal impact of partisan sorting on affective polarization among U.S. voters. Prior studies explored factors such as policy disagreements, elite polarization, partisan identity, and media exposure, but have rarely isolated the distinct influence of sorting itself. Consequently, the independent contribution of partisan sorting remains unclear. To address this gap, this paper utilizes the Target Trial Emulation framework, a methodological approach designed to approximate randomized controlled trials using observational data. Leveraging panel data from the American National Election Studies (ANES) covering 2016–2020, this paper emulates a hypothetical trial where individuals are assigned to partisan sorting in 2016 and their affective polarization is observed in 2020. Augmented inverse probability weighting and debiased machine learning is used to explore the causal impact, alongside comprehensive sensitivity analyses. The results indicate that partisan sorting causally increases affective polarization by approximately 2.8 points on a 0-100 scale over four years, when all confounding are adjusted. These findings highlight sorting's independent role in intensifying inter-party hostility, supporting predictions from social identity theory. However, sensitivity analyses reveal that the model's estimates remain sensitive to potential unobserved confounding, highlighting an important limitation.

I. Introduction

Affective polarization, characterized by escalating hostility, mistrust, and animosity toward opposing political parties, poses an increasingly critical challenge to democratic societies (Iyengar et al., 2019). Over recent decades, affective polarization in the United States (the U.S) has intensified significantly, mostly observed in deeply divided ethnic or religious groups (Mason and Wronski, 2018). Scholars have sought to uncover the root causes of affective polarization, emphasizing factors such as policy disagreements, elite polarization, partisan identity, and media dynamics (Druckman, Green and Iyengar, 2023; Huddy, Mason and Aaroe, 2015; Iyengar and Westwood, 2015; Levendusky, 2013; Mason, 2015; Rogowski and Sutherland, 2016).

Despite these efforts, previous literature has struggled to disentangle these interconnected mechanisms within the causes of affective polarization (Dias and Lelkes, 2022). Notably, the precise causal effect of partisan sorting, defined as the alignment of ideological orientation with partisan identification, remains inadequately explored. Most research examines sorting alongside other factors, complicating efforts to isolate its distinct contribution to affective polarization. Additionally, studies on the mechanisms of partisan identity have paid less attention to the time-varying nature of social identity.

This study addresses these gaps by explicitly isolating and assessing the causal impact of partisan sorting on affective polarization. Utilizing a Target Trial Emulation (TTE) framework, an innovative methodological approach designed to approximate randomized controlled trial (RCT) conditions within observational data (Hernán and Robins, 2016), this study leverages panel data from the American National Election Studies (ANES) from 2016 to 2020. The analysis emulates a randomized experiment in which participants are hypothetically assigned to partisan sorting. To estimate the causal effect, I employ Augmented Inverse Probability Weighting (AIPW) to assess the average treatment effect and further applies Debiased Machine Learning (DML) to mitigate the effect of unobserved confounding and possible biases, thereby complementing the AIPW results.

This study makes significant theoretical and methodological contributions. Theoretically, by isolating the effect of partisan sorting independently from other influencing factors, social identity theory's claim regarding the central role of identity alignment in intensifying political animosity is substantiated. Methodologically, by applying the TTE framework to observational data, researchers can draw robust causal inferences and open new avenues for political science inquiry.

The remainder of this paper is structured as follows. The first section reviews existing literature, synthesizing previous findings into a coherent causal framework by utilizing Directed Acyclic Graphs (DAGs) (Pearl and Mackenzie, 2019). The second section explains the need for a new approach regarding existing literature and introduces the TTE framework and its components. The third section details the emulating process of a target trial and causal analysis methods. The final section presents empirical findings and discusses their implications.

II. Literature Review

1. Causes of Affective Polarization in the United States

Although affective polarization has been widely studied, this review aims specifically to integrate existing findings into a coherent causal framework and highlight how prior work leaves unresolved the direct, “pure” effect of partisan sorting on affective polarization. In other words, after summarizing the main drivers identified in the literature, I will show that these mechanisms are deeply intertwined. By making those interconnections explicit through constructing DAGs of previous studies, I will set the stage for constructing a DAG that ultimately justifies using Target Trial Emulation (TTE) to isolate the causal impact of partisan sorting.

Affective polarization refers to the increasing animosity, mistrust, and hostility that individuals feel toward members of the opposing political party, above and beyond ideological disagreements. This

phenomenon encompasses emotional reactions and social identity-based divisions, which have significantly intensified in recent decades, reaching levels comparable to entrenched conflicts seen in societies deeply divided by ethnicity or religion (Iyengar and Westwood, 2015; Iyengar et al., 2019). Unlike traditional ideological polarization that emphasizes policy or issue-based divergence (Abramowitz and Saunders, 2008), affective polarization focuses on the emotional and identity-driven aspects of political rivalry, manifesting in behaviors such as reduced inter-party social interactions, increased intergroup hostility, and stronger negative stereotypes (Iyengar et al., 2019).

Scholars have offered various explanations for the rise of affective polarization, with research primarily coalescing around three broad categories: policy differences, partisan identity, and other contextual factors such as media exposure, campaign rhetoric, and social media dynamics. Despite considerable empirical investigation, the literature remains divided regarding the relative importance and causal dynamics of these factors.

One prominent explanation for affective polarization emphasizes substantive policy differences between political parties. Advocates of this perspective argue that heightened ideological conflict and significant disagreements over critical issues drive emotional polarization among partisans (Bougher, 2017; Orr and Huber, 2020; Rogowski and Sutherland, 2016; Webster and Abramowitz, 2017). In this view, affective polarization is not just blind group bias. It is partly a rational, or at least issue-based response to substantive disagreement. These scholars argue that when one learns about a person's partisan affiliation, they may update their beliefs about that person's policy positions, shifting their views of counterpart in response to partisan information even if they only care about policy and not about partisan identity itself. In other words, if the other party's stance on important issues is anathema to someone's values, that person may develop genuine dislike or moral distrust toward supporters of that stance through moral conviction, perceived threat, and ideological stereotyping (Cahmbers, Schlenker and Collisson, 2013; Graham, Nosek and Haidt, 2012).

Lelkes (2018) demonstrated experimentally that ideology extremity and policy differences often exert stronger effects on inter-personal evaluations than partisan labels alone. Similarly, Orr, Fowler and Huber (2023) found that policy alignment significantly affects how partisans evaluate each other, with substantive disagreements often overshadowing partisan identity when the two are in conflict. On the other hand, scholars also pointed out that interplay between elites and the mass public acts as another driver of affective polarization (Druckman, Green and Iyengar, 2023). They suggest that as elite Democrats and Republicans become more ideologically extreme and divergent, they send clearer signals to voters that the other party represents a starkly different, even threatening, out-group. Rogowski and Sutherland (2016) manipulated the policy positions taken by elites in their experiments and found that respondents' evaluations of candidates are responsive to elite ideological polarization. Skytte (2021) also showed that elite-level issue polarization increased affective polarization. Webster and Abramowitz's (2017) empirical findings support this by showing that the public's social welfare policy preferences are strongly related to evaluations of elites and the parties.

Overall, these scholars suggest that gradual diffusion of issue positions within mass public and ideological extremity from elites to the mass public contributes to increased out-party hostility. However, Dias and Lelkes (2022) nuanced these findings by highlighting that policy-driven affective polarization is particularly pronounced when the policy positions are clearly partisan-branded, suggesting an intertwined relationship between policy disagreements and partisan identity. By experimentally separating identity from issue positions, they found that knowing someone's party alone causes a big jump in affective bias. In other words, policy preferences and views of ideological extremity do influence hostility toward the opposing party, but it is partisan identity that ultimately underlies this effect.

Then what makes partisan identity important? The partisan identity and associated psychological mechanisms have been highlighted by several scholars as core drivers of affective polarization (Huddy, Mason and Aaroe, 2015; Iyengar, Sood and Lelkes 2012; Iyengar et al., 2019; Mason, 2015; 2016; 2018). These scholars root their explanations in social identity theory (Tajfel and Turner, 1979). Social

identity theory argues that strengthening individuals' identities will engender an affectively polarized public that reflects more positive evaluations of one's own side and more negative evaluations of the opposing side.

Specifically, social identity theory rests upon three processes; social categorization, social identification, and social comparison (Tajfel and Turner, 1979). People categorize themselves and others into social groups to organize and simplify the world around us. Then, they embrace the identity of the group to which they belong and attach their self-esteem to group membership. Finally, to protect self-esteem, they engage in social comparisons in such a way that in-group compares favorably to out-groups. Acting in concert, these psychological processes tend to create an in-group bias or prejudice in individuals' social attitudes and behaviors that reflect a blend of in-group favoritism and out-group hostility (Rudolph and Hetherington, 2021).

This perspective argues that political parties can function similarly as social groups, eliciting strong in-group favoritism and out-group derogation independent of substantive policy differences (Iyengar, Sood and Lelkes, 2012). Experimental research by Iyengar and Westwood (2015) provided compelling evidence showing that partisan identity alone significantly biases trust and resource allocation behaviors, sometimes surpassing even racial biases. Additionally, Huddy, Mason, and Aaroe (2015) have found that when the status of a party is threatened, the strongest partisans react with extreme levels of anger. They predict that when an individual is strongly attached to a partisan identity, the victory of their supporting party generates positive emotional reactions, while a loss generates negative, particularly angry, emotional reactions.

The alignment of social identities such as ideology, race, and religion to partisanship also strengthens in-group bias and out-group hostility (Levendusky, 2009; Mason, 2015; 2016; 2018; Mason and Wronski, 2018). Levendusky (2009) finds that in the last 50 years, the percentage of sorted partisans, i.e., partisans who identify with the party that most closely reflects their ideology, has steadily increased. Due to this increase, the parties have become more homogeneous internally and more distinct from each other in ideology and in demographics. This sorting makes it easier for partisans to view the opposing party as different from their own and perhaps as threatening. Building upon this, Mason (2015; 2016) argues that partisan identities have contributed to the rise of negative attitudes between Democrats and Republicans in the mass public. Findings suggest that individuals who are consistent partisans exhibit much higher out-party hostility than those with cross-cutting identities. Furthermore, she showed that this pattern of reinforcing cleavages turns partisanship into a "mega identity", reducing cross-cutting identities (Mason, 2018). Overall, as social sorting increased in the American electorate, the portion of the cooler heads inspired by cross-cutting identities in American electorate decreased, resulting in less cross-pressure to treat the other side with respect and tolerance (Mason, 2016).

However, recent work by Lelkes (2018) suggests that sorting alone cannot fully explain the rise in affective polarization. He provides evidence that affective polarization increased almost equally among the unsorted as among the sorted, suggesting that the overall societal trend of affective polarization cannot be attributed solely to the increase in sorted individuals. Other findings provide that partisan sorting is only weakly related to affective polarization except among the most politically knowledgeable citizens, suggesting other forces are pushing even apolitical or moderate individuals into dislike. Importantly, Lelkes (2018) argues that reciprocal effect, that affective polarization itself can lead people to become more sorted.

Beyond policy and identity, additional factors significantly contribute to affective polarization. Media exposure, particularly partisan media consumption, is recognized as a potent amplifier of polarization (Huddy and Yair, 2021; Iyengar, Sood and Lelkes, 2012; Lelkes, Sood and Iyengar, 2017; Levendusky, 2013; Levendusky and Malhotra, 2016). These scholars address that partisans are now immersed in congenial information environments. Partisan media sources frequently reinforce biased views, heighten negative stereotypes, and perpetuate negative portrayals of opposing partisans and thus effectively deepen existing divisions (Garrett et al., 2014). Additionally, modern political campaigns, characterized by increasingly negative and confrontational rhetoric further exacerbate partisan

animosity (Iyengar, Sood and Lelkes, 2012). Empirical evidence supports the claim that toxic elite messages can cause partisan animosity. For example, partisans with access to broadband internet as well as those more heavily exposed to campaign messaging have significantly colder feelings toward their rival party (Lelkes, Sood and Iyengar, 2017).

Social media usage can also intensify affective polarization, as platforms facilitate the rapid dissemination of polarizing content, foster echo chambers, and magnify emotional polarization through algorithmically driven content exposure and peer reinforcement of partisan biases (Alcott et al., 2020; Kubin and von Sikorski, 2021; Lorenz-Spreen et al., 2023). Settle (2018) conducted multiple surveys and experiments and found that social media platforms like Facebook can exacerbate affective polarization, even among users who are not interested in politics. Alcott et al. (2020)'s experiment showed that deactivating participants from social media reduced out-group hostility among users who got news from social media fairly often or often. Other research also shows that talking negatively about out-party or expressing moral outrage increases attention on social media (Brady et al., 2017; Rathje et al., 2021). However, Beam, Hutchens and Hmielowski (2018) suggests that exposure to differing points of view, facilitated by social media, may curb polarization.

In summary, the literature reveals multiple, interwoven causes of affective polarization in the United States. Policy preferences and ideological disagreements provide substantive grievances that fuel animosity. Elite polarization sharpens conflict by sending extreme cues. Partisan identity transforms political differences into social schisms through in-group bias out-group hostility while social sorting aligns multiple social cleavage lines along party divides, magnifying “us vs. them” perceptions. Media and campaign rhetoric then amplify these biases, reinforcing negative stereotypes and heightening emotional responses. Yet, although scholars have identified these factors, they have often studied them separately or in loosely related clusters rather than as an integrated system.

Critically, the specific causal pathways among these mechanisms remain underspecified. For instance, do policy disagreements strengthen partisan identity, or does identity heighten sensitivity to policy differences? Does elite signaling operate primarily by increasing ideological sorting, or by directly stoking emotional reactions irrespective of sorting? Existing studies such as Orr, Fowler, and Huber (2023) and Dias and Lelkes (2022) adopt different identifying assumptions—one holding identity constant to isolate a “substance” effect, the other orthogonally varying identity and policy to disentangle both effects—but no work to date has systematically mapped all these relationships in a unified causal graph to identify and isolate the causal impact of variables in focus. For these reasons, despite evidence that partisan sorting is a key correlation of affective polarization (Levendusky, 2009; Mason, 2015), literature lacks research that isolates the direct effect of sorting itself among other factors that influence the mechanism. Lelkes (2018) even suggests a reciprocal relationship in which affective polarization can itself drive greater sorting. In other words, sorting may be both the cause and consequence of affective bias. This causal relationship has not been fully addressed.

To fill this gap, I aim to address this limitation by systematically employing DAGs to elucidate causal relationships among the mechanisms identified in prior studies to isolate the direct causal impact of partisan sorting on affective polarization. A DAG is a graphical representation of causal assumptions in which each node denotes a variable, and each directed edge (arrow) denotes a direct causal effect from one variable to another, with the key constraint that no series of arrows can form a cycle (Pearl, 1995). By encoding causal structure explicitly, DAGs allow researchers to identify which variables must be measured and adjusted to block “back-door” paths via the concept of d-separation and thus obtain unbiased effect estimates (Digitale, Martin and Glymour, 2022). A DAG makes conditional independence transparent, guiding the selection of appropriate confounding variables and clarifying assumptions about confounding, mediation and collider bias (See Pearl and Mackenzie, 2019 for details).

By integrating existing theoretical and empirical findings on policy preferences, partisan identity, sorting, elite polarization, and media exposure into a comprehensive DAG framework, this study seeks to provide a clearer, unified depiction of how these factors collectively contribute to affective

polarization. Ultimately, this approach enables holistic understanding of the causal pathways underpinning affective polarization.

2. Directed Acyclic Graphs

Figure 1. Policy Difference DAG

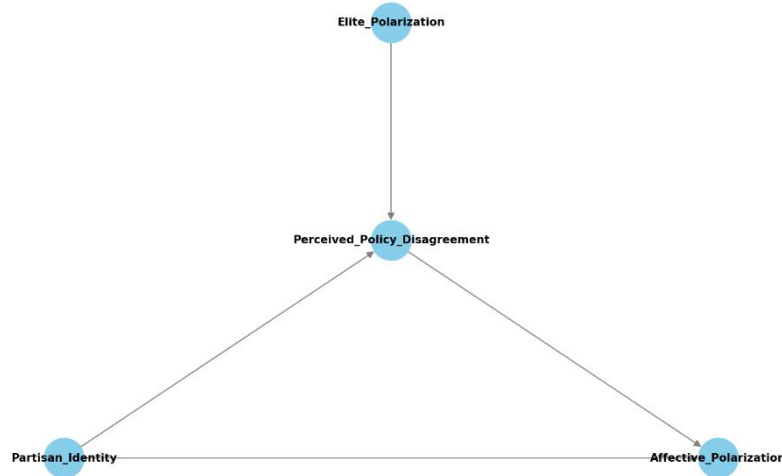


Figure 1 represents the causal mechanism of policy difference and affective polarization. Elite polarization signals individuals perceived policy disagreement or differences between in-group and out-group, thus resulting in affective polarization. Additionally, as Dias and Lelkes (2022) notes, policy preferences drive affective polarization, in large part, by signaling partisan identity.

Figure 2. Partisan Identity

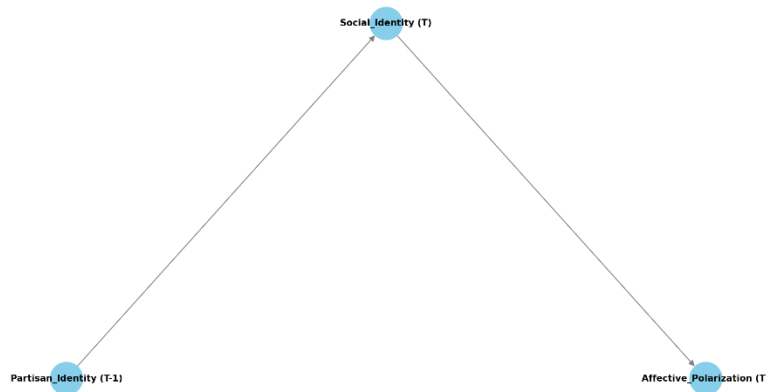


Figure 2 shows causal mechanisms of partisan identity and affective polarization. The crucial point is that it takes time to form what previous studies have defined as social identity. A wealth of empirical evidence indicates that social identities do not emerge instantaneously but instead solidify only after repeated experiences and prolonged exposure to group-related cues. Several studies demonstrated that individuals' sense of belonging to a political party often requires participation in campaign events, community meetings, and other sustained political activities before a stable social identity takes hold (Sears and Funk, 1999; Yates and Youniss, 1998). Similarly, Green, Palmquist, and Schickler (2002) show that partisan attachments develop gradually over multiple election cycles, as voters internalize

party norms and narratives over time.

Figure 3. Partisan Sorting

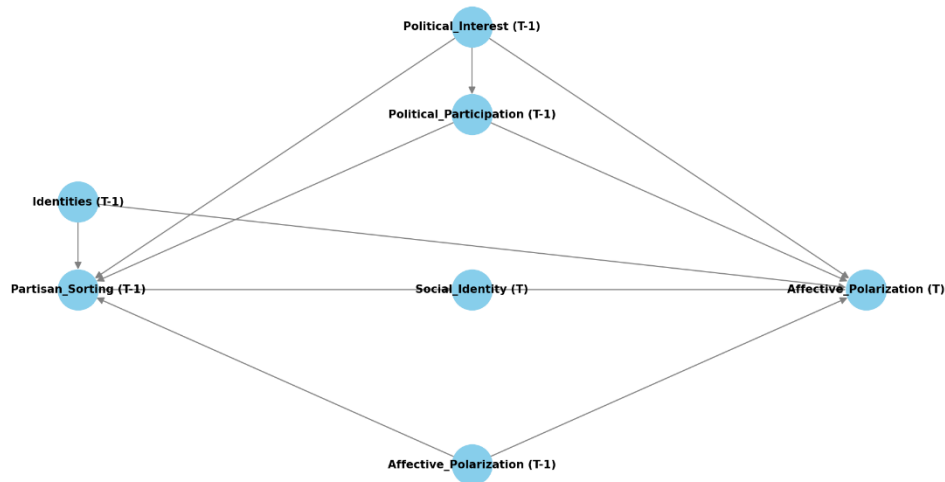
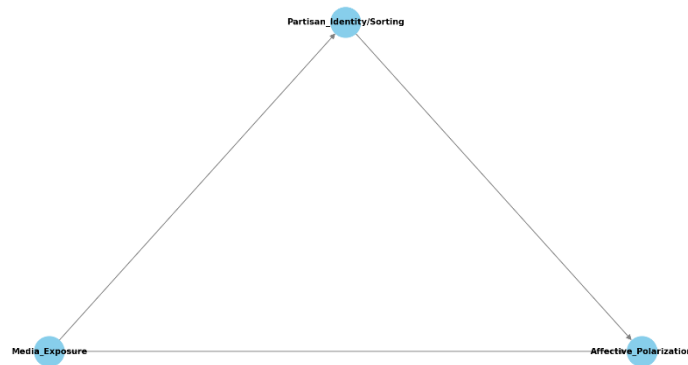


Figure 3 depicts the causal mechanism that reflects the time to form social identity. Different identities align with partisan identity to form the state of partisan sorting at t-1. Political interest, as well as political participation both affects partisan sorting at t-1, and affective polarization at t. Most importantly, as Lelkes (2018) emphasizes, affective polarization at t-1 itself also serves as a confounder. Lastly, figure 4 represents the causal relationship between media exposure(campaign) and social media usage and affective polarization.

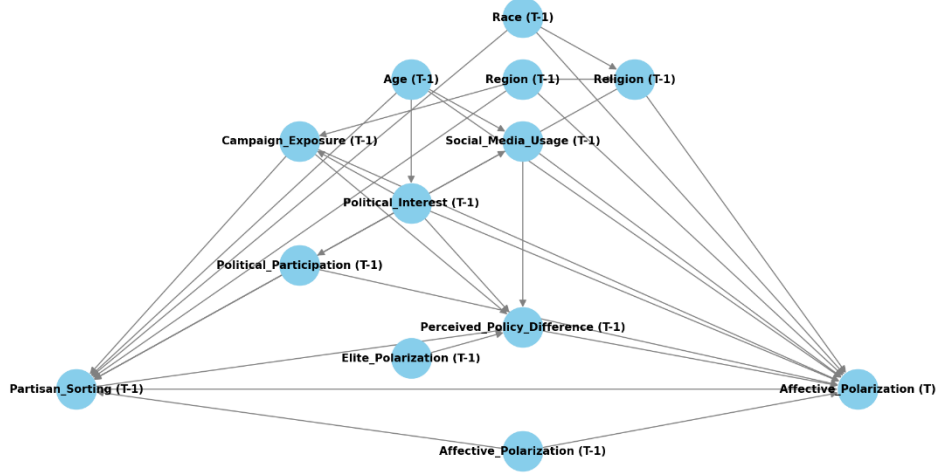
Figure 4. Media Exposure



Through examining the DAGs of previous literature, it is certain that existing explanations of affective polarization are intertwined. Furthermore, previous research has undermined the time it takes for social identity to solidify. To clarify the causal mechanisms, the intertwined relationships need to be combined.

This study tries to capture the effect of partisan sorting which only captures the alignment of ideological identity and partisan identity. It assumes that partisan sorting includes both causal effects of party identification and ideological identification. To not eliminate the possibility of social sorting (Mason, 2018), it will also include race, region and religion as a variable. Figure 5 represents the combined causal mechanisms of previous literatures to capture the effect of partisan-ideological sorting on affective polarization.

Figure 5. Partisan Sorting on Affective Polarization



As mentioned above, the impacts of partisan sorting on affective polarization are unlikely to appear instantaneously, necessitating an extended observation period to capture the full emergence and reinforcement of social identity. Ideally, a RCT designed to evaluate these long-term causal effects would randomly assign individuals to treatment and control conditions at a clearly defined baseline ("time-zero") and systematically track these same individuals over an extended follow-up period (Hernán and Robins, 2016). Such a design ensures balance between groups on both measured and unmeasured confounders through randomization and can effectively capture evolving or delayed causal effects that unfold gradually. However, implementing a long-term follow-up RCT is challenging due to logistical constraints, participant attrition, and significant resource demands. Moreover, partisan sorting describes a "state" of alignment between one's party identification and ideological or other social identifications—making random assignment to such a state practically infeasible. Thus, experimental studies that try to study differences between affective polarization of two groups when partisan sorting is given (treated) have not been conducted.

Due to these challenges, researchers frequently resort to observational methods that attempt to approximate RCT conditions, such as matching frameworks (Mason, 2015), setting partisan sorting as if it is a result of a treatment in observational data. Yet, these observational methods remain constrained by significant methodological shortcomings (Hernán and Robins, 2020; Imbens and Rubin, 2015). Specifically, these methods inherently depend on balancing or controlling only observable characteristics, leaving biases from unmeasured or unobservable confounders unresolved (Fu, 2023). Furthermore, observational strategies frequently struggle with accurately managing time-varying confounding, as confounders influencing both treatment and outcomes may evolve dynamically after baseline measurement (Hernán, 2018). Lastly, many observational studies lack a clearly defined baseline or "time-zero," creating ambiguity between pre-existing baseline differences in polarization and causal effects directly induced by partisan sorting itself (Hernán and Robins, 2016).

To overcome these methodological hurdles, this study introduces TTE, a methodological framework explicitly designed to replicate the rigorous structure of RCT within observational data, thereby enabling a more robust and rigorous examination of the long-term causal effects of partisan sorting on affective polarization.

III. Introducing Target Trial Emulation

Prior research on affective polarization relied on observational survey data and survey-based experimental methods, which focused on the mechanisms that explain how individuals become affectively polarized. These approaches have greatly enriched our understanding of the phenomena but

inherently face methodological limitations, especially concerning the establishment of causal relationships. Experimental designs often grapple with constraints related to feasibility, ethical concerns, or generalizability, while observational studies commonly confront biases stemming from confounding variables and flawed temporal frameworks. Due to this, few - if any - studies have attempted to isolate the causal effect of being sorted itself, independent of the downstream pathways. In other words, existing work has not assessed what affective consequences arise when partisan sorting is treated as a given condition, rather than a process.

In addressing these methodological challenges, this study introduces TTE, an advanced causal inference framework developed to systematically enhance observational study designs by explicitly emulating RCTs. First introduced (formally) by Hernán and Robins (2016), TTE framework has been widely implemented across a diverse array of domains such as medical, epidemiology, and policy research where traditional RCTs are infeasible due to ethical, logistical, or financial constraints. For example, Antoine et al. (2024) employed TTE to assess the real-world effectiveness of 8 major metastatic breast cancer treatments by emulating pivotal RCTs using data from the French Epidemiological Strategy and Medical Economics Metastatic Breast Cancer Cohort (ESME-MBC), which included over 32,000 patients. This approach was necessary because traditional RCTs often exclude patients encountered in real-world settings and cannot capture long-term, post-approval treatment dynamics.

Similarly, Kutcher et al. (2021) reviewed the application of TTE in cardiovascular medicine, including scenarios such as antiplatelet therapies for myocardial infarction, where logistical and ethical barriers to randomization exist. They emphasized how emulated trials based on observational cohorts can yield causal estimates comparable to RCTs when key assumptions (e.g., time-zero alignment, confounding control) are met. Beyond clinical medicine, Seewald et al. (2024) extended the framework to policy evaluation—illustrating its use in assessing health policy effects such as medical cannabis legalization or state-level interventions. Since policy implementation is never randomly assigned, observational studies emulating randomized designs are the only viable method to estimate causal impacts of such interventions.

TTE operates by first clearly articulating the causal question through the specification of a hypothetical randomized trial (target trial) that, if conducted, would ideally answer the question at hand. This hypothetical trial is comprehensively detailed in a protocol that specifies essential design elements, including eligibility criteria, treatment strategies, assignment methods, the precise timing of treatment initiation (time-zero) along follow-up period, outcome measures, causal contrasts, and a statistical analysis plan (Hernán and Robins, 2016). Specific explanations (and example from Hernán and Robins, 2016) are given in Table 1. The second step involves aligning the observational analysis to this protocol by selecting eligible individuals, assigning them to treatment strategies compatible with their observed data, initiating follow-up concurrently with treatment assignments to avoid immortal time bias (explained later), and applying statistical adjustments for confounding variables.

<Table 1> Protocol Elements

Eligibility Criteria	Defines which individuals are considered appropriate for inclusion in the trial, ensuring the selected population closely represents those who would ideally participate in the actual RCT if conducted <i>E.g. Postmenopausal women within 5 years of menopause between the years 2005 and 2010 and with no history of cancer and no use of hormone therapy in the past 2 years</i>
Treatment Strategy	A hypothetical treatment which investigator would apply if the RCT is conducted. Must specify interventions or exposures whose effects are being investigated <i>E.g. Refrain from taking hormone therapy during the follow-up. Initiate estrogen plus progestin hormone therapy at baseline and remain on it during the follow-up unless you are diagnosed with deep vein thrombosis, pulmonary embolism, myocardial infarction, or cancer</i>

Assignment Method	The approach for assigning eligible participants to different treatment strategies, ensuring comparability between groups. In observational data, this step involves assigning individuals based on their observed treatment history, adjusted statistically to mimic randomization <i>E.g. Participants will be randomly assigned to either strategy at baseline and will be aware of the strategy to which they have been assigned</i>
Time-zero and Follow-up period	Clearly defining the point at which eligibility is assessed, treatment assignment is finalized, and follow-up begins/ends. Proper alignment of these elements at time-zero prevents biases such as immortal time bias, where participants erroneously appear to be immune to outcomes due to misalignment in timing <i>E.g. Starts at randomization and ends at diagnosis of breast cancer, death, loss to follow-up, or 5 years after baseline, whichever occurs first</i>
Outcome Measures	Clearly specified primary and secondary outcomes at end of follow-up period that the trial aims to assess. These outcomes must be explicitly defined, measurable, and relevant to the causal question posed <i>E.g. Breast cancer diagnosed by an oncologist within 5 years of baseline</i>
Causal Contrast	Explicitly defined comparisons between treatment groups, specifying the exact estimands such as intention-to-treat effects or per-protocol effects, which reflect either the impact of assigned treatment or adherence to treatment protocols.
Statistical Analysis Plan	A detailed methodological framework outlining how the data will be analyzed, including methods for addressing confounding (e.g., propensity score matching, inverse probability weighting), handling missing data, and performing sensitivity analyses to assess the robustness of results. <i>E.g. Intention-to-treat effect estimated via comparison of 5-year cancer risks among individuals assigned to each treatment strategy. Per-protocol effect estimation requires adjustments for pre- and postbaseline prognostic factors associated with adherence to the strategies of interest. All analyses will be adjusted for pre- and postbaseline prognostic factors associated with loss to follow-up</i>

Defining the protocol for target trial before aligning it with the observational analysis is crucial as it prevents investigators from introducing biases, explained in Table 2. These biases can be prevented through specifying the protocol and aligning them with observational data. Immortal time bias can be prevented through precisely aligning eligible individuals, treatment assignment and start of follow-up at a single time-zero (Hernán et al., 2016). This ensures that only post time-zero events contribute to causal estimates, while blocking any affection that can manipulate the effect of intervention that happened before the start of target trial. Selection bias occurs when inclusion or continuation in the study depends on events or outcomes that occur after baseline. In other words, when information that was observed after the baseline affects the selection of eligible individuals, selection bias happens. It induces a non-random subset of the original population in the analysis, enrolling individuals that are systematically different to the target population. This can be prevented through clearly defining pre-specified eligibility criteria applied at time-zero, irrespective of later events, and avoiding adjustments of variables that were observed after time-zero.

Depletion of Susceptible Bias and Lead-time bias can be avoided through Time-zero specification (Hernán and Robins, 2016). By designing target trials which all participants receive intervention or treatment at baseline, it blocks the possibility of experiencing intervention or treatment before experiment begins. Specifying treatment strategy is important as it prevents Specification bias. If treatment strategy is vaguely defined, the selected observational results can have inconsistent exposure measurements. For example, suppose the intervention is defined simply as “exercise” without specifying type, intensity, or duration. In the observational data, participants may engage in vigorous jogging for 30 minutes, moderate cycling for 45 minutes, or light walking for 10 minutes. Lumping these heterogeneous activities together misclassifies exposure, as differences in outcomes may stem from intensity and duration rather than a uniform exercise protocol. By clearly defining detailed, protocol-based definitions of treatment strategies, investigators can avoid misclassification.

<Table 2> Common Biases in Observational Studies

Immortal Time Bias	Occurs when periods during which an outcome cannot occur (e.g., before treatment initiation) are incorrectly attributed to the treatment group, making the treatment appear artificially beneficial.
Selection Bias	Arises when eligibility or follow-up depends on post-baseline events or outcomes, leading to non-comparability between groups.
Depletion of Susceptible Bias	Happens when individuals most susceptible to an event experience it early, leaving a “healthier” subgroup over time
Lead-Time Bias	Often seen in screening studies, where earlier detection of intervention or treatment lengthens observed survival
Specification Bias	Results from vague or heterogeneous definitions of interventions.

While TTE effectively addresses these design-related biases, regarding the prerequisites of RCT, one must still fulfill three more assumptions: Consistency, Exchangeability and Positivity (Dahabreh et al., 2024). In the context of a TTE, the consistency assumption requires that definition and implementation of the treatment in the observational data correspond exactly to the version of the treatment one would administer in the hypothetical target trial. In other words, for each individual, the observed outcome under the treatment they actually (hypothetically) received must coincide with the potential outcome that would occur under the same treatment in the target trial. This entails that there must be no meaningful “multiple versions” of the intervention—differences in how, when, or at what intensity the treatment is delivered must be either absent or irrelevant to the outcome. Violations such as existence of different implementation versions can introduce bias within causal estimates by conflating disparate interventions into one indistinct category. Consistency assumption is related with specification bias mentioned above and can be prevented by clearly defining treatment strategy in the TTE protocol.

Exchangeability in TTE mirrors the randomization principle of an actual RCT. It requires the treated and untreated groups to be comparable when measured baseline confounders are conditioned(adjusted), such that any differences in outcomes can be attributed solely to the treatment itself rather than to pre-existing imbalances (Pearce and Vandenbroucke, 2023). In practice, achieving exchangeability requires measuring and adjusting for all relevant confounders, so that the distribution of both measured and unmeasured factors is as if treatment had been randomized. Exchangeability must be met because failure to account for an important confounder leads to residual bias, undermining the validity of the emulated trial’s causal effect estimate.

Positivity, also known as overlap assumption, demands that, for every combination of baseline confounder that is observed in the data, there is a non-zero probability of receiving each level of treatment under consideration. Specifically, if we stratify the sample by, for example, age, education, sex, and any other confounder in the target model, then within each stratum one must observe both individuals who experienced the treatment and individuals who did not. When positivity holds, we can meaningfully compare outcomes across treated and untreated groups everywhere in the confounder variable space. If, however, in some subgroups everyone is treated (or everyone is untreated), then we cannot estimate what would have happened under the alternative intervention in that subgroup because there will be no data to examine that alternative intervention, and any causal contrast there becomes undefined. This is usually done by checking the propensity score overlaps (Gomes et al., 2022).

Despite these limitations, TTE provides several advantages. It aligns observational research designs and data closely with formal counterfactual reasoning, bridging gaps between the internal validity of randomized experiments and the pragmatism of real-world data, yielding causal estimates that closely approximate what one would obtain in an idealized RCT. Moreover, by emulating a hypothetical RCT, TTE allows researchers to estimate the causal effects of interventions that would be impossible -or ethically, financially, or logistically prohibitive- to evaluate in a conventional RCT. By explicitly defining the “protocol” of the hypothetical trial and then applying appropriate statistical methods to observational data, TTE reconstructs the design features of an RCT without ever randomizing

participants. This means we can assess the long-term impact of exposures that cannot be manipulated, explore intervention effects over extended horizons, or study rare outcomes in large populations—scenarios in which a traditional trial would either be infeasible or outright impermissible.

In the context of affective polarization and partisan sorting, employing TTE allows for robust causal inference from observational survey data, providing insights that are otherwise challenging to achieve through traditional observational or purely experimental methodologies, which was conducting an long-term RCT that assigns partisan sorting as a treatment. This approach not only enhances methodological rigor but also significantly strengthens the validity and interpretability of causal claims drawn from observational analyses. Thus, I will follow the steps of TTE framework to emulate a hypothetical RCT which tries to answer the following causal question:

“What is the effect of assigning United States Voters to a condition of partisan sorting—defined as ideological consistency with party identification—in 2016, on their affective polarization—defined as difference between the feeling towards in-party and out-party—in 2020, relative to what would have happened had they not been assigned to be sorted?”

IV. Research Design: Emulating a Target Trial

This study draws on the ANES 2016–2020 panel to emulate a hypothetical randomized controlled trial in which respondents at the 2016 pre-election wave are “assigned” to either a partisan-sorted or not-sorted condition. It then follows these same individuals through to the 2020 pre-election wave to measure their affective polarization and estimate the average causal effect of partisan sorting. The emulation process will follow the TTE protocol, defining and operationalizing each variable, and concludes with a systematic presentation of all TTE protocol elements.

1. Measures

Figure 5 represents the DAG that this paper will use. Perceived policy differences, elite polarization and religion are not adjusted nor measured. As mentioned earlier, this paper tries to isolate and identify the direct effect of partisan sorting on affective polarization. In this case, perceived policy difference works as a mediator as well as a collider, 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. Furthermore, the unit of analysis of elite polarization does not match with other variables. Lastly, although religion as a demographic feature works as a confounding variable, it is also a collider. Thus, these three variables are excluded from the analysis.

Age, Race, Region, Political Interest, Political Participation works as a confounder. Although Campaign Exposure and Social Media Usage serves as a collider, by adjusting all the confounding mentioned, possible backdoor paths are blocked, so it is included in the measures (See Pearl and Mackenzie, 2019 for details). Campaign exposure and Social media usage are included because it serves as a confounding, and importantly, in the previous literature it is mentioned as a core driving factor that heightens affective polarization.

Treatment Variable-Partisan Sorting: This study solely focuses on alignment of partisan identity and ideological identity. Using the seven-point ideology scale (0 = Strong Liberal to 7 = Strong Conservative) and the seven-point party ID scale (0 = Strong Democrat to 7 = Strong Republican), I code partisan sorting as 1 for respondents whose scores align—that is, those scoring 1–3 on both scales (liberal alignment) or 5–7 on both scales (conservative alignment)—and as 0 for all others, including neutrals (score = 4) or any mismatched ideology–party combinations.

I treat partisan sorting as a binary indicator, coded simply as “sorted” versus “unsorted”, because my goal is to estimate the causal effect of being “treated” with partisan sorting, not to model how varying intensities of sorting might differentially impact affective polarization. By defining the treatment as the

mere presence of sorting, I cleanly emulate a trial contrast between sorted and unsorted conditions and recover the average treatment effect.

Outcome Variable (Confounder for 2016)-Affective Polarization: Affective polarization is measured using the 0–100 feeling-thermometer ratings. For respondents with a clear party ID, I calculate the absolute difference between their in-party and out-party thermometer scores. For those without a clear party affiliation, I take the absolute difference between their Democratic and Republican thermometer scores. This procedure yields an affective polarization index ranging from 0 to 100.

Confounders

Political Participation: Political participation is constructed by combining respondents’ actual turnout in past elections -presidential, gubernatorial, Senate, and House of Representatives (coded 1 if voted; else 0), with their self-reported likelihood of voting in the upcoming presidential election (0: Not registered, does not intend to register or vote; 3: Registered and voted early).

Political Interest: Political Interest is operationalized by respondents’ attention towards politics, election and news on media, each z-scored to form a common metric, and were averaged.

Social Media Usage: Social media use is defined by respondents’ self-reports of how many days per week they engage with social media.

Campaign Exposure through media: Campaign Exposure through media is measured through questions about whether they encountered campaign news via each of the following five outlets: Television news programs; Newspapers; Television talk shows or public-affairs/news-analysis programs; Internet sites, chat rooms, or blogs; Radio news or talk shows. Each item is coded 1 if the respondent reports having seen or heard campaign information through that outlet and 0 otherwise. We then sum these five binary indicators and divide by 5 to create a normalized exposure index ranging from 0 (no exposure) to 1 (exposed via all outlets).

Demographic variables: Race, age, region of respondents were included as dummy variables.

2. Target Trial Emulation Protocol

Table 3 illustrates the Target Trial Emulation protocol for the hypothetical RCT conducted in this paper.

Table 3. Protocol Elements of the Target Trial

Eligibility Criteria	Participants include respondents from the American National Election Studies (ANES) 2016 pre-election wave. Eligible respondents must provide complete and valid baseline data on partisan sorting status, political interest, political participation, media consumption habits, age, race, region, and baseline affective polarization. Respondents who lack any essential baseline data or are unavailable for the follow-up wave in 2020 are excluded from the analysis
Treatment Strategies	<ul style="list-style-type: none"> - Treatment (Partisan sorted): Defined as respondents whose political party identification aligns consistently with their ideological orientation at baseline. - Control (Not Partisan Sorted): Defined as respondents whose party identification and ideological orientation are not consistently aligned at baseline, or neutral - Hypothetically assigned after the completion of ANES 2016 pre-election survey.
Treatment assignment	Participants are assigned observationally based on their actual sorting status at baseline, as recorded in the ANES 2016 pre-election data. Randomization emulated through propensity score inverse probability weighting.

Time Zero and Follow-up	<ul style="list-style-type: none"> - Time Zero (Baseline): Clearly defined as the time of the ANES 2016 pre-election survey wave. All the variables except the outcome variable must be measured in line with the baseline. - Follow-up: Follow-up begins on the day 2016 pre-election survey is completed and ends after the completion of 2020 pre-election survey. Outcomes are assessed after a 4-year period, specifically using data collected in the ANES 2020 pre-election survey wave.
Outcome measures	The primary outcome is affective polarization, measured in the ANES 2020 pre-election survey wave. This measure is operationalized through differences in party feeling thermometer scores, which compare evaluations of one's own party (in-party) versus the opposing party (out-party).
Causal Contrasts	<p>The causal contrast of interest is an intention-to-treat (ITT) estimand, comparing the expected average affective polarization in 2020 between respondents initially classified as partisan sorted versus those not sorted at baseline.</p> $E[Y^{2020} Sorted_{2016} = 1] - E[Y^{2020} Sorted_{2016} = 0]$

Intention-to-treat (ITT) estimand assesses the causal effect of initial treatment assignment (partisan sorting) regardless of any changes in sorting status during the follow-up period. ITT was chosen as it reflects the average causal effect of being initially sorted on the outcome, capturing the relevant scenario of assigning sorting without conditioning on subsequent adherence or changes. As ANES 2016-2020 Panel data does not capture the changes of respondents in their initial treated status within the follow-up period, for example 2018, ITT is more suitable for this emulated RCT than Per-protocol estimand, which considers only participants who maintain their initial treatment status throughout the entire follow-up period.

3. Statistical Analysis Plan

This section addresses how observational data will be analyzed through following the TTE protocol and its assumptions. First, all the missing data that does not meet the requirement of the eligibility criteria, treatment assignment, follow-up and outcome measures will be excluded from the population. By excluding the missing data, it fulfils the consistency assumption of the TTE. The number of excluded observations will be reported at the end through Consolidated Standards of Reporting Trials (CONSORT)-like chart. Second, adjustment of confounder will be operationalized through generating propensity score. To eliminate immortal time bias, all variables except the outcome will be measured in 2016 ANES pre-wave to align with the beginning of the treatment, assuring exchangeability assumption. Propensity score overlap will be reviewed to exclude observations that do not meet positivity assumption. Final population for target trial is concluded here.

Third, I will use Augmented Inverse Probability Weighting (AIPW) to calculate the causal contrast. AIPW combines propensity score weighting with outcome regression to yield a doubly robust estimator of the average treatment effect (ATE). In the first IPW step, each unit is weighed by the stabilized inverse of the probability of receiving its observed treatment, shown in equation (1), where

$$SW_i = \frac{\hat{\pi}}{e(X_i)}, \frac{1-\hat{\pi}}{1-e(X_i)} \quad (1)$$

$e(X_i) = \Pr(A = 1|X_i)$ is the propensity score, $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n A_i$ is the marginal probability of (binary) treatment, and X_i is the confounders. Stabilized IPW is used to anchor each observed unit's weight to the overall sample distribution and mitigate the undue influence of extreme propensity score which can bias the causal estimate (Harder, Stuart and Anthony, 2010).

The augmentation step then fits an outcome regression model $\hat{\mu}(A_i, X_i) = E(Y_i | A_i, X_i)$ and adds a bias-correction term to the stabilized IPW estimator, shown in equation (2).

$$\widehat{AIPW} = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i) + \frac{\hat{\pi} A_i}{e(X_i)} [Y_i - \hat{\mu}(1, X_i)] - \frac{1-\hat{\pi}(1-A_i)}{1-e(X_i)} [Y_i - \hat{\mu}(0, X_i)]] \quad (2)$$

Because it leverages both a correctly specified propensity score model and an outcome regression, AIPW is doubly robust, meaning if either model is correct, the estimator remains consistent (Band and Robins, 2005). The use of stabilized weights further reduces variance without sacrificing unbiased balancing. Standardized Mean Difference (SMD) will be checked to determine whether extreme weighting has occurred or not (see Stuart, 2010 for details). Bias-corrected and accelerated (BCa) Bootstrapping is conducted for robustness check.

Sensitivity analysis for unmeasured confounders will be conducted through calculating E-value. The E-value is a sensitivity measure that quantifies the minimum strength of association that an unmeasured confounder would need to have with both treatment and outcome, conditional on the measured confounders, to fully explain away an observed exposure–outcome association (VanderWeele and Ding, 2017). The interpretation is that any unmeasured confounder would need to be associated with both the exposure and the outcome by risk ratios of at least the E-value (above and beyond measured confounders) to reduce the observed effect estimate to the null. Larger E-values therefore indicate greater robustness to unmeasured confounding, while small E-values (close to 1) suggest that modest unmeasured bias could overturn the observed association.

Finally, to guard against misspecification of complex, unknown relationships between confounders and the treatment or outcome, I complement AIPW with a Debiased Machine-Learning (DML) estimator that allows for nonparametric estimation of nuisance functions (Chernozhukov et al., 2018). Whereas AIPW depends on researcher-specified parametric models such as logistic or linear regressions for the propensity score and outcome regressions, DML leverages flexible, nonparametric learners to approximate these nuisance functions without imposing restrictive functional forms. By performing cross-fitting (See Chernozhukov et al. for details), DML fits each nuisance model on one subset of the data and then computes residuals or orthogonalized scores on a separate subset, thereby mitigating overfitting and ensuring that any small estimation error in the nonparametric regressions does not bias the final treatment-effect estimate (Molak and Jaokar, 2023). The resulting orthogonalized moment conditions remove first-order bias, guaranteeing that the DML estimator of the average treatment effect remains consistent and asymptotically normal provided at least one of the nonparametric nuisance estimators is sufficiently accurate.

Moreover, by harnessing flexible machine-learning algorithms to capture complex, nonlinear associations and interactions among confounders, DML provides an additional safeguard against residual confounding arising from omitted higher-order terms or misspecification. In contrast, AIPW’s reliance on confounder adjustment guided strictly by the researcher’s DAG leaves open the possibility that unmodeled interactions or functional-form errors could bias results. Thus, by subjecting the treatment-effect estimates to both AIPW and DML and observing congruent point estimates and confidence intervals, this paper obtains dual protection: AIPW enforces transparency through explicit confounder control, while DML delivers a data-driven safeguard against specification error, together ensuring that the causal conclusions reflect robust, genuine effects rather than artifacts of model misspecification.

Implementation of DML is done through the DoWhy and econml package in Python. It uses GradientBoostingRegressor for the outcome model, logistic regression for the treatment model, and a least absolute shrinkage and selection operator cross validation (LassoCV) for final model. Gradient boosting is used as it flexibly estimates $E[Y | A, X]$ without assuming a specific functional form (Friedman, 2001). My treatment is binary, so logistic regression is used as it remains a standard, well-calibrated choice for estimating the propensity score (Imbens and Rubin, 2015), ensuring stable probability estimates even in moderately high dimensions. Finally, using LassoCV in the final step enforces sparsity and guards against overfitting in the process of projecting residualized outcomes onto residualized treatments (Chernozhukiv et al., 2018). Cross-fitting these three components by raining the GradientBoostingRegressor and logistic model on one-fold and then fitting LassoCV to the orthogonalized scores on held-out data yields a doubly robust estimator that is consistent and asymptotically normal as long as at least one nuisance model is well estimated. Five-fold cross-validation was implemented both for Lasso hyperparameter and for the outer DML sample splits.

To assess the robustness of the estimated ATE against model misspecification and unobserved confounding, I employ DoWhy’s suite of refuters, which systematically tests whether spurious associations or data artifacts drive our causal estimates (Sharma and Kiciman, 2020). First, the random common cause refuter injects a synthetic confounder U (uniformly distributed between 0 and 1) into adjustment set. If inclusion of this irrelevant variable materially shifts the ATE, it suggests that the identification strategy is overly sensitive to even random noise, indicating possible residual bias. Second, the random outcome (negative outcome) refuter replaces the true outcome with an independent random variable and re-estimates the effect. A well specified model should yield an ATE near zero, while any non-negligible effect signals that the estimator is capturing spurious patterns rather than a genuine causal relationship.

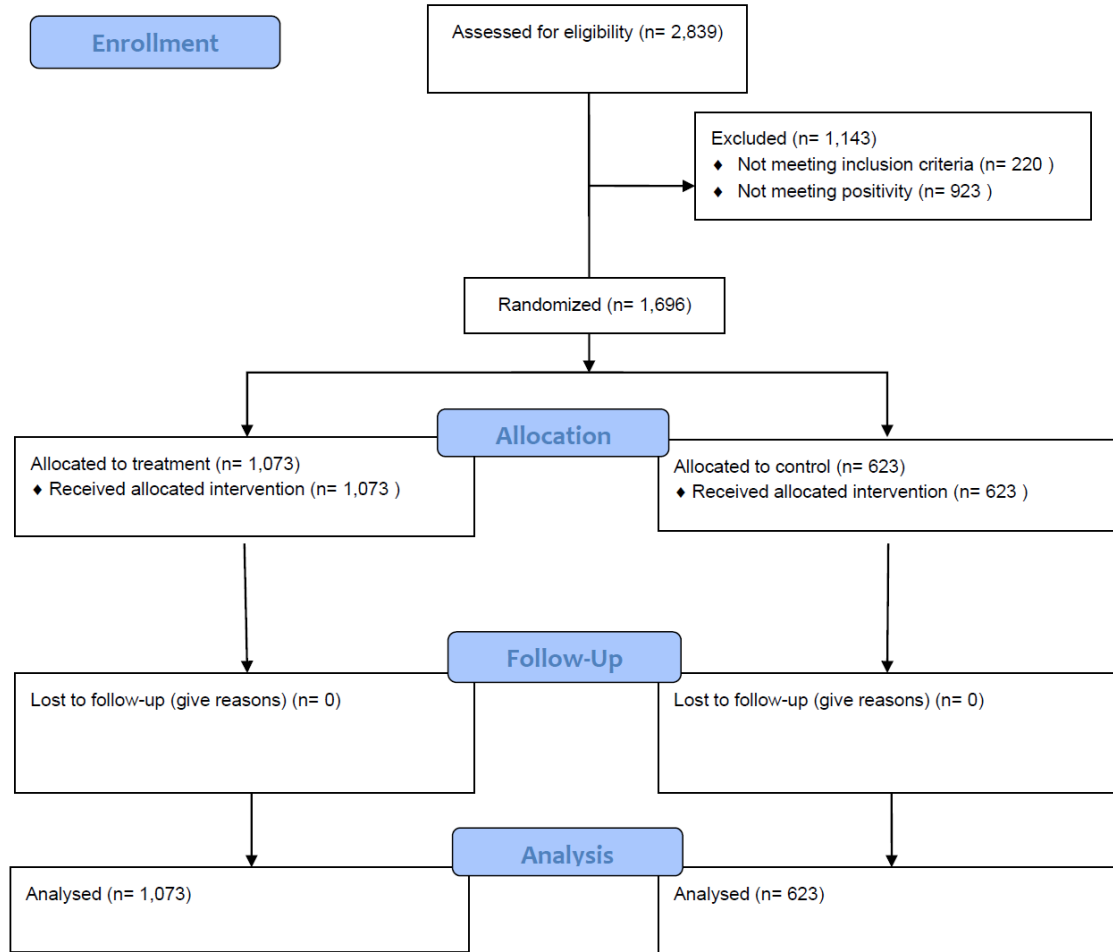
Third, the placebo treatment refuter constructs a new, purely random treatment variable and re-runs the estimation using the newly generated treatment. As placebo treatment is by design independent of both treatment and outcome, a correctly specified model should produce an ATE close to zero. In other words, a non-zero ATE indicates structural misspecification or unaccounted confounding. Fourth, the placebo permute refuter is used to randomly permute the original treatment assignment to break any true association with the outcome while preserving the empirical distribution of the treatment and re-runs the estimation. If the permuted treatment still yields a nonzero ATE, this suggests that the model is picking up artifacts of the data or model rather than a true causal effect. Finally, the subset refuter repeatedly recomputes the ATE on random subsamples (e.g., 50% of observations) to gauge stability. If the estimated effect varies dramatically across different subsamples, it suggests that a few influential units may be driving the result. Together, these refuters serve as a battery of falsification checks. All refuters were repeated 50 times for validity.

Beyond refuter-based validation of the DML model, this study conducts a formal sensitivity analysis to quantify how strong an unmeasured confounder U would need to be to explain away the estimated causal effect (Cinelli and Hazlett, 2020). Using DoWhy, it will calculate the smallest partial R^2 values for unobserved confounder U , using Riesz-based partial R^2 , that would drive the effect estimate to zero, or to the edge of its confidence interval. The nonparametric Riesz-based partial R^2 analysis complements refuter-based checks by quantifying how powerful an unobserved confounder would need to be. If no such confounder is conceivable, I can be confident that our causal estimate is unlikely to be fully explained away by omitted variables.

V. Result

Figure 6 shows CONSORT-like chart for the target trial. ANES Panel 2016-2020 data had a total of 2,839 observations. 220 observed data were excluded from the analysis as it did not meet the eligibility criteria. After excluding the missing data, the propensity score of the confounders were calculated. Figure 7 represents the overlapping areas of the propensity score.

Figure 6. CONSORT-like Chart for the Target Trial

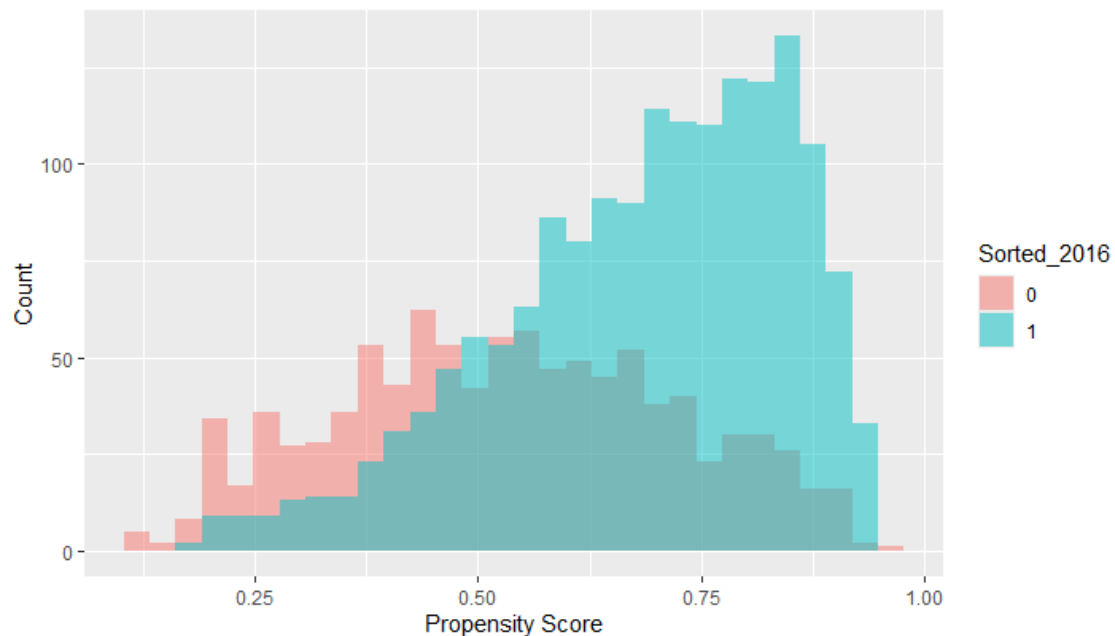


Propensity scores $\hat{e}(X)$ for each respondent was calculated using logistic regression (Appendix Table A1 displays result of logistic regression). Although the raw overlap between treated and controlled propensity score distributions spanned approximately $[0.1, 0.9]$, there were very few respondents of treated group existing in $\hat{e}(X) < 0.4$, while very few respondents of controlled group existing in $\hat{e}(X) > 0.8$. To satisfy the positivity assumption and to avoid large IPW and improve both the validity and precision of the IPW estimates (Crump et al., 2009; Stürmer et al., 2010), trimming was done over the $0.4 < \hat{e}(X) < 0.8$. This trimming retained 1,696 respondents, excluding 923 extra respondents after the eligibility assumption. The remaining respondents had a reasonable probability of being both treated

and controlled. Through this process, the final number of respondents to this paper studies are 1,696, where 1,073 respondents in treatment group and 623 respondents in controlled group.

1. Augmented Inverse Probability Weighting

Figure 7. Propensity Score Distribution by Treatment Group



Based on the trimmed population, propensity score was recalculated to generate stabilized IPW (Appendix Table A2 displays result of logistic regression of trimmed sample). Re-estimating the propensity score on the trimmed population is necessary because the act of discarding observations with extreme estimated probabilities changes the joint distribution of treatment and confounders. A logistic model fitted to the full sample yields coefficient estimates that reflect relationships in regions of poor overlap. If those extreme-support observations are removed, the original coefficients may no longer produce calibrated scores for the remaining units. Using full-sample propensity scores to weight a trimmed subsample can leave residual imbalances and, in some cases, still assign very low or very high weights within the retained range, thereby undermining the positivity assumption and inflating variance (Crump et al., 2009). By refitting the logistic regression after trimming, one ensures that each retained unit's estimated probability truly lies in a stable support region, yielding inverse-probability weights that minimize variance and achieve better confounder balance. (Appendix Figure A3 displays Propensity Score Overlap for trimmed sample, Figure A4 displays IPW distribution).

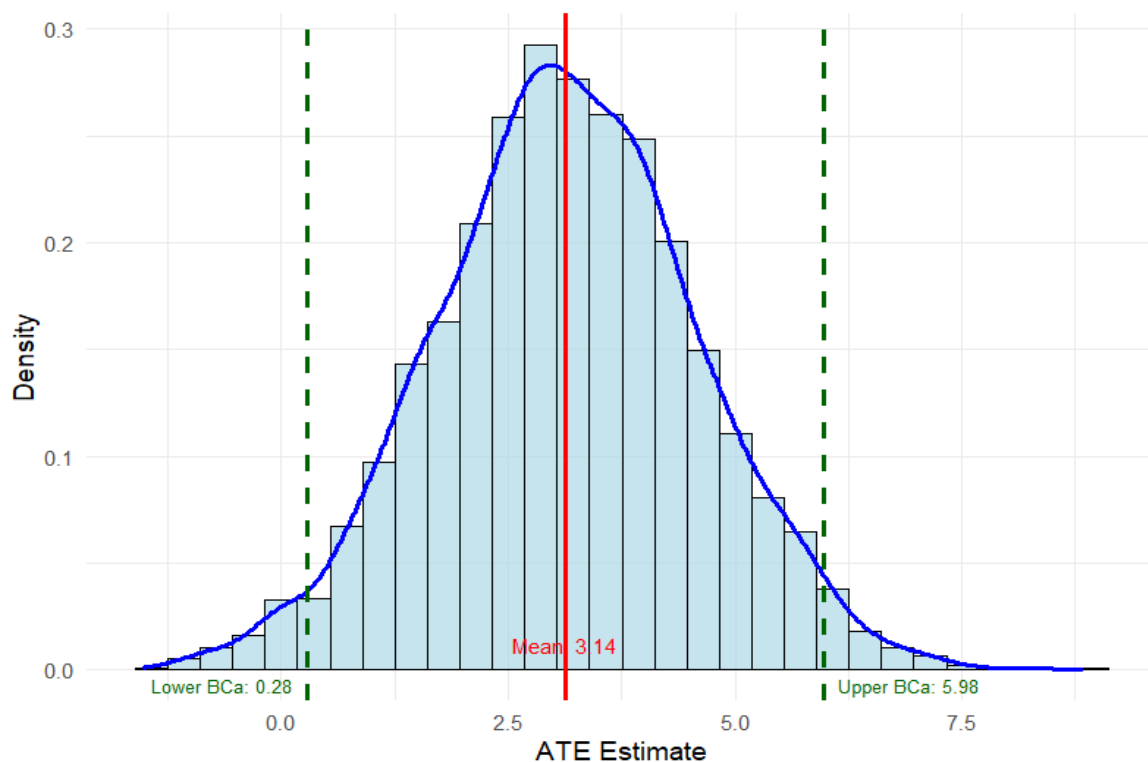
Confounder balance is assessed using the SMD. This is done to check the two groups resemble one another on baseline characteristics, assuring the RCT like environment. Weighted balance diagnostics after trimming confirmed that all SMDs fell below (absolute value) 0.10. (Appendix Table A5 displays SMD values before and after adjustment, Figure A6 displays loveplot of calculated SMD). After applying standardized IPW weights, the effective sample size dropped to 548.22 in the control arm and 1,024.79 in the treated arm. This reduction reflects the variance inflation introduced by weighting. Although it still retain over 80% of the original control sample and over 95% of the treated sample in “effective” terms, the precision of the estimates is governed by these weighted (effective) sample sizes rather than by the raw counts.

As mentioned above, AIPW needs two results, regression of standardized IPW adjustment and normal regression. The standardized IPW-adjusted coefficient on treatment (partisan sorting in 2016) is 2.78 with 95 % robust confidence interval of [-0.0209, 5.5858]. This indicates that after accounting for measured confounding, individuals who are sorted exhibit, on average, a 2.78-unit higher outcome than

those who are not sorted. However, the robust confidence interval includes zero, implying that this effect does not reach conventional statistical significance at the 0.05 level when using heteroskedasticity-consistent standard errors. By contrast, the non-robust 95% confidence interval gives [0.1670, 5.3978], reflecting smaller standard errors are obtained under homoskedasticity assumption. The coefficient of an ordinary regression model without stabilized IPW weights is 2.97 (Appendix Table A7 displays the result of regression).

The stabilized-weight AIPW estimator yields a point estimate of the ATE equal to 2.78. To assess sampling variability without relying solely on asymptotic normality, this paper conducted 5,000 nonparametric bootstrap replicates and constructed BCa confidence intervals. The resulting 95% BCa interval is [0.285, 5.982]. Graphically (see Figure 8), the bootstrap distribution of the ATE is approximately bell-shaped and centered near 3.14, indicating that after adjusting for measured confounders, being partisan sorted in 2016 has a causal effect on at least a +0.29-unit increase and up to +5.98 units in the 2020 affective-polarization score. The bootstrap BCa interval is slightly wider than the normal-theory interval [0.0098, 5.5473] obtained from the analytic SE, reflecting the BCa procedure's adjustment for both bias and skewness in the ATE's empirical sampling distribution. In sum, the bootstrap results corroborate the analytic finding of a positive ATE and provide robust evidence that the effect is unlikely to be explained by chance alone.

Figure 8. Bootstrap Distribution of ATE (AIPW, BCa CI 95%)



E-value was calculated to quantify the robustness of AIPW estimate to potential unmeasured confounding. After re-expressing the continuous-outcome AIPW estimate of 2.78 (SE = 1.41) as an approximate risk ratio (RR = 1.0917, 95% CI = 1.0005–1.1913), the E-value of 1.4082 for the point estimate and 1.0225 for the lower 95% confidence-interval bound was obtained. An E-value of 1.4082 implies that an unmeasured confounder would need to be associated with both being “sorted” in 2016 and the 2020 affective-polarization score by a risk ratio of at least 1.408, above and beyond all measured confounders, to entirely explain away the observed positive effect. This reflects a moderately strong confounding influence, meaning only a factor that increases the odds of treatment and outcome by

roughly 40% each could shift the point estimate to the null. By contrast, the lower-CI E-value of 1.0225 indicate that, at the edge of our confidence interval of risk ratio, only a very weak confounder that can change treatment and outcome by about 2% would suffice to render the effect statistically nonsignificant. In conclusion, while the AIPW point estimate demonstrates moderate resistance to unmeasured bias, the weakest plausible effect remains sensitive to even small unobserved confounding.

2. Debiased Machine Learning

Using the DML procedure with five-fold cross-fitting and Lasso-tuned nuisance models (Appendix Figure B1-B4 displays model compatibility of logistic regression used in DML), this paper used a nonparametric bootstrap (1,500 draws) to characterize the sampling distribution of the causal effect. A point estimate of ATE is not reported because due to the characteristics of ML, the DML model creates slightly differing results when initiated. Out of 1,500 bootstrap samples, 1,247 converged successfully (Appendix Figure B5 shows Bootstrap distribution with model failure). The remaining 253 were discarded because, in those replicates, one or more nuisance-model fits failed (Figure 9). Among the 1,247 valid replicates, the average ATE was 3.340, reflecting minor finite-sample bias and the 95% confidence interval is [1.471, 5.661]. In conclusion, with 95% confidence, sorting in 2016 increases the 2020 affective polarization score by at least 1.47 points and at most 5.66 points.

Figure 9. Bootstrap Distribution of ATE (DML, failures removed)

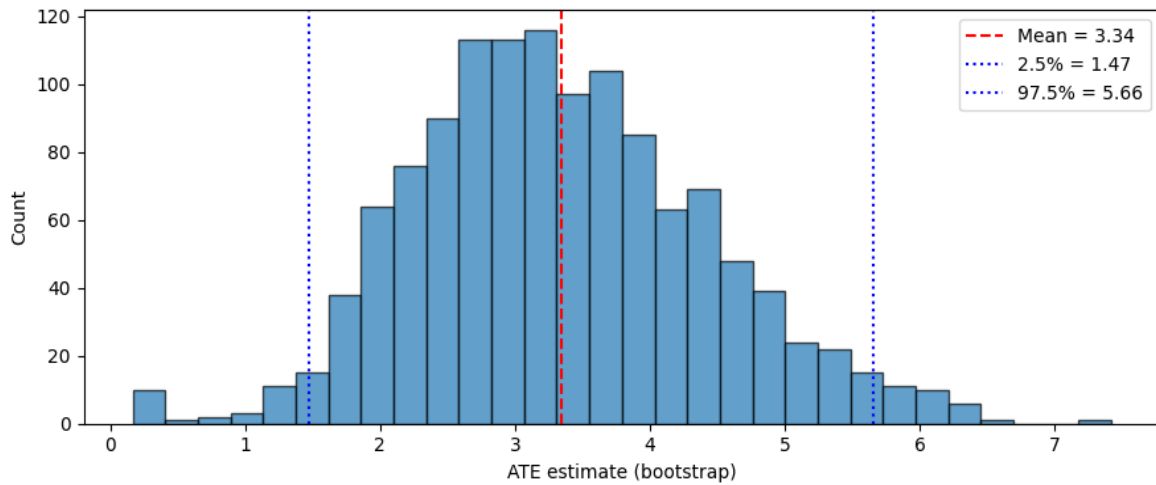


Table 4. Model Refuter Results

Refuter	New effect (Original estimate: 2.78)	p-value (CI for random outcome)
Random cause	2.8193	0.78
Placebo treatment	0.1245	0.97
Placebo 'permute'	-0.0418	0.93
Subset (0.5)	2.5208	0.88
Random outcome	-0.033	[-0.143, 0.000]

Table 4 displays refutation analysis on a DML model that obtained the ATE result of 2.78. For random causes, introducing a purely random confounder did not meaningfully ($p=0.78$) alter the original estimate, this suggests that the DML model is not unduly sensitive to the inclusion of irrelevant predictors. Placebo refuters showed similar trends. Both placebo treatment tests gave an ATE collapsing toward zero and became indistinguishable from null. Subset test revealed that although sampling variability caused a modest downward shift, the direction and magnitude remained broadly consistent with the full-sample effect, indicating that model's findings are not driven solely by any small group of influential observations. As a final falsification, bootstrapping 1,500 replicates of DML on random outcome was conducted. Entire interval is effectively centered at zero, confirming that it does not systematically produce non-zero estimates when the true effect is null. Taken together, these sensitivity checks provide strong evidence that the observed positive ATE of 2.78 is unlikely to be an artifact of misspecification, overfitting, or unmodeled data anomalies.

Sensitivity analysis to quantify how strong an unobserved confounder would have to be to explain away the estimated ATE using nonparametric partial r square showed 2 findings (Appendix B6). The resulting robustness value of 0.05 indicates that unobserved confounder would have to explain at least 5 % of the residual variance in both treatment and outcome to reduce the point estimate to zero. In most observational settings, it is rare for any single omitted variable to explain 5 % of residual variance after adjusting for key demographic and political predictors, which suggests that the estimated effect size is reasonably robust. By contrast, the robust value at $\alpha=0.05$ is 0.01, meaning that an unmeasured confounder explaining only 1 % of the residual variance in treatment and outcome would suffice to render the lower bound of the 95 % confidence interval exactly zero. As 1 % is a relatively small proportion of residual variance, this indicates that the inference about statistical significance is sensitive to even a weak omitted variable. In summary, while it is unlikely that any plausible confounder could fully eliminate the estimated ATE of 2.783 ($RV = 0.05$), the narrow margin by which the confidence interval excludes zero implies that a modest unmeasured bias could overturn our conclusion of significance.

VI. Conclusion

This study employed a TTE framework to robustly assess the causal impact of partisan sorting on affective polarization among U.S. voters over a four-year period (2016–2020). Leveraging panel data from the ANES, observational data was designed to approximate RCT. Through rigorous adjustments for confounding variables by AIPW and DML methods, the result of analysis consistently indicates that partisan sorting in 2016 causally increased affective polarization by approximately 2.8 units on a 0-100 scale by 2020.

Theoretically, these findings significantly advance our understanding of causal impact that partisan sorting has on affective polarization. Prior literature has extensively documented the complex interdependencies between policy disagreements, elite polarization, partisan identities, and media influences, yet few studies have explicitly isolated the "pure" effect of partisan sorting itself. This study fills this critical gap by demonstrating that sorting plays a distinct and consequential role in intensifying affective polarization, independent of other drivers. These results corroborate social identity theory's emphasis on alignment across social identities, reinforcing perceptions of out-group threat and deepening inter-party animosity (Mason, 2018).

Methodologically, this study makes a notable contribution to political science by demonstrating how TTE can enhance causal inference within observational data. By explicitly following the protocol necessary for the TTE framework, it provided a model for future political science research. Adoption of doubly robust estimators, complemented by machine learning based robustness checks, illustrates a rigorous approach to handling potential bias from both observed and unobserved confounding, setting a high standard for observational research in the field.

Despite these robust findings, this study carries several critical limitations. First, although the sensitivity analyses provide confidence against moderate unmeasured confounding, the inference regarding statistical significance remains sensitive to even relatively small, omitted variables. Second, trimming respondents based on propensity score overlap ensures internal validity but limits external generalizability primarily to individuals with moderate probabilities of sorting, as the methodology creates a ‘pseudo-population’. Third, operationalization of sorting and affective polarization through summary indices could obscure underlying heterogeneity in individual attitudes and behaviors.

Future research should therefore address these limitations by exploring different causal mechanisms regarding the interdependence of drivers. Exploring heterogeneity in the sorting effect across different subgroups is also crucial. Furthermore, longitudinal studies that incorporate more frequent observation could capture the causal relationship between sorting and affective polarization. Experimental interventions manipulating sorting cues could also provide complementary evidence about underlying psychological mechanisms. Moreover, comparative analyses across different democratic contexts might reveal whether similar causal dynamics prevail in varied political environments, thus enhancing the generalizability of these findings.

In conclusion, this study demonstrates the significant, independent causal effect of partisan sorting on affective polarization using causal inference methods. It underscores the critical role sorting plays in fueling political division based on rich theoretical backgrounds that previous research has provided. By bridging methodological rigor and theoretical insight, this research contributes substantially to our understanding of affective polarization and provides a robust foundation for designing strategies aimed at fostering healthier democratic dialogue and social cohesion.

References

- Abramowitz, A. I., & Saunders, K. L. (2008). Is Polarization a Myth? *The Journal of Politics*, 70(2), 542-555.
- Aleksander, M., & Ajit, J. (2023). *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. Packt Publishing.
- Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The Welfare Effects of Social Media. *American Economic Review*, 110(3), 629-676.
- Amit Sharma & Emre Kiciman, 2020. "DoWhy: An End-to-End Library for Causal Inference," Papers 2011.04216, arXiv.org.
- American National Election Studies. 2024. ANES 2016-2020 Panel Study Merged File [dataset and documentation]. May 15, 2024 version. www.electionstudies.org
- Antoine, A., Pérol, D., Robain, M., Bachelot, T., Choquet, R., Jacot, W., Ben Hadj Yahia, B., Grinda, T., Delaloge, S., Lasset, C., & Drouet, Y. (2024). Assessing the real-world effectiveness of 8 major metastatic breast cancer drugs using target trial emulation. *European Journal of Cancer*, 213, 115072.
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962-973.
- Beam, M. A., J., H. M., & Hmielowski, J. D. (2018). Facebook news and (de)polarization: reinforcing spirals in the 2016 US election. *Information, Communication & Society*, 21(7), 940-958.
- Bougher, L. D. (2017). The Correlates of Discord: Identity, Issue Alignment, and Political Hostility in Polarized America. *Political Behavior*, 39(3), 731-762.
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proc Natl Acad Sci U S A*, 114(28), 7313-7318.
- Chambers, J. R., Schlenker, B. R., & Collisson, B. (2013). Ideology and Prejudice: The Role of Value Conflicts. *Psychological Science*, 24(2), 140-149.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Dahabreh, I. J., Matthews, A., Steingrimsson, J. A., Scharfstein, D. O., & Stuart, E. A. (2024). Using Trial and Observational Data to Assess Effectiveness: Trial Emulation, Transportability, Benchmarking, and Joint Analysis. *Epidemiologic Reviews*, 46(1), 1-16.
- Dias, N., & Lelkes, Y. (2022). The Nature of Affective Polarization: Disentangling Policy Disagreement from Partisan Identity. *American Journal of Political Science*, 66(3), 775-790.
- Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *J Clin Epidemiol*, 142, 264-267.
- Druckman, J. N., Green, D. P., & Iyengar, S. (2023). Does Affective Polarization Contribute to Democratic Backsliding in America? *The ANNALS of the American Academy of Political and Social Science*, 708(1), 137-163.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189-1232.
- Fu, E. L. (2023). Target Trial Emulation to Improve Causal Inference from Observational Data: What, Why, and How? *J Am Soc Nephrol*, 34(8), 1305-1314.
- Garrett, R. K., Gvirsman, S. D., Johnson, B. K., Tsfati, Y., Neo, R., & Dal, A. (2014). Implications of pro- and Counterattitudinal Information Exposure for Affective Polarization. *Human Communication Research*, 40(3), 309-332.
- Gomes, M., Latimer, N., Soares, M., Dias, S., Baio, G., Freemantle, N., Dawoud, D., Wailoo, A., & Grieve, R. (2022). Target Trial Emulation for Transparent and Robust Estimation of Treatment Effects for Health Technology Assessment Using Real-World Data: Opportunities and Challenges. *PharmacoEconomics*, 40(6), 577-586.
- Graham, J., Nosek, B. A., & Haidt, J. (2012). The Moral Stereotypes of Liberals and Conservatives: Exaggeration of Differences across the Political Spectrum. *PLOS ONE*, 7(12), e50092.
- Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods*, 15(3), 234-249.
- Hernán, M. A. (2018). The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *Am J Public Health*, 108(5), 616-619.
- Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*, 183(8), 758-764.
- Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC
- Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*, 79, 70-75.

- Huddy, L., Mason, L., & AarØe, L. (2015). Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity. *American Political Science Review*, 109(1), 1-17.
- Huddy, L., & Yair, O. (2021). Reducing Affective Polarization: Warm Group Relations or Policy Compromise? *Political Psychology*, 42(2), 291-309.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(Volume 22, 2019), 129-146.
- Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly*, 76(3), 405-431.
- Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, 59(3), 690-707.
- Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3), 188-206.
- Kutcher, S. A., Brophy, J. M., Banack, H. R., Kaufman, J. S., & Samuel, M. (2021). Emulating a Randomised Controlled Trial With Observational Data: An Introduction to the Target Trial Framework. *Can J Cardiol*, 37(9), 1365-1377.
- Lelkes, Y. (2018). Affective Polarization and Ideological Sorting: A Reciprocal, Albeit Weak, Relationship. *The Forum*, 16, 67-79.
- Lelkes, Y., Sood, G., & Iyengar, S. (2017). The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect. *American Journal of Political Science*, 61(1), 5-20.
- Levendusky, M. S. (2009). *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. University of Chicago Press.
- Levendusky, M. S. (2013). Why Do Partisan Media Polarize Viewers? *American Journal of Political Science*, 57(3), 611-623.
- Levendusky, M. S., & Malhotra, N. (2015). (Mis)perceptions of Partisan Polarization in the American Public. *Public Opinion Quarterly*, 80(S1), 378-391.
- Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7(1), 74-101.
- Mason, L. (2015). "I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science*, 59(1), 128-145.
- Mason, L. (2016). A Cross-Cutting Calm: How Social Sorting Drives Affective Polarization. *Public Opinion Quarterly*, 80(S1), 351-377.
- Mason, L. (2018). Ideologues without Issues: The Polarizing Consequences of Ideological Identities. *Public Opinion Quarterly*, 82(S1), 866-887.
- Mason, L., & Wronski, J. (2018). One Tribe to Bind Them All: How Our Social Group Attachments Strengthen Partisanship. *Political Psychology*, 39(S1), 257-277.
- Orr, L. V., Fowler, A., & Huber, G. A. (2023). Is Affective Polarization Driven by Identity, Loyalty, or Substance? *American Journal of Political Science*, 67(4), 948-962.
- Orr, L. V., & Huber, G. A. (2020). The Policy Basis of Measured Partisan Animosity in the United States. *American Journal of Political Science*, 64(3), 569-586.
- Pearce, N., & Vandenbroucke, J. P. (2023). Are Target Trial Emulations the Gold Standard for Observational Studies? *Epidemiology*, 34(5), 614-618.
- Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4), 669-688.
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc.
- Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), e2024292118.
- Rogowski, J. C., & Sutherland, J. L. (2016). How Ideology Fuels Affective Polarization. *Political Behavior*, 38(2), 485-508.
- Rudolph, T. J., & Hetherington, M. J. (2021). Affective Polarization in Political and Nonpolitical Settings. *International Journal of Public Opinion Research*, 33(3), 591-606.
- Sears, D. O., & Funk, C. L. (1999). Evidence of the Long-Term Persistence of Adults' Political Predispositions. *The Journal of Politics*, 61(1), 1-28.
- Seewald, N. J., McGinty, E. E., & Stuart, E. A. (2024). Target Trial Emulation for Evaluating Health Policy. *Ann Intern Med*, 177(11), 1530-1538.

- Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge: Cambridge University Press.
- Skytte, R. (2021). Dimensions of Elite Partisan Polarization: Disentangling the Effects of Incivility and Issue Polarization. *British Journal of Political Science*, 51(4), 1457-1475.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci*, 25(1), 1-21.
- Tajfel H, Turner J. 1979. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, ed.WG Austin, SWorchel, pp. 33–47. Monterey, CA: Brooks Cole
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*, 167(4), 268-274.
- Webster, S. W., & Abramowitz, A. I. (2017). The Ideological Foundations of Affective Polarization in the U.S. Electorate. *American Politics Research*, 45(4), 621-647.
- Yates, M., & Youniss, J. (1998). Community Service and Political Identity Development in Adolescence. *Journal of Social Issues*, 54(3), 495-512.

Appedix

A1. Logistic Regression Result for Propensity Score before Trimming

	Model 1
Political Interest	1.350*** (0.247)
Political Participation	0.234*** (0.050)
Social Media Usage	0.049 (0.108)
Campaign Exposure	0.351* (0.160)
age_group_30-44	0.027 (0.143)
age_group_45-59	-0.458** (0.147)
age_group_60+	-0.338* (0.151)
race_Hispanic	0.424* (0.210)
race_Other	0.436 [†] (0.224)
race_White	0.574*** (0.160)
regionNortheast	0.093 (0.137)
regionSouth	0.081 (0.115)
regionWest	0.181 (0.130)
Affective Polarization(2016)	0.023*** (0.002)
(Intercept)	-2.280*** (0.262)
Observations	2,619
Null deviance	3455.8
Residual deviance	3045.6
AIC	3075.6
Wald Chi ²	334.45***
AUC	0.7294

Note: standard errors in parentheses

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed)

A2. Logistic Regression Result for Propensity Score after Trimming

	Model 1
Political Interest	1.567*** (0.319)
Political Participation	0.236*** (0.071)
Social Media Usage	0.045 (0.129)
Campaign Exposure	0.387* (0.196)
age_group_30-44	0.107 (0.174)
age_group_45-59	-0.552** (0.182)
age_group_60+	-0.382* (0.188)
race_Hispanic	0.336 (0.258)
race_Other	0.495 [†] (0.278)
race_White	0.623** (0.199)
regionNortheast	0.000 (0.160)
regionSouth	0.013 (0.137)
regionWest	0.226 (0.156)
Affective Polarization(2016)	0.028*** (0.003)
(Intercept)	-2.486*** (0.412)
Observations	1,696
AIC	2139.4
Wald Chi ²	112.27***
AUC	0.6573

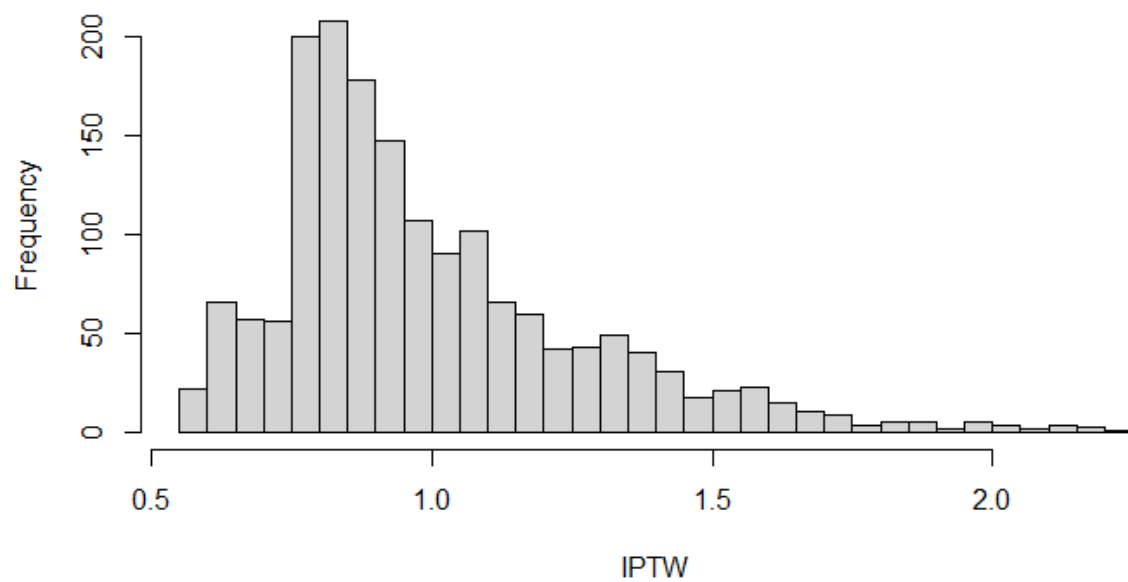
Note: standard errors in parentheses

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed)

A3. Propensity Score Distribution by Treatment Group after Trimming



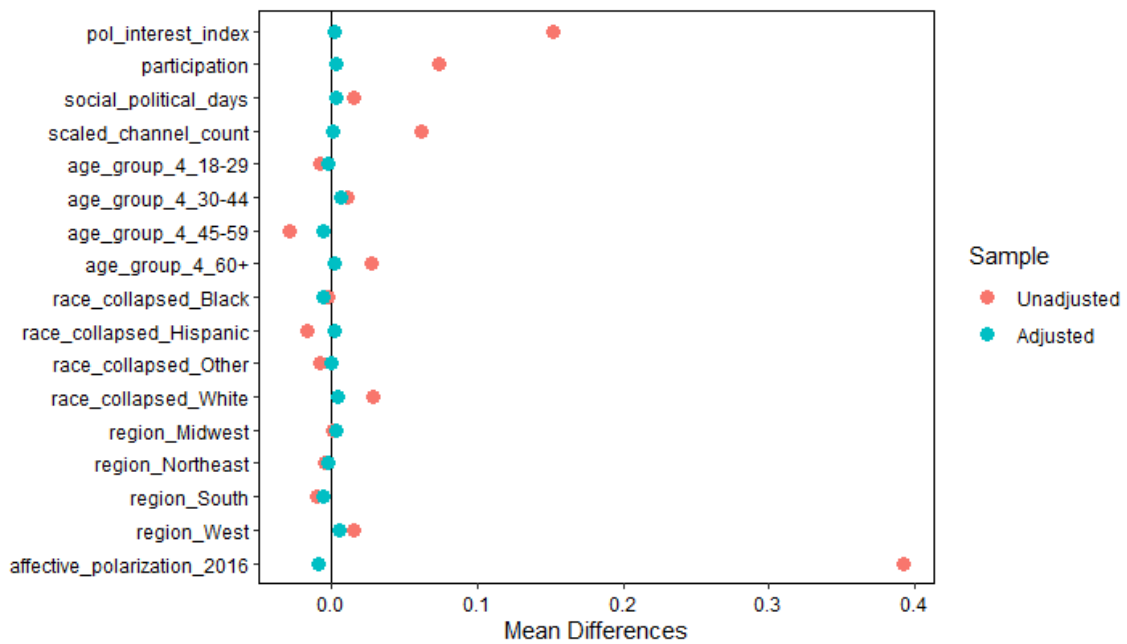
A4. Stabilized Inverse Probability Weights Distribution



A5. Standard Mead Difference Before and After adjustment

Variable	Diff.Un	Diff.Adj
Political Interest	0.1514	0.0014
Political Participation	0.0739	0.0027
Social Media Usage	0.0150	0.0035
Campaign Exposure	0.0616	0.0010
age_group_18-29	-0.0083	-0.0023
age_group_30-44	0.0102	0.0063
age_group_45-59	-0.0286	-0.0064
age_group_60+	0.0267	0.0024
race_Black	-0.0028	-0.0056
race_Hispanic	-0.0168	0.0016
race_Other	-0.0082	-0.0006
race_White	0.0278	0.0045
region_Midwest	0.0002	0.0030
region_Northeast	-0.0047	-0.0023
region_South	-0.0108	-0.0060
region_West	0.0153	0.0052
Affective Polarization(2016)	0.3920	-0.0090

A6. Loveplot of Standard Mean Difference



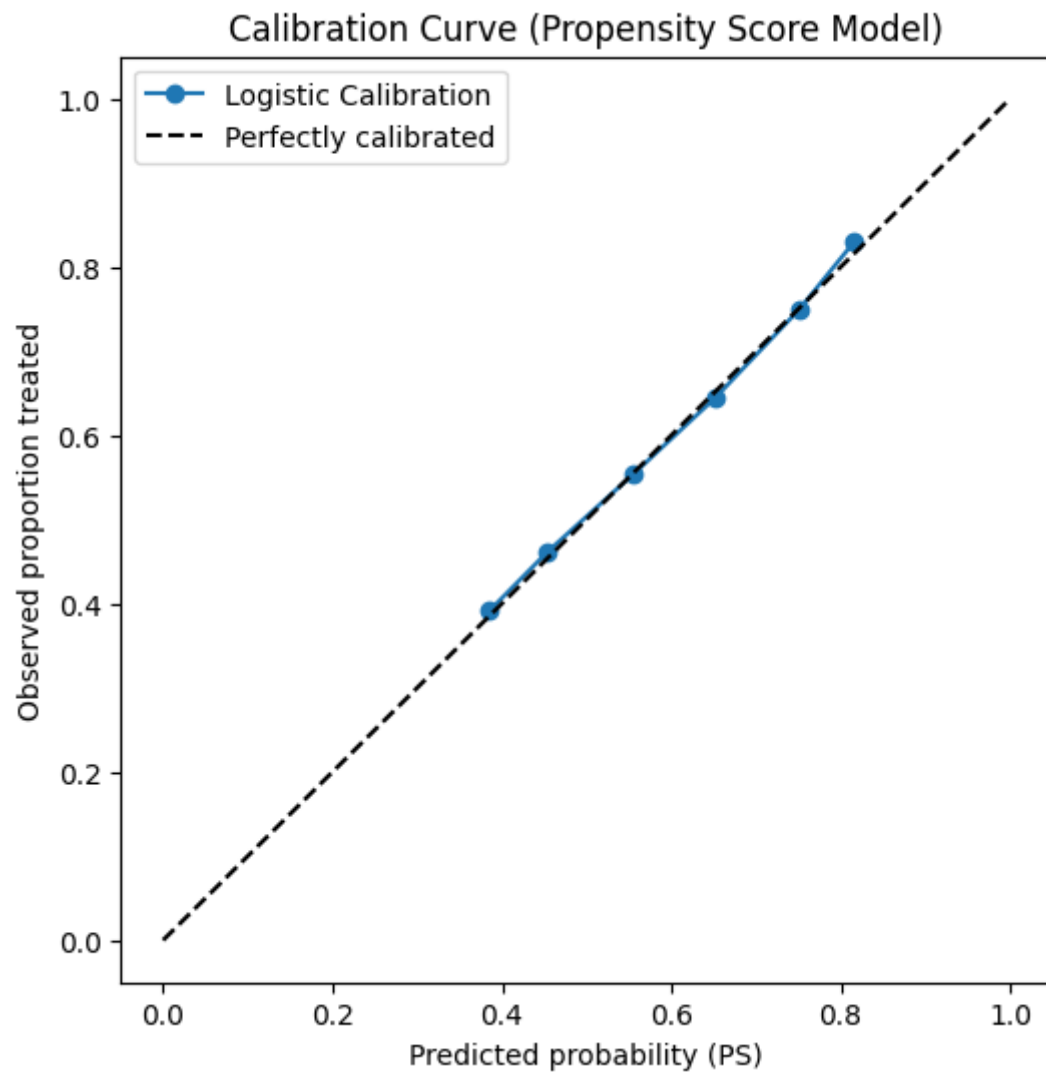
A7. Regression Result (without Inverse Probability Weighting)

	OLS Model
sorted_2016	2.967* (1.381)
Political Interest	8.878* (3.932)
Political Participation	2.578** (0.877)
Social Media Usage	0.048 (1.581)
Campaign Exposure	-2.304 (2.424)
age_group_30-44	2.237 (2.144)
age_group_45-59	4.090 [†] (2.251)
age_group_60+	7.281** (2.308)
race_Hispanic	3.699 (3.206)
race_Other	7.371* (3.445)
race_White	8.311*** (2.467)
regionNortheast	3.175 (1.967)
regionSouth	4.124* (1.688)
regionWest	0.735 (1.911)
Affective Polarization(2016)	0.499*** (0.035)
(Intercept)	6.317 (5.067)
Observations	1,696
F	22.15***
R^2	0.1651
Adjusted R^2	0.1577
Root MSE	26.436

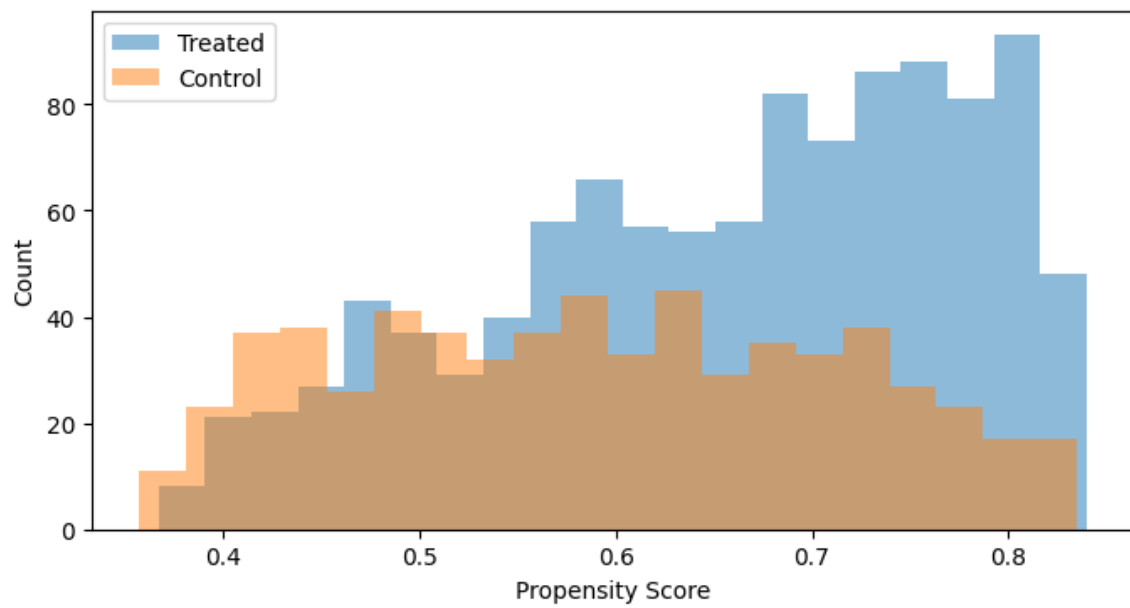
Note: standard errors in parentheses

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed)

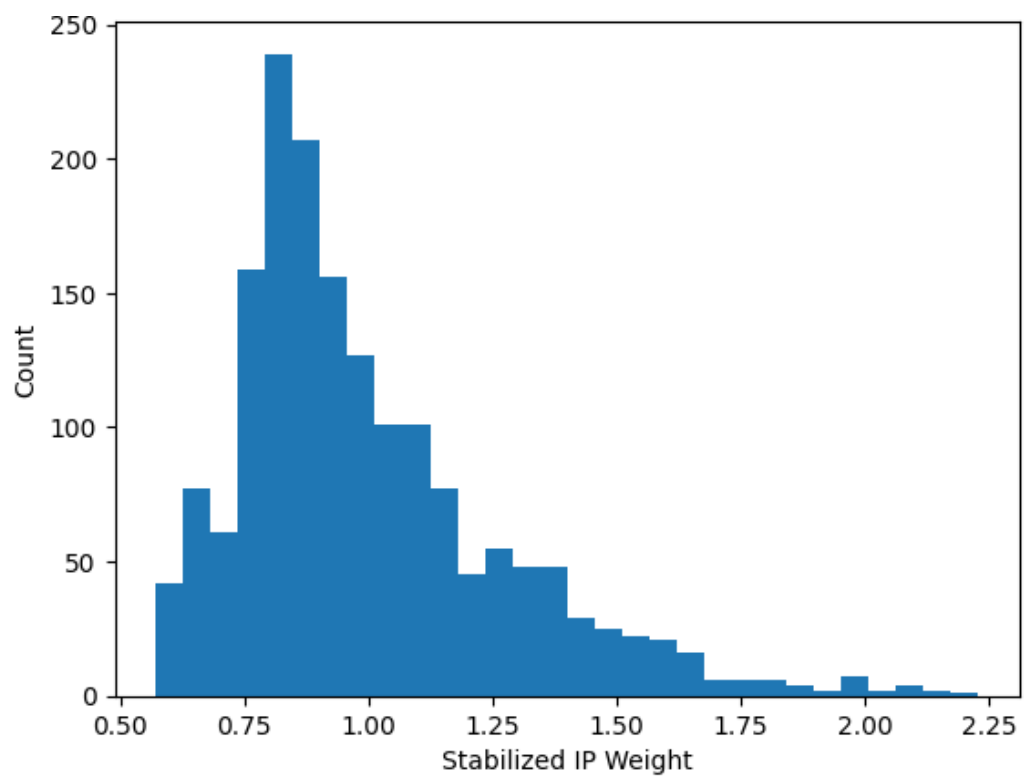
B1. Calibration Curve for Model_T (Logistic Regression)



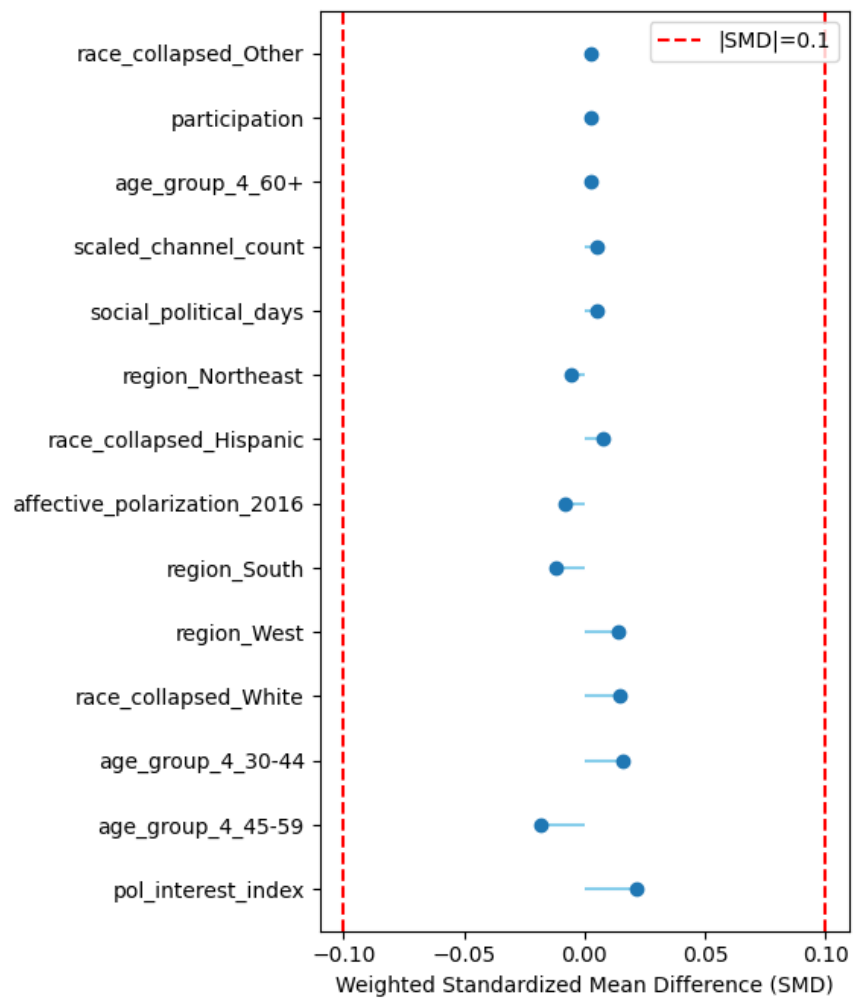
B2. Propensity Score Distribution by Group



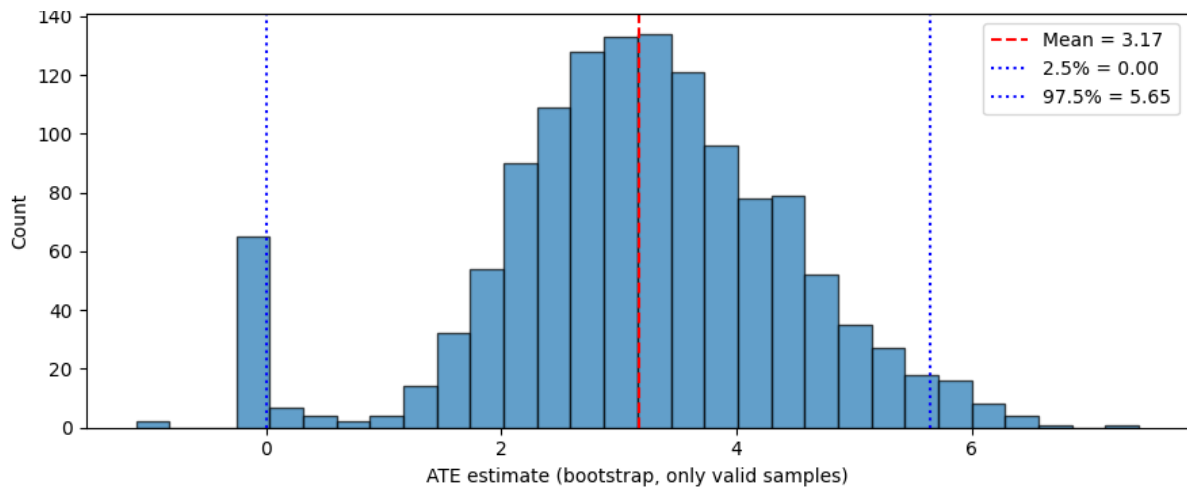
B3. Stabilized Inverse Probability Weights Distribution



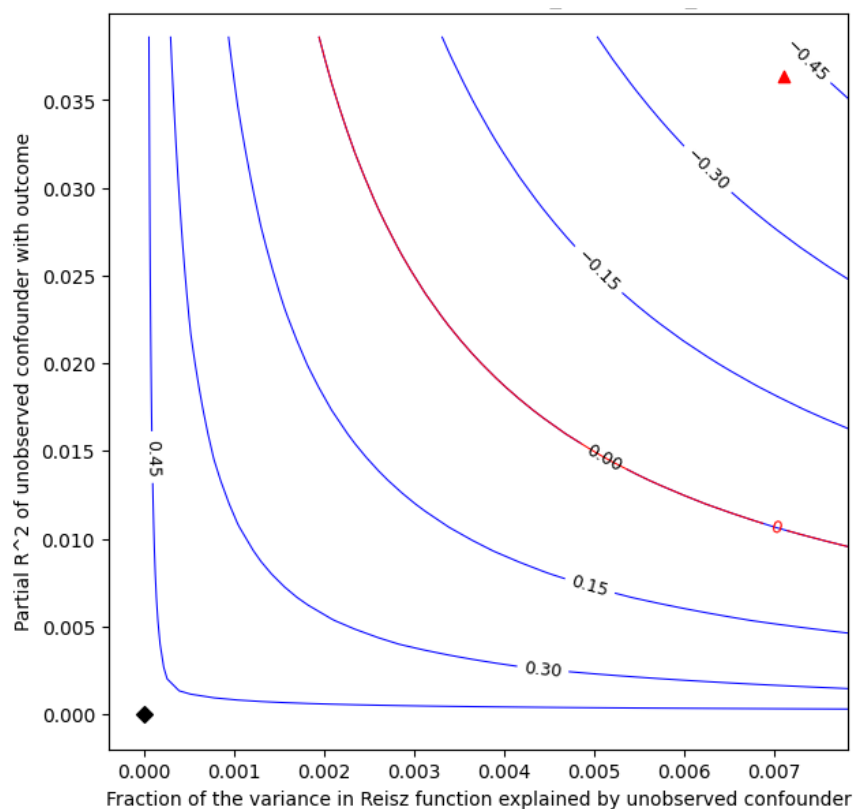
B4. Loveplot of Standard Mean Difference



B5. Bootstrap of Debiased Machine Learning with Failure



B7. Nonparametric Partial R squared sensitivity analyses



Original Effect Estimate : 2.783030340615291

Robustness Value : 0.05

Robustness Value (alpha=0.05) : 0.01

Interpretation of results

- Any confounder explaining less than 5.0% percent of the residual variance of both the treatment and the outcome would not be strong enough to explain away the observed effect i.e bring down the estimate to 0
- For a significance level of 5.0%, any confounder explaining more than 1.0% percent of the residual variance of both the treatment and the outcome would be strong enough to make the estimated effect not 'statistically significant'