UC Davis

UC Davis Electronic Theses and Dissertations

Title

Advances in Machine Learning for Experimental Research

Permalink

https://escholarship.org/uc/item/8hx988zj

Author

Fuller, Samuel Joseph

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Advances in Machine Learning for Experimental Research

By

SAMUEL FULLER

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Political Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Christopher Hare, Chair

James Adams

Lauren Young

Alexa Bankert

Committee in Charge

2024

© Samuel Fuller, 2024. All rights reserved.

To my friends and family who always supported me in my endeavors.

Acknowledgments

I would like give a special thanks to my advisor, mentor, and (most importantly) friend Chris Hare for helping me navigate this whole grad school process. Your help, motivation, and friendship have been invaluable—I couldn't have done it without you. I also would like to thank Jack Rametta for his friendship and collaboration. We've both taken a weird and circuitous route to get here, but I'm sure we're on to something.

Thank you to all of my coauthors including Hilary Izatt, Carlos Algara, Tzu-Ping Liu, Jim Adams, and Alexa Bankert. I have learned so much from each of you and look forward to more collaborations in the future. And thank you to my committee for their thoughts and their flexibility with the defense, it was all very much appreciated.

My friends and family were invaluable through this whole grad school process. Ray Brandt and Paige Pellaton were wonderful roommates and friends, especially in the early years. The COVID movie nights I had with Jonathan Colner, Zach Durec, Jonathan Dahlsten, and Jack Rametta kept me sane—though if I ever have to watch Atlantis again I may, in fact, go insane. And all of the conversations and debates with Mieke Miner kept me honest and, somehow, made me smile too. I'd like to thank Angelita Repetto and Paola Langer for keeping me optimistic and joining me on this rollercoaster that is academia. And finally, my mom, dad, and brother were always so supportive of anything and everything I've wanted to do. I wouldn't be who I am today without that kind of unconditional support.

I would be remiss to not thank both Stella and Chumbi too, they're the best dogs I've ever had the pleasure to know.

A final thanks to the Galois Group at UC Davis for providing a public service by creating and posting the latex template that I used for this dissertation.

Dissertation Chair: Christopher D. Hare Sam Fuller

Advances in Machine Learning for Experimental Research

Abstract

How do individuals conceptualize and operationalize psychological and political phenomena to inform their opinions, beliefs, and behavior? How do we best interrogate these questions and how do new machine learning methodologies inform and alter this research? In three chapters I offer a first attempt at leveraging these methods to answer substantive questions, introducing a new method, and providing guidelines for researchers on how to use these methods for their own research. In the first chapter, I and my coauthors explore the psychological foundations of policy responses to partisan endorsements/cues. Using the combination of a novel survey experiment, multidimensional scaling, and the causal forest method we evaluate the competing importance of motivated bias and affect transfer. This intersection of design, unsupervised machine learning, and causal machine learning leads us to divergent results, compared to the previous literature. We find that affect, not ideology, is the sole moderator of out-party partisan cueing processes. In my second chapter, we introduce a new machine learning method, with accompanying software, for detecting randomization issues in experiments. Leveraging the predictive power of machine learning in combination with the inferential power of permutations, we provide an approach that vastly outperforms standard methods (e.g., balance tables and F-tests). We illustrate this superior performance across simulated, real, and even fabricated data. Importantly, our method requires no ex-ante model-specification and successfully identifies multi-dimensional covariate imbalance where standard methods do not. While chapters one and two highlight the distinct benefits of adopting and integrating new causal machine learning methods, there exists little-to-no guidance on how and when any given researcher should use these methods. Consequently, in my third chapter I introduce principled guidelines for the application of causal machine learning to experiments. This includes design and analysis considerations, especially in relation to the detection and calculation of heterogeneous treatment effects. The goal of this chapter is to reduce the confusion surrounding these methods and their black-box nature and to illustrate the significant benefits they can provide to researchers when used in a principled, transparent way. Overall, these three chapters contribute to the literature on both political behavior and experimental research by introducing, integrating, and applying causal machine learning to experiments in political science.

Contents

Introduction 1

Chapter 1. Affect, Not Ideology:
The Psychological Mechanisms of Partisan Information Processing 3

Bibliography 29

Chapter 2. The Balance Permutation Test:
A Machine Learning Replacement for Balance Tables 33

Bibliography 55

Chapter 3. Causal Machine Learning for Experiments:
Guidelines for Effective and Transparent Research 58

Bibliography 88

Discussion 92

Appendix A. Supplemental Material for Affect, Not Ideology
The Psychological Mechanisms of Partisan Information Processing 95

Appendix. Bibliography 107

Appendix B. Supplemental Material for The Balance Permutation Test:
A Machine Learning Replacement for Balance Tables 108

Appendix. Bibliography 122

Introduction

Experiments have long been a cornerstone of social science research, and consequently there is a deep well of literature on how they should be designed and administered. While research on the design of experiments is well-established—there is no remedy to a poorly-designed experiment—new research on the estimation of effects recovered from experiments has blossomed. Despite the growth in these methods, their use has not spread in a similar way in any of the social sciences.

We are in the midst of a paradigm shift in the way we design, conduct, and execute research. Researchers, in both observational and experimental studies, are increasingly taking advantage of the growth of non-parametric, flexible machine learning methods to enhance inference into data. These methods are enabling the investigation of research questions in robust and novel ways, often impossible or impractical using standard methodologies. This includes detailed, nuanced, and rigorous subgroup analyses as well as the detection of randomization issues. The potential of these methods, simply put, is profound. However, the accessibility of these methods is often poor and their adoption across the discipline has suffered as a consequence.

Far too many methods, while being useful in their own right, fail to reach many practitioners who would benefit significantly from their use. Even when methods are relatively approachable, guidelines and recommendations for how to use these methods in a rigorous and reliable manner are rare or nonexistent. This is of particular concern given the significant differences between these models and standard methods (e.g., sample-splitting and tuning procedures). Furthermore, above and beyond accessibility, principled and clear guidelines for how these methods should be used and reported are nearly nonexistent and this absence continues to fuel legitimate concerns about the "black box" nature of ML methods.

Outline

This dissertation seeks to illustrate the benefits and potential of these methods for experimental social science when used in a transparent, principled manner. To do this, I introduce three chapters:

"Affect, Not Ideology," "The Balance Permutation Test (BPT)," and "Causal Machine Learning for Experiments." These chapters, respectively: illustrate the benefits of causal machine learning by applying it to the study of American political psychology; introduce a new causal ML method for the detection of randomization issues; and provide detailed guidelines for how causal ML methods should be used and incorporated into experimentalists' research.

My first chapter leverages the capabilities of causal machine learning to interrogate whether ideology or affect are more important in determining responses to partisan policy cues. Leveraging a novel survey experiment alongside causal forest, we are able to disentangle the competing mechanisms of motivated bias and affect transfer. The combination of the experimental design and the ability of causal forest to identify and estimate heterogeneous treatment effects represents the main contribution of this paper. Contrary to previous findings, we find that we find that, for only out-party cues, affect is the only covariate that influences partisan cueing processes. This suggests that affect transfer dominates partisans' bias and that ideology has little-to-no influence.

My second chapter introduces a new method that vastly outperforms standard methods, including difference-in-means and F-tests, at identifying randomization issues in experiments. Using causal forest, the BPT is able to non-parametrically predict treatment assignment using the entire space of pretreatment covariates. Even if randomization issues manifest along two dimensions (e.g., PID and gender) the BPT is able to detect these issues. What is more is that any minor-to-moderate randomization issue is easily addressed by using the recovered treatment propensity scores in an estimation framework called doubly-robust machine learning (DRML). In short, DRML adjusts estimates based on recovered treatment propensities, leading to more robust estimates (as we also illustrate in this chapter). Additionally, as a side-benefit, BPT is also able to easily detect data fabrication and fraud.

Finally, given the lack of guidance for how one should use these methods to design and analyze experiments, my third chapter provides detailed guidelines. These include the measurement of pretreatment covariates, randomization issue detection, treatment effect estimation, subgroup analysis, and the presentation of results. While the use and application of these methods can be daunting, the aim of this chapter is to both reduce the barrier to entry and to ensure that researchers do not make (significant) mistakes or misinterpretations when using causal ML methods.

CHAPTER 1

Affect, Not Ideology:
The Psychological Mechanisms of Partisan Information Processing

Sam Fuller,* Nicolás de la Cerda,† & Jack T. Rametta*

*University of California, Davis

†University of North Carolina, Chapel Hill

How do individuals process political information? What psychological mechanisms drive partisan bias? In this paper, we evaluate the extent to which partisan bias is driven by affect or ideology in a three-pronged approach informed by both psychological theories and recent advances in methodology. First, we use a novel survey experiment designed to disentangle the competing mechanisms of motivated bias and affect transfer. Second, we leverage multidimensional scaling methods for latent variable estimation for both partisan affect and ideology. Third, we employ a principled machine learning method, causal forest, to detect and estimate heterogeneous treatment effects. Contrary to previous literature, we find that affect is the *sole moderator* of partisan cueing processes, and only for out-party cues, suggesting that affect transfer dominates partisans' bias. These findings not only contribute to the literature on political behavior, but underscore the importance of careful measurement and robust subgroup analysis.

How do individuals process political information? This question has been foundational to the study of public opinion, including understanding attitude (in)stability and mass-elite linkages, since the dawn of the earliest voter surveys (Berelson, Lazarsfeld, and McPhee 1954; Campbell et al. 1960). Within this space, there is significant research on understanding simple elite-level policy cues (for a summary of most of the previous literature on partisan cues see Bullock 2020). Overwhelmingly, research has found that partisans respond positively to in-party cues and negatively to out-party cues. However, there is significant variation in treatment effects depending on the policy (Tappin 2022), an individual's need for cognition (Bullock 2011), their political knowledge/awareness (Kam 2005), and both partisanship strength and ideology (Barber and Pope 2019).

While research in psychology and political science consistently shows that citizens process information in biased ways, it is still unclear *how* and *why* individuals are influenced by political, partisan information. In general, we know that our motivations determine the kind of information we pay attention to and the way in which we incorporate and evaluate it. Reasoning is always motivated in the sense that it is directed towards a goal (Kunda 1990). Specifically, prior information plays a key role in guiding how we evaluate new information in at least two ways. First, when faced with new evidence, individuals are more likely to accept information that is consistent with their previous beliefs while being less likely to accept information that challenges their priors (Ditto and Lopez 1992). Second, cognition is often affectively charged: concepts stored in long-term memory are associated with positive or negative affect. These affectively charged concepts shape the way in which we process new information in that, similar to priors, information from sources we favor is more likely to be incorporated whereas information from sources we dislike is instead less likely to be incorporated (Lodge and Taber 2013).

In politics, these two mechanisms often lead to similar outcomes. Partisan reasoning—reasoning driven by the need to reach a conclusion that is consistent with one's own party (Lodge and Hamill 1986)—can be explained by both motivated bias and affect transfer. For example, Republicans may support newly introduced legislation because it is consistent with traditional conservative values (motivated bias) or because the legislation was introduced by a group of Republican House members, triggering positive affect (affect transfer). However, because these two mechanisms often work together, at an empirical level it is very difficult to disentangle their unique contribution to attitude bias.

As a consequence, we know much less about the psychological mechanisms that underlie information processing and how individual characteristics moderate it. If partisan information processing is driven by motivated bias, we should expect ideology to play an important role in shaping it. Conversely, if affect transfer is a more relevant causal path, then partisan affect should be a key moderator. Because both ideology and affect vary across the entire population and within each party, we would expect that not all partisans would have the same propensity to follow party cues.

However, evaluating the relative importance of motivated bias and affect transfer involves significant empirical challenges. Particularly, partisan ideological sorting has led to a strong association between partisan affects and ideology which, in turn, hinders researchers' capacity to separate them at the empirical level. Furthermore, in order to establish the relative relevance of each mechanism, we need fine-grained measures that allow us to accurately capture the micro-dynamics of motivated bias and affect transfer.

As we summarized above, there is significant work on partisan bias. However, much of it has been limited by two primary issues: measurement and methodology. Generally speaking, standard practice for exploring the moderators of partisan cues—or nearly any moderator in political science experiments—has been two: First, measure potential moderators before the treatment (pretreatment covariates); second, estimate conditional average treatment effects (CATEs) using an interaction term between the treatment indicator and covariates of interest in a series of regressions including or excluding other control variables; and third, examine the statistical significance of the interaction term, often through coefficient plots.

This approach, however, is *necessarily* limited by the nature of simple interactive regression models run on an entire experimental sample. These models can produce spurious conclusions due to omitted interactive bias (Blackwell and Olson 2022), a lack of statistical power (Arel-Bundock et al. 2022), specification searching, a lack of sample-splitting, and other issues (Ratkovic 2021). Furthermore, a significant portion of the previous literature uses unadjusted, self-reported measures, particularly for strength of party-identification and ideology. Research suggests that these unadjusted measures are inaccurate or incomplete (e.g., Hare et al. 2015; Bankert 2020). Simply put, as is the case with any behavioral research, it is critical to robustly measure and estimate latent variables.

Our empirical strategy is uniquely tailored to address 1) the significant complications with exploring the underlying mechanisms of motivated reasoning and 2) the measurement and methodological issues that plague experimental political behavior research.

First, we use a cueing experiment to generate instances of partisan bias in which motivated bias and affect transfer should lead to divergent outcomes. Participants in our study were exposed to three hypothetical legislative proposals. While the bills were selected to cover the liberal-conservative ideological spectrum, we randomly assigned a Republican, Democratic, or non-partisan cue to each one of the bills. This allowed for the combination of conservative bills supported by Democratic legislators and liberal bills supported by Republicans. In both cases, we should expect affect transfer and motivated bias to lead to opposite outcomes (for both Democrats and Republicans), thus providing us with an opportunity to disentangle the moderating capacity of affect and ideology.

Second, we rely on robust latent estimates of partisan affect and ideology. To measure ideology, we leverage Bayesian Aldrich-McKelvey scaling to correct for differential-item functioning (DIF) and recover a robust estimate of individuals' ideology. For affect, we use ratio unfolding on both a set of partisan feeling thermometers and partisan trust scores as well as a graded response model (GRM) on a set of social distance questions to recover thermometer-, trust-, and social distance-level affect. Importantly, by using scaled measures of both affect and ideology, we are able to recover more reliable estimates, thus, allowing us to better compare the relative effect of each.

Finally, we use a recently developed method to study heterogeneous treatment effects: causal forest. Causal forest is a machine learning method built to detect treatment effect heterogeneity, and subsequently calculate conditional average treatment effects (CATEs) using a doubly robust estimation procedure. This method—combined with a principled estimation procedure and subsample splitting building on state-of-the-art guidance for subgroup analysis (Ratkovic 2021)—has numerous advantages over traditional interactive regression models (Wager and Athey 2018). In brief, our approach is both more agnostic *and* more conservative for detecting treatment effect heterogeneity while allowing for arbitrarily complex interactions and nonlinear relationships between our covariates of interest. Additionally, causal forest not only allows us to directly compare the relative importance of affect and ideology on information processing, but also, to compare them to other covariates such as interest in politics (Kam 2005) and spatial knowledge (Palfrey and

Poole 1987). Last, the "saturated" nature of causal forest models belies the possibility of omitted interactive bias (Blackwell and Olson 2022), a serious concern for other research that attempts to uncover treatment effect heterogeneity with simple interactive regression models.

In line with previous research, we find that partisanship has a strong effect on information processing. Both Democratic and Republican cues strongly shape policy support. However, contrary to both recent and seminal work on cueing effects, we find that while in-party cues do have positive effects, they are not conditioned by affect, ideology, or any other covariates. We believe that these findings, which run against core findings in the literature, are the result of using fine-grained measures of affect, ideology, and political knowledge as well as our use of finely tuned causal forest models. Additionally, we find that the effect of out-party cues is conditioned by several covariates, with partisan affect being the most important moderator and ideology being relatively unimportant. We take this as evidence that affect transfer plays a more relevant role than motivated bias on information processing in contemporary America, at least with relatively low-salience, moderate policy proposals. Finally, contrary to what previous research suggests, we find little evidence of political interest or spatial knowledge as a source of treatment effect heterogeneity.

These findings are concerning but in line with much of the literature on affective polarization in the United States. In an ideal world, policy support would be driven by ideology, not partisan affect. However, our findings suggest that policy support is mostly driven by in- and out-group distinctions and affect. Contrary to previous work, not partisan identity, but partisan affect fuels responses to *out*-party cues but has no role in partisans responding to cues from their own party.

This has significant ramifications for mass-elite linkages. Despite ideological congruence between individuals and policies, support was both increased and decreased by in- and out-party cues, respectively. Furthermore, it appears that parties and elites can easily decrease support for a policy—despite ideological congruence and especially for the most affectively charged partisans—by highlighting out-party endorsements. This suggests that, despite recent literature that shows that mass opinion informs elite political behavior (Canes-Wrone 2015; Rogowski 2023), within-party elites have significant power in influencing behavior, particularly by informing supporters of the out-party's support.

Forming Political Opinions

(Non)Partisan Information How do partisans form or shift their political opinions? The "issue evolution" literature on new issues is well-understood, where initial opinions on a new issue are informed by non-partisan information (Carmines and Stimson 1980; Cochran and Sanders 2009). However, with solidified issues, previous literature suggests that partisans and ideologues use motivated reasoning to form their own opinions (Leeper and Slothuus 2014). Recent research supports this idea, with both motivated bias and affect transfer influencing how individuals integrate information to form or shift their political opinions, for both newer (e.g., the COVID pandemic) and existing (e.g., healthcare and welfare) issues (Ditto et al. 2019; Druckman et al. 2021; Slothuus and Bisgaard 2021).

However, other research does suggest that factual information plays a greater role in informing political opinions than we may think, based on the literature summarized above. Specifically, Wood and Porter (2019) find that there is no evidence of a backfire effect, across 52 different issues. Further research has confirmed these findings, suggesting that factual information, presented in a nonpartisan way, plays a large role in opinion formation and change (For a summary of the previous literature see Swire-Thompson, DeGutis, and Lazer 2020). However, other research suggests that framing information, even in an ideologically amenable way, on certain issues like universal basic income (UBI) fails to influence partisans (Yeung 2022). However, other research finds the exact opposite impact on the same issue of UBIs, when framing with morality (Thomas et al. 2023). In general, it seems, information *can* have an impact, but framing and quality heavily influence its potential impact when provided from a nonpartisan source.

Partisan Cueing Despite this, political information often comes from partisan sources. Furthermore, partisan cues are often devoid of factual information. Instead, partisans are often simply informed of political elites and parties endorsements, not why they endorse a given policy. This brings us to the cueing literature, summarized below. Before we address that, however, what evidence exists for elites following mass-opinion rather than elites shaping mass-opinion.

Previous research has consistently identified the large influence of elite partisan cues on public opinion and policy support. In general, in-party policy cues (e.g., an elite Democratic cue to a Democratic Party supporter) increase support for a given policy whereas out-party policy cues

(e.g., an elite Republican cue to a Democratic Party supporter) decrease support (Bullock 2011, 2020). However, there is important variation regarding effect sizes, driven by the type of policy issue and individual-level characteristics.

On one side, Tappin (2022) shows significant variation in elite partisan cueing effects across 34 policy issues, finding that differences in effect sizes across studies can be largely explained by between-issue variation. Additionally, Clifford, Leeper, and Rainey (2023) identifies differences in cueing effects between social, foreign, and economic issues. Both the category of policy and the average level of awareness respondents had for the policy influenced the effect sizes, with low-awareness, economic policy cues having the largest effects and high-awareness, social policy cues having smaller effects. On the other hand, at least three broad categories of individual-level moderators have been examined in the American cueing literature: 1) political sophistication (often operationalized as political knowledge), 2) "need for cognition," or how much individuals prefer to think about issues, and 3) issue salience, that is, the relative importance a respondent gives to the particular issue (Bullock 2020).

Prior research provides strong evidence for the relevance of individual-level moderators on partisan cueing. First, studies consistently find that as political sophistication increases the effect of cues decreases (Kam 2005; Boudreau and MacKenzie 2014; Barber and Pope 2019). Second, more recent work, particularly Bakker and Lelkes (2018), finds that higher levels of need for cognition correspond to larger cueing effects. And third, the evidence is mixed on the influence of issue salience on cueing effects. While some research finds party cues to have stronger effects on low-salience issues (e.g., Arceneaux and Vander Wielen 2017; Mummolo, Peterson, and Westwood 2021), others find that either salience plays no role (Boudreau and MacKenzie 2014) or that, in fact, high-salience issues have greater cueing effects (Slothuus and De Vreese 2010). However, across the cueing literature, there are significant effects on high-salience issues (Berinsky 2009; Bullock 2020).

Although the previous literature has demonstrated the importance of these individual-level moderators, there are two serious gaps in previous research: 1) affect as a moderator of partisan cues and 2) methodological approaches to assessing treatment effect heterogeneity. First, partisan affect has been understudied in previous research. While much attention has been given to understanding the structure, nature, and implications of partisan affect and identity (e.g., Bankert 2020; Iyengar

et al. 2019; Druckman and Levendusky 2019), particularly its rise over time (Iyengar, Sood, and Lelkes 2012), it has yet to be examined as an individual-level moderator for cueing effects. This should be surprising given the dual-process model's emphasis on not only previous considerations, but individuals' *affect* towards the source of the cue or information (Lodge and Taber 2013). So far, much of the work has used individual's self-reported partisan identity and strength as proxies for affect. Although affect and identity are undoubtedly related, identity does not accurately predict in- and out-party affect (Bankert 2020). Accordingly, the consistent link between partisan identity strength and the effect sizes of cues is likely arising from this strength being a rough measure of in- and out-party affect.

Second, based on recent research in machine learning for experimental analyses, the methodology used by previous work for determining the moderators of partisan cues is limiting. Simply put, interactive regression models without sample splitting are not sufficient for, first, identifying treatment effect heterogeneity and, second, calculating CATEs (Ratkovic 2021).[1] While many results are likely to be confirmed even when using these advanced methods, many are also likely to be contradicted. As a result, we should be more critical of experimental conclusions—in this literature and in experimental research generally—that hinge on conditional treatment effects estimated using interaction terms.

Who Follows Whom? Mass-Elite Linkages Simply put, elites can heavily influence partisan opinions. Given this, however, we must also ask if elites respond to mass-level opinions. While some connection between voters, and specifically partisan constituencies, and elites has long been theorized in the American politics literature (see Mayhew 1974; Fenno 1977), more recent literature has attempted to quantify the degree to which mass opinion shapes elite's policy stances. This literature is mixed on the degree to which mass-opinion informs elite-opinion, and a significant portion of the literature highlights the feedback effects that arise from existing policies influencing policy support (see Campbell 2012, for a summary of much of the previous literature here). However, the literature generally agrees that there is *some* influence from the bottom-up and that the 1960s to 70s were a marked high-point in mass-influence (Canes-Wrone 2015).

[1]For interested readers, please see our methods section for details on the limitations of standard methods and the benefits of using ML.

Recent research has begun to push back on this idea of low responsiveness, however. Very recent work by Caughey and Warshaw (2022) makes a sweeping argument that, while often slow, state-level responsiveness is significant to mass-preferences. Furthermore, at the national level, recent literature even finds significant influence in the foreign policy realm, a space largely considered to be dominated by elite opinions (Heinrich, Kobayashi, and Long 2018).

Affect, Ideology, and Cognition

As we have shown above, the literature on political cognition has systematically shown that partisanship plays a fundamental role in shaping information processing. Both Democrats and Republicans are more likely to accept information that is congruent with their partisan loyalties (Ditto and Lopez 1992), leading to sharp gaps in perceptions of political events (Bartels 2002).

However, it is not clear what key mechanisms drive these perceptual differences. According to the dual-process model of information processing (Lodge and Taber 2013), perceptual differences arise from unconscious affective processes that then shape conscious and cognitive reactions to political events. The way in which we evaluate political events is deeply related to affectively charged, difficult-to-control processes. Lodge and Taber (2013) highlight two channels through which information processing becomes affectively charged: (1) hot cognition and motivated bias, and (2) affect priming and affect contagion. Both of these channels operate at an unconscious level and are affectively driven, but their theoretical underpinnings are distinct. While the first channel is related to feelings that are intrinsic to the stimulus, the second channel is related to affect that is merely incidental. In simple terms, on one side, hot cognition and motivated bias arise from the content of the message: the nature of a given policy or the stance regarding a particular issue or debate. On the other, affective priming and contagion are more related to the way in which a message is conveyed and who the messenger is.

This means that partisan perceptual gaps can arise through two commonly related but distinct channels, the content of a particular message and the characteristics of the messenger. We expect these two channels to have distinct effects on the way in which citizens process information.

Prior attitudes play a fundamental role in the first of these two channels: motivated bias. The way in which citizens feel about an issue can positively or negatively charge new information. Individuals are more likely to accept information that is consistent with their prior views and to reject

inconsistent information. While consistent information is quickly incorporated, individuals tend to use more cognitive resources interrogating (and often arguing against) inconsistent information. That is, when people feel strongly about an issue, they tend to engage in both confirmation and disconfirmation biases (Taber and Lodge 2016). Consistent with this theoretical model, previous literature on partisan cueing has found that when individuals receive information about the policy positions of either in- or out-party elites, they adjust their positions (Bullock 2011, 2020).

The second channel, affect transfer, is driven by elements that are incidental to a particular stimulus, particularly the messenger's identity and characteristics, and how the message is delivered. In this case, positive or negative affects towards whoever delivers a message can increase or decrease the likelihood that the information contained in the message is incorporated or rejected. This means that individuals can negatively charge a message, even in cases when they agree with the policy content. In fact, previous research has shown that individuals' partisan affect can increase the probability of accepting in-party congruent misinformation and decrease the probability of accepting out-party congruent misinformation (Jenke 2023).

Because these two channels have different cognitive underpinnings, we expect each one of them to operate independently and to have distinct effects on the way in which individuals process political information:

- Hypothesis 1: An (out-)in-party cue will (decrease) increase support for a given policy regardless of the ideological bent of the policy.
- Hypothesis 2: Liberal and conservative ideological positions increase the support for ideologically congruent policies, regardless of what partisan group supports the policy.
- Hypothesis 3a: Positive in-party affect increases support for policies that are supported by a co-partisan group, regardless of the particular content of the policy.
- Hypothesis 3b: Negative out-party affect decreases support for policies that are supported by an out-party group, regardless of the particular content of the policy.

Finally, it is worth highlighting that the two channels through which information becomes affectively charged commonly work in tandem. Because the content of a message is usually related to the characteristics of the messenger, when citizens are exposed to information that is congruent with their priors, it usually comes from a source that is associated with positive affect. In practical terms, Democrats tend to support liberal policies and Republican tend to support conservative

ones. Accordingly, separating the effects of these two channels is particularly challenging at an empirical level. In the next section, we present an experimental design that is uniquely tailored to confront these challenges.

Experimental Design

Our experiment follows a relatively standard cueing experiment, where individuals are presented with three bills (randomly ordered):

{Cue} has proposed a bill to increase the federal minimum wage to $12 an hour from its current level of $ 7.25 an hour. This change is intended to increase incomes for low-wage workers while limiting related costs for affected businesses.

{Cue} has proposed to increase the US Customs and Border Protection budget by 20%. The stated purpose is to increase efficiency and safety at US border crossings.

{Cue} has proposed a bill to increase subsidies for American farmers due to rising competition from markets abroad. This is intended to both protect American farmers and to keep the cost of food low for Americans.

Where "{Cue}" is block-randomized within the party identification blocks of our respondents, specifically Democrats (plus leaners), Republicans (plus leaners), and Independents. Respondents were block randomized into three different groups for each one of the bills: control, Republican treatment, and Democratic treatment. For all three bills respondents were asked "Considering this information, do you favor or oppose this proposal?" and were given approval options "Strongly oppose, somewhat oppose, neither favor nor oppose, somewhat favor, strongly favor." This results in a 5-point scale of policy support for each bill for every respondent.

We selected these proposals based on their general ideological bent, and made sure to have them be relatively "moderate" proposals to ensure that the proposals and the groups that supported them seemed believable. Specifically, the minimum wage bill is generally expected to be seen as liberal, the Border Protection budget increase as conservative, and the farming subsidy as nonpartisan/non-ideological. Because the cues are randomized for each one of the policy proposals, some respondents were presented with a conservative proposal alongside a Democratic cue, and a liberal proposal with a Republican cue.

Figure S1 in the appendix illustrates the baseline policy support of our three policies in the control (nonpartisan group) condition by PID. Our range of support, and the granularity of our support variable, across all three of our policies and within each PID, gives us reason to believe that our experiment is not deeply subject to ceiling or floor effects.

Block Randomization Before the experiment, respondents were first asked for their party identification. This was done to allow for block-randomization in our sample, where individuals within each PID (Democrat, Republican, Independent) were then randomly assigned to treatments for each cue. This was done to ensure that we had adequate power to estimate the effects of each cue, by each policy, by each PID. If standard randomization were used, it could lead to imbalanced PID by cues and policies.

To ensure the success of this randomization approach, we first simulated data before fielding and, after the survey was completed, we checked the distribution of respondent assignments to each cue by each policy and PID, reported in Table S1 in the appendix. The largest difference between any PID, cue, and policy combination is 4% with the vast majority of combinations maintaining a less than 1% difference. Moreover, as simple balance tests can be insufficient to detect systematic attrition or imbalance, we also use a machine learning balance test, the "balance permutation test," introduced by Rametta and Fuller (2024), to further test for randomization. We find no evidence that randomization failed significantly, but we do find some minor class imbalance among Republicans responding to two Democratic cues and one Republican cue. Consequently, we use doubly-robust treatment estimation (detailed later) as a robustness check. The detailed results of this balance test are reported in Figure S2 in the appendix.

Data

Our data comes from a nationally-representative survey of American adults collected online by the survey firm Lucid in September and October of 2022. The dataset contains 3,359 observations, and was collected with quotas based on Census data for gender, race, region, and age. Along with this data, we also collect data on other covariates: demographics, religion, political interest and identification, political attitudes, affect and ideology.

To ensure quality responses from our respondents, we include a single attention check and consent to provide honest and accurate answers. Respondents who failed the attention check or

refused to consent to providing good answers, were immediately prevented from completing the survey and thus entering our dataset. Furthermore, we also check for "straightliners" (individuals who answer quickly and select the same response for all stimuli in a grid-question) among those who did complete the entire survey (Kim et al. 2019). We identify and remove 261 individuals who straightlined in various parts of our survey resulting in a complete dataset of 3,098 respondents.[2]

Methodology

Measurement: Scaling Ideology & Affect We begin by collecting standard measures of both affect and ideology from the American National Election Study (ANES). Specifically, we collect self-reported ideology (on a 7-point Likert scale from "Extremely liberal" to "Extremely conservative") and feeling thermometers for both the Democratic and Republican Parties (0–100 with 0 being coldest and 100 being warmest feelings). Traditionally, a standard partisan affect measure is then calculated by taking the difference between an individual's feeling thermometer towards their in-party and out-party (Iyengar et al. 2019), resulting in a -100–100 point scale.

While the previous literature has tended to use these measures in their raw format (e.g., Barber and Pope 2019), simply taking an individual's self-reported ideology and, in the rare cases partisan affect is measured, raw affect measures, we instead opt to leverage three different scaling methodologies to recover more robust measures of ideology, thermometer-level affect, social distance, and partisan trust. We opt to do this given previous findings suggesting that simple thermometer-level affect does not capture the entire phenomenon of partisan affect (Bankert 2020; Druckman and Levendusky 2019; Kingzette 2020). Furthermore, research has consistently shown that scaling multiple measures of some latent trait is more robust than raw measures and addresses significant issues with measurement error (Ansolabehere, Rodden, and Snyder 2008). These measures and the methodology used to estimate them are explained below for interested readers and summarized in Table 1.1.

Using a set of 15 thermometer ratings spread across both parties, for each respondent we leverage ratio unfolding to recover a measure of thermometer-level affect (Mair, Groenen, and Leeuw 2022).[3] Using this method and these measures we recover a one-dimensional representation

[2]This sample size is still more than sufficient to satisfy 90% power for the average treatment effects reported in this study.

[3]For a list of these thermometers, see Section 2 of the appendix.

of affect, recovering both locations of individuals and stimuli. These respondent locations are used as our thermometer measure of affect.

To measure partisan trust, we follow Druckman and Levendusky (2019) and ask respondents "How much of the time do you think you can trust these groups to do what is right for the country?" for six stimuli, the Democratic (Republican) Party, Democratic (Republican) Party candidates and elected officials, and Democratic (Republican) Party voters on a five point scale: "Almost never, once in a while, about half of the time, most of the time, and almost always." We then use the same scaling method used above with our thermometers, ratio unfolding, on all six stimuli to recover a single dimension scale of partisan trust. Furthermore, similar to above, we also generate two additional scales on just Democratic and Republican stimuli, respectively, and consider them to be in- and out-party measures of affect for both Democratic and Republican respondents.

Again, to measure the level of social distance individuals feel towards not only their out-party, but their in-party as well, we adapt questions from Druckman and Levendusky (2019), and ask the question "How comfortable would you be having a Democratic (Republican) Party supporter as a (stimulus) for 5 relationships; "Fellow citizen, neighbor, close personal friend, child in-law, and romantic partner;" using a 4-point scale: "Not at all comfortable, not too comfortable, somewhat comfortable, and extremely comfortable." This gives us five responses per-party for each respondent, ranging from most distant, fellow citizen, to closest, romantic partner.

Given that we can conceive of each of these stimuli as being more socially closer/distant, we can liken these sets of responses to a difficulty parameter, which directly motivates us to utilize a graded response model, a type of item response theory (IRT) model that can handle ordinal data. Using this model on each set of party-level questions, we recover two single-dimensional scales of social distance feeling toward Democratic and Republican Party supporters, respectively. We then take the difference between these two measures (specifically, Republican social distance minus Democratic social distance) to recover a single dimension of differential social distance affect, with those who feel socially close to Democrats and not to Republicans being located on the far left of the scale, and those who feel socially close to Republicans, but not Democrats, being located to the far right of the scale.

Finally, alongside self-reported ideology, respondents were also asked to ideologically place other stimuli on the same 7-point scale. Specifically, we asked respondents to place 6 stimuli:

- The Democratic (Republican) Party
- Democratic (Republican) Party Supporters
- Joseph R. Biden
- Donald J. Trump

Given this data, along with individuals' self-placements, we leverage Bayesian Aldrich-McKelvey (BAM) scaling (Hare et al. 2015) to recover more robust ideological estimates that account for differential-item functioning (individuals having different perceptions of what a given scale means). These result in ideology estimates of every respondent bounded from -11–11. Furthermore, following Palfrey and Poole (1987) we calculate the correlation between the recovered estimates of the ideological positions of our stimuli and each individual's placements to generate a spatial knowledge/information index from -1 to 1. Simply put, the higher the correlation the more informed the respondent is about the ideological locations of our stimuli and thus the American ideological spectrum.

Overall, and especially compared to the previous literature, these robust and various measures of affect and ideology give us the ability to more confidently calculate conditional treatment effects in our sample. Furthermore, using continuous, scaled measures of all of these phenomena allow us to compare their relative importance on a more level playing field. This, combined with our causal forest analysis, detailed below, give us a great deal of confidence in our results.

Causal Forests In general, machine learning methods have been employed for the task of prediction, and not necessarily for inference. However, recently a variety of machine learning algorithms have been adapted specifically for different causal inference problems (Athey and Imbens 2016; Athey and Wager 2019; Chernozhukov et al. 2022).[4] In this paper, we employ one of these new methods, causal forests, a method designed to detect and estimate treatment effect heterogeneity.

Causal forests adapt the random forest estimator (Breiman 2001). Random forests are an ensemble of individual decision trees optimized to minimize prediction error (e.g., MSE, AUC, etc.). In contrast to these random forests, causal trees instead optimize for maximum treatment effect heterogeneity. Additionally, the causal forest employs honest subsample splitting: half of the data is used to choose splits in the causal trees, and the other half is used to estimate the leaves. Subsample splitting ensures that the causal trees used to build the forest are not overfitting

[4]See Grimmer, Roberts, and Stewart (2021) for more details.

TABLE 1.1. Description of Conditioning Variables and Scaling Procedures

Variable	Raw Measurement	Scaling Procedure
Ideology	7-category Likert scale ("Extremely liberal" to "Extremely conservative")	Bayesian Aldrich-McKelvey Scaling (BAM)
Thermometer Affect	0–100 scale for both parties (multiple stimuli in each)	Ratio Unfolding
Social Distance Affect	Multiple social distance questions for different potential acquaintances	Graded Response Model (GRM)
Partisan Trust	5-category Likert scale for both parties ("Almost Always" to "Almost Never"; multiple stimuli in each)	Ratio Unfolding
Spatial Knowledge	Ability to place stimuli	Derived from Bayesian Aldrich-McKelvey Scores

on the training data. Moreover, forms of sample splitting are generally advisable to avoid bias in subgroup analysis (Ratkovic 2021). Once the causal forest is estimated, to test for treatment effect heterogeneity we employ a heteroskedasticity robust, omnibus test developed by Chernozhukov et al. 2022. In brief, the test takes an input causal forest model and generates a best linear fit of the CATEs on held-out data. It then uses the results of this linear fit to evaluate the quality of the CATE estimates. This test acts as both an omnibus heterogeneity test and also returns calibration information such that the fit of the causal forest model can be evaluated.

While we believe this estimation procedure is superior to traditional subgroup analysis, in that it is more robust to model specification and more difficult to game, it does come at a cost. Because the causal forest employs sub-sample splitting, it does not use all of the training data to estimate the CATEs. As such, this approach does add some noise relative to subgroup analysis, and can thus be viewed as conservative in some contexts. However in our view, this is a price worth paying, as simple interactive regression models are only superior to causal forest when the model is very well specified, and moreover, that the researcher knows this perfect specification ex ante. This is hardly ever the case, and poorly specified interactive methods can be subject to omitted interactive bias (Blackwell and Olson 2022).

For the purposes of this article, we execute separate causal forest models across the three different policy cues, subsetting by party ID. We also run models that pool across all policy cues. These different approaches have benefits and drawbacks. The pooled models increase available power but assume the cueing effects are similar across policies, while the separate models allow the cueing effects to vary but lose power relative to the pooled approach.

Results

Baseline Results First, using simple difference-in-means, with unadjusted standard errors, we calculate conditional average treatment effects (CATEs) by respondent PID, reporting our results in Figure 1.1.[5] Figure 1.1 is broken down into two panels, with the left representing the pooled Democratic cues and the right representing the pooled Republican cues. Effects for these cues (Y-axis) are presented with 95 and 80% confidence-intervals—thin and thick lines, respectively—and are broken down within each panel by a respondent's self-reported PID (X-axis; leaners are included in weak PIDs). Overall, we find results that are highly consistent with previous research. In general, partisans are influenced by both in- and out-party cues across the pooled policies. Furthermore, the effect sizes and significance increase by the strength of PID, for both Democrats and Republicans. Independents, as we may expect when not controlling for ideology, consistently have null effects across both Democratic and Republican treatments.

One notable trend among these results is the general size of the effects of out-partisan cues (e.g., Republican treatments among Democratic respondents). In general, the effect sizes tend to be larger and the confidence bands further away from zero. These results clearly suggest that PID or, as is more likely, PID as some rough substitute for some dimension of underlying ideological and/or affective beliefs is influencing the effects of these cues. Consequently, we move to further investigate these potential influences.

While these results are promising, in that they are consistent with previous findings surrounding simple partisan policy cues, there are limitations to using standard OLS for calculating CATEs. Specifically, there is no clear test for detecting real treatment effect heterogeneity above and beyond checking the statistical significance of the conditioning variable. Furthermore, the specification of

[5]For a table of these same results, please see section 4.1 in the appendix. Furthermore, we also calculate these effects using a causal forest and bootstrapped standard errors, finding similar effect sizes and confidence intervals. These results are also included in section 4.2 of the appendix.

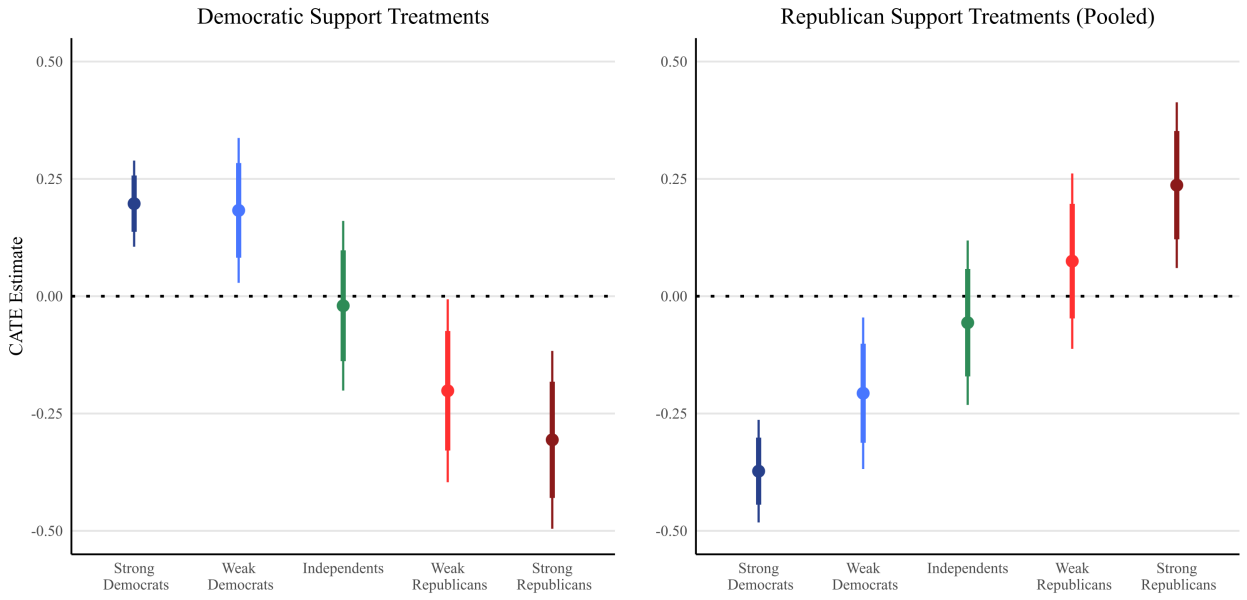


FIGURE 1.1. Pooled Cue CATE Estimates by Respondent and Cue PID

the regression used to calculate any CATEs has significant influence on whether or not these variables do appear statistically significant. Specifically, this manifests itself in two ways: 1) the choice of a simple interaction between the treatment assignment and the conditioning variable of interest or in calculating treatment effects by subsetting each regression to the conditioned variable group; and 2) in the decisions of both what variables to condition on and what covariates should be included in the calculation of any given CATE. Finally, as with any estimation technique, OLS necessarily assumes additive, linear effects, which significantly constrains the estimation of any CATEs.

Given these limitations, and our distinct interest in accurately estimating CATEs for respondents' ideology, affect, and spatial knowledge, we opt to use a causal forest to not only calculate our CATEs, but to first even determine if treatment effect heterogeneity exists in our experiment. The following section proceeds by first detailing the causal forest methodology; second, reporting the results of our causal forest heterogeneity tests and subsequent variable importance score calculations; and finally, reporting our CATE estimations, including interaction effects, for important conditioning variables.

Treatment Effect Heterogeneity and Variable Importance Based on the test detailed above, we check to see if treatment effect heterogeneity exists in our experiment based on our

full set of covariates of interest (see Table 1.1) and other standard demographic variables. These other variables include: political interest, age, sex, race, partisan strength, education, and religiosity. We run this test for each policy separately, subsetting also by PID, reporting the results below in Table 1.2. Across all cues we find distinct evidence for treatment effect heterogeneity, but only for *out-party cues*. This suggests that the effect of an in-party cue on support, for any of these policies, is generally consistent across all of our measured covariates. However, for these out-party cues, for both Democrat and Republican respondents, there appears to be significant heterogeneity.

TABLE 1.2. Treatment Effect Heterogeneity Checks by Policy Cue and Pooled

Treatment	Resp. PID	Min. Wage	Border Funding	Farm Subsidy	Pooled
Democratic Cue	Democrats Republicans Independents				***X***
Republican Cue	Democrats Republicans Independents	*X*	*X*	*X*	***X*** ***X***

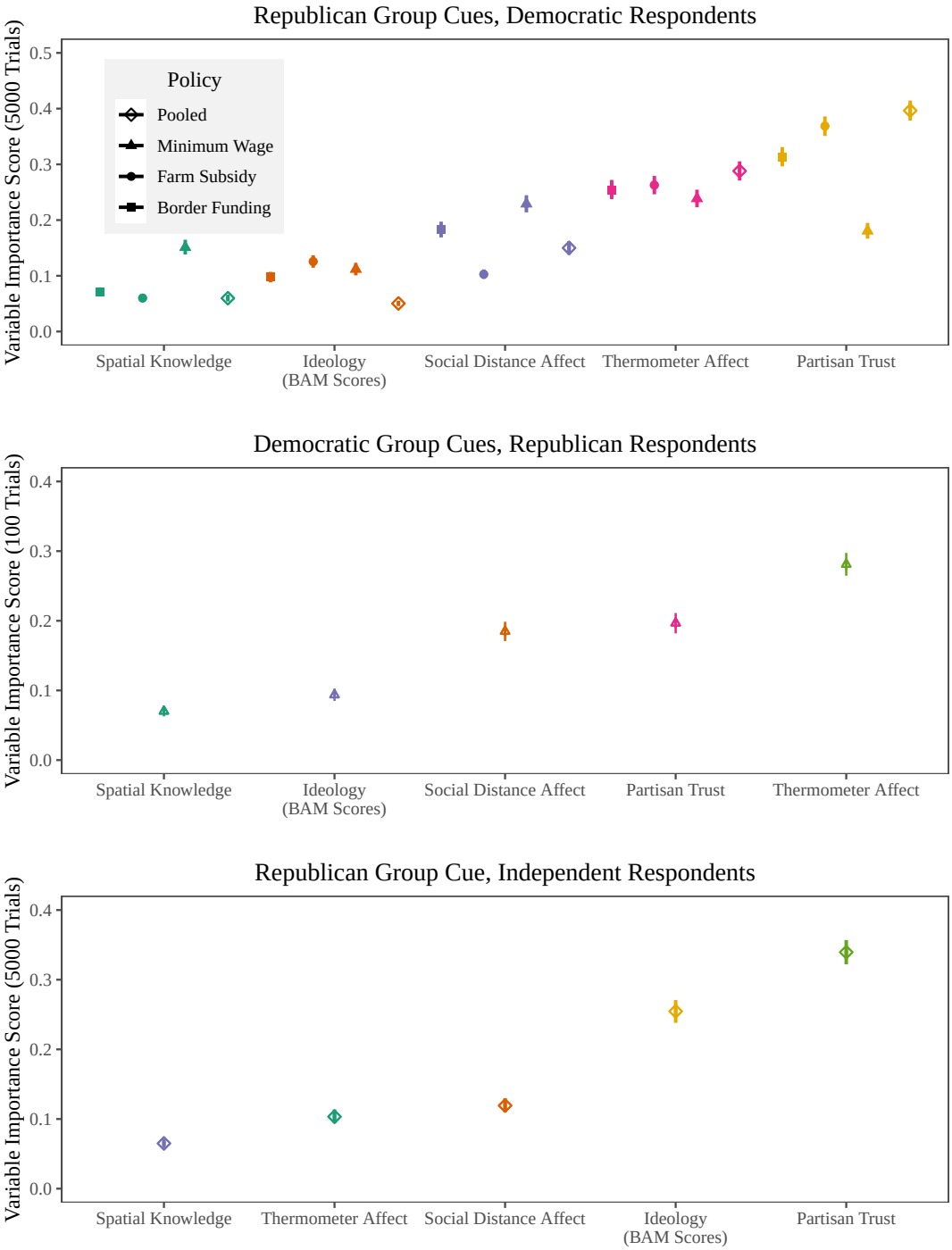
Consequently, we further employ causal forests to determine *which* variables are leading to this heterogeneity, reporting the bootstrapped (5000 trial) calculation of variable importance for predicting heterogeneous treatment effects below in Figure 1.2.[6]

Importance is calculated for each variable as a weighted sum of how often, and at what depth, a variable appears in all of the trees of a given causal forest model. The logic of this is that splits earlier on in any given tree are necessarily more important for prediction than those later, and since each tree is a random selection of variables it is important to sum across all trees in the forest. Finally, the scores themselves are relative, in that the sum of all scores is equal to 1. The higher a variable's score, the higher its relative importance is in accurately predicting heterogeneity in our experiment. For a detailed explanation of the calculation of these variable importance scores, see section 3.4 of the appendix.

As seen in Figure 1.2, consistent across all three policies (and pooled policies), for out-party cues and for Democratic and Republican respondents, our measures of affect, specifically thermometer

[6]We employ bootstrapping here for robustness as changes in the random number generator employed can lead to slightly different variable importance estimates.

FIGURE 1.2. Variable Importance for Estimating Heterogeneous Treatment Effects



affect and partisan trust, are the most important variables for capturing treatment effect heterogeneity. Interestingly, ideology is also a highly important predictor for independent respondents.

Those who are on the right (conservative) part of the scale have positive reactions to Republican cues, whereas those on the left (liberal) part of the scale have negative reactions. This makes sense, in that it appears that ideology for independent respondents acts as almost a proxy for party identification.

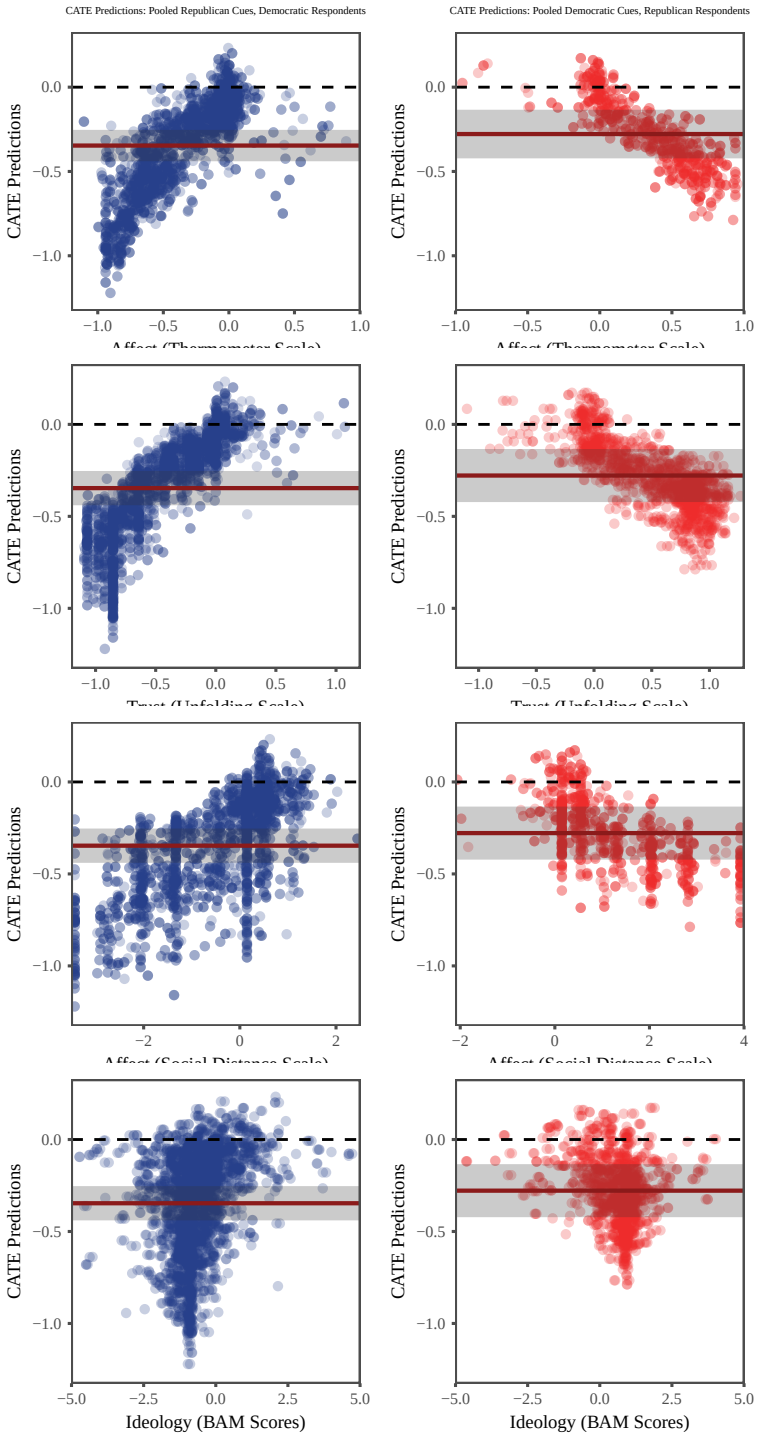
CATE Estimation for Affect and Ideology We proceed in this section by calculating CATEs for affect (thermometer-level affect and social distance) and ideology by both Democratic and Republican respondents for out-party cues. To do this, we map the causal forest CATE predictions over different pre-treatment covariates. These CATE estimates are derived from separate causal forest models trained on the pooled policy cues and subsetting respondents by party ID.

One benefit of using this method, is that we are able to calculate CATEs while including all three of our moderating variables of interest (and our other covariates as well). Consequently, the significant results below are not the effects of these moderating variables in a vacuum, rather they are the effects of these variables when including all other covariates. For all of our CATE estimations we include ideology, thermometer-level affect, social distance, partisan trust, and spatial knowledge as well as all of our demographic covariates. Figure 1.3 shows out-of-bag predictions[7] for individual CATEs for each of our affect variables and ideology (rows) by PID (columns) for pooled, out-party cues. Each panel also shows the ATE (with 95% confidence bands) for the given PID. To put these results in context, recall our dependent variable (policy support) is measured on a five point likert scale from "strongly support" to "strongly oppose." As such, a one-unit movement in either direction reflects a change from, say, support to strongly support, or oppose to strongly oppose.

Clearly, there is significant treatment effect heterogeneity across thermometer-level affect, social distance, and partisan trust. Respondents with the lowest predicted CATEs (i.e., the strongest negative cueing effects) tend to distrust the out-party, rate out-partisans very poorly on thermometers, and report high levels of social distance from out-partisans. In other words, the most affectively polarized respondents have the most extreme response to partisan cues. Conversely, respondents who have relatively high levels of trust in out-partisans, rate out-partisans highly on thermometers, and report low levels of social distance from out-partisans tend to have very small or null cueing effects. While affect across multiple dimensions clearly conditions the cueing effects in our experiment, there is no clear conditioning role for ideology for either Democratic or Republican

[7]For a description of out-of-bag prediction see appendix section 3.2.

FIGURE 1.3. CATEs for All Respondents by Variable and PID

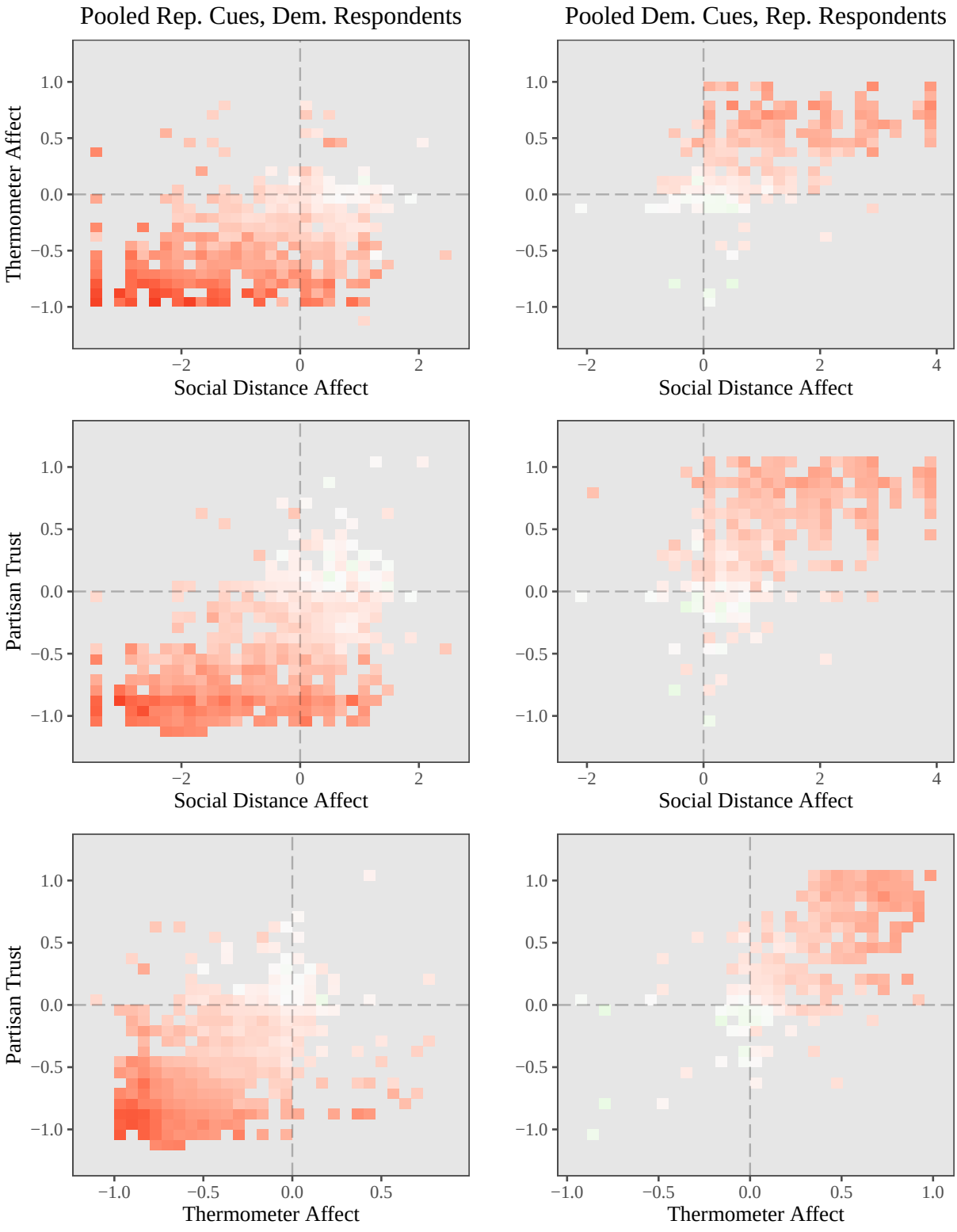


respondents. As the CATE predictions display, respondents that are close in ideological space can have very different response to cues.

Based on the results above, we opt to further explore the interaction CATEs for thermometer-level affect, social distance, and partisan trust. We do this by calculating the partial dependence between these three variables for out-party cues with pooled policies for Democratic and Republican respondents. This is done simply by calculating the CATEs by, for example, a given level of thermometer *and* social distance affect from both pooled, out-party cue causal forests. We choose to utilize partial dependence calculations and plots because they enable, under certain conditions, causal interpretations of interactive, multidimensional CATEs (Zhao and Hastie 2021). Specifically, they enable us to explore the interactions between these variables that have large conditioning effects on ATEs. Figure 1.4 reports the partial dependence plots for thermometer (Y-axis) and social distance (X-axis) affect for out-party cues by PID (columns) for the pooled policies. We estimate and report CATEs only in regions of parameter space for which we have real observations. As such, some cells in the Figure 1.4 where no observations exist are blank (gray).

Clearly, those who are at the extreme end of both affect scales have greater negative treatment effects than those who are less extreme on one or both of the scales. Simply put, those who have higher partisan affect as a sum of both of our measures of affect, have greater negative treatment effects than their counterparts. While most respondents exist within a relatively negative affective space (bottom-left for Democrats, top-right for Republicans), there is significant variation across these measures, with a good deal of observations near the center. These respondents are, in affective terms, very moderate and, consequently, respond very little, or not at all, to partisan cues. This comports with our variable importance results, where ideology is essentially irrelevant to cue responses.

FIGURE 1.4. Partial Dependence Plots for Thermometer & Social Distance Affect



Discussion

In this paper we explore the role of affect and ideology as moderators of information processing. Based on the dual-process model of information processing (Lodge and Taber 2013), we argue that affect and ideology play a distinct role in shaping the way in which political information is processed. While the model assumes that both processes are unconscious and lead to affectively charged processing, they have different theoretical underpinnings. We theorize that perceptual gaps can arise through intrinsic and extrinsic elements of a particular message. In concrete terms, this means that affect and ideology can shape information processing in different ways.

In contrast to the previous literature, we find no evidence of the conditioning effect of partisan strength, spatial knowledge, nor ideology. Furthermore, in our dual-process model framework, we find that affect, *not* ideology, influences the effect of partisan policy cues. It is possible that these results are, at least in large part, dependent on the relatively moderate bent of the tested policies. In a different experiment with more salient, ideological policies, we might expect ideology to play more of a role (at least with predicting baseline policy support). However, our results still indicate that partisan cues have a large and significant effect on policy support. Additionally, we only find significant heterogeneity in cueing effects for out-party cues, whereas for in-party cues cueing effects are relatively constant. These constant effects are in contrast with the previous literature, which suggests in-party cueing effects are moderated by political knowledge. In fact, we find that all of our measures of affect—thermometer-level, social distance, *and* partisan trust—have highly significant, interactive effects on predicting treatment effects. Those who have greater differences in their in- minus out-party affect react much more negatively to out-party cues than their other, less extreme copartisans.

Our contributions to this literature are three-fold: First, we directly evaluate the extent to which two well-theorized mechanisms of biased, partisan information processing shape individual responses to partisan policy cues in the US electorate. Second, we do so by estimating individuals' partisan affect and ideology utilizing scaling methods that better capture these concepts and adjust for differential item-functioning. Previous studies on cueing effects have tended to use rough measures which can lead to less reliable estimates. Third, whereas previous literature relies on simple interactive regression models to estimate CATEs, we exploit a methodology designed explicitly for detecting treatment effect heterogeneity and calculating CATEs: causal forest. This agnostic,

and relatively conservative, approach to detecting heterogeneity, which incorporates all measured covariates, significantly increases our confidence.

Taken together, these results reframe our understanding the psychological and behavioral mechanisms that constitute the effects of partisan cues. They suggest that partisans with high degrees of in-party affect and out-party dislike are heavily influenced by party cues. This in an of itself is troubling given the growth in partisan affect across the United States (West and Iyengar 2022). Additionally, this top-down influence could have severe implications for representation and mass-elite linkages. If individuals can so easily be persuaded about policy support, even on policies that individuals should ostensibly support ideologically, the masses' ability to move elites' policy stances could be severely limited.

Finally, there are promising directions for future research that arise from this work. For instance, additional attention should be paid to understanding how ideology and partisan affect influence opinions on both more salient and more ideologically crystallized policies, such as abortion or immigration reform. Furthermore, researchers could attempt to de-emphasize partisan affect and instead emphasize ideological considerations. Regardless, our findings suggest that there is much more work to be done on understanding the nuances and effects of partisan cues on policy attitudes and support.

Bibliography

Ansolabehere, Stephen, Jonathan Rodden, and James M Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102 (2): 215–232.

Arceneaux, Kevin, and Ryan J Vander Wielen. 2017. *Taming intuition: How reflection minimizes partisan reasoning and promotes democratic accountability.* Cambridge University Press.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D Stanley. 2022. *Quantitative political science research is greatly underpowered.* Technical report. I4R Discussion Paper Series.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Athey, Susan, and Stefan Wager. 2019. "Estimating treatment effects with causal forests: An application." *Observational Studies* 5 (2): 37–51.

Bakker, Bert N, and Yphtach Lelkes. 2018. "Selling ourselves short? How abbreviated measures of personality change the way we think about personality and politics." *The Journal of Politics* 80 (4): 1311–1325.

Bankert, Alexa. 2020. "Negative and Positive Partisanship in the 2016 US Presidential Elections." *Political Behavior,* 1–19.

Barber, Michael, and Jeremy C Pope. 2019. "Does party trump ideology? Disentangling party and ideology in America." *American Political Science Review* 113 (1): 38–54.

Bartels, Larry M. 2002. "Beyond the running tally: Partisan bias in political perceptions." *Political behavior* 24:117–150.

Berelson, Bernard R, Paul F Lazarsfeld, and William N McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign.* University of Chicago Press.

Berinsky, Adam J. 2009. "In Time of War." In *In Time of War.* University of Chicago Press.

Blackwell, Matthew, and Michael P Olson. 2022. "Reducing model misspecification and bias in the estimation of interactions." *Political Analysis* 30 (4): 495–514.

Boudreau, Cheryl, and Scott A MacKenzie. 2014. "Informing the electorate? How party cues and policy information affect public opinion about initiatives." *American Journal of Political Science* 58 (1): 48–62.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45:5–32.

Bullock, John G. 2011. "Elite influence on public opinion in an informed electorate." *American Political Science Review* 105 (3): 496–515.

Bullock, John G. 2020. "The Oxford Handbook of Electoral Persuasion." Chap. Party Cues. Oxford University Press, June. https://doi.org/10.1093/oxfordhb/9780190860806.001.0001. https://doi.org/10.1093/oxfordhb/9780190860806.001.0001.

Campbell, Andrea Louise. 2012. "Policy Makes Mass Politics." *Annual Review of Political Science* 15:333–351.

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes. 1960. *The American Voter.* University of Chicago Press.

Canes-Wrone, Brandice. 2015. "From Mass Preferences to Policy." *Annual Review of Political Science* 18:147–165.

Carmines, Edward G, and James A Stimson. 1980. "The Two Faces of Issue Voting." *American Political Science Review* 74 (1): 78–91.

Caughey, Devin, and Christopher Warshaw. 2022. *Dynamic Democracy: Public Opinion, Elections, and Policymaking in the American States.* University of Chicago Press.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2022. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments.* arXiv: 1712.04802 [stat.ML].

Clifford, Scott, Thomas J Leeper, and Carlisle Rainey. 2023. "Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues." *Political Behavior,* 1–24.

Cochran, John K, and Beth A Sanders. 2009. "The Gender Gap in Death Penalty Support: An Exploratory Study." *Journal of Criminal Justice* 37 (6): 525–533.

Ditto, Peter H, Brittany S Liu, Cory J Clark, Sean P Wojcik, Eric E Chen, Rebecca H Grady, Jared B Celniker, and Joanne F Zinger. 2019. "At Least Bias is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives." *Perspectives on Psychological Science* 14 (2): 273–291.

Ditto, Peter H, and David F Lopez. 1992. "Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions." *Journal of personality and social psychology* 63 (4): 568.

Druckman, James N, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. "Affective pPolarization, Local Contexts and Public Opinion in America." *Nature Human Behaviour* 5 (1): 28–38.

Druckman, James N, and Matthew S Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83 (1): 114–122.

Fenno, Richard F. 1977. "US House Members in Their Constituencies: An Exploration." *American Political Science Review* 71 (3): 883–917.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2021. "Machine learning for social science: An agnostic approach." *Annual Review of Political Science* 24:395–419.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll, and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59 (3): 759–774.

Heinrich, Tobias, Yoshiharu Kobayashi, and Leah Long. 2018. "Voters Get What They Want (When They Pay Attention): Human Rights, Policy Benefits, and Foreign Aid." *International Studies Quarterly* 62 (1): 195–207.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22:129–146.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–431.

Jenke, Libby. 2023. "Affective Polarization and Misinformation Belief." *Political Behavior,* 1–60.

Kam, Cindy D. 2005. "Who toes the party line? Cues, values, and individual differences." *Political Behavior* 27:163–182.

Kim, Yujin, Jennifer Dykema, John Stevenson, Penny Black, and D Paul Moberg. 2019. "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys." *Social Science Computer Review* 37 (2): 214–233.

Kingzette, Jon. 2020. "Who Do You Loathe? Feelings toward Politicians vs. Ordinary People in the Opposing Party." *Journal of Experimental Political Science,* 1–10.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108 (3): 480.

Leeper, Thomas J, and Rune Slothuus. 2014. "Political Parties, Motivated reasoning, and Public Opinion Formation." *Political Psychology* 35:129–156.

Lodge, Milton, and Ruth Hamill. 1986. "A partisan schema for political information processing." *American Political Science Review* 80 (2): 505–519.

Lodge, Milton, and Charles S Taber. 2013. *The rationalizing voter.* Cambridge University Press.

Mair, Patrick, Patrick JF Groenen, and Jan de Leeuw. 2022. "More on Multidimensional Scaling and Unfolding in R: smacof Version 2." *Journal of Statistical Software* 102:1–47.

Mayhew, David R. 1974. *Congress: The Electoral Connection.* Yale University Press.

Mummolo, Jonathan, Erik Peterson, and Sean Westwood. 2021. "The limits of partisan loyalty." *Political Behavior* 43:949–972.

Palfrey, Thomas R., and Keith T. Poole. 1987. "The Relationship between Information, Ideology, and Voting Behavior." *American Journal of Political Science* 31 (3): 511–530.

Rametta, Jack, and Sam Fuller. 2024. *MLbalance.* V. 0.1, January. https://github.com/samjfuller/MLbalance.

Ratkovic, Marc. 2021. "Subgroup Analysis: Pitfalls, Promise, and Honesty." In *Advances in Experimental Political Science,* edited by James N. Druckman and Donald P.Editors Green, 271–288. Cambridge University Press. https://doi.org/10.1017/9781108777919.020.

Rogowski, Jon C. 2023. "Public Opinion and Presidents' Unilateral Policy Agendas." *American Journal of Political Science.*

Slothuus, Rune, and Martin Bisgaard. 2021. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* 65 (4): 896–911.

Slothuus, Rune, and Claes H De Vreese. 2010. "Political parties, motivated reasoning, and issue framing effects." *The Journal of Politics* 72 (3): 630–645.

Swire-Thompson, Briony, Joseph DeGutis, and David Lazer. 2020. "Searching for the Backfire Effect: Measurement and Design Considerations." *Journal of applied research in memory and cognition* 9 (3): 286–299.

Taber, Charles S, and Milton Lodge. 2016. "The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning." *Political Psychology* 37:61–85.

Tappin, Ben M. 2022. "Estimating the between-issue variation in party elite cue effects." *Public Opinion Quarterly* 86 (4): 862–885.

Thomas, Catherine C, Gregory M Walton, Ellen C Reinhart, and Hazel Rose Markus. 2023. "Mitigating Welfare-Related Prejudice and Partisanship Among US Conservatives With Moral Reframing of a Universal Basic Income Policy." *Journal of Experimental Social Psychology* 105:104424.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

West, Emily A, and Shanto Iyengar. 2022. "Partisanship as a Social Identity: Implications for Polarization." *Political Behavior* 44 (2): 807–838.

Wood, Thomas, and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41:135–163.

Yeung, Eddy SF. 2022. "Can Conservatives Be Persuaded? Framing Effects on Support for Universal Basic Income in the US." *Political Behavior,* 1–27.

Zhao, Qingyuan, and Trevor Hastie. 2021. "Causal interpretations of black-box models." *Journal of Business & Economic Statistics* 39 (1): 272–281.

CHAPTER 2

The Balance Permutation Test: A Machine Learning Replacement for Balance Tables

Jack T. Rametta & Sam Fuller*

University of California, Davis

Balance tests are standard for experiments despite their inability to detect randomization issues and covariate imbalance. There is no consensus on how randomization and balance should be checked, and how failures and imbalances should be addressed, if they should be addressed at all. In this article we provide clear guidelines and implement a new method, the "balance permutation test," designed to detect arbitrarily complex randomization failures. Our approach leverages a combination of permutation inference and machine learning to accomplish this task. First, we show how the balance permutation test is able to detect complex imbalance in real, simulated, and even fabricated data. Second, using an extensive set of Monte Carlo simulations, we demonstrate the advantages of doubly-robust estimation over existing methods. Finally, we introduce an `R` package, `MLbalance` implementing our approach. Our hope is this method resolves the debate over how to detect and adjust for randomization issues in experiments.

*Both authors contributed to the writing, analysis, and package construction equally. This is one of several papers by the authors on the development of machine learning tools for social science. The ordering of names reflects a principle of rotation.

In experiments, causal inference depends on random assignment (Fisher 1925). Unfortunately the ways in which randomization issues can arise—either randomization failures and/or covariate imbalance[1]—are numerous and range in severity (e.g., Imai 2005). If the assignment process results in significant imbalance along one or more relevant pre-treatment covariates, then any effect estimation is likely to be biased and could even lead to spurious conclusions. As such, detecting randomization issues is critical. Furthermore, these issues are a potential sign of data fabrication, an increasingly salient issue in light of recent cases of fraudulent research (Berenbaum 2021; Simonsohn, Nelson, and Simmons 2023). So if randomization issues are so dangerous, how should researchers go about detecting and adjusting for them?

The most prolific approach for assessing randomization and covariate (im)balance is the simple "balance test," see Table 2.1 for an example, where descriptive statistics for measured covariates are displayed and assessed for imbalance across treatment and control groups, often with simple statistical tests (e.g., t- or F-tests). These balance tests are widely used in the social sciences and are required for experimental submissions to many journals (see Mutz, Pemantle, and Pham 2019 for a representative sample of articles that use and journals that require balance tests and Gerber et al. 2014 for an example of a journal requirement). Often, if imbalance over a single covariate is detected—which increases in likelihood as the number of measured covariates increases—researchers are instructed to make post hoc adjustments to their analysis to adjust for said imbalance. However, this "Balance Test and Adjust" process is rife with problems that can *generate* bias, regardless of whether or not randomization was truly achieved (Mutz, Pemantle, and Pham 2019).

On one hand, Mutz, Pemantle, and Pham (2019) argue that using balance tests, generally, can result in the introduction of significant bias or, in the worst case, lead to p-hacking/model specification searching. Furthermore, they argue that researchers are better off providing standard information about the experiment: if imbalance is significant enough to justify an adjustment, then it should be apparent from the details of the experiment. In a similar vein, Freedman (2008) argues that unmotivated, ad hoc regression adjustments can lead to unexpected, and unintentional, bias in treatment effect estimates. He thus advocates the presentation of unadjusted treatment effect estimates and discourages regression adjustments on experimental data.

[1]Specifically, by randomization failure we refer to issues in the randomization process that leads to nonrandom or biased experimental groups and covariate imbalance as when pre-treatment covariates are not equally balanced across experimental groups.

TABLE 2.1. Example Balance Table

	Control		Treatment		Diff-in-Means
	Mean	SD	Mean	SD	
Female	0.49	0.50	0.43	0.50	-0.06
Black	0.26	0.44	0.19	0.39	-0.07
Income	51318	10787	49345	9820	-1973*
Democrat	0.47	0.50	0.43	0.50	-0.04
Age	53.10	17.79	49.78	19.49	-3.32*

Note: Data simulated. $*$ corresponds to $\alpha \leq .05$. $N = 1,000$.

However, on the other hand, while there are *severe* problems with the balance test and adjust approach, real randomization issues *do* occur. Successful randomization is an identifying assumption that should be tested, not taken for granted. While some extant research develops methods to verify successful randomization, the literature is disparate and mostly limited to regression-based approaches that do not leverage the predictive power of machine learning (Mize and Manago 2022). Additionally, there is evidence to suggest that well-motivated and well-specified regression adjustments can improve statistical efficiency in a wide variety of settings (Lin 2013).

In this article we ask if there exists a way to square these circles. In other words, can we have our cake and eat it too? Specifically, can a balance test procedure accurately detect and adjust for randomization issues while also avoiding the pitfalls of specification searching and ad-hoc adjustments? We argue yes.

We introduce the "balance permutation test": an intuitive procedure that exploits the predictive power of machine learning (ML) to interrogate random assignment and covariate balance. Furthermore, we provide clear guidelines for researchers to employ this test and steps to adjust for minor to moderate randomization issues. This adjustment process culminates with doubly-robust machine learning (DRML) treatment effect estimation using augmented inverse probability weighting (Robins, Rotnitzky, and Zhao 1994; Zhong et al. 2021), which has the distinct benefits of increasing precision (via powerful ML curve fitting algorithms) while simultaneously restricting researcher degrees of freedom. Our method is also dual purpose: useful for detecting unintentional failures of randomization *and* checking for the fabrication of experimental data. While the fabrication of experimental data in published research is (hopefully) rare, the existence of simple tools to detect fabrication, like those provided in this article, should further deter such activity. Finally, we

provide an accompanying, easy-to-use `R` package, `MLbalance`, that implements this process. Overall, our goals are two-fold: 1) to eliminate the current confusion surrounding what researchers should, and should not, do to assess randomization and covariate (im)balance in "clean" experimental data, and 2) to illustrate the usefulness of causal ML methods for detection and adjustment.

The paper proceeds in three major parts. First, we review the previous literature on balance tests, regression adjustments, and advancements in experimental machine learning methods. Second, we introduce and detail our new method with motivating examples. Third, we illustrate the utility of our approach on real and fabricated data and report the results from an extensive set of Monte Carlo simulations demonstrating the advantages of DRML to existing adjustment methods. We conclude with our suggestions for the application and use of this approach and potential areas for future work in this space.

Unifying balance tests, treatment effect adjustments, & machine learning

"Regression adjustments are often made to experimental data. Since randomization does not justify the models, *almost anything can happen*."

–Freedman (2008)

"If the actual random assignment yields substantial covariate imbalance, it is hardly reassuring to be told that the difference in means is unbiased over all possible random assignments."

–Lin (2013)

The trouble with balance tests and (un)adjusted treatment effects The previous literature enumerates significant issues with the current state of balance tests and regression adjustments. First, balance tests are unfit for detecting real randomization issues (Mutz, Pemantle, and Pham 2019). Second, ad hoc adjustments to the treatment effect estimation procedure can, and do, introduce bias under common experimental conditions (Freedman 2008). In other words, these ad hoc adjustments can ruin treatment effect estimates by introducing bias. In addition, we note that these tests are only useful, if ever, at detecting *unidimensional* covariate imbalance. In truth, a balance table could make an experiment appear to be balanced across some set of covariates, even if there is multidimensional imbalance (e.g., the treatment and control groups have

roughly the same proportion of female respondents and Black respondents, but the treatment group has significantly more Black, female respondents). We point interested readers to Mutz, Pemantle, and Pham (2019) for a detailed breakdown of the numerous flaws with ad-hoc adjustments from balance tests. In sum, balance tests are insufficient for detecting real randomization issues, and can encourage "messy" analyses of "clean" experimental data (in the language of), which can erode researchers' ability to come to causally identified conclusions about treatment effects. To quote Mutz, Pemantle, and Pham (2019), "Absolutely any result can be obtained when there are sufficiently many covariates" when employing the balance test and adjust process.

However, despite these significant concerns with the use of balance tests and ad-hoc adjustments, a large number of authors argue for the benefits of regression adjustments (Lin 2013; List, Muir, and Sun 2022; Negi and Wooldridge 2021; Zhao and Ding 2022). Lin (2013) makes the point for regression adjustment quite simply: if we have reason to believe that there is significant covariate imbalance, taking the naive difference in means average treatment effect (ATE) as the ground truth is "hardly reassuring." In this space, methods and researchers are balancing two, sometimes competing, concerns: efficiency and credibility (Mutz, Pemantle, and Pham 2019). Efficiency refers broadly to the precision of treatment effect estimates, and is often increased by regression adjustments where the adjusting variable is predictive of the outcome variable. On the other hand, credibility is concerned more generally with the believability of the experimental design, its resulting data, and finally the estimated treatment effects. The conflict between these two concerns arises from the ability to obtain more precise estimates with regression adjustments, but in so doing potentially changing what effect the researcher is estimating. In many situations, if these adjustments are implemented, the researcher is no longer recovering a simple, unbiased treatment effect. Instead, they are estimating an effect that may only exist under a very particular, sample-specific condition. As a result, despite their methodological rigor and significant efficiency gains, these adjustment techniques are likely to decrease the credibility of a large swathe of treatment effect estimates.

There are also several disparate threads of literature that attempt to verify successful random assignment and covariate balance in experimental designs. One existing method recommends testing the validity of randomization by modeling treatment assignment as a function of some set of pre-treatment covariates using simple regression models (Mize and Manago 2022). Some applied work

uses bespoke ML models to detect randomization failures (Hübert and Copus 2022). Last, the closest analog to our approach is the "classification permutation test" introduced by Gagnon-Bartsch and Shem-Tov (2019). In brief, their approach proposes applying permutation inference to test for covariate (im)balance in observational designs but still advocates for the inclusion of balance tables in addition to their test results. While these methods represent a marked improvement over simple balance tests, they do not directly address experiments nor do they provide a clear remedy for estimating treatment effects when imbalance is detected.

In sum, the literature stresses two major, if not conflicting, points. 1) The balance test and adjust process specifically, and ad hoc regression adjustment generally, can (and often does) introduce significant bias to the estimation of treatment effects. 2) Doing nothing in the face of significant covariate imbalance is deeply unsatisfying: we should verify that random assignment was successful to avoid estimating biased and/or inefficient ATEs. This conflict puts researchers squarely between a rock and a hard place. Should they report ATEs they believe to be biased or should they attempt adjustments that very well may introduce just as much or even more bias to their estimates? Luckily, a solution that satisfies both concerns is hidden among recent research on machine learning analyses for experiments.

Exploiting advances in machine learning for experiments Supervised machine learning algorithms are incredibly powerful for prediction tasks, but are often derided in the social sciences given their "black box" inscrutability. Despite this derision, their predictive power has a variety of useful applications for causal inference (for an overview see Grimmer, Roberts, and Stewart 2021). In this context, the predictive capacity of ML is incredibly suited for detecting randomization issues. Simply put, a well-tuned ML model should be able to detect contamination between the assignment variable and pre-treatment covariates—whether it is across a single dimension (e.g., there are more men in the control group relative to the treatment) or multiple (e.g., there are more high school educated, white, Democratic men in the control group relative to the treatment). These methods, as we show below, are much more powerful than standard balances tests and should increase our confidence in whatever inferences we may draw from a given experiment.

A different thread of the machine learning literature examines the utility of ML for estimating more precise and unbiased treatment effects, culminating in the development of doubly robust treatment effect estimators (Robins, Rotnitzky, and Zhao 1994; Chernozhukov et al. 2018). In brief,

popular doubly-robust methods (e.g., augmented inverse probability weighting, targeted maximum likelihood estimation, debiased machine learning, etc.) work by employing arbitrary ML algorithms to estimate flexible models of both treatment propensity and, separately, the outcome variable. Taken together, this composite estimator outputs semi-parametrically efficient average treatment effect estimates in general settings (Chernozhukov et al. 2018). Given this desirable property, we suggest researchers estimate treatment effects using some flavor of DRML for all experimental analyses, generally, and particularly in cases where randomization fails. Unlike the balance test and adjust process, DRML estimators can model the outcome variable across a large number of covariates, reducing the potential for researchers to make arbitrary model specification choices after the fact. In other words, researchers can simply include all appropriate pre-treatment covariates rather than choose a semi-arbitrary subset for treatment effect estimation. Estimates from a DRML model may also be useful as a robustness check, as on clean data the estimates should be nearly identical to, or even more precise than, those derived from standard methods, assuming successful randomization and constant effects. Of course, if the failure is so severe that there is no overlap in treatment propensities, or the data is otherwise unsalvageable, it may become necessary to change the estimand of interest (e.g., from the average treatment effect to the average treatment effect on the treated (ATT)) or resample altogether.

The balance permutation test process

Given the distinct utility of using ML to test and adjust for randomization failures, we propose new guidelines for experimentalists in this section that describe the balance permutation test process. This can be summarized as follows:

(1) Simulate the *correct* assignment process for the experiment, or alternatively, randomly permute the real treatment assignment values.

(2) Train black box models separately on the simulated (or permuted) and real assignment values to generate treatment propensities, using pre-treatment covariates as predictors.

(3) Compare the simulated (or permuted) and real treatment propensity distributions.

(4) If no significant difference between the distributions is observed, proceed as usual for treatment effect estimation.

(5) If a significant difference is observable, summary statistics and variable importance metrics should be examined to investigate the discrepancy.

(6) If the issue is not severe, treatment effects may be obtained via DRML. If the issue is severe (i.e., little or no overlap in propensity distributions), consider changing the estimand of interest or re-field the experiment with a corrected randomization procedure.

(7) Transparency over simplicity: in all cases, both unadjusted and DRML effect estimates should be presented.

We continue in this section by explaining this process in detail. We also include a flowchart of the process in Figure 2.1. First, researchers should simulate the assignment process of the experimental design to generate simulated treatment assignment values.[2] Alternatively, researchers can randomly permute the real treatment assignment vector for the same purpose. Simply randomly permuting the real treatment assignment vector may be preferable in circumstances where the assignment process is exotic, or otherwise difficult to simulate. An advantage of simulation, over permutation, is the ability to detect class imbalance. These simulated, or permuted, assignment values represent the null assignment vector to be used for comparison in the following steps.[3]

Next, we train a black box model using the real treatment assignment vector as the outcome variable, and all measured non-treatment, non-outcome covariates as independent variables. Any pre-treatment covariate that should not be systematically related to treatment assignment is appropriate here. The predictions from this model represent treatment propensities for the actual observations in the experiment. Our specific approach implements honest (Wager and Athey 2018), n-step boosted random forests (Ghosal and Hooker 2020) with out-of-bag predictions for treatment propensities.[4] This method combines the predictive power of boosting, the scrutability of random forest, and the important debiasing technologies of honest estimation and out-of-bag prediction. While other algorithms and splitting techniques could be used in place of this approach, regardless, it should be stressed that researchers *must* "tune" the propensity models based on an extensive set of tuning parameters and non-outcome covariates to credibly attempt to detect nonrandom

[2]While this might sound time-consuming, it can be done very efficiently for most standard, and many exotic, experimental designs using *DeclareDesign* (Blair et al. 2019). Furthermore, the sample code included with this article includes simple examples for binary, multi-arm, and continuous designs.

[3]The basic logic of this method is derived in part from the permutation test for feature importance from Breiman (2001) and a similar test is proposed in Gagnon-Bartsch and Shem-Tov (2019) for observational studies.

[4]Implemented by the generalized random forests (`grf`) package in `R` (Athey, Tibshirani, and Wager 2019).

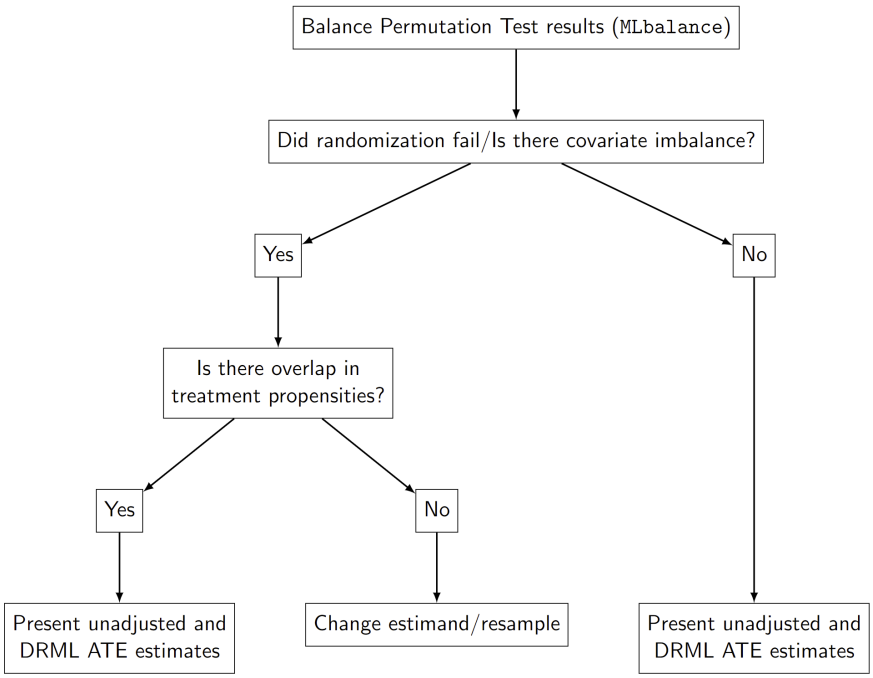


FIGURE 2.1. Balance permutation test and adjust process.

assignment. Untuned models may fail to detect randomization issues merely due to being poorly trained (Arnold et al. 2023). In that vein, the code and tuning procedure used to generate any given treatment propensity model should be reported to ensure transparency. Moreover, the modeling approach should account for possible interactions and nonlinearities that might not be evident ex ante. This qualification makes simple regression models undesirable for this task.

With the real distribution in hand, we then run the same black box model estimated on the real data using the same set of covariates and modeling strategy as in the previous step. However, instead of using the real treatment assignment vector as our dependent variable, we instead use the simulated or permuted "null" treatment assignment vector. The resulting predictions from this model generate a null distribution of treatment propensities, in which we know randomization was successful, to which compare the real treatment propensities for the observations in the actual experimental data. In other words, this null distribution is how the treatment propensities *should* be distributed if random assignment was successfully achieved.[5]

[5]The variance of the propensity score distributions will depend on the specific ML algorithm and tuning parameters. However, our approach sidesteps this issue by locking both models' algorithm and tuning parameters to settings used to build the real propensity score model.

In a general sense, if a researcher is unable to predict treatment assignment using a well-tuned ML model, they can likely conclude, as in a proof by contradiction, that randomization was successfully achieved. In that situation, one can simply proceed as normal for their treatment effect estimation using whatever method is appropriate for the corresponding research design. However, given the desirable properties of DRML treatment effect estimates, we recommend reporting both standard and DRML estimates, even if there is no evidence for a randomization failure. If the model results indicate that nonrandom assignment *has* occurred and that through whatever mechanism covariates have contaminated treatment assignment the data should be considered observational and thus simple unadjusted treatment effect estimates cannot be trusted. In other words, a researcher should be very concerned if the treatment assignment *can* be deterministically predicted from covariates for some, or all, observations (e.g., when treatment propensities are distributed bimodally around zero and one).

One issue with this approach is determining whether a model *sufficiently* predicts treatment assignment to conclude randomization failed. This is a difficult problem. While one could imagine rule-of-thumb cutoffs for accuracy, area under the curve (AUC), root mean squared error (RMSE) in the case of continuous treatments, or other model performance criteria, these would necessarily be arbitrary. Would a model just below or just above these cutoffs be sufficient evidence of (non)random assignment? These problems of interpretability, particularly regarding p-value rules of thumb (e.g., $p < 0.05$ is "statistically significant"), already plague political science. For example, meta-analyses have consistently found that these thresholds, particularly p-values, significantly bias published research in political science, specifically (Gerber, Malhotra, et al. 2008), and the social sciences, generally (Franco, Malhotra, and Simonovits 2014).

There are several other reasons we choose to eschew a test statistic. For example, what if only a few observations, or a small cluster, have extreme treatment propensity scores? In this case, testing for differences in the location, or scale, of the propensity distributions might falsely suggest there is no randomization issue. A similar problem exists for testing the difference in means of the treatment propensity distributions. Two distributions can have equal means but drastically different shapes (e.g., a distribution narrowly centered around 0.5 and a bimodal distribution with peaks at zero and one might have nearly the same mean). The same is true for difference in variance tests. In

all of these instances, a simple visual examination of the treatment propensity distributions would suffice, whereas bespoke test statistics might fail.

Once both propensity score models are estimated, we recommend plotting the null and real propensity score distributions and associated means, allowing for a simple "ocular test." If randomization was successful, we should expect these distributions to be similar in location (i.e., central tendency) and scale (i.e., variance). Importantly, propensity scores should be non-deterministic, located away from zero or one. In most situations, as illustrated in the next section, one should be able to immediately tell whether randomization was achieved by this simple graphical comparison. This recommendation diverges from both simple balance tests and related permutation inference tests that culminate in a test statistic (e.g., Gagnon-Bartsch and Shem-Tov 2019). While having a single test statistic has intuitive and parsimonious appeal, we argue, as above, that a singular statistic, and a corresponding arbitrary threshold, produces more issues than it may address.

There are significant "free lunch" benefits to our approach. By presenting *both* the unadjusted and doubly robust ATE estimates, researchers can have the best of both worlds: improved efficiency via DRML ATE estimates, and forced honesty about the simple unadjusted effects. While a simple difference in means estimator will recover an unbiased ATE estimate when randomization is successful, using a machine learning adjustment, like DRML, will generally provide significant efficiency improvements in small samples (List, Muir, and Sun 2022). Moreover, as we demonstrate in a later section, DRML treatment effect estimates tend to have similar, or better, efficiency and similar coverage relative to both simple unadjusted ATE estimates, regression adjusted ATEs based on balance table adjustments, and the popular Lin (2013) ATE estimator, but smaller absolute errors. Lastly, given the largely automatic nature of the doubly robust estimation procedure, the potential for specification searching is minimized relative to custom model specification.

A note on false positives One reasonable concern with any statistical test is the threat of false-positives. For many tests this threat can be evaluated using Monte-Carlo simulations where false-positive (and negative) rates are reported across various data generating processes. However, given the lack of a test statistic in our method, this approach is infeasible. For us to even consider running and evaluating these simulations, we would need to decide upon some sort of (arbitrary) threshold where our test correctly or incorrectly detected imbalance. In fact, early iterations of this method toyed with the idea of leveraging location and scale tests to compare the null and

real distributions. However, on top of the issues we noted with statistical tests and their cutoffs above, these specific tests are ill-suited for our method. This arises from the simple fact of how the null distribution is calculated: it will almost always be more centered with lower variance than the real distribution of propensity scores.[6] Additionally, simple tests, such as difference in means (or medians) would fail in numerous cases (e.g., comparing a bimodal and normal distribution that are centered in the same location).

Critically, however, there is very little to lose, if anything, from leveraging DRML, *regardless* of if randomization issues are present. Previous research overwhelmingly illustrates the benefits of estimating ATEs with DRML over standard approaches and we also demonstrate these benefits using Monte Carlo simulations later in this article (Zhong et al. 2021; Colnet et al. 2024). This is all to say that we need not worry about both false positives and what could be described as "muddy" cases, where it is unclear if the experiment has randomization issues. Even if there is a false positive, the effect estimates recovered from DRML estimation will most often, as we show in ∼99% of cases in our simulations, be superior to standard estimation. In sum, we recommend researchers report both doubly robust treatment effects and standard estimates in all cases.

Demonstrating the utility of the balance permutation test

We proceed in this section by highlighting the effectiveness of this approach on three simulated data generating processes. We also provide details about the results of applying the balance permutation test to real-world and fabricated data. Lastly, we offer an expansive set of simulation results that illustrate the benefits of doubly robust treatment effect estimation over popular estimators (i.e., simple regression adjustment and regression adjustment following Lin 2013).

Motivating illustrations Looking back at the balance test in Table 2.1, recall that there were no significant differences between the control and treatment groups across female nor Black respondents. Despite this appearance, these data were generated in such a way that *no Black female respondents* are in the treatment group. This simple simulation is meant to demonstrate a fundamental limitation of balance tables: they test marginal covariate balance, *not* joint covariate balance. Joint imbalance may manifest in marginal imbalance across a single or multiple covariates, but as we show in this simulation, that is not necessarily the case.

[6] One could alter our test to arbitrarily induce increased variance in simulated assignment probabilities, however, how would one decide the amount of variance to include? Again this introduces more issues than it solves.

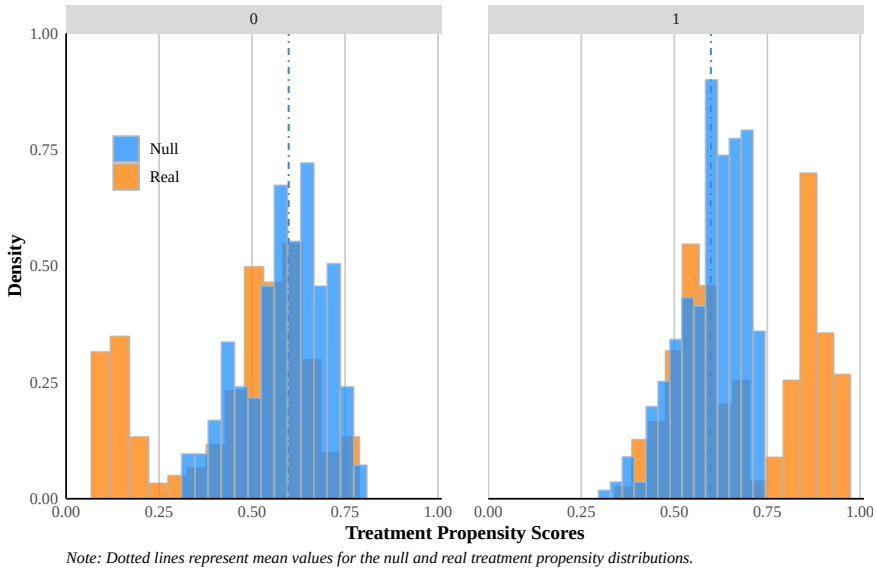


FIGURE 2.2. Balance permutation test result for joint imbalance example

Unlike simple balance tables, the balance permutation test checks for both marginal and joint imbalance across included covariates. This is true even if the joint imbalance is over an even higher dimensional space (e.g., 3 or 4 dimensions). In contrast to the balance test, our method, as shown in Figure 2.2, immediately reveals the imbalance in this simulation. The null distribution of treatment propensities in both the treatment and control groups are centered around 0.5 matching expectation. Conversely, the distribution of real treatment propensities approaches zero and one, indicating deterministic propensity scores for some observations.

Importantly, identifying the offending covariates in this context is incredibly easy. Table 2.2 reports the ten observations with the lowest treatment propensities in the dataset. What we clearly see here is that every one of these respondents are Black women. Moreover, simple summary statistics reveal the bottom quartile of treatment propensities are exclusively Black and female. We will note, however, that variable importance scores from this model also point the finger squarely at these two variables. Regardless, our method vastly outperforms the simple balance test in this context.

Additionally, we execute the balance permutation test on a single-armed experiment with simple randomization. In this simulated experiment, the sample size is 1,000, treatment assignment (W_b) is binary, and there are twenty covariates (X_N where $N \in [1, 20]$) aside from the outcome variable. Random assignment has failed and treatment assignment is in reality conditioned by two

Female	Black	Income	Dem.	Age	W	Propensity
1	1	44308	1	55.11	0	0.080
1	1	45320	1	73.52	0	0.081
1	1	39735	1	54.22	0	0.083
1	1	47080	1	62.19	0	0.088
1	1	44725	1	45.83	0	0.095
1	1	58138	1	53.52	0	0.082
1	1	60283	0	62.75	0	0.090
1	1	41606	1	50.68	0	0.079
1	1	46838	1	53.41	0	0.069
1	1	69806	0	61.37	0	0.095

TABLE 2.2. Ten lowest treatment propensity observations.

covariates together (X4 and X8). The specific data generating process (DGP) for the real treatment assignment is:

$$W_b = \text{Binomial}\left(n, \begin{cases} 0.021 + \mid 0.4X_4 - 0.5X_8 \mid & \text{if } .021 + \mid 0.4X_4 - 0.5X_8 \mid < 1 \\ 1 & \text{otherwise} \end{cases}\right)$$

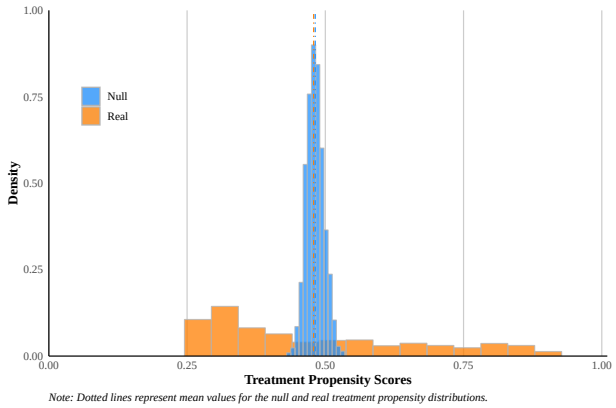
By design, the raw counts of observations in each group are roughly per expectation, and simple balance tests do not clearly reveal the problem (see Table 2.3). Even worse, the balance table reveals imbalance by chance on covariates that are not actually contaminating treatment assignment in the simulation (X1, X11, and X18).

In contrast, the results of the balance permutation test are displayed in Figure 2.3a. This figure compares the null distribution of treatment propensity scores to the real distribution of treatment propensities as generated by the boosted random forest models. From this comparison it becomes immediately clear that randomization has failed, and that contamination between the covariates and the treatment assignment has occurred. This example also highlights the importance of visual examination of the propensity score distributions. The means of the two distributions are quite similar, but the shape of the distributions are divergent. Additionally, using a measure of variable importance (or VIP score) displayed in Table 2.3,[7] we can see precisely over which dimensions the contamination has occurred: $X4$ and $X8$.

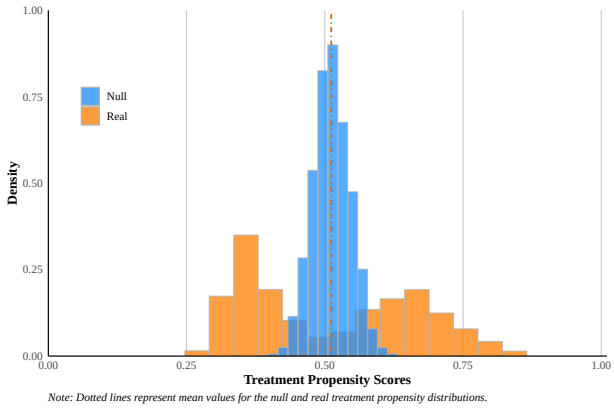
[7]The variable importance measure employed here is detailed in the supporting information. In brief, variable importance across all covariates is constrained to sum to one. Higher values indicate relatively greater importance to predicting the outcome variable.

TABLE 2.3. Simple balance table and variable importance scores for binary DGP

		Control		Treatment				
	Pred Treat?	Mean	SD	Mean	SD	Diff in Means	p Val.	VI
X1	No	0.0	1.0	0.1	1.0	**0.1**$^{+}$	0.06	0.01
X2	No	0.0	1.0	0.0	1.0	-0.1	0.21	0.01
X3	No	0.0	1.0	0.0	1.0	0.0	0.89	0.01
X4	**Yes**	0.0	0.8	-0.1	1.1	-0.1	0.22	**0.19**
X5	No	0.0	1.0	0.1	0.9	0.1	0.23	0.01
X6	No	0.0	1.0	0.1	1.0	0.1	0.13	0.02
X7	No	0.0	1.0	0.0	1.0	-0.1	0.20	0.01
X8	**Yes**	0.0	0.7	0.1	1.2	**0.1**$^{+}$	0.08	**0.62**
X9	No	0.0	1.0	-0.1	1.1	0.0	0.65	0.01
X10	No	0.0	0.9	0.0	1.0	0.0	0.63	0.01
X11	No	0.0	1.0	0.1	1.0	**0.1**$^{+}$	0.09	0.01
X12	No	0.0	1.0	0.0	1.0	0.0	0.84	0.01
X13	No	0.0	1.0	0.0	1.0	0.0	0.99	0.03
X14	No	0.0	1.0	0.0	1.0	0.0	0.93	0.01
X15	No	0.0	1.0	0.0	1.0	0.0	0.50	0.01
X16	No	0.0	1.0	0.0	1.0	0.0	0.68	0.02
X17	No	0.0	1.0	0.0	1.0	-0.1	0.41	0.01
X18	No	0.1	1.0	-0.1	1.0	**-0.2****	0.01	0.01
X19	No	0.0	1.0	0.0	1.0	0.1	0.39	0.01
X20	No	0.1	0.9	0.0	1.0	0.0	0.69	0.01



(a) Binary Treatment
(Multidimensional Contamination)



(b) Continuous Treatment
(Unidimensional Contamination)

FIGURE 2.3. Balance permutation test results for both a binary treatment (panel a) and continuous treatment simulation (panel b).

To reinforce the results of this simulation, we further apply the balance permutation test to a simulated example with a continuous treatment. In particular, we simulate a continuous treatment

assignment vector (W_c) with unidimensional contamination. We include 40 simulated pre-treatment covariates with a sample size of 4,000 observations. The specific DGP for the real treatment assignment is:

$$W_c = \text{Binomial}(n, \frac{1}{1 + e^{-X_{16}}}) + \mathcal{N}(0, 1)$$

The results of the balance permutation test are presented in Figure 2.3b. The test straightforwardly indicates contamination, and this contamination is correctly localized to X_{16} by the included variable importance measure.

Real-world & fabricated data examples The balance permutation test is not only useful for detecting contamination in simulated data, but is also useful for detecting issues in real-world experiments. To illustrate this, we apply the test to Gerber and Green (2000), a large, complex field experiment where randomization went partially awry (noted and corrected by Imai 2005) and Moehler and Conroy-Krutz (2016), a paper that utilizes a multi-arm experiment where the assignment process went wrong (noted by Mutz, Pemantle, and Pham 2019). We also apply the test to a case where data was fabricated: Shu, Mazar, Gino, Ariely, and Bazerman (2012) (retraction: Berenbaum 2021).

First, however, we apply the test to an experiment with no randomization failure, Bailard et al. (2023), to illustrate what the test would look like under many normal circumstances. We also apply the test to an experiment with minor to moderate imbalance, Karim (2020), to illustrate the benefits of using this approach over the balance test and adjust process. The result of the test on Bailard et al. (2023), reported in Figure 2.4, shows no evidence of imbalance with the real and permuted propensity scores completely overlapping.

In contrast, Figure 2.5 shows significant imbalance between the treatment and control groups for one of their treatment arms (the female police treatment). A visual inspection of the treatment propensity distributions reveals unusually wide variance in the real propensity distribution relative to the null. However, none of the real treatment propensity scores are deterministic or near-deterministic (i.e., very close to zero or one). In this case, we present doubly-robust average treatment effect estimates and compare them to those calculated in the original article (from Table 2, regressions 4–6). In this experiment, the treatment we examine was whether individuals were

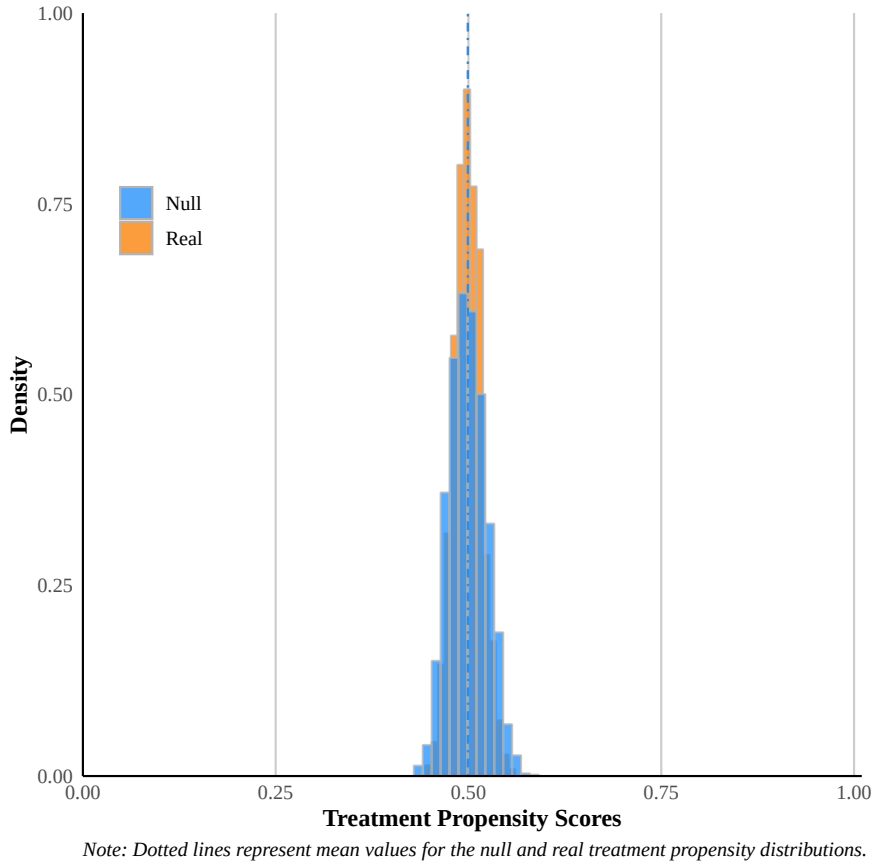
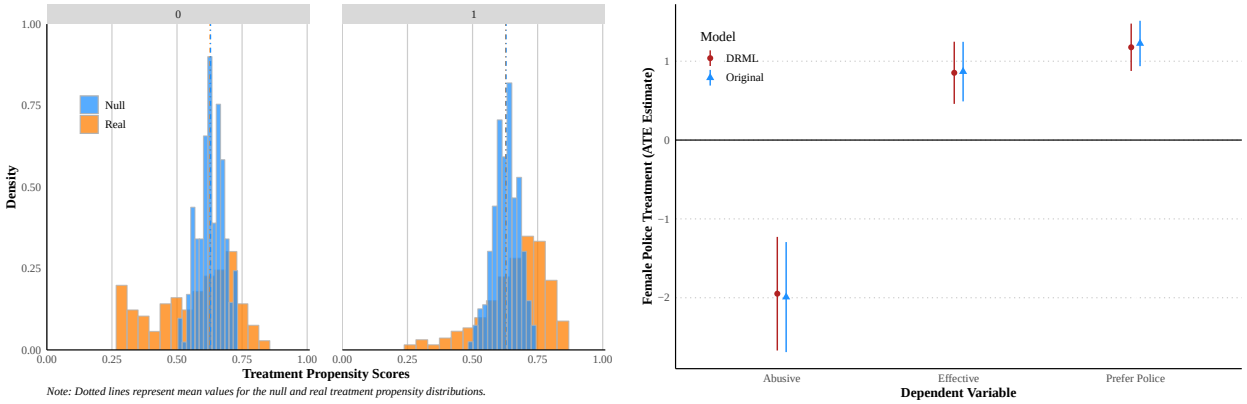


FIGURE 2.4. Balance permutation test result for Bailard et al. (2023).

exposed to a female or male police officer in a given village in Liberia. The three outcome variables were individuals' reports of whether the police were abusive, effective, and if the individual preferred the police.

In this instance, treatment assignment is not truly orthogonal to measured pre-treatment covariates: some combination of pre-treatment covariates are related to treatment assignment. However, in this instance these covariates, such as number in household, are largely unrelated to the outcome variable, and thus do not confound the relationship between treatment and outcome. Because of this, we find very little difference between the unadjusted and adjusted effects. This highlights an important consideration when assessing successful randomization/covariate balance: a randomization issue is only problematic insofar as it can confound the effect of the treatment. Imbalance over covariates unrelated to the outcome variable is not a cause for concern.

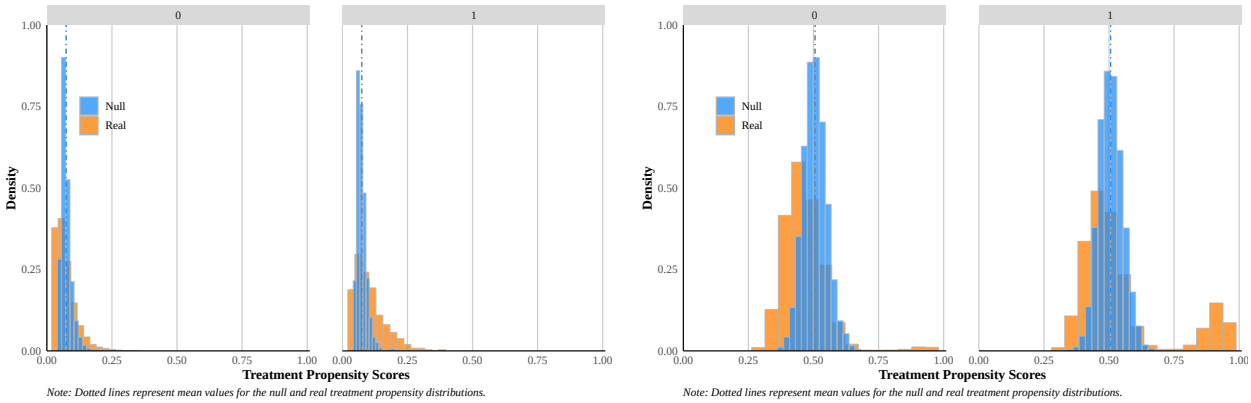
Here we report the results of the Gerber and Green and the Shu, Mazar, Gino, Ariely, and Bazerman analysis in Figure 2.6a and 2.6a, respectively. For brevity, we choose to report the



(a) Balance Test Results

(b) DRML ATE Results

FIGURE 2.5. Balance permutation test result (panel a) and resulting doubly robust machine learning ATEs (panel b) for Karim (2020).



(a) Gerber and Green (2000)
Randomization Issue

(b) Shu et al. (2012)
Data Fabrication

FIGURE 2.6. Balance permutation test results for a randomization issue (panel a) and data fabrication (panel b).

Moehler and Conroy-Krutz results in the supporting information. These two randomization failure examples illustrate the ability of our method to detect slight to moderate contamination that is missed in a standard balance table; whereas the Shu, Mazar, Gino, Ariely, and Bazerman example shows how our method is able to easily detect data fabrication.

For Gerber and Green (2000), due to the complexity of the experimental design which contains multiple treatment assignments (yielding more than 20 experimental conditions), we focus exclusively on a single assignment variable and a relatively small subset of pre-treatment covariates

from Imai (2005). Our balance permutation test results indicate significant imbalance over pre-treatment covariates and the model estimates many treatment propensity scores are dangerously close to zero. In contrast to Moehler and Conroy-Krutz (2016), this situation is highly unusual, particularly given the complexity of the experiment itself, but is representative of a relatively severe failure in randomization. Having discovered this failure, it would be important for any researcher to investigate why this may have occurred and to potentially make adjustments to their estimand.

Our approach is also useful for detecting a common form of data fabrication. The test exploits a common feature of many fabricated articles: the fabrication creates a suspicious relationship between treatment assignment and some set of pre-treatment covariates, which *should* be orthogonal in an experimental context. This is the case in Shu, Mazar, Gino, Ariely, and Bazerman (2012) (retraction: Berenbaum 2021) for example. Applied to this dataset, see the supporting information, our approach easily detects this manipulation, even though it is relatively subtle. As we show with simulations in Table 2.3, we are also able to detect contamination that results in a multidimensional relationship between covariates and treatment assignment, among dozens that may or may not be implicated in the fabrication. Thus, our approach offers a significant post facto auditing benefit in addition to the standard use case of prospective checks. Moreover, this test can be more easily applied by researchers with heterogeneous skill sets: the associated `R` package does not require a specialization in statistical methodology nor data sleuthing.

Our test on Moehler and Conroy-Krutz (2016) indicates that the failure is caused by some moderate, subtle differences in the age of subjects between treatment groups, particularly for treatment arms three and four. However, failure could also be due in part to simple numerical imbalance between the treatment arms. Regardless, given that the failure is not severe, it is a relatively straightforward process to estimate DRML treatment effects. We expect most randomization issues to exist somewhere in this space, so we think this example is representative of what the application of the balance permutation test would look like in cases of moderate imbalance.

Doubly robust treatment effect tests In this section we interrogate the usefulness of DRML for estimating ATEs relative to simple unadjusted estimates, regression adjustments following the balance test and adjust approach, and the popular Lin estimator (Lin 2013). We simulate

480 sets of 2,000 simulations over a wide range of simulation parameters. The effect size for all simulations is 1, treatment assignment is simple complete randomization, and the parameters varied are listed in Table 2.4.

TABLE 2.4. DRML performance test simulation parameters

Parameter	Possible Settings
Sample Size	100, 500, 1000, 5000
Number of Covariates	5, 10, 20, 50, 100
Heterogeneous Effects	True or False
Error Variance	Low (1), Medium (5), High (10)
Linear Effects	True or False
Interactions	True or False

This results in an extensive set of 960,000 tests demonstrating the performance of DRML over standard methods. We evaluate performance using two metrics, reduction in absolute and standard error, averaging each metric across all simulations within each set. This results in 1,440 comparisons for each metric: 480 per method ($480 \times 3 = 1{,}440$). Given the quantity of these comparisons, we opt to report the full results in the supporting information, and instead report an aggregated version in the main text. Specifically, our results are summarized in Figure 2.7 by the number of covariates (y-axis), sample size (x-axis), and error variance (panel). Each grid location represents the average absolute error reduction of all our comparisons across all other parameter combinations.

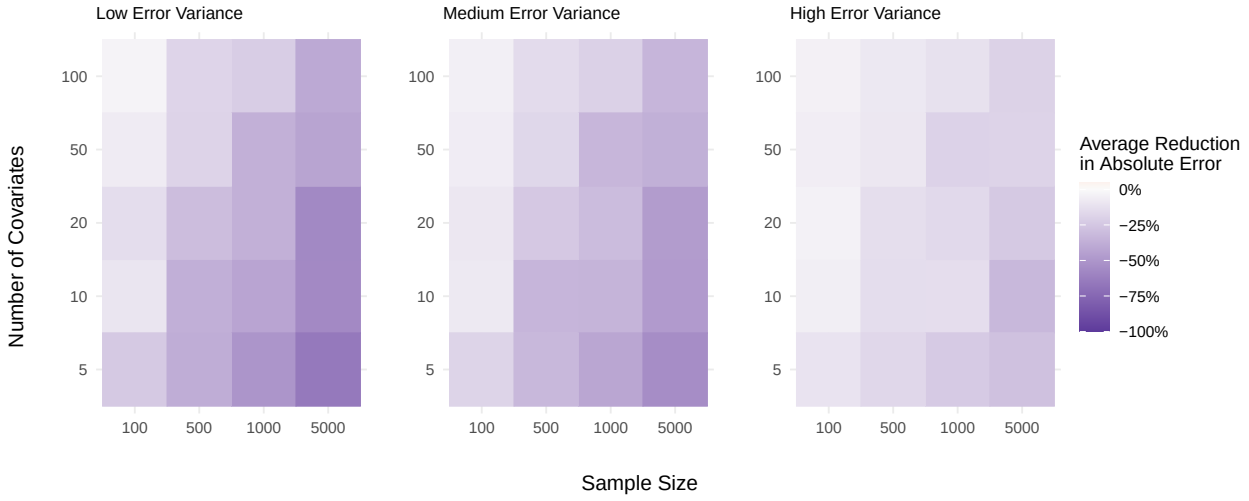


FIGURE 2.7. Average reduction in absolute error of the DRML estimated effects versus unadjusted difference in means, regression-adjusted, and Lin-adjusted effects.

Overall, we find that the DRML estimates recover vastly more fine-grained, accurate effect estimates across nearly every single set of simulations while maintaining mostly proper coverage. We do find DRML suffers a marginal decrease in coverage, 93% across all simulations rather than the desired 95%. This undercoverage could have many causes. For instance, in the simulation study we use the default tuning parameters in our models rather than employing an extensive tuning procedure, and further, we force the estimator to model the propensity score rather than taking the assignment probability as given. In practice, model tuning and setting the known assignment probability when appropriate should further improve coverage.

The possibility of slight undercoverage also motivates our recommendation that researchers present both the simple unadjusted treatment effect estimates alongside the DRML estimates. Additionally, if perfect coverage is a requirement, the DRML model's default uncertainty bands could be replaced by conformal intervals (Lei and Candès 2021), which guarantee coverage at a pre-specified rate. For a further discussion of coverage rates for all methods see the supplementary information.

Despite the marginal decrease in coverage for DRML, its performance is essentially unrivaled by any of the comparisons. Specifically, DRML is outperformed in only 19 out of 1,440 comparisons, only ∼1% of the time (the specific instances are reported in the supporting information). Importantly, these 19 instances are largely idiosyncratic parameter combinations and exist exclusively in extremely small samples (100 observations).[8]

Discussion

In this article, we outline the balance permutation test, a simple method that exploits the predictive power of ML to detect randomization failures in experimental data. We illustrate its utility on real datasets as well as through Monte Carlo simulations, where our method vastly outperforms standard balance test procedures. We provide clear guidelines and an accompanying `R` package, `MLbalance`, for researchers to employ this method, as well as steps to take if randomization failures are detected. Moreover, the testing procedure outlined in this paper can be employed not only as a diagnostic check, but also post facto by reviewers, or future replicators to investigate possible data fabrication.

[8]The 19 instances are detailed specifically in the supporting information.

While we believe that our method is superior to previous approaches, it should be said that even by using ML we can only be more confident that randomization was achieved, not completely certain. Any contamination of the assignment process that does not yield either numerical imbalance in the assigned groups relative to ex ante expectations, or is not detectable on observable covariates, will go unnoticed. However, *any* model or approach would fail in this situation, so some degree of epistemic humility is generally warranted.

Additionally, there remain significant opportunities to further incorporate ML into the experimental design and analysis process. For example, ML could be used to detect systematic attrition in panel surveys and experiments as well as adjust for the bias that attrition may create. Furthermore, there is a need for clear guidelines for how ML should be used in a principled way for both subgroup analyses and experimental analysis, in general.

Bibliography

Arnold, Christian, Luka Biedebach, Andreas Küpfer, and Marcel Neunhoeffer. 2023. "The role of hyperparameters in machine learning models and how to tune them." *Political Science Research and Methods,* 1–8.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–1178. https://doi.org/10.1214/18-AOS1709. https://doi.org/10.1214/18-AOS1709.

Bailard, Catie Snow, Matthew H Graham, Kimberly Gross, Ethan Porter, and Rebekah Tromble. 2023. "Combating Hateful Attitudes and Online Browsing Behavior: The Case of Antisemitism." *Journal of Experimental Political Science,* 1–14.

Berenbaum, May R. 2021. "Retraction for Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end." *Proceedings of the National Academy of Sciences* 118 (38): e2115397118. https://doi.org/10.1073/pnas.2115397118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2115397118. https://www.pnas.org/doi/abs/10.1073/pnas.2115397118.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113 (3): 838–859.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45:5–32.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. *Double/debiased machine learning for treatment and structural parameters.*

Colnet, Bénédicte, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. "Causal inference methods for combining randomized trials and observational studies: a review." *Statistical science* 39 (1): 165–191.

Fisher, Ronald Aylmer. 1925. *Statistical Methods for Research Workers.* Edinburgh [Scotland]: Oliver / Boyd.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345 (6203): 1502–1505.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40 (2): 180–193.

Gagnon-Bartsch, Johann, and Yotam Shem-Tov. 2019. "The Classification Permutation Test." *The Annals of Applied Statistics* 13 (3): 1464–1483.

Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R Biggers, and David J Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1 (1): 81–98.

Gerber, Alan, Neil Malhotra, et al. 2008. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quarterly Journal of Political Science* 3 (3): 313–326.

Gerber, Alan S, and Donald P Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American Political Science Review* 94 (3): 653–663.

Ghosal, Indrayudh, and Giles Hooker. 2020. "Boosting random forests to reduce bias; one-step boosted forest and its variance estimate." *Journal of Computational and Graphical Statistics* 30 (2): 493–502.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2021. "Machine learning for social science: An agnostic approach." *Annual Review of Political Science* 24:395–419.

Hübert, Ryan, and Ryan Copus. 2022. "Political Appointments and Outcomes in Federal District Courts." *The Journal of Politics* 84 (2): 908–922.

Imai, Kosuke. 2005. "Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments." *American Political Science Review* 99 (2): 283–300.

Karim, Sabrina. 2020. "Relational state building in areas of limited statehood: Experimental evidence on the attitudes of the police." *American Political Science Review* 114 (2): 536–551.

Lei, Lihua, and Emmanuel J Candès. 2021. "Conformal Inference of Counterfactuals and Individual Treatment Effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (5): 911–938.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318. https://doi.org/10.1214/12-AOAS583. https://doi.org/10.1214/12-AOAS583.

List, John A, Ian Muir, and Gregory K Sun. 2022. *Using Machine Learning for Efficient Flexible Regression Adjustment in Economic Experiments.* Technical report. National Bureau of Economic Research.

Mize, Trenton D, and Bianca Manago. 2022. "The Past, Present, and Future of Experimental Methods in the Social Sciences." *Social Science Research* 108:102799.

Moehler, Devra C, and Jeffrey Conroy-Krutz. 2016. "Partisan media and engagement: A field experiment in a newly liberalized system." *Political Communication* 33 (3): 414–432.

Mutz, Diana C, Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1): 32–42.

Negi, Akanksha, and Jeffrey M Wooldridge. 2021. "Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects." *Econometric Reviews* 40 (5): 504–534.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89 (427): 846–866.

Simonsohn, Uri, Leif Nelson, and Joe Simmons. 2023. *[109] Data Falsificada (Part 1): "Clusterfake".* https://datacolada.org/109.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Zhao, Anqi, and Peng Ding. 2022. "To Adjust or not to Adjust? Estimating the Average Treatment Effect in Randomized Experiments with Missing Covariates." *Journal of the American Statistical Association* 0 (0): 1–11. https://doi.org/10.1080/01621459.2022.2123814. eprint: https://doi.org/10.1080/01621459.2022.2123814. https://doi.org/10.1080/01621459.2022.2123814.

Zhong, Yongqi, Edward H Kennedy, Lisa M Bodnar, and Ashley I Naimi. 2021. "AIPW: an r package for augmented inverse probability–weighted estimation of average causal effects." *American Journal of Epidemiology* 190 (12): 2690–2699.

CHAPTER 3

Causal Machine Learning for Experiments: Guidelines for Effective and Transparent Research

Researchers have produced a large literature on adapting machine learning methods for experimental analysis. However, this field's growth has not corresponded to increases in use among social scientists. This is due to both a lack of principled guidance on how to apply these methods and the esoteric nature of their current implementations. While previous research has focused on *some* aspects of machine learning applications to experiments, like subgroup analysis, a holistic approach to how researchers should utilize these methods is missing. Furthermore, previous guidance unnecessarily limits the unique ability of these methods to uncover complex relationships. Therefore, this article aims to provide a complete framework for how machine learning methods should be incorporated into experimental analysis. This framework and guidelines should significantly improve social scientists' ability to examine the internal validity of any experiment, improve experimental diagnostics, and increase the resolution of experimental results.

Experiments are a cornerstone of modern social science research. The ability to isolate causal effects continues to make experiments generally, and the randomized control trial (RCT) specifically, the gold standard for causal inference. As a result, experimental work in the social sciences has blossomed across disciplines. Above and beyond applications of experiments and their resulting substantive findings, the methodology of conducting and analyzing experiments is also widespread (e.g., Druckman et al. 2011; Jackson and Cox 2013; Kagel and Roth 2015; Schinka, Velicer, and Weiner 2013). This includes work dedicated to improving research design and diagnostics (Blair et al. 2019a), sampling and balancing (Imai, King, and Stuart 2008), and pre-registration (Nosek et al. 2018) among other topics. However, some aspects of experimental data analysis (i.e, the calculation of average treatment effects) continue to rely largely on standard methodologies (e.g., difference-in-means, linear regressions, maximum-likelihood estimation, etc.). Simply put, there has been limited attention paid to incorporating these new methods into standard practice for the analysis of experimental data.

Despite this, there *is* a large and growing literature applying machine learning (ML) methodologies to social science applications, generally, and RCTs, in particular (Grimmer, Roberts, and Stewart 2021).[1] With the development of causal forests (CF) (Athey, Tibshirani, and Wager 2019; Athey and Wager 2019), research in this sub-literature has accelerated with new estimation techniques coming forward at a rapid clip (e.g, Dorie et al. 2022). The application of ML in these situations, while at first seeming anathema to causal inference, is useful for experimental analysis in part when predictive accuracy, rather than the identification of a causal effect, is the primary goal. For example, the use of ML, with its flexibility and general lack of parametric assumptions, to determine whether randomization was achieved or if experimental groups are balanced is *necessarily* better than standard, parametric models, like linear probability models or maximum likelihood estimation (MLE) (Rametta and Fuller 2024). This is because we want to ensure that even the *best* predictive algorithms cannot predict treatment assignment, so, as in a proof by contradiction, we can be relatively well-assured that randomization and balance were achieved (at least on observables). While applications of ML for experiments span a much broader set of functions and goals than just checking randomization and balance, as illustrated below, this example is indicative of the general ability of ML to enhance the analysis of experimental data. If deployed properly, causally

[1]What methods technically count as machine learning, or statistical learning, techniques is a subject of debate (Hastie et al. 2009).

oriented machine learning methods enable researchers to more confidently analyze experiments than standard methods.

However, while these ML methodologies have been developed for the express purpose of analyzing experimental data—and are in fact, as shown later in this paper, incredibly useful for just that—there still exists no principled, step-by-step guidelines for applying these methodologies and how to report related results. This paper aims to remedy this gap in the literature by providing both general guidelines for analysis for any experimentalist, but also by providing additional tools above and beyond those that are currently available, to streamline this entire process. Importantly, the application of ML methods, in the ways outlined below, should be the standard for analyzing experimental data. In that vein, this paper aims to put forward practical, actionable guidance for improving inference and increasing the resolution of experimental results by using machine learning.

This article proceeds in two major parts: The first outlines this paper's theoretical argument and proposes a general checklist that researchers should use when analyzing experiments, including the application of machine learning for the purposes of checking randomization and balance as well as identifying accurate (conditional) average treatment effects (CATEs). The second part of this paper proceeds by applying the checklist to two previous social science experiments to illustrate their utility. The paper then concludes with a discussion of the importance of these methods and the potential for future work in this field.

How Should Machine Learning be Used in Experiments?

Previous methodological research on experiments has focused primarily on research design (e.g., Druckman et al. 2011; Druckman and Green 2021; Druckman 2022; Jackson and Cox 2013; Kagel and Roth 2015; Schinka, Velicer, and Weiner 2013), and more recently on ethics and reporting standards (e.g., Brodeur, Cook, and Heyes 2020; Druckman and Green 2021; Gerber et al. 2014; Mutz and Pemantle 2015; Scheel, Schijen, and Lakens 2021). The latter of which, rightfully so, has dominated recent discussions of experimental research, with multiple journals across disciplines adopting clear rules regarding the presentation of experimental research and some also requiring the preregistration of experiments.

Likewise, research on new ML methodologies for analyzing experimental data has blossomed (Chernozhukov et al. 2023). This research has primarily focused on applying machine learning

methods inside causal frameworks, with causal ML methods like generalized random forests (Athey, Tibshirani, and Wager 2019) and various flavors of causal Bayesian additive regression trees (BART) (Chipman, George, and McCulloch 2010; Robinson and Duch 2022), growing in popularity in the social sciences. These methods leverage the predictive capabilities of machine learning to better estimate both nuisance parameters and actual estimands of interest like conditional average treatment effects.

Work on *how* to use these new ML methods in a principled way for experimental data, however, is nascent. This research, so far, is mainly limited to two book chapters and an article, respectively, by political scientists Ratkovic (2021); Green and White (2023); and Zheng and Yin (2023), as well as an article by epidemiologists Jawadekar et al. (2023). First, Jawadekar et al. (2023) and Zheng and Yin (2023) both illustrate the usefulness of these methods in their respective disciplines, though they differ in their approach. Jawadekar et al. (2023) provide an explanation of causal forests and its advantages and limitations and illustrate these using simulated data, essentially distilling this information for their discipline. In contrast, Zheng and Yin (2023) only briefly explains the methodology. Instead, their paper focuses on illustrating the usefulness of causal forests in estimating CATEs by outlining the methodological process recommended by Athey and Wager (2019) and applying it to two previous studies. Green and White (2023) provides a detailed and rigorous introduction to these methods and also provides examples of implementations. However, this chapter, and the articles mentioned above, do not provide sufficient guidance on how these methods should be incorporated in both preregistrations and analyses.

While not directly about machine learning methods, Ratkovic (2021) does provide the clearest guide for using ML for subgroup analysis, highlighting how these new ML methods can be used to better estimate conditional effects. Specifically, Ratkovic argues for the following:

(1) Preregister the covariates, along with the level of interaction, that are going to define subgroups targeted for inference.

(2) Preregister the method and, if applicable, the random seed with which it will be initialized. Utilize a method that estimates subgroup effects directly, to maximize power, and implements a sample splitting in this step, to maintain honesty of inference.

(3) Preregister the marginal plots that will be used to look for heterogeneity. Every plot should be thought of as a degree of freedom, such that the p-value threshold should be Bonferroni adjusted. These uncovered effects can be discussed using the language of significance, but with the understanding that the process is conditional on the particular implementation of the experiment.

(4) In order to define discovered effects, pre-register an acceptable false discovery rate and implement a method that achieves this rate. No inferential claims about statistical significance can be made about these, but interesting discoveries should be noted for future experimentation.

Ratkovic 2021, pp. 8

While this checklist provides a solid foundation for analyzing subgroups specifically, and experiments generally, I argue for significant alterations.

First, it is unclear why the level of interactions between covariates necessarily need to be preregistered, particularly when implementing forest based methods (like causal forest) which naturally model unspecified high-order interactions. One of the distinct benefits of using ML in detecting heterogeneity and estimated CATEs is its ability to explore potential interactions agnostically, without the need for strong priors.

Second, for methods that employ sub-sample splitting like causal forests, rather than preregistering a particular random seed, practitioners should instead cross-validate results over a large number of seeds. In the case of variable importance metrics, this allows researchers to generate point estimates and confidence bands for variable importance that are independent of the arbitrary sub-samples used to construct the single-shot metrics. As is well known, the results of an analysis based on a single, arbitrary seed could lead to spurious conclusions (Colas, Sigaud, and Oudeyer 2018).

Third, while it is reasonable to think of each marginal plot as a degree of freedom, researchers should primarily be concerned with avoiding the creation of marginal plots for variables that fail

to achieve significant importance in predicting treatment effect heterogeneity. Simply put, over-analyzing noisy predictions over these unimportant covariates is unlikely to lead to clear inferences.

Finally, Ratkovic's checklist omits two important considerations that can directly affect the results of any subgroup analysis: feature design and tuning. Poor feature design, for example one-hot encoding an ordered categorical variable with a large number of categories, can lead researchers to spuriously conclude that features are unimportant when they might otherwise be if better coded (Johannemann et al. 2019). Thus it is important that researchers think carefully about variable coding and measurement prior to including them in an analysis. A related issue is the way forest based estimators handle colinearity: if one includes two colinear variables in the same estimation the algorithm will often drop one of the two, or split between them. This poses obvious problems for inference over features included in a subgroup analysis and should be closely considered. Moreover, machine learning methods that might be employed in subgroup analysis have various tuning parameters that should also be carefully tested and chosen. In particular, researchers should preregister the tuning procedure they plan to follow, in particular what fit statistic, or other test, is being targeted for tuning. For example, a subgroup analysis could tune for the best linear projection heterogeneity test proposed by Chernozhukov et al. (2023).

In general, while Ratkovic (2021) contributes significantly to the literature by proposing an initial way of incorporating ML into analyzing experiments by focusing on the important considerations of preregistration and the possibility of false discovery, there is significant room for improvement. In contrast to this previous work, this paper contributes to the literature by proposing an *integrated* approach to experimental data analysis, somewhat analogous to Chernozhukov et al. (2023). Specifically, testing for the existence of heterogeneous treatment effects and estimating CATEs should *mainly* be done through the use of an experimental ML methodology. Ex ante theoretical expectations and preregistration are not necessarily sufficient to prevent spurious CATE estimates using standard estimation procedures. Furthermore, limiting ML analysis generally and CATE estimation specifically to preregistered analyses limits the usefulness of ML in fully exploring treatment effects over the covariate space. Given all of this, this paper proposes a full, integrated framework for applying machine learning to analyzing experimental data, with a focus on subgroup analysis, and illustrates its usefulness in the following sections. The checklist is as follows:

The Experimentalist's Checklist

(1) Design a clear experiment and check expectations

(2) Liberally and accurately measure the relevant covariate space

(3) Test for assignment failures and/or test for systematic attrition using ML

(4) Create and tune an accurate model

(5) Test for and identify treatment effect heterogeneity using the *entire* covariate space
   (a) For each treatment with identified heterogeneity report variable importance
   (b) Only present CATEs if the variable is identified as important

(6) Present results from DRML estimation, both ATEs and CATEs, alongside standard ATE estimates

This next section proceeds by detailing each of these steps, why they should be followed, and how researchers should go about accomplishing each.

The Experimentalist's Checklist

Experimental Design Research on experimental design is a deep and detailed literature (for a handbook of design, see either Druckman et al. 2011; Kagel and Roth 2015; or Schinka, Velicer, and Weiner 2013). While giving a comprehensive summary on best practices for research design is out of the scope of this paper, it is important to emphasize the importance of this step of the process. Simply put, no amount of innovative or effective methods can remedy a poorly designed experiment. Given that, researchers should use resources, such as `Declare Design` (Blair et al. 2019b), to not only design their experiments, but to be able to generate data to test the validity of said design in basic analyses. And, as always, researchers should preregister their experiment to ensure transparency and to present a set of clear, principled expectations before any data is gathered, let alone analyzed (Nosek et al. 2018). The discussion around transparency in experimental research is still ongoing, and recommendations are sure to evolve, but a good summary of current, common practices can be found in Boudreau (2021).

Given the importance of transparency in analyzing experimental data, this paper clearly explains and give examples of what should be reported from the recommended exploratory analyses, which should minimize perceived issues with manipulating data and analyses ("p-hacking") when doing exploratory work. Furthermore, in this vein of transparency, researchers should also include the analyses described below in their registration's analysis plan. And finally, while exploring the

breadth of your data, including covariates, is very important for capturing accurate treatment effects and possible effect heterogeneity, having clear and specific hypotheses registered about your general expectations for the design is still necessary. In general, it is important to be meticulous with designing an experiment, testing said design using data generating resources, and then being clear with both expectations and an analysis plan for an experiment in its preregistration.

Testing for Randomization Issues and Systematic Attrition Machine learning's greatest strength is likely its ability to give highly accurate predictions from a non- or semi-parametric model. Consequently, its most straightforward use-cases in the social sciences are when predictive accuracy is valued over easy interpretation. While this is certainly not an exhaustive list, I note two areas where ML can be particularly useful, particularly in the context of causal ML: 1) in detecting randomization issues and 2) in detecting systematic attrition. The former is detailed in Rametta and Fuller (2024) with the introduction of the "balance permutation test." In essence, the logic of the modelling strategy is that if we fail to accurately predict treatment assignment using our pretreatment covariates in a ML model, like causal forest, then we can be more certain that there were no randomization issues in our experiment. This method represents a marked improvement over the "balance test and adjust" approach detailed in Mutz, Pemantle, and Pham (2019), especially because any minor-to-moderate randomization issues can be directly addressed using the recovered treatment propensities in a doubly robust machine learning (DRML) estimation framework.[2] Additionally, this method reduces researcher's degrees of freedom in that the functional form of the model is not directly specified. In sum, instead of using a balance table or F-test, experimental researchers should always use the balance permutation test (or a similar model) to assess if there are any randomization issues.

Relatedly, machine learning can also be applied to detecting non-random or systematic attrition in panel surveys/experiments. As it stands, descriptive statistics and maximum-likelihood estimation is often used to identify systematic attrition. However, similar to the issues detailed by Mutz, Pemantle, and Pham (2019), these methods are limited by their parametric assumptions and give researcher's too many degrees of freedom in their modelling choices. The logic of an "attrition permutation test" is highly similar to the balance permutation test, in that if we are unable to predict attrition using first-wave/pretreatment covariates then we can be more assured

[2]Details on this approach are provided below.

that there is no systematic attrition biasing our estimates. And importantly, just as above, if there are minor-to-moderate issues they can be addressed directly by using DRML.

Measuring Covariates Similar to experimental design, choosing which covariates to measure is both a theoretical and design driven process. The theoretical process is well-established in the social sciences, with researchers needing to understand the nuances of the previous, related literature to determine potentially important covariates to measure (Druckman et al. 2011). The latter process, however, has significantly limited the measurement of a broad range of covariates in previous work. While ideally, with randomization, average treatment effects can simply be calculated using difference-in-means between the treatment and control groups, conditionality is almost the norm in social science research. Given the limitations of standard methods of analysis, mainly ordinary least squares regressions (OLS) and maximum likelihood estimation, choosing covariates to condition on (or control for) has been highly limited. Variables that are the most established in having a relevant effect are very likely to be measured (and tested). This is because if a researcher were to condition on an under-researched covariate, it could be reasonably construed as being arbitrary, this is often referred to as "specification searching." This concern could be raised, however, even if this variable is important for determining treatment effect heterogeneity.

Simply put, if the previous literature, particularly experiments, indicates that a variable is important to measure and to use for conditioning or as a control, then a given researcher would naturally be inclined to measure that variable. However, this process necessarily leads to the exclusion of under-researched covariates in analyses, particularly ones that have not been measured in previous experiments. As a result, I argue that, given the ability of causal machine learning methods to identify heterogeneous treatment effects generally, and conditioning variables specifically, more covariates should be measured and tested. This is not to say, however, that researchers should just include a random smattering of covariates in their analyses. Instead, it is to say that researchers should include covariates that, based on tangentially related (or observational) literature, appear to be relevant for conditioning treatment effects. Doing so will, ideally, enable researchers to test many more potential conditioning variables and result in a better, broader understanding of how treatment effects are heterogeneously distributed. Of course, one should not condition on post-treatment variables. As Imai and Yamamoto (2013) among many others point out, conditioning on a post-treatment variable can induce collider bias, which can lead to incorrect inferences about the

presence or magnitude of heterogeneous effects. Consequently, researchers must carefully consider the timing and nature of the variables they use to estimate these effects.

Multidimensional scaling techniques (which are fundamentally a type of unsupervised ML) should be more widely used to better measure various concepts, such as partisan affect or political ideology. While single-variable, self-reported feelings and preferences can be useful, they often either fail to represent well an entire dimension of a concept or are subject to differential-item functioning (DIF). For example, while differences in party-level feeling thermometers (e.g., Democratic Party feeling thermometer − Republican Party feeling thermometer) have been consistently used as a measure of partisan affect/affective polarization (Iyengar et al. 2019), it has been well-established that they fall prey to DIF (Wilcox, Sigelman, and Cook 1989). However, using scaling on a large set of partisan feeling thermometers can remedy this issue. Similarly, DIF is well-present in self-placements for ideology, which has led to implementations of scaling methods that remedy this concern (Bayesian Aldrich-Mckelvey scaling, Hare et al. 2015).

Furthermore, even with ML methods, but particularly with OLS and MLE, poor measurement of variables (including multicollinearity), can lead to incorrect conclusions about treatment effect heterogeneity. While ML methods like CF and BART can handle a large covariate space in the determination of treatment effect heterogeneity and the calculation of conditional average treatment effects, they cannot directly address measurement issues.

So in general, the recommendations in this section are two-fold. First, by using these methods, researchers can and should measure more covariates that may be important in conditioning treatment effects. While there is legitimate concern for specification searching, these ML methods, when used correctly, heavily reduce this concern if all measured covariates are included in estimated models. Second, researchers should deeply engage with the question of how to measure their variables of interest, including the possibility of using robust scaling methods. This is because accurate or inaccurate measurements can lead to significantly different results and conclusions regardless of the methods used to estimate effects.

Doubly Robust Machine Learning Treatment Effect Estimation In this section, I note the benefits of doubly robust machine learning (DRML) treatment effect estimation (Chernozhukov et al. 2022). To do this, I will also explain the specific benefits of causal forest (Wager and Athey 2018) in this framework.

DRML treatment effect estimation works within the standard linear model framework. Specifically, it uses a partially linear model:

$$Y = D\theta_0 + g_0(X) + \zeta, \quad \mathbb{E}(\zeta|D, X) = 0,$$

$$D = m_0(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

where Y is our outcome variable and $D = \{0, 1\}$ is an indicator variable of treatment.[3] The vector $X = (X_1, ..., X_p)$ is composed of all pretreatment covariates, and ζ and V are random errors for each equation.

In plain language, we have two equations we are working with here. The first, $Y = D\theta_0 + g_0(X) + \zeta$, models the outcome variable as a function of treatment (D) and pretreatment covariates (X) as well as random error (ζ) that, in expectation, is equal to zero (and can be ignored) when controlling for treatment and pretreatment covariates ($\mathbb{E}(\zeta|D, X) = 0$). Here, θ_0 is the causal effect of treatment (the coefficient for D) and $g_0(\cdot)$ represents the functional relationship between pretreatment covariates and the outcomes ($g_0(X)$). This first equation, however, is just a simple regression framework for modelling the effect of some treatment. The "D" or "Doubly" of DRML, and why this approach enables us to recover causal estimates, arises from the use of the second equation for treatment effect estimation.

This second equation, $D = m_0(X) + V$, models treatment as a function of pretreatment covariates and random error (V). Again, the random error (V) in expectation is zero given pretreatment covariates ($\mathbb{E}(V|X) = 0$), so we can ignore that term. In an RCT context where randomization was successful, we would essentially be able to ignore this equation because treatment assignment would, ideally, be orthogonal (unrelated) to any covariates, even if they are related to our outcome of interest. Specifically, we attempt to estimate $m_0(\cdot)$ which is the functional relationship between pretreatment covariates and treatment assignment.

DRML also has important desirable properties: 1) it is semiparametrically efficient and, most importantly, 2) it is robust to model misspecification in either the outcome or treatment propensity models (i.e., we only need to correctly specify one of these models to recover robust, causal effects). The first property relates to the fact that the model is efficient (recovers the real treatment effect with a limited number of observations) even though the functional form of some of the components

[3]CF can also handle multi-arm and continuous treatments.

of the model (namely $g_0(\cdot)$ and $m_0(\cdot)$) are unknown. The second property is that DRML will provide robust treatment effects as long as we model the outcome and/or treatment assignment correctly. This component can be evaluated theoretically—are we including variables that we have good reason to believe are related to the outcome and/or treatment assignment?—and empirically—based on model statistics, are we accurately specifying these models between training (data used to build the models) and testing sets (data held out from the training set)?

In general, to use DRML we want a model strategy that gives us highly accurate predictions of both the outcome and treatment. You could simply use a linear model, like OLS, or a maximum-likelihood estimator, like a logit or probit model, but given that our sole goal is to get an accurate model, rather than understand the constituent relationships between IVs and the DV, researchers should opt to use an ML model like causal forest.

Causal forest models adapt the standard random forest model from Breiman (2001). First, random forests work through a "wisdom of the crowds" effect, where a large number of "dumb" predictions result in a "smart" prediction when compiled together. The model is a called a random forest because it uses a "forest," or set, of decision trees to provide predictions. Specifically, a decision tree operates quite simply, given a set of variables/predictors $X = (X_1, ..., X_p)$ it will create decision-rules that maximize/minimize some criteria, normally prediction/accuracy. The random forest makes each tree "dumb" through a few different *possible* options (these are also known as tuning parameters and not all of these parameters have to be used): 1) limiting the number of variables (e.g., 3 out of 10 variables), by randomly selecting some subset, available to any given tree; 2) requiring a minimum number of observations to be in any given "bin," or prediction value/category, for each tree; 3) requiring a minimum number of observations to be located down a split in the tree (regardless of the numbers in the eventual prediction bins); and/or 4) limiting the maximum number of splits for any given tree. Regardless of the minutiae of how random forests are constructed, they are incredibly flexible and tend to outperform most linear models in terms of raw predictive power.

Causal forests build upon the general random forest concept by introducing another tree-building parameter: honest subsample splitting. Essentially, the honesty parameter ($\alpha \in (0, 1)$) determines what portion of the sample is used to determine splits in each tree and what portion of the sample is used to estimate the effects of each bin/leaf. This specific addition that CF contributes

to random forests has numerous benefits, enumerated in Athey and Imbens (2016), but suffice it to say that it provides better coverage than "nonhonest" models. A subsequent article by Wager and Athey (2018) prove that CF provides valid statistical inferences for treatment effects. Overall, this approach reduces the bias of individual tree estimates and improves the overall capability and stability of CF estimates relative to classic random forests. Moreover, given the usual selection on observables assumptions (i.e., unconfoundedness), CF estimates converge on the truth (i.e., are consistent) and are asymptotically normal.

Now, looking back at the DRML framework, we can see how CF can be incorporated. Specifically, we can use it to model the relationship between the outcome and pretreatment covariates ($g_0(X)$) and treatment assignment and pretreatment covariates ($m_0(X)$). With these models in hand, particularly the treatment assignment model, we can then estimate treatment effects (θ_0) in the first equation $Y = D\theta_0 + g_0(X) + \zeta$. If treatment is nonrandom, researchers can opt to calculate treatment effects using an estimand from Li, Morgan, and Zaslavsky (2018), which should be used in the case there there is poor overlap in treatment propensities (i.e., when the propensities between the treatment and/or control groups are very close to 0 and/or 1 and do not completely overlap). Essentially, treatment effects are only aggregated from overlapping treatment propensities between the treatment and control groups. The specific DRML calculation of treatment effects is done using augmented inverse-propensity weighting (AIPW) Glynn and Quinn (2010).

Overall, as shown in Rametta and Fuller (2024), there are *essentially* no drawbacks to estimating ATEs using DRML. Consequently, I recommend all researchers leverage these benefits and report both their DRML and standard ATE estimates. It is also critical to note that DRML in combination with CF have significant benefits for the calculation of CATEs as well, detailed below in the next section. Ideally, and often, DRML estimates will be largely similar to standard ATE estimates in experiments (see Figures 3.3 and 3.4). However, DRML CATE estimation *often* leads to significantly divergent results from standard estimation approaches, like OLS.[4]

[4]See Fuller and Rametta (2024) for examples of highly divergent results.

Models and Treatment Effect Heterogeneity for Subgroup Analysis

> ...restricting the heterogeneity analysis to pre-registered subgroups amounts to throwing away a large amount of potentially valuable information, especially now that many researchers collect large baseline data sets. It should be possible to use the data to discover ex post whether there is any relevant heterogeneity in treatment effect by covariates.
>
> Chernozhukov et al. 2023, pp. 2

This section discusses the application of ML for detecting heterogeneity in treatment effects and, if detected, for estimating CATEs. In general, and in partial contrast to Ratkovic (2021), this paper agrees with Chernozhukov et al. (2023) in their view of subgroup analysis: heterogeneity analysis should be applied to the full sample and with a saturated set of covariates, *so long as* some form of sub-sample splitting/honest estimation is implemented to guard against overfitting. Details on when and how CATEs should be reported are provided in the section below.

Many theories throughout the social sciences posit the existence of heterogeneous treatment effects. Unfortunately, in the context of most finite-sample experiments, the estimation of heterogeneous treatment effects through regression interaction terms is highly problematic. Despite this, for experiments with binary treatments, estimating CATEs using a simple multiplicative interaction term in a linear regression on the entire sample is an incredibly frequent occurrence. There are a number of potential issues with this approach to estimating interaction effects discussed at length in recent literature (Blackwell and Olson 2022). Moreover, given the complexity of relationships in the social world, our level of epistemic confidence that simple preconceived models are correct and unconfounded by other observables should be low. In some cases, researchers employ more advanced tools; the ensemble learning method proposed by Grimmer, Messing, and Westwood (2017) is one example. However, as noted by Ratkovic (2021), standard ML tools and ensemble methods target prediction by default. In order to make them useful for inference, they need further modification (as in Wager and Athey 2018; Chernozhukov et al. 2023).

As such, researchers should employ one of the causal machine learning models specifically designed to target treatment effect heterogeneity. While this literature is moving quickly, the current landscape has four main competitors: causal forests (Wager and Athey 2018), flavors of causal BART (Hahn, Murray, and Carvalho 2020; Dorie et al. 2022), X-learner (Künzel et al. 2019), and very recently, weighted split conformal quantile regression (or CQR) (Lei and Candès 2021).

While it is out of the scope of this paper to delve deeply into the specific benefits and drawbacks of these different approaches here, any of them will outperform standard ML or parametric methods for detecting treatment effect heterogeneity and estimating CATEs. Both weighted split CQR and Bayesian approaches go further, allowing for the estimation of individual treatment effects, while X-learner and causal forests do not. However, the `R` implementation of causal forest allows users to estimate average treatment effects, conditional average treatment effects for any particular subgroup, and implements several tests for heterogeneity.

Illustrating and Reporting Results Given that the analysis explained above is exploratory in nature, it is important to report the use of these methods and their results in a transparent and principled manner. While the use of these methods should, in fact, increase our confidence in our ability to accurately calculate ATEs and CATEs and ensure that spurious CATEs are not calculated, the hesitation surrounding these types of analyses is well-justified (Nosek et al. 2018). Given the importance of transparency, and likely initial hesitation surrounding the use of these methods, this section recommends what results and details should be reported.

First and foremost, all tuning parameters tested and used for any machine learning models, whether for randomization checks or for detecting heterogeneity, should be reported. As mentioned above, and detailed more below, testing multiple tuning parameters to ensure that any of these given models are well constructed is necessary to ensure that results from said model are accurate. Second, and relatedly, heterogeneity tests should be reported with both coefficient estimates and statistical significance for both the estimation of the ATE and the heterogeneity test itself. If there are multiple treatments, arms, or subsetting necessary to calculate ATEs, each of these should have their own heterogeneity test reported as well (if power is sufficient to allow confident testing in the first place). If heterogeneity is detected, then variable importance should be reported for each heterogeneous treatment effect identified.

As an example of principled reporting of variable importance, I detail here the method introduced by Wager and Athey (2018). To construct this measure, first a matrix is generated that, for each node from 1 to the maximum tree-depth, sums the amount of times each variable appears as a split from any tree in that node. This results in a matrix with rows equal to the number of variables and columns equal to the maximum tree depth and each cell being the number of times that given variable appears as a split in that node depth for all trees. Importantly, this matrix is

limited to 4 columns/nodes (the default for this function), excluding subsequent splits. Second, each row is divided by its sum, to calculate the node incidence proportions for every variable (e.g., variable 1 appears 40% of the time in node 1, 20% in node 2, 30% in node 3, and 10% in node 4 resulting in a vector $\vec{V_1} = [.40, .20, .30, .10]$).

The weights for the matrix (vector $\vec{W}$) are determined by a tuning parameter called "decay" which, in our iteration, is set to 2 and leads to an exponential discounting of a variable being used for a split at each depth. So variables that split first are fully weighted, those that are split second are weighted by $\frac{1}{4}$, third $\frac{1}{9}$, fourth $\frac{1}{16}$ (i.e., $\vec{W} = [1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}]$). The weights are then divided by the sum of weights (in this instance, divided by $1\frac{61}{144}$) resulting in relative weights being applied to each node column. Then the sum is finally calculated for each variable row vector, leading to a *relative* variable importance score for each variable.

Now this measure of variable importance does vary somewhat conditional on the random seed used in the generation of the causal forest (Chan and Ratkovic 2020), particularly with a large number of pretreatment covariates. This issue arises because with different random seeds, different subsamples of the total sample are used to define the splits and leafs on the causal trees. To address this issue, and generate uncertainty intervals around the variable importance scores, I generate cross-validated variable importance scores (CV-VIs) by replicating the causal forest estimation process over a large number of random seeds.

Subsequently, CATEs should be calculated for any variable whose variable importance is above $\frac{1}{p}$, where p is the number of pretreatment covariates included, and be reported in a plot with the ATE also highlighted. Based on these plots, researchers can discern partial dependencies, and it should be clear which variables contribute to conditionality. While these methods can also be used to calculate results and generate plots similar to those resulting from standard CATE estimation procedures, this should not be done for any variables that are not clearly categorical. This is because it is necessary to draw relatively arbitrary boundaries on these continuous variables, so it is better to report individual-level, continuous CATE plots for these variables.

Furthermore, there may be interaction effects between multiple important variables as well. If there is reason to believe that two important covariates are interactive, then a partial dependence plot (PDP) should be reported for these variables. These plots should clearly illustrate the combined influence of any two covariates, and give both the researcher and reader an understanding if any

interaction effect is occurring from these two variables. Ideally, PDPs for all combinations of important variables should be reported in the appendix.

Overall, if these results and figures are reported, it should be clear to both the researcher and any reader or reviewer what conditionality, if any, is occurring in the given analyzed experiment. Furthermore, as illustrated below, this will also help researchers avoid any spurious conclusions resulting from standard CATE estimation approaches.

Vignette: Using the Checklist

This section proceeds by using data from Barber and Pope (2019) and a working paper by Fuller, De la Cerda, and Rametta (2024) to illustrate the usefulness of following the proposed checklist. To quickly summarize the structure of these papers, both are testing for the effects partisan cues have on policy support. Specifically, Barber and Pope (2019) examines whether "liberal" and "conservative" policy support cues from, at the time, President Donald Trump influence policy support across a large set of issues, particularly among Republicans. In general, they find that both conservative and liberal cues have greater effects for stronger partisans and, particularly, for individuals with low political knowledge (as measured by an additive scale). Fuller, De la Cerda, and Rametta (2024) follow a similar structure, but instead look at whether cues from Democratic and Republican groups influence support for three issues ($12 minimum-wage, increasing Border Patrol funding, and increasing farm subsidies), separately and combined. This paper finds that treatment effect heterogeneity only exists, generally, for out-party cues and is primarily conditioned on an individual's level of positive and negative partisanship.

Somewhat obviously, because these experiments have already been fielded and these papers already written it is impossible to illustrate the first few components of the list, specifically the design-phase. However, I am able to discuss the importance of liberally and accurately measuring the covariate space. As noted by Fuller and Rametta (2024), Barber and Pope does not use any methods to robustly measure ideology nor partisan affect, instead they use coarse measures. Specifically, they measure: Trump approval, seven-category party identification, political knowledge, age, income, political interest, self-identified political ideology, education, race and gender. In contrast, Fuller, De la Cerda, and Rametta use scaling methods including Bayesian Aldrich-McKelvey (BAM) scaling (Hare et al. 2015) (for ideology), ratio unfolding (for feeling thermometers and trust scores),

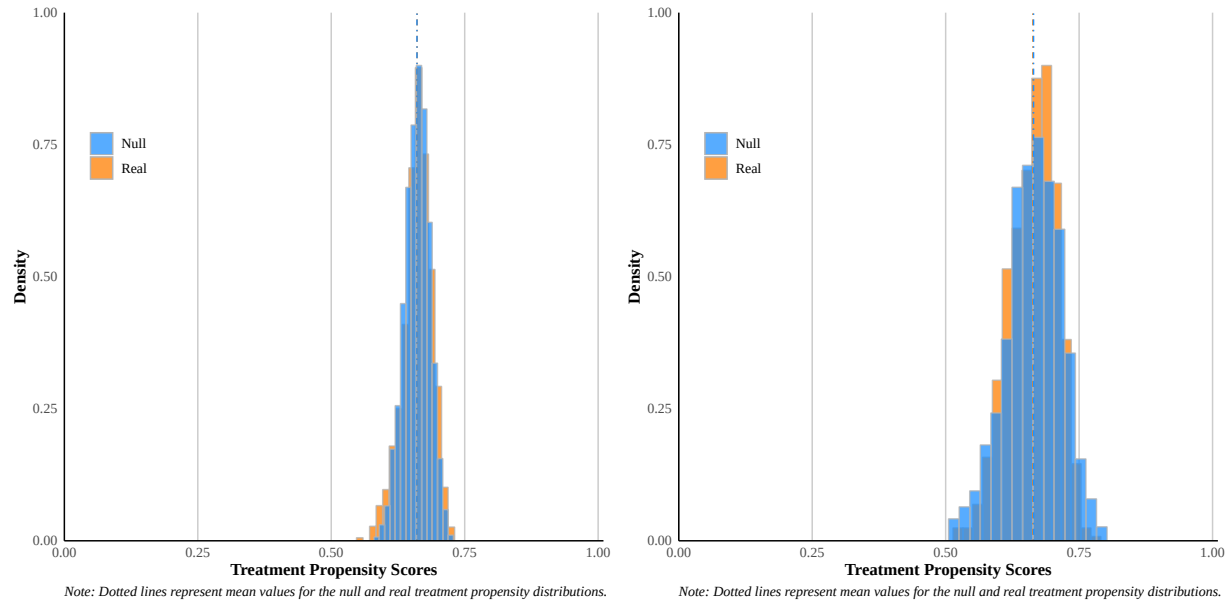
and a Graded Response Model (for social distance; similar to item response theory models). Overall, they measure age, education, race, gender, political interest, thermometer-level affect, social distance, political ideology, spatial knowledge (recovered from the BAM estimation), as well as in and out-party trust scores. As will be made evident below, the robust or coarse measurement of latent concepts, like partisan affect and ideology, alongside ML methods like CF can lead to significantly divergent results.

Moving onto the next step, I use a method introduced by Rametta and Fuller (2024), called the Balance Permutation Test to determine if there are any randomization issues in either experiment. The specifics of this test are detailed in Rametta and Fuller (2024), but suffice to say, this method leverages a combination of permutation inference and machine learning to determine if there are any randomization issues (either a failure or covariate imbalance). It does this by using covariates to attempt to predict treatment assignment and then comparing the recovered treatment propensities to what they *should* be. The results of these tests are reported below in Figures 3.3 and 3.4, for Fuller, De la Cerda, and Rametta and Barber and Pope respectively. Based on these figures, indicated by the overlapping blue and orange distributions in each panel, there is no evidence of randomization issues in either experiment in any treatment arm.

We now move to the next part of the checklist. Here, I illustrate the use of CF to estimate average treatment effects and to detect treatment effect heterogeneity for both experiments, using the covariates available from each dataset. I follow the model specification set forth by Barber and Pope, with all policies pooled and respondents in any given analyzed treatment also being pooled together, *regardless* of their own party identification.[5] For Fuller, De la Cerda, and Rametta, I instead specify the model to analyze in- versus out-party cues by party identification. Specifically, I examine for both Democratic and Republican respondents whether there is any treatment effect heterogeneity for in- or out-party cues, for every policy individually and with all policies pooled as well. To be clear, this difference in model specification *does* have implications for heterogeneity checks, which will be clearly shown below.

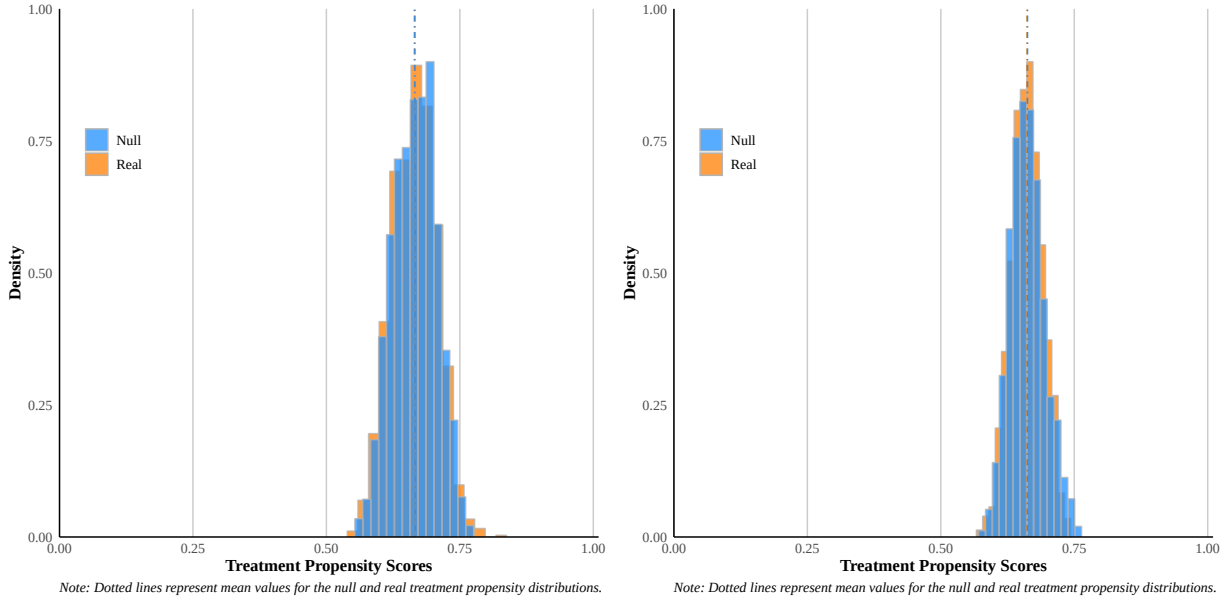
The standard and doubly robust average treatment effects are reported by PID and cue for Fuller, De la Cerda, and Rametta and Barber and Pope in Figures 3.3 and 3.4, respectively.

[5]Fuller and Rametta (2024) notes significant issues with this specification and reanalyzes this experiment by PID rather than pooling across all respondents.



(a) Democratic Treatment, Democratic Respondents

(b) Democratic Treatment, Republican Respondents

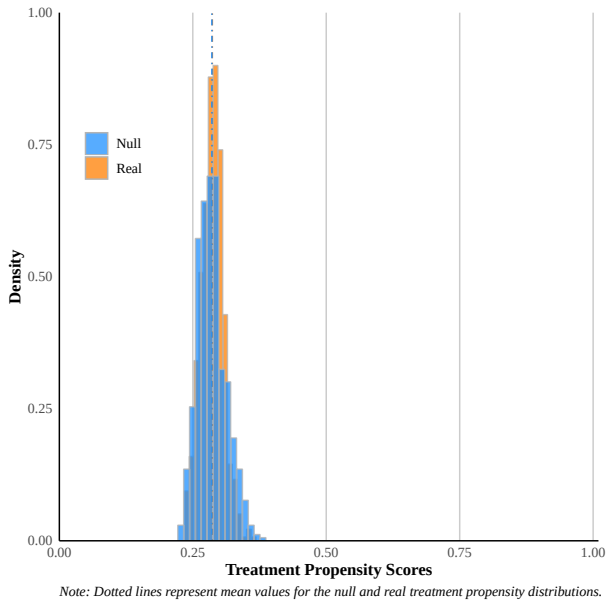


(c) Republican Treatment, Democratic Respondents

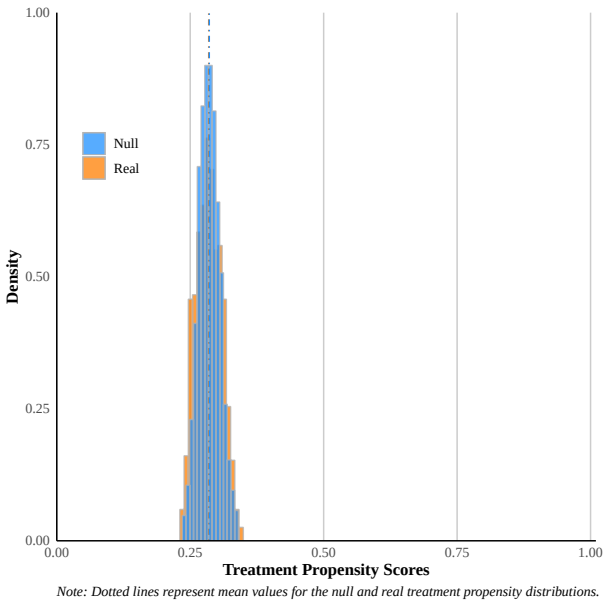
(d) Republican Treatment, Republican Respondents

FIGURE 3.1. Fuller, De la Cerda, and Rametta (2024): Permutation Balance Test Results

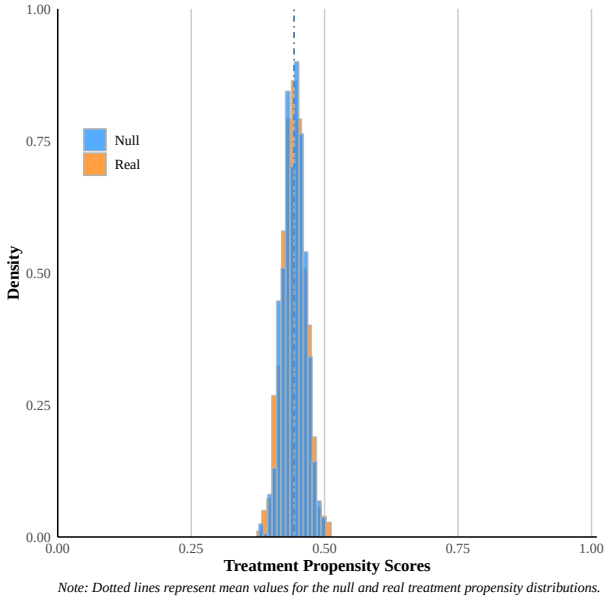
Looking at Figures 3.3 and 3.4 we clearly see that the doubly robust estimation of ATEs by PID is not significantly different from the standard OLS approach for either Fuller, De la Cerda, and Rametta nor Barber and Pope. This is expected in most contexts, assuming that treatment



(a) Conservative Trump Treatment



(b) Liberal Trump Treatment

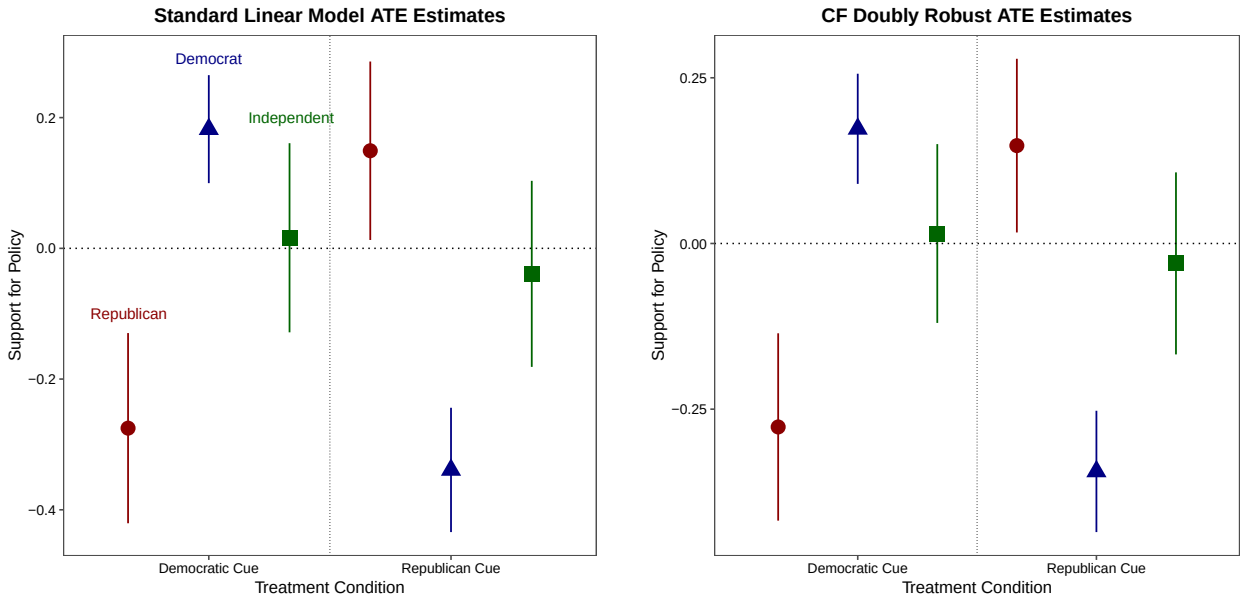


(c) Congressional GOP Treatment

FIGURE 3.2. Barber and Pope (2019): Permutation Balance Test Results

assignment was successful. The smallest difference comes in Figure 3.3, which is due to the fact that this experiment used block randomization by PID.

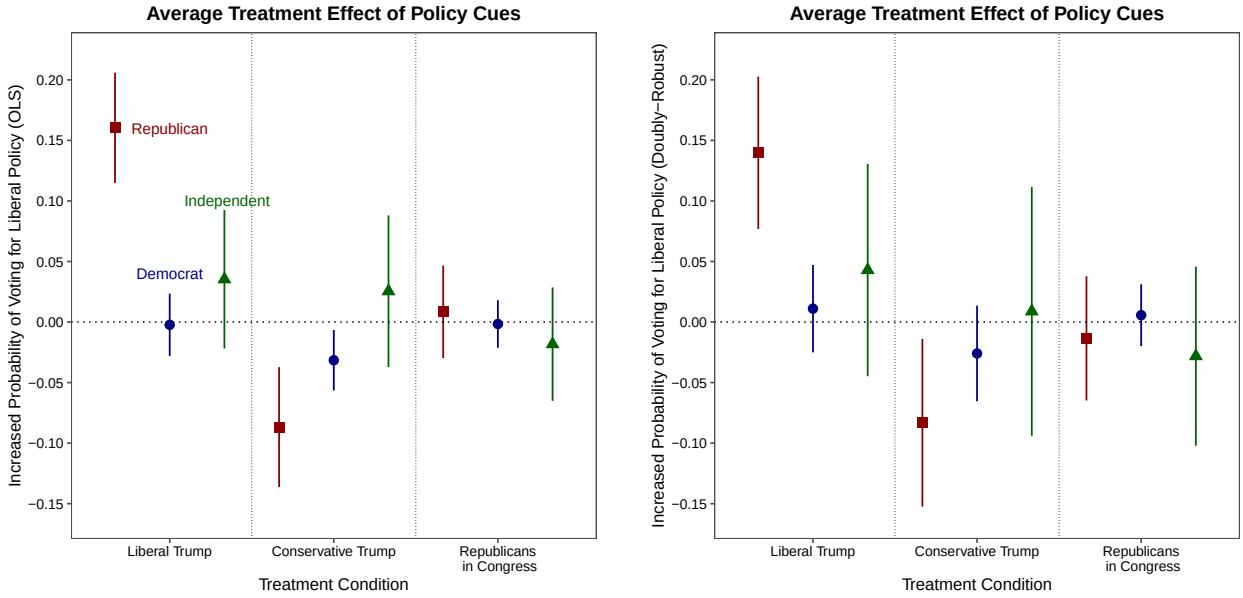
Moving on to the estimation of conditional effects, Tables 3.1 & 3.2 report the results of these heterogeneity tests for each analysis, within each experiment. Based on the standard specification set forth by Barber and Pope, treatment effect heterogeneity is identified for both the liberal and



(a) Standard Results

(b) DRML Results

FIGURE 3.3. Fuller, De la Cerda, and Rametta (2024): ATEs for All Respondents by PID and Treatment Condition



(a) Standard Results

(b) DRML Results

FIGURE 3.4. Barber and Pope (2019): ATEs for All Respondents by PID and Treatment Condition

conservative Trump cues. For Fuller, De la Cerda, and Rametta, there is evidence of heterogeneity for out-party cues, among pooled policies. For specific policies, only the border funding policy for

Republican respondents is important for an in-party cue, and out-party cues are consistently important for all three policies for Democratic respondents. Clearly, there appears to be heterogeneity for out-party cues, so I explore those below with variable importance calculations and plots.

TABLE 3.1. Barber and Pope (2019):
Treatment Effect Heterogeneity Checks

Treatment	Party ID	All Policies
Conservative Trump Cue	Pooled	***X***
	Republicans	
	Democrats	
	Independents	
Liberal Trump Cue	Pooled	***X***
	Republicans	
	Democrats	
	Independents	

TABLE 3.2. Fuller, De la Cerda, and Rametta (2024):
Treatment Effect Heterogeneity Checks by Policy Cue and Pooled

Treatment	Party ID	Min. Wage	Border Funding	Farm Subsidy	Pooled
Democratic Cue	Democrats				
	Republicans		*X*		***X***
	Independents				
Republican Cue	Democrats	*X*	*X*	*X*	***X***
	Republicans		*X*		
	Independents				***X***

First, however, let us turn back to the results from Barber and Pope. Given that I identify effect heterogeneity for both treatments, variable importance is calculated for both and is reported in Table 3.3. It appears that, for both liberal and conservative treatments, Trump Approval is important for predicting heterogeneous treatment effects. Similar to their published findings, political knowledge is also very important for the conservative Trump treatment, but in contradiction, it is not important at all for the liberal Trump treatment.[6]

[6]It is important to remember here that that with the correct model specification, there is no evidence of treatment effect heterogeneity.

TABLE 3.3. Barber and Pope (2019):
All Respondents Variable Importance

	Trump Treatment	
Variable	Conservative	Liberal
Trump Approval	0.24	0.31
Political Knowledge	0.21	0.04
Age	0.14	0.08
Income	0.13	0.15
PID-7	0.07	0.17
Interest	0.06	0.02
Ideology	0.06	0.05
Education	0.04	0.14
Race	0.04	0.00
Gender	0.02	0.03

Based on both the general heterogeneity checks and the variable importance from the table above, it would appear that CF confirms the findings from Barber and Pope, at least for the conservative Trump cue, albeit with Trump Approval being identified as just as important as political knowledge. However, when we proceed to the next step in the checklist, checking conditionality, and delve into the CATEs for Trump Approval, a different picture emerges. Figure 3.5 reports individual CATEs (Y-axis, positive values correspond to an increased likelihood of supporting a liberal policy) across Trump Approval (X-axis) and conservative versus liberal cues (panels), with the colors of each respondent's CATE corresponding to their self-reported PID (red, Republican; blue, Democrat).

What becomes immediately clear is that Trump Approval is operating as a continuous proxy for PID, where whether an individual has a positive, null, or negative response to a given cue is directly determined by their party identification. This makes intuitive sense, as people who identify with Trump's party, Republicans, are going to be more likely to approve of him, and vice-versa for Democrats. Obviously there are exceptions to this rule, where some Republican respondents do not approve of Trump, and some Democrats instead approve, but in general this PID proxy trend appears to hold for both treatments. This result motivates the respecification of the model for CF, where I instead look at whether the cues have heterogeneous treatment effects *within* a given party. Across all PIDs, and most importantly for Republicans, there is no evidence of treatment effect heterogeneity (as summarized above in Table 3.1).

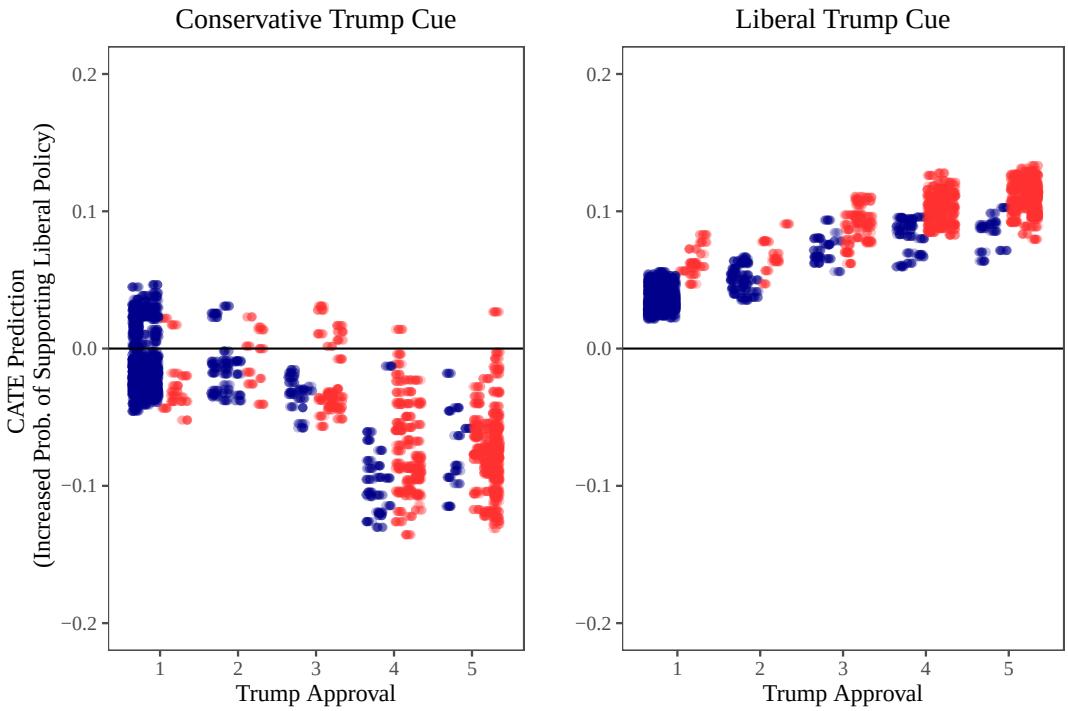


FIGURE 3.5. Barber and Pope (2019):
Trump Approval CATEs by Treatment and Respondent PID

In general, these results highlight the importance of using these heterogeneity checks. Simply put, these tests, one of the best in class for detecting heterogeneity, do not identify nearly *any* treatment effect heterogeneity in this experiment, even though the main conclusions of that paper are derived directly from conditional average treatment effects. Overall this example highlights two potential pitfalls that the checklist seeks to help researchers avoid: 1) model misspecification and 2) calculating CATEs when treatment effect heterogeneity has not been confidently shown to exist. Simply put, model specification is incredibly important for calculating accurate ATEs and even more so for detecting heterogeneity. And relatedly, calculating CATEs through standard estimation techniques, even when these calculations are preregistered in an analysis plan, can lead to spurious or at least low-confidence conclusions.

Additionally, I suspect that the lack of heterogeneous treatment effects in the correctly specified models is, at least in part, due to rough measures of ideology, a lack of partisan affect measures (e.g., social distance questions or feeling thermometers), and a coarse measure of support for Donald Trump. As mentioned above, the accurate measurement of latent concepts and variables is critical for good science and reliable and trustworthy conclusions. Simply put, if the authors had

better measures of these underlying concepts, I would expect to identify significant treatment effect heterogeneity.

To show a contrast to this paper, I continue to show how to report conditionality checks and CATE estimations once a model is well-specified and treatment effect heterogeneity has been identified, using Fuller, De la Cerda, and Rametta (2024). Variable importance for all treatments with well-identified heterogeneity are reported below in Figure 3.6. Consistent across all three policies (and pooled policies), for out-party cues and for Democratic and Republican respondents, the measures of affect, specifically thermometer and social distance affect, are the most important variables for capturing treatment effect heterogeneity (with the exception of the border funding Republican cue and respondents). Interestingly, ideology is by far the most important predictor for independent respondents. Those who are on the right (conservative) part of the scale have positive reactions to Republican cues, whereas those on the left (liberal) part of the scale have negative reactions. This makes sense, in that it appears that ideology for independent respondents acts as almost a proxy for party identification.

Based on these results, I proceed to explore treatment effect conditionality by both the social distance and thermometer scales, broken down by PID, in Figure 3.7. Specifically, this figure shows individual CATEs for each variable (rows) by PID (columns) for pooled, out-party cues. Each panel also shows the ATE (with 95% confidence bands) for the given PID, as the red-line across the plot.

Clearly, there is significant treatment effect heterogeneity across both thermometer-level affect and social distance. For Democratic (Republican) respondents, those on the left (right) side of both scales have very large negative treatment effects whereas those on the right (left) side of the scale have essentially null effects.

Based on the importance of both these measures of affect, I now explore the potential interaction CATEs for both thermometer-level affect and social distance. To do this, I calculate the partial dependence between thermometer-level affect and social distance for out-party cues with pooled policies for Democratic and Republican respondents. This is done simply by calculating the CATE by *both* a given level of thermometer and social distance affect from both pooled, out-party cue causal forests. Figure 1.4 reports the partial dependence plots for thermometer (Y-axis) and social distance (X-axis) affect for out-party cues by PID (columns) for the pooled policies.

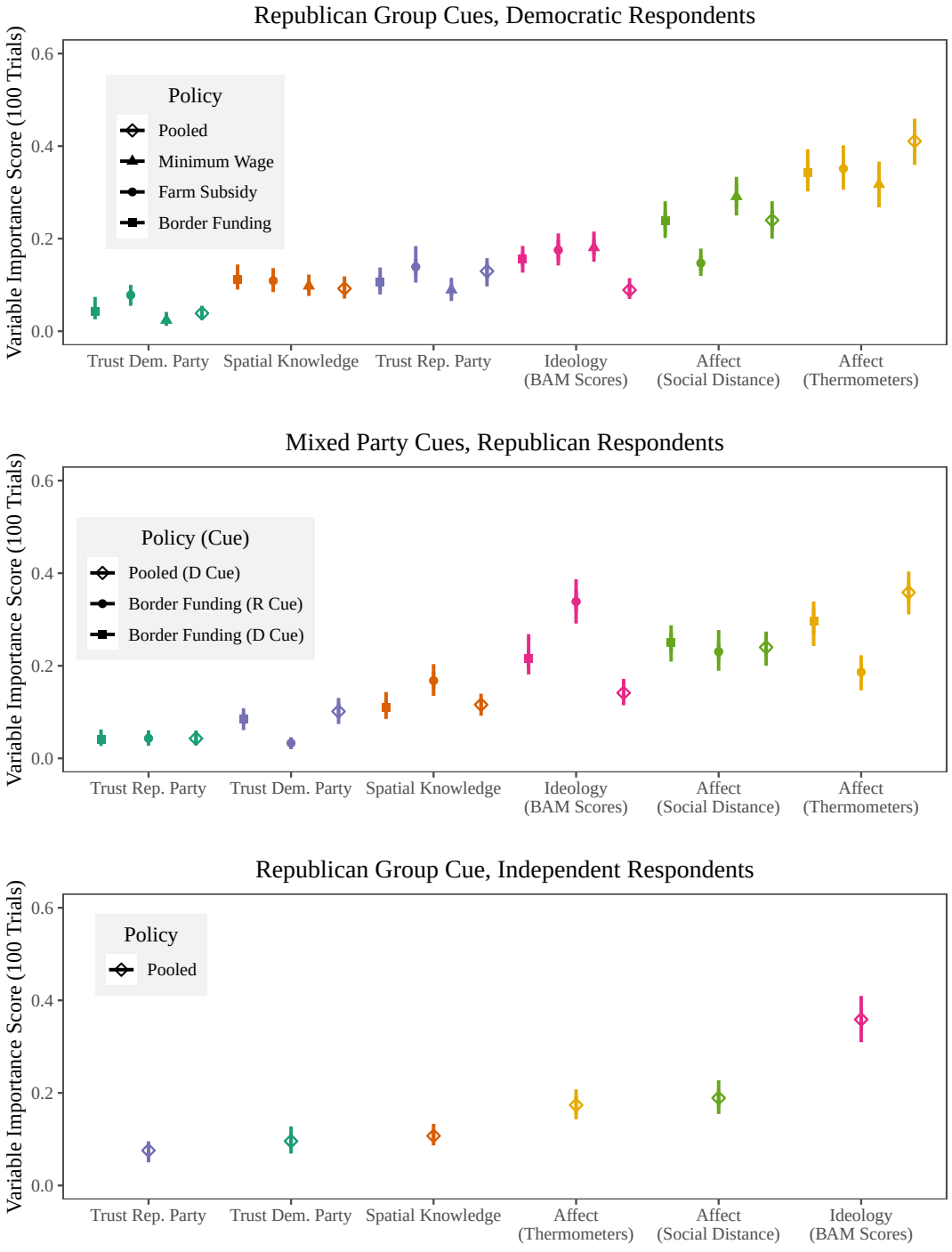
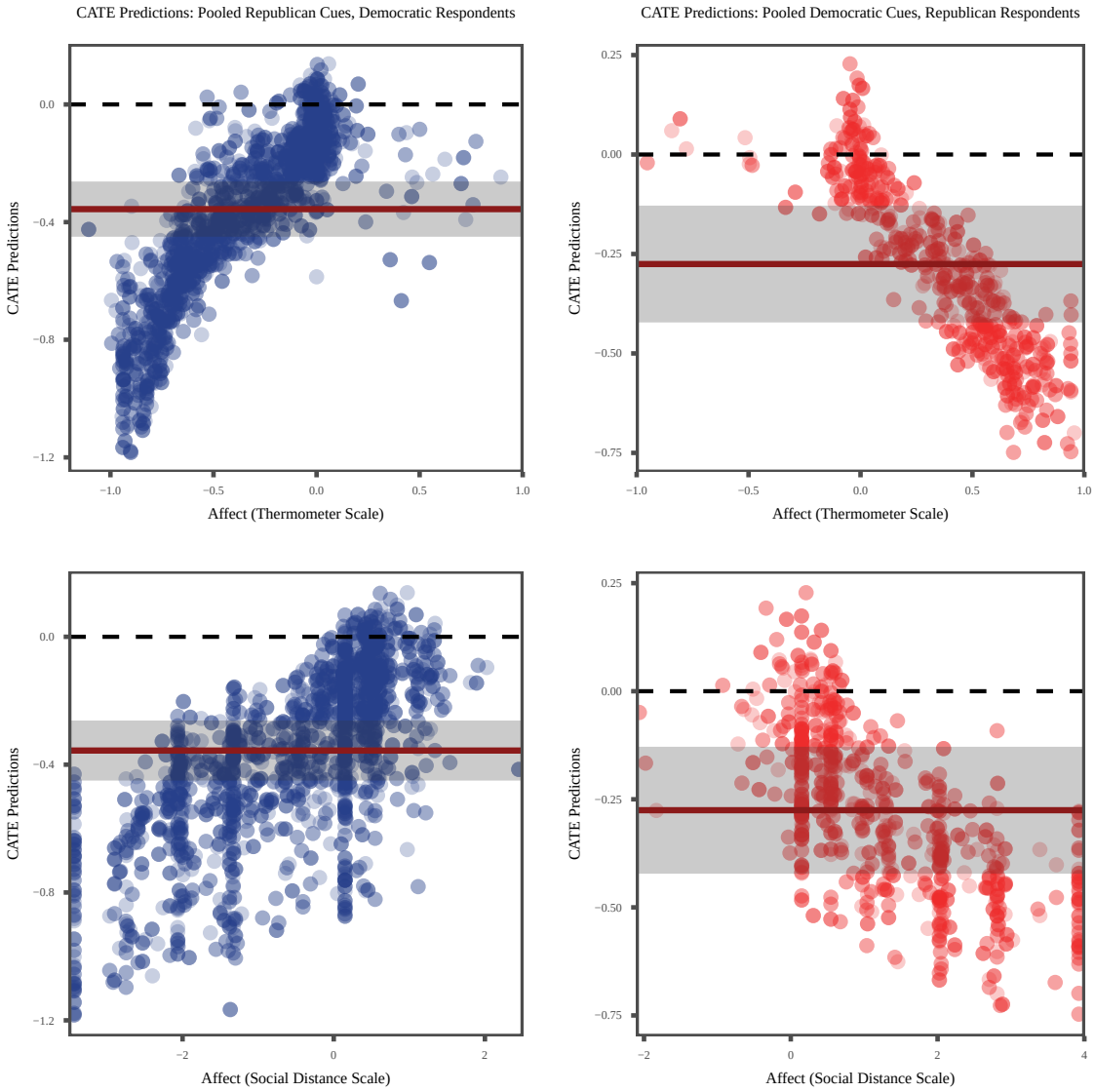


FIGURE 3.6. Fuller, De la Cerda, and Rametta (2024):
Variable Importance for Estimating Heterogeneous Treatment Effects

Clearly, Figure 3.8 illustrates a significant interaction effect between these two measures of affect. Specifically, those who are at the extreme end of both affect scales have greater negative

FIGURE 3.7. Fuller, De la Cerda, and Rametta (2024):
CATEs for Democratic Respondents by Variable and PID



treatment effects than those who are less extreme on one or both of the scales. Simply put, those who have higher partisan affect as a sum of both measures of affect, have greater negative treatment effects than their counterparts.

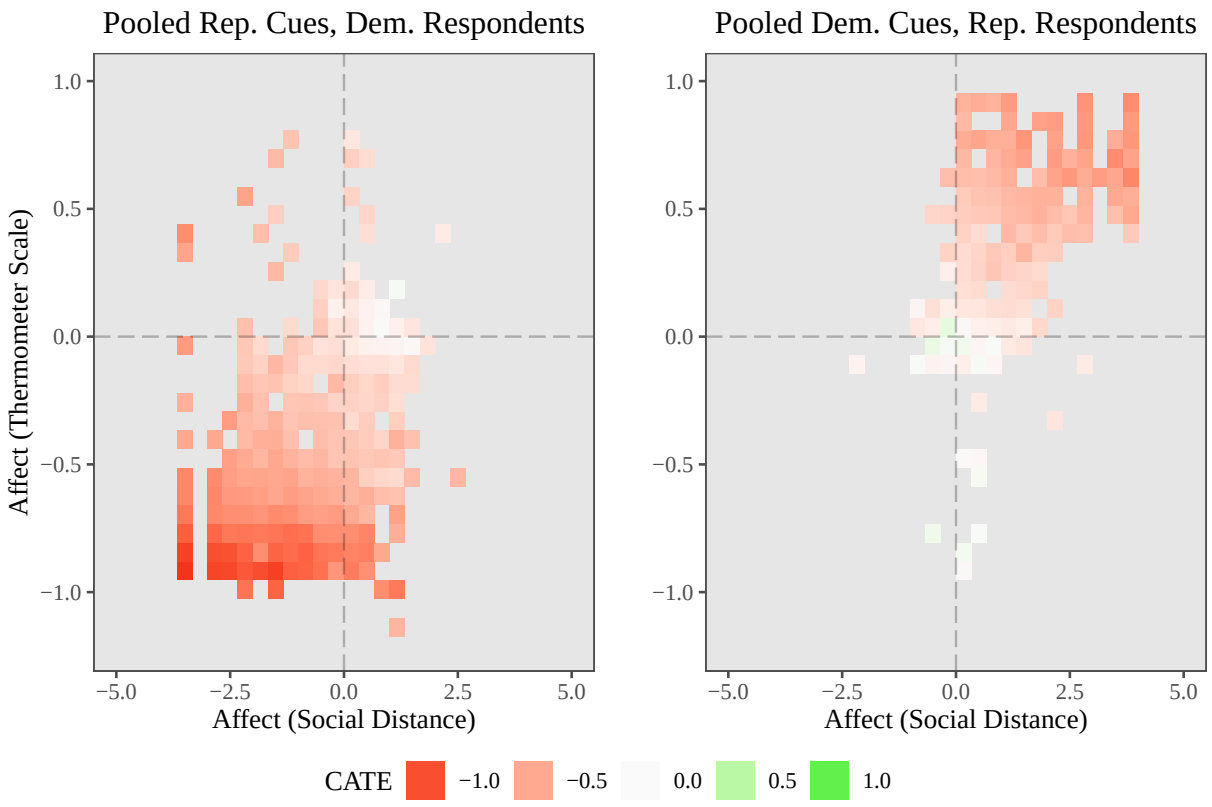


FIGURE 3.8. Fuller, De la Cerda, and Rametta (2024):
Partial Dependence Plots for Thermometer & Social Distance Affect

Discussion and Future Work

Overall, this paper outlines a general framework for researchers to conduct experimental analyses exploiting ML techniques and applies this framework to two existing experimental datasets to illustrate its utility. As evidenced by these two applications, standard methods can either obscure interesting treatment effect heterogeneity or, even worse, falsely indicate the existence of conditionality in treatment effects when there is none. Based both on the theoretical arguments and vignettes presented above, I argue that this framework should be applied in all cases when analyzing experimental data. For all intents and purposes, using advanced methods to accomplish these goals has many benefits but a few drawbacks.

As noted in previous sections of this paper, there is significant space for future work. Broadly, I would divide these into four categories: variable importance, new applications of ML, usability, and proliferation and reevaluation.

Advances are being made at a breakneck pace for determining variable importance, particularly interactive importance. The next step is to both implement these methods in usable ways and determine their effectiveness. Generally, however, these VI measures will be particularly important in the context of social science research because they are both better able to deal with multicollinearity and identify important and interesting interactions between variables.

New applications of ML, particularly in cases where prediction is the primary goal, will also be a fruitful research area. Already, Rametta and Fuller (2024) have implemented a test that leverages ML for identifying randomization issues in experiments and future work should also explore areas like detecting non-random attrition from panel surveys and experiments.

One major issue with the causal ML literature is their usability in practical scenarios. Additionally, implementing many of the variable importance measures mentioned above require both bespoke knowledge of the method and significant coding knowledge and skills. Even if one is able to implement these methods to their specific use-case, the possibility for making methodological or coding errors is significant. Already, literature is being published that uses ML methods, like CF, in naive, if not irresponsible, ways (See Zheng and Yin 2023, for an example of a poorly tuned and haphazardly applied CF model see). This is not because these authors or users have malintent, rather the methods themselves and the corresponding software/code are esoteric, needlessly complicated, or even broken.

Finally, as this is directly related to many of the areas noted above, there needs to be more work on spreading these methods in a responsible manner. Fundamentally, these new methods when used with DRML estimation have been shown to far surpass standard competitors (e.g., OLS or difference-in-means estimation) by having lower power requirements, higher average accuracy, better coverage guarantees, and the ability to identify treatment effect heterogeneity.[7] There are two specific avenues in which this could happen: Monte Carlo simulations and meta analyses. The former should clearly illustrate the benefits noted above by comparing DRML methods against standard methods using simulated data from known data generating processes (DGPs). While some of these simulations already exist (see Rametta and Fuller 2024), there should be more tests across numerous, diverse DGPs. The latter should reevaluate ATEs and CATEs from a diverse set of the experimental literature (Fuller and Rametta 2024, is an example of this kind of reanalysis).

[7]Some of these benefits are illustrated in Rametta and Fuller (2024).

Ideally, a combination of these two approaches, along with creating more accessible methods and software, would increase the use of these methods across political science.

Bibliography

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–1178. https://doi.org/10.1214/18-AOS1709. https://doi.org/10.1214/18-AOS1709.

Athey, Susan, and Stefan Wager. 2019. "Estimating treatment effects with causal forests: An application." *Observational Studies* 5 (2): 37–51.

Barber, Michael, and Jeremy C Pope. 2019. "Does party trump ideology? Disentangling party and ideology in America." *American Political Science Review* 113 (1): 38–54.

Blackwell, Matthew, and Michael P Olson. 2022. "Reducing model misspecification and bias in the estimation of interactions." *Political Analysis* 30 (4): 495–514.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019a. "Declaring and diagnosing research designs." *American Political Science Review* 113 (3): 838–859.

———. 2019b. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–859. https://declaredesign.org/paper.pdf.

Boudreau, Cheryl. 2021. "Transparency in Experimental Research." In *Advances in Experimental Political Science,* edited by James N. Druckman and Donald P. Green, 339–353. Cambridge University Press. https://doi.org/10.1017/9781108777919.024.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45:5–32.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–3660.

Chan, Zenobia, and Marc Ratkovic. 2020. "Improving Variable Importance Measures." In *Annual Meeting of the Society for Political Methodology (PolMeth XXXVII), University of Toronto, Toronto, CA,* 14–17.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2023. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.* arXiv: 1712.04802 [stat.ML].

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4): 1501–1535.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4 (1): 266–298. https://doi.org/10.1214/09-AOAS285. https://doi.org/10.1214/09-AOAS285.

Colas, Cédric, Olivier Sigaud, and Pierre-Yves Oudeyer. 2018. "How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments." *arXiv preprint arXiv:1806.08295,* https://arxiv.org/abs/1806.08295.

Dorie, Vincent, George Perrett, Jennifer L Hill, and Benjamin Goodrich. 2022. "Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning." *Entropy* 24 (12): 1782.

Druckman, James. 2022. *Experimental Thinking.* Cambridge University Press.

Druckman, James N, and Donald P Green. 2021. *Advances in Experimental Political Science.* Cambridge University Press.

Druckman, James N, Donald P Greene, James H Kuklinski, and Arthur Lupia. 2011. *Cambridge Handbook of Experimental Political Science.* Cambridge University Press.

Fuller, Sam, Nicolás De la Cerda, and Jack T Rametta. 2024. "Affect, Not Ideology: The Psychological Mechanisms of Partisan Information Processing." Working Paper. https://osf.io/preprints/osf/4k38a.

Fuller, Sam, and Jack T Rametta. 2024. "The Dangers of Calculating Conditional Effects: A Reevaluation of Barber and Pope (2019)." Working Paper. https://osf.io/preprints/osf/hwzsu.

Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R Biggers, and David J Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1 (1): 81–98.

Glynn, Adam N, and Kevin M Quinn. 2010. "An introduction to the augmented inverse propensity weighted estimator." *Political analysis* 18 (1): 36–56.

Green, Jon, and Mark H. White II. 2023. "Machine Learning for Experiments in the Social Sciences." In *Elements in Experimental Political Science.* Cambridge University Press. https://doi.org/10.1017/9781009168236.

Grimmer, Justin, Solomon Messing, and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–434.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2021. "Machine learning for social science: An agnostic approach." *Annual Review of Political Science* 24:395–419.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho. 2020. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)." *Bayesian Analysis* 15 (3): 965–1056.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll, and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59 (3): 759–774.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Vol. 2. Springer.

Imai, Kosuke, Gary King, and Elizabeth A Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

Imai, Kosuke, and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21 (2): 141–171.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22:129–146.

Jackson, Michelle, and David R Cox. 2013. "The Principles of Experimental Design and Their Application in Sociology." *Annual Review of Sociology* 39:27–49.

Jawadekar, Neal, Katrina Kezios, Michelle C Odden, Jeanette A Stingone, Sebastian Calonico, Kara Rudolph, and Adina Zeki Al Hazzouri. 2023. "Practical Guide to Honest Causal Forests for Identifying Heterogeneous Treatment Effects." *American Journal of Epidemiology,* kwad043.

Johannemann, Jonathan, Vitor Hadad, Susan Athey, and Stefan Wager. 2019. "Sufficient Representations for Categorical Variables." *arXiv preprint arXiv:1908.09874,* https://arxiv.org/abs/1908.09874.

Kagel, John H, and Alvin E Roth. 2015. *The Handbook of Experimental Economics, volume 2.* Princeton university press.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–4165.

Lei, Lihua, and Emmanuel J Candès. 2021. "Conformal Inference of Counterfactuals and Individual Treatment Effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (5): 911–938.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Mutz, Diana C, and Robin Pemantle. 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2 (2): 192–215.

Mutz, Diana C, Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1): 32–42.

Nosek, Brian A, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606.

Rametta, Jack T, and Sam Fuller. 2024. "The balance permutation test: A machine learning replacement for balance tables." Working Paper. https://osf.io/preprints/osf/xcwt9.

Ratkovic, Marc. 2021. "Subgroup Analysis: Pitfalls, Promise, and Honesty." In *Advances in Experimental Political Science,* edited by James N. Druckman and Donald P. Green, 271–288. Cambridge University Press. https://doi.org/10.1017/9781108777919.020.

Robinson, Thomas S, and Raymond M Duch. 2022. "How to Detect Heterogeneity in Conjoint Experiments." Working Paper, https://raymondduch.com/files/how-to-detect-heterogeneity-in-conjoint.pdf.

Scheel, Anne M, Mitchell RMJ Schijen, and Daniël Lakens. 2021. "An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports." *Advances in Methods and Practices in Psychological Science* 4 (2): 1–12.

Schinka, John A, Wayne F Velicer, and Irving B Weiner. 2013. *Handbook of Psychology: Research Methods in Psychology, Vol. 2.* John Wiley & Sons, Inc.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Wilcox, Clyde, Lee Sigelman, and Elizabeth Cook. 1989. "Some like it hot: Individual differences in responses to group feeling thermometers." *Public Opinion Quarterly* 53 (2): 246–257.

Zheng, Li, and Weiwen Yin. 2023. "Estimating and Evaluating Treatment Effect Heterogeneity: A Causal Forests Approach." *Research & Politics* 10 (1): 20531680231153080.

Discussion

At the risk of being redundant, the implications of these chapters can be divided into two broad categories: substantive and methodological. First, the substantive findings of my first chapter are novel in comparison to the previous literature on partisan policy cues and information processing. We find that only affect, not ideology nor any other covariate, influence responses to partisan policy cues and only for out-party cues. These results hold across the ideological spectrum of policies and have substantively significant effects on the level of support.

Methodologically, the first chapter illustrates the application of causal ML methods and also highlights the benefits from leveraging them for subgroup analyses and CATE estimation. While not reported directly in the paper, the use of CF leads to significantly divergent results compared to standard methods. In fact, both ideology and affect (among other covariates) appear to be important when using OLS. Ideally, this chapter not only contributes to the literature on partisan information processing but also represents an illustrative example of the benefits of causal ML.

My second chapter introduces a new method, the balance permutation test, that represents a marked improvement over standard methods for detecting randomization issues in experiments. Using the predictive capability of ML methods, and their non-parametric nature, we are able to confidently determine whether any randomization issues exists in our experiment. Not only that, any minor-to-moderate randomization issues can be directly addressed by using treatment propensities in the DRML treatment effect estimation framework. I hope that this method becomes standard in the identification and adjustment of randomization issues in social science experiments.

Finally, my third chapter provides guidelines for how researchers should use these methods in their own research. Given the lack of literature on this topic, this information is highly necessary. While some of the methods have been available for a significant amount of time, some for over five years, their use has not proliferated in the social sciences, generally, and political science, specifically. These methods are not particularly easy to understand or to use, and although their capabilities are significant, it is not clear how to bring about their benefits. I aim for a future

iteration of this chapter to be the initial guidance for researchers hoping to leverage these methods for their own work.

Overall, this dissertation not only introduces causal ML and a new method for detecting randomization issues, but it also shows how the use of these methods in a principled way improves the quality of experimental research. I hope this work represents my first set or batch of research within this space and represents the promise that this entire branch of literature has.

Future Work

This dissertation points to multiple avenues of future work. First, these causal ML methods enable much more robust and fine-grained experimental analyses, especially of subgroups. Their ability to detect and identify treatment effect heterogeneity enables, as in my first chapter, the interrogation of more nuanced questions. Researchers should more liberally measure pre-treatment covariates and explore potential (unexpected) heterogeneity. In my own work, I plan to continue to leverage these methods to explore more questions relating to political behavior, particularly related to determinants of anti-democratic attitudes.

Second, as I mention in my third chapter, more research should be done on introducing these methods and integrating them into experimentalists' research. I plan to expand my third chapter into a publication that helps with just that. Additionally, I also plan to conduct a meta-analysis of existing experimental research to re-evaluate subgroup analyses and to test for randomization issues using the balance permutation test. Ideally, similar to my work in Fuller and Rametta (2024), I would identify significant discrepancies between standard methods and causal ML and highlight the importance of using causal ML.

Additionally, there are doubtlessly numerous novel applications in which machine learning may outperform standard methods (like for detecting systematic attrition). Broadly, these are often cases when prediction, rather than inference, is the primary goal of a model. However, as illustrated in my first chapter, machine learning can also be adapted for use in inference. This space, in particular, will likely grow significantly in the future, especially with the adaptation and application of interpretable machine learning methods for unpacking the "black-box" nature of these models. In particular, work on variable importance measures has the potential to provide a much more detailed and nuanced understanding of the relationship between any dependent variable (or set of DVs) and

an outcome. In truth, the gap in worse interpretation/inference and higher computing time between machine learning methods and standard parametric approaches, like OLS and MLE, is rapidly decreasing. Additionally, the costs and limitations of using parametric methods, in comparison to semi- and non-parametric methods, is critically underappreciated in the social sciences today.

These issues include model misspecification (including restrictive functional form assumptions), a high number of researcher's degrees of freedom, and the curse of dimensionality. All of these issues *in isolation* are significant threats to model validity, let alone when considered together. MLE models, like logistic regression, are particularly bad. In fact, they combine the worst parts of ML models and the worst parts of parametric models: they are incredibly difficult to interpret correctly and they have incredibly restrictive functional form assumptions. While much work, mostly in econometrics, is dedicated to remedying or mitigating many of the issues that plague these parametric models, particularly OLS, there are certainly diminishing marginal returns: it is like putting lipstick on a pig.

This is all to say that I also believe that there should be a reevaluation, and continuing comparisons and investigations of, standard, parametric methods along with new and existing ML methods. While I believe that simple parametric models will continue to have a place in social science research, their dominance should be questioned. I hope to contribute to this reevaluation in future work.

APPENDIX A

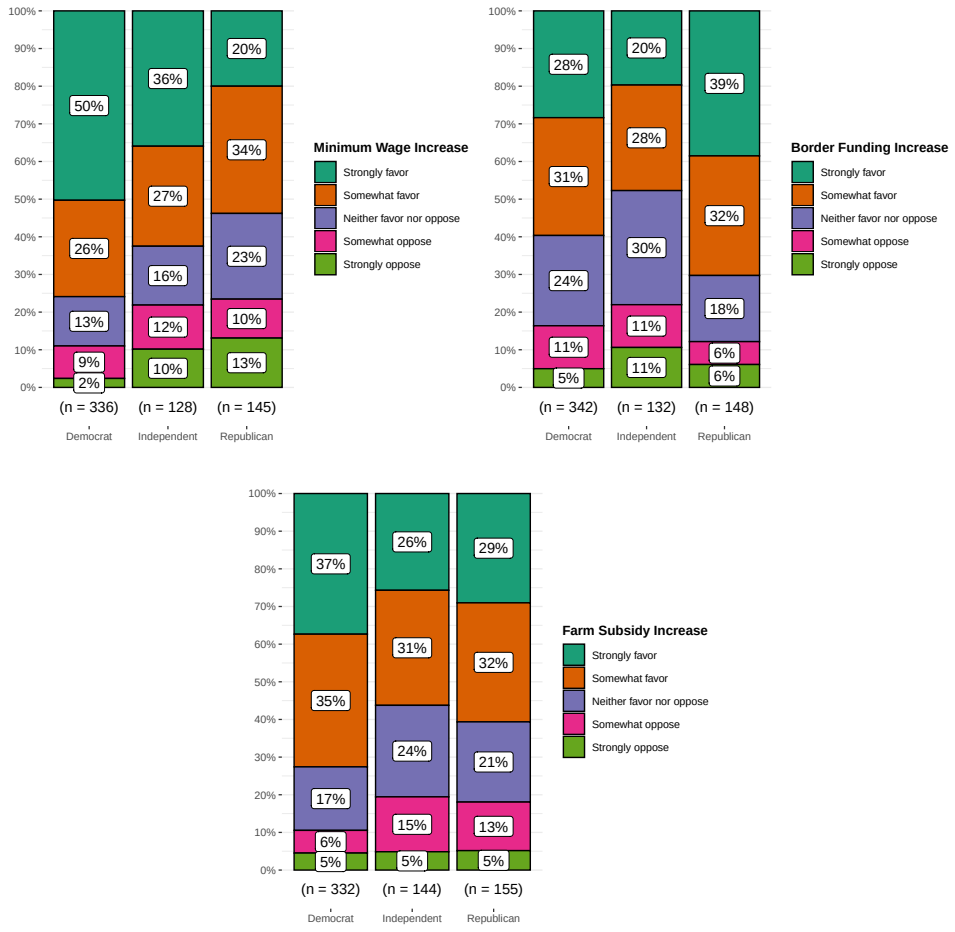
Supplemental Material for Affect, Not Ideology

The Psychological Mechanisms of Partisan Information Processing

Experimental Details

Baseline Policy Support Here we check for baseline policy support across all PIDs and policies. We find significant variation within every PID and across all policies.

FIGURE A.1. Baseline Policy Support by PID



Randomization Tests Here we report an assignment table as well as a standard balance table. Furthermore we also present the results from the "balance permutation test" introduced and detailed in Rametta and Fuller (2024).

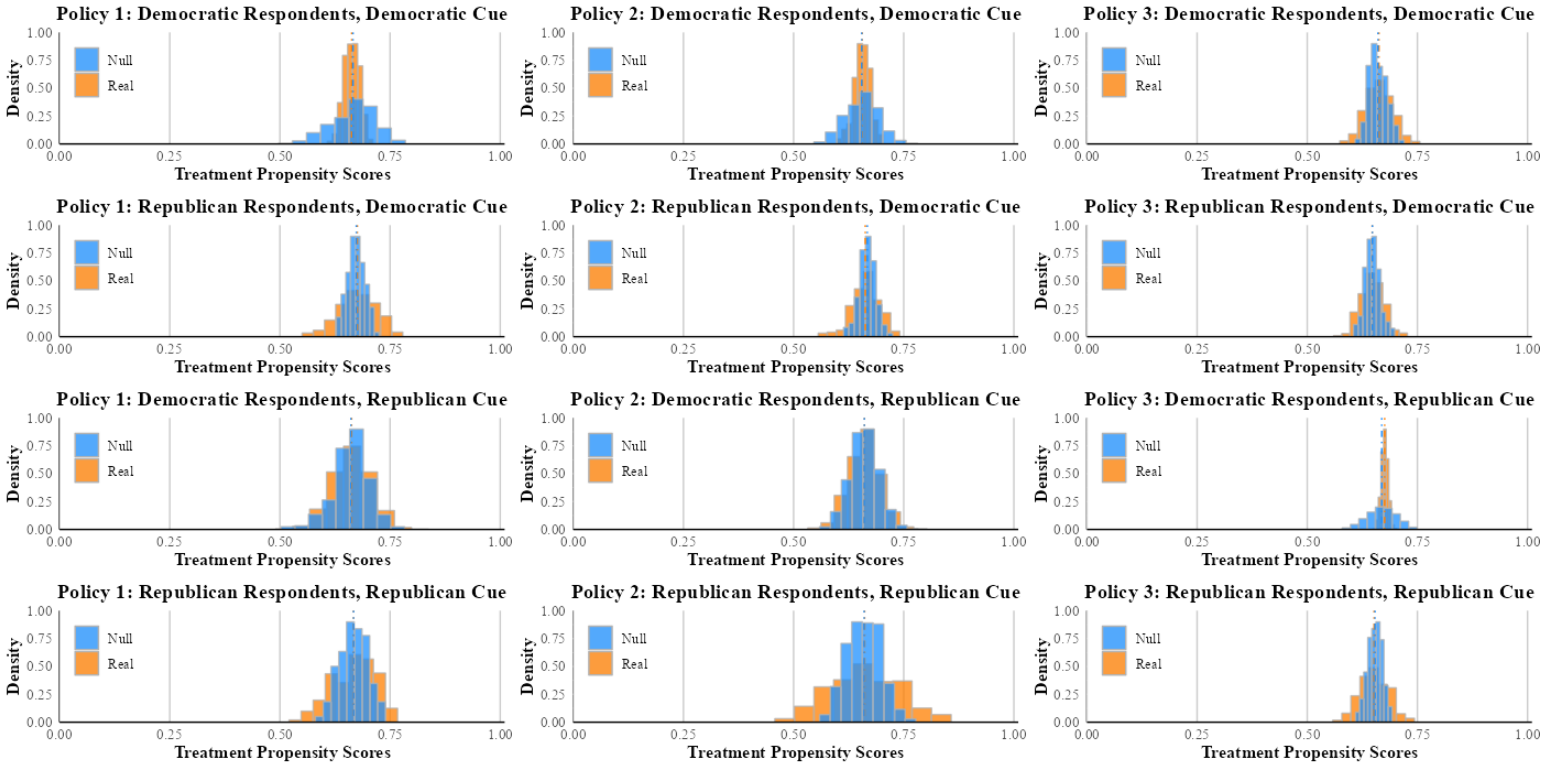
TABLE A.1. Assignment Table for Treatment Conditions

		T1		T2		T3	
PID	Group	Count	%	Count	%	Count	%
Democrat	Nonpartisan Group	336	0.203	342	0.206	332	0.200
Democrat	Democratic Masses	326	0.197	323	0.195	332	0.200
Democrat	Republican Masses	331	0.200	330	0.199	341	0.206
Democrat	Democratic Elites	340	0.205	326	0.197	317	0.191
Democrat	Republican Elites	324	0.196	336	0.203	335	0.202
Republican	Nonpartisan Group	145	0.198	148	0.202	155	0.211
Republican	Democratic Masses	145	0.198	153	0.208	136	0.185
Republican	Republican Masses	151	0.206	145	0.198	143	0.195
Republican	Democratic Elites	157	0.214	141	0.192	152	0.207
Republican	Republican Elites	136	0.185	147	0.200	148	0.202
Independent	Nonpartisan Group	128	0.181	132	0.187	144	0.204
Independent	Democratic Masses	145	0.205	130	0.184	146	0.207
Independent	Republican Masses	152	0.215	153	0.216	134	0.190
Independent	Democratic Elites	145	0.205	156	0.221	139	0.197
Independent	Republican Elites	137	0.194	136	0.192	144	0.204

TABLE A.2. Balance Table Across Pretreatment Covariates

	C (N=609)		D (N=1258)		R (N=1231)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Ideology (BAM)	-0.3	1.1	-0.2	1.3	-0.2	1.2
Trust (Scale)	-0.2	0.6	-0.1	0.6	-0.1	0.6
Social Distiance (IRT)	0.0	1.4	0.0	1.4	0.0	1.4
Affect (Therm. Scale)	-0.2	0.4	-0.1	0.5	-0.2	0.5
Spatial Knowledge	0.5	0.6	0.5	0.7	0.5	0.7
Age	3.7	1.8	3.8	1.9	3.8	1.8
Female	0.5	0.5	0.5	0.5	0.5	0.5
White	0.7	0.5	0.6	0.5	0.6	0.5
College Degree	0.5	0.5	0.5	0.5	0.5	0.5
Political Interest	3.7	1.7	3.6	1.7	3.6	1.6
Religious Importance	3.2	1.0	3.1	1.0	3.2	1.0
Born Again	0.5	0.5	0.5	0.5	0.5	0.5

FIGURE A.2. ML Balance Test Results



Testing Mass Versus Elite Cue Differences First, we run a series of tests to determine whether individuals who were given a mass or elite-level cue, but from the same party, had substantively different outcomes. In other words, we are interested in whether these mass- versus elite-level cues had differential effects in increasing or decreasing support for any of our three policies.

To do this, we run three separate tests. First, we do a simple difference-in-means test between the average treatment effect (ATE) for the mass and elite cues, by cue PID (e.g., Democratic Senators versus Democratic grassroots group) and respondent PID.[1] We subset by respondent PID given our different expectations of effect direction. In other words, we expect positive cueing effects when the cue and respondent PID match and negative effects when they differ. Consequently, if we pooled across respondent PID, we would likely recover null effects for all ATEs. Second, we run Mann-Whitney U tests following the same format.

The results across these three tests are consistently null, indicating that, at least in the context of this cuing experiment, respondents are not substantively differentiating between mass- and elite-level cues, regardless of cue and respondent PID. The details of all these tests are included below. Consequently, we opt to pool our mass- and elite-level cues by the cue's PID, i.e., we treat the mass- and elite-level cues as the same for Democratic and Republican cues, respectively. This increases our experiment's power and simplifies our analysis.

Preregistration and Survey Details Given the presence of identifying information we are not able to include the full preregistration plan in this appendix. There is a preregistration of this experiment which can be provided later, however, and more information about this plan can be included in subsequent versions, if necessary.

[1] In this context we calculate the ATEs by subtracting the average policy support from a given treatment-policy combination, by cue and respondent PID, from the control condition of the same policy, by respondent PID.

Scaling Details

Alongside self-reported ideology, respondents were also asked to ideologically place other stimuli on the same 7-point scale. Specifically, we asked respondents to place 6 stimuli:

- The Democratic (Republican) Party
- Democratic (Republican) Party Supporters
- Joseph R. Biden
- Donald J. Trump

Given this data, along with their self-placements, we leverage Bayesian Aldrich-McKelvey (BAM) scaling (Hare et al. 2015) to recover more robust ideological estimates that account for differential-item functioning (individuals having different perceptions of what a given scale means). These result in ideology estimates of every respondent bounded from -11–11. Furthermore, following Palfrey and Poole (1987) we calculate the correlation between the recovered estimates of the ideological positions of our stimuli and each individual's placements to generate a spatial knowledge/information index from -1 to 1. Simply put, the higher the correlation the more informed the respondent is about the ideological locations of our stimuli.

For our feeling thermometers, we also asked individuals to rate 15 stimuli:

- Democratic (Republican) Party Voters
- Democratic (Republican) Party Candidates and Elected Officials
- The Democratic (Republican) Party
- Joseph R. Biden
- Nancy Pelosi
- Bernie Sanders
- Kamala Harris
- Joe Manchin
- Donald J. Trump
- Mike Pence
- Ron DeSantis
- Susan Collins

Using this set of thermometer ratings, for each respondent, we leverage ratio unfolding for recovering a measure of thermometer-level affect (Mair, Groenen, and Leeuw 2022). Using this method we recover a one-dimensional representation of affect based on these measures, recovering both locations of individuals and stimuli. These respondent locations are then used as our thermometer measure of affect in our subsequent analyses. We also calculate two additional scales of thermometer-level affect on just Democratic and Republican stimuli, respectively. We consider these scales to be in- and out-party measures of affect for Democrats and Republicans. Importantly, these scales should give us, generally, more robust measures of affect, using more granular data than simply taking the difference between an individual's in-party and out-party thermometers.

To measure partisan trust, we follow Druckman and Levendusky (2019) and ask respondents "How much of the time do you think you can trust these groups to do what is right for the country?" for six stimuli, the Democratic (Republican) Party, Democratic (Republican) Party candidates and elected officials, and Democratic (Republican) Party voters on a five point scale: "Almost never, once in a while, about half of the time, most of the time, and almost always." We then use the same scaling method used above with our thermometers, ratio unfolding, on all six stimuli to recover a single dimensional scale of partisan trust. Furthermore, similar to above, we also generate two additional scales on just Democratic and Republican stimuli, respectively, and consider them to be in- and out-party measures of affect for both Democratic and Republican respondents.

Finally, to measure the level of social distance individuals feel towards not only their out-party, but their in-party as well, we again adapt questions from Druckman and Levendusky (2019), and ask the question "How comfortable would you be having a Democratic (Republican) Party supporter as a" for 5 stimuli; "Fellow citizen, neighbor, close personal friend, child in-law, and romantic partner;" using a 4-point scale: "Not at all comfortable, not too comfortable, somewhat comfortable, and extremely comfortable." This gives us five responses per-party for each respondent, ranging from most distant, fellow citizen, to closest, romantic partner.

Given that we can conceive of each of these stimuli as being more socially closer/distant, we can liken these sets of responses to a difficulty parameter, which directly motivates us to utilize a graded response model (GRM), a type of item response theory (IRT) model that can handle ordinal data. Using this model on each set of party-level questions, we recover two single-dimensional scales of social distance feeling toward Democratic and Republican Party supporters, respectively. We

then take the difference between these two measures (specifically, Republican social distance minus Democratic social distance) to recover a single dimension of differential social distance affect, with those who feel socially close to Democrats and not to Republicans being located on the far left of the scale, and those who feel socially close to Republicans, but not Democrats, being located to the far right of the scale.

Causal Forests Details

Honesty Honest estimation, or sub-sample splitting, is a data splitting technique wherein some fraction of the training data is used to estimate the model (in the case of random forest, estimate the splits in the decision trees) and some held-out fraction of the training data is used to estimate the effects (in the case of random forest, filling in the leaves on the decision trees). Any leaves in a decision tree that are left empty are also pruned in the estimation procedure. The idea of honest splitting is an important debiasing technique popularized by Athey and Imbens (2016) and Wager and Athey (2018).

Out-of-bag Predictions To generate "out-of-bag" predictions, the predictions for each observation are generated using only the trees in the forest that did not use that observation as a training unit. In so doing, we reduce bias in the model predictions stemming from overfitting without the need for computationally expensive cross-validation (Breiman 1996).

Tuning Procedure We employ the cross-validation (CV) procedure developed in the `grf` R package (Tibshirani et al. 2023). In brief, the CV procedure first randomly draws 100 sets of parameter values and for each set, a forest is trained and out-of-bag error estimates are estimated. Specifically, "mini forests" are trained using a small number of trees, for computational speed. However, given these small forests a debiasing process is used through a variance decomposition. Finally, a smoothing function is used to determine optimal parameter values. Also note that CFs use an error measure developed by Nie and Wager (2021). To verify the CV procedure is producing sensible results, we employ the omnibus heterogeneity test developed by (Chernozhukov et al. 2023), adjusting tuning parameter values as necessary to improve performance on this test. For more information see the `grf` package materials.

Variable Importance For the purpose of this paper, we employ the measure of variable importance introduced by Wager and Athey (2018). To construct this measure, first a matrix is generated that, for each node from 1 to the maximum tree-depth, sums the amount of times each variable appears as a split from any tree in that node. This results in a matrix with rows equal to the number of variables and columns equal to the maximum tree depth and each cell being the number of times that given variable appears as a split in that node depth for all trees. Importantly, we limit this matrix to 4 columns/nodes (the default for this function), excluding subsequent splits. Second,

each row is divided by its sum, to calculate the node incidence proportions for every variable (e.g., variable 1 appears 40% of the time in node 1, 20% in node 2, 30% in node 3, and 10% in node 4 resulting in a vector $\vec{V_1} = [.40, .20, .30, .10]$).

The weights for the matrix (vector $\vec{W}$) are determined by a tuning parameter called "decay" which, in our function, is left on the default value of 2 and leads to an exponential discounting of a variable being used for a split at each depth. So variables that split first are fully weighted, those that are split second are weighted by $\frac{1}{4}$, third $\frac{1}{9}$, fourth $\frac{1}{16}$ (i.e., $\vec{W} = [1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}]$). The weights are then divided by the sum of weights (in this instance, divided by $1\frac{61}{144}$) resulting in relative weights being applied to each node column. Then the sum is finally calculated for each variable row vector, leading to a *relative* variable importance score for each variable. There are, of course, alternative measures of variable importance, and these metrics should *not* be interpreted causally in the context of a propensity score model. However, for the application of nonrandom assignment detection and simple identification of predictors that contribute contaminating assignment, this simple variable importance metric is suitable.

Treatment Effect Estimates

Unconditional Treatment Effect Estimates Here we report standard treatment effect estimates within PID, pooled across policies and cues.

TABLE A.3. Unconditional Treatment Effects: Pooled Democratic Cues

	Strong D.	Weak D.	Indep.	Weak R.	Strong R.
Intercept	3.96***	3.71***	3.49***	3.64***	3.63***
	[3.89, 4.03]	[3.59, 3.84]	[3.34, 3.64]	[3.48, 3.80]	[3.47, 3.78]
Dem. Cue	0.20***	0.18*	−0.02	−0.20*	−0.31**
	[0.11, 0.29]	[0.03, 0.34]	[−0.20, 0.16]	[−0.40, −0.01]	[−0.50, −0.12]
Num.Obs.	2351	869	830	633	888
R2	0.008	0.006	0.000	0.006	0.011

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

TABLE A.4. Unconditional Treatment Effects: Pooled Republican Cues

	Strong D.	Weak D.	Indep.	Weak R.	Strong R.
Intercept	3.96***	3.71***	3.49***	3.64***	3.63***
	[3.87, 4.05]	[3.58, 3.85]	[3.35, 3.64]	[3.48, 3.79]	[3.48, 3.77]
Rep. Cue	−0.37***	−0.21*	−0.06	0.07	0.24**
	[−0.48, −0.26]	[−0.37, −0.05]	[−0.23, 0.12]	[−0.11, 0.26]	[0.06, 0.41]
Num.Obs.	2378	864	842	656	845
R2	0.018	0.007	0.000	0.001	0.008

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Doubly Robust Treatment Effect Estimates The popular "DoubleML" approach works by employing arbitrary ML algorithms to estimate both a propensity score model and separately, a model of the outcome. Taken together, the composite estimator outputs consistent average treatment effect estimates so long as one of the two models is correctly specified (Chernozhukov, Newey, and Singh 2022). DoubleML models incorporate treatment propensity information across a large number of covariates, reducing the potential for researchers to make arbitrary model specification choices after the fact. We present these estimates as a robustness check on our unconditional treatment effect estimates, similar to the presentation of regression adjusted treatment effects presented in the subsequent experimental literature.

TABLE A.5. Doubly Robust Treatment Effects: Pooled Democratic Cues

	Strong D.	Weak D.	Indep.	Weak R.	Strong R.
Dem. Cue	0.19*	0.16	−0.04	−0.20*	−0.31*
	[0.10, 0.28]	[−0.002, 0.32]	[−0.21, 0.13]	[−0.39, −0.01]	[−0.50, −0.12]
Num.Obs.	2351	869	830	633	888

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

TABLE A.6. Doubly Robust Treatment Effects: Pooled Republican Cues

	Strong D.	Weak D.	Indep.	Weak R.	Strong R.
Rep. Cue	−0.38*	−0.21*	−0.07	0.12	0.22*
	[−0.48, −0.28]	[−0.37, −0.05]	[−0.25, 0.10]	[−0.07, 0.31]	[0.05, 0.39]
Num.Obs.	2378	864	842	656	845
R2	0.018	0.007	0.000	0.001	0.008

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Bibliography

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Breiman, Leo. 1996. "Out-of-bag estimation," https://www.stat.berkeley.edu/pub/users/breiman/OOBestimation.pdf.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2023. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.* arXiv: 1712.04802 [stat.ML].

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2022. "Automatic debiased machine learning of causal and structural effects." *Econometrica* 90 (3): 967–1027.

Druckman, James N, and Matthew S Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83 (1): 114–122.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll, and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59 (3): 759–774.

Mair, Patrick, Patrick JF Groenen, and Jan de Leeuw. 2022. "More on Multidimensional Scaling and Unfolding in R: smacof Version 2." *Journal of Statistical Software* 102:1–47.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108 (2): 299–319.

Palfrey, Thomas R., and Keith T. Poole. 1987. "The Relationship between Information, Ideology, and Voting Behavior." *American Journal of Political Science* 31 (3): 511–530.

Rametta, Jack, and Sam Fuller. 2024. *MLbalance.* V. 0.1, January. https://github.com/samjfuller/MLbalance.

Tibshirani, Julie, Susan Athey, Erik Sverdrup, and Stefan Wager. 2023. *Algorithm Reference.* https://grf-labs.github.io/grf/REFERENCE.html#parameter-tuning.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

APPENDIX B

Supplemental Material for The Balance Permutation Test: A Machine Learning Replacement for Balance Tables

Balance permutation test illustration

To generate the example in the main text that results in Table 1 we simulate a simple Bernoulli random experiment where $N = 1,000$ and an uncontaminated treatment vector W_i. We then contaminate the treatment assignment vector with the following modifications resulting in W_c:

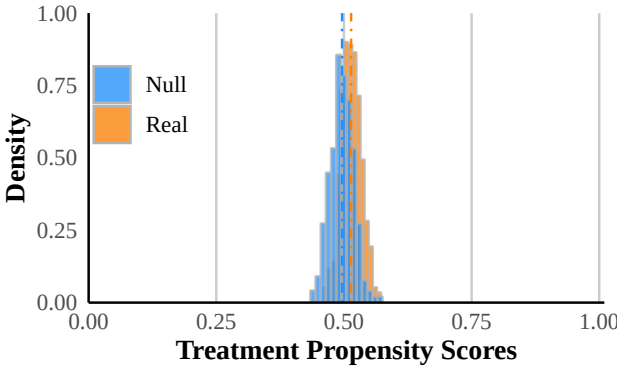
$$\mathrm{W}_c = \begin{cases} 0 & \text{if female}_i = 1 \text{ and black}_i = 1, \\ 1 & \text{if female}_i = 0 \text{ and black}_i = 1, \\ \mathrm{W}_i & \text{if female}_i = 0 \text{ and black}_i = 0, \\ \mathrm{W}_i & \text{if female}_i = 1 \text{ and black}_i = 0 \text{ and income}_i > 50000, \\ \text{Bernoulli}(0.75) & \text{if female}_i = 1 \text{ and black}_i = 0 \text{ and income}_i \leq 50000. \end{cases}$$

The code used to generate this data is also available in our replication materials.

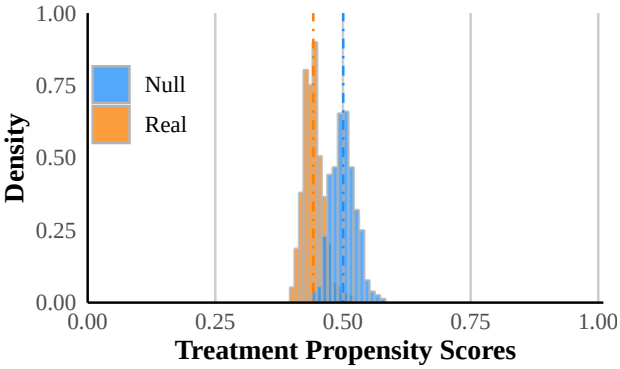
Additional balance permutation test application

Moehler & Conroy-Krutz (Moehler and Conroy-Krutz 2016) In this subsection, we demonstrate the balance permutation test on Moehler and Conroy-Krutz (2016), a real example of randomization failure as noted by Mutz, Pemantle, and Pham (2019). To execute the test on this multi-arm data, we subset the observations to isolate each treatment arm and the control group and simply run the test three times over. Importantly, we simulate the assignment process for this example as the randomization failure here occurs simply in terms of numerical count (i.e., the intention was full random assignment without replacement). For predictors we include a large set of measured pretreatment covariates. Our test clearly detects the failure in arms 3 and 4, at least relative to the ex ante expectations as described in the original article. Of course in this case, a

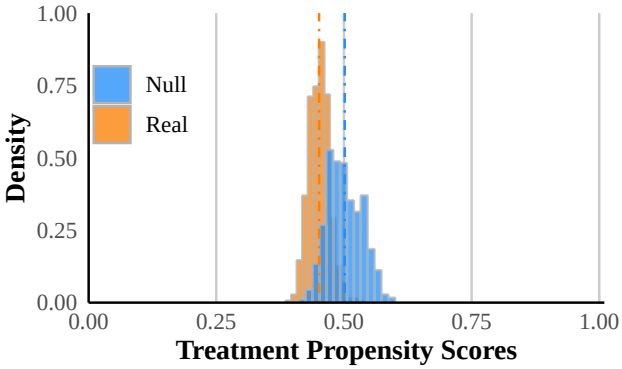
simple count table would suffice to see the numerical imbalance; count tables are also included in the output of the balance permutation test. However, further examination of variable importance metrics for treatment arms 3 and 4 indicates age appears to have a slight relationship with treatment assignment. The results of the test are summarized below.



Note: Dotted lines represent mean values for the null and



Note: Dotted lines represent mean values for the null and



Note: Dotted lines represent mean values for the null and real treatment propensity distributions.

Details of honest n-step boosted random forests

Honesty Honest estimation, or sub-sample splitting, is a data splitting technique wherein some fraction of the training data is used to estimate the model (in the case of random forest, estimate the splits in the decision trees) and some held-out fraction of the training data is used to estimate the effects (in the case of random forest, filling in the leaves on the decision trees). Any leaves in a decision tree that are left empty are also pruned in the estimation procedure. The idea of honest splitting is an important debiasing technique popularized by Athey and Imbens (2016) and Wager and Athey (2018).

n-step boosted random forests We employ an n-step Boosted Random (Regression) Forest (Ghosal and Hooker 2020) to estimate treatment propensities. The procedure for estimating the boosted model is as follows: First, estimate a single random forest model. Second, estimate a new, small random forest model targeting the out-of-bag error from the first forest. If the new forest can improve on the out-of-bag error from the first forest, a new full random forest model is estimated. This procedure iterates until a new model can no longer explain a large fraction of the error from the previous forest. Thus, the final boosted random forest is an ensemble of n individual forests, or n-steps.

Variable importance For the purpose of this paper, we employ the measure of variable importance introduced by Wager and Athey (2018). To construct this measure, first a matrix is generated that, for each node from 1 to the maximum tree-depth, sums the amount of times each variable appears as a split from any tree in that node. This results in a matrix with rows equal to the number of variables and columns equal to the maximum tree depth and each cell being the number of times that given variable appears as a split in that node depth for all trees. Importantly, we limit this matrix to 4 columns/nodes (the default for this function), excluding subsequent splits. Second, each row is divided by its sum, to calculate the node incidence proportions for every variable (e.g., variable 1 appears 40% of the time in node 1, 20% in node 2, 30% in node 3, and 10% in node 4 resulting in a vector $\vec{V_1} = [.40, .20, .30, .10]$).

The weights for the matrix (vector $\vec{W}$) are determined by a tuning parameter called "decay" which, in our function, is left on the default value of 2 and leads to an exponential discounting of a variable being used for a split at each depth. So variables that split first are fully weighted, those

that are split second are weighted by $\frac{1}{4}$, third $\frac{1}{9}$, fourth $\frac{1}{16}$ (i.e., $\vec{W} = [1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}]$). The weights are then divided by the sum of weights (in this instance, divided by $1\frac{61}{144}$) resulting in relative weights being applied to each node column. Then the sum is finally calculated for each variable row vector, leading to a *relative* variable importance score for each variable. There are, of course, alternative measures of variable importance, and these metrics should *not* be interpreted causally in the context of a propensity score model. However, for the application of nonrandom assignment detection and simple identification of predictors that contribute contaminating assignment, this simple variable importance metric is suitable.

Out-of-bag predictions To generate "out-of-bag" predictions, the predictions for each observation are generated using only the trees in the forest that did not use that observation as a training unit. In so doing, we reduce bias in the model predictions stemming from overfitting.

Tuning procedure We employ the cross-validation (CV) procedure developed in the `grf` R package (Tibshirani et al. 2023). In brief, the CV procedure first randomly draws 100 sets of parameter values and for each set, a forest is trained and out-of-bag error estimates are estimated. Specifically, "mini forests" are trained using a small number of trees, for computational speed. However, given these small forests a debiasing process is used through a variance decomposition. Finally, a smoothing function is used to determine optimal parameter values. Also note that CFs use an error measure developed by Nie and Wager (2021). For more information see the `grf` package materials.

Full simulation results

This section reports the complete results of all simulations. Specifically, each plot (and subplot) compares either the reduction in absolute or standard error from using DRML estimation over each alternative method: unadjusted difference in means, simple regression adjusted, and Lin regression adjustment. We also include the average across all three methods as the final plot in the section.

Coverage Results The coverage rates for each method are presented in Table B.1 below. Results for the Lin estimator are presented in two different ways: one where NAs (i.e., where the estimator failed) are included in the calculation as "uncovered," and the other simply drops the NAs from the calculation. In general, all four methods achieve or closely approach 95% coverage. Unfortunately, the DRML coverage dips slightly below the desired 95% rate. This slight undercoverage could have many causes. For instance, in the simulation study we use the default tuning parameters for the DRML models, and further, we force the estimator to model the propensity score rather than taking the known assignment probability as given (i.e., setting it to .5). Regardless, given the potential for slight undercoverage, we recommend researchers include both the DRML results and the simple unadjusted results. Another remedy for undercoverage would be to implement conformal prediction intervals, which guarantee coverage at a pre-specified rate. This approach could be employed for any of the estimators mentioned, one useful software implementation is `cfcausal` (Lei and Candès 2021). A primer on conformal prediction can be found here (Angelopoulos and Bates 2021).

DRML	Unadjusted DIM	Regression-Adjust	Lin (NAs=0)	Lin (Drop NAs)
93%	95%	95%	86%	95%

TABLE B.1. Coverage rates averaged over all simulations

Simulations where DRML is outperformed The following list indicates the 19 simulation comparisons where DRML is beaten, either in terms of absolute error or standard error length, relative to one of the other methods.

- Sample Size: 100, Covariates: 20, Error Variance: 5, Linear: True, Interactions: False, Heterogeneity: False
- Sample Size: 100, Covariates: 20, Error Variance: 5, Linear: True, Interactions: False, Heterogeneity: True

- Sample Size: 100, Covariates: 50, Error Variance: 1, Linear: False, Interactions: True, Heterogeneity: False
- Sample Size: 100, Covariates: 50, Error Variance: 1, Linear: True, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 100, Error Variance: 1, Linear: False, Interactions: False, Heterogeneity: False
- Sample Size: 100, Covariates: 100, Error Variance: 2, Linear: False, Interactions: True, Heterogeneity: False
- Sample Size: 100, Covariates: 100, Error Variance: 2, Linear: False, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 100, Error Variance: 2, Linear: True, Interactions: True, Heterogeneity: False
- Sample Size: 100, Covariates: 100, Error Variance: 2, Linear: True, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 100, Error Variance: 5, Linear: False, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 20, Error Variance: 1, Linear: False, Interactions: False, Heterogeneity: False
- Sample Size: 100, Covariates: 20, Error Variance: 2, Linear: False, Interactions: False, Heterogeneity: False
- Sample Size: 100, Covariates: 20, Error Variance: 5, Linear: False, Interactions: False, Heterogeneity: False
- Sample Size: 100, Covariates: 20, Error Variance: 5, Linear: False, Interactions: False, Heterogeneity: True
- Sample Size: 100, Covariates: 50, Error Variance: 2, Linear: True, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 10, Error Variance: 2, Linear: True, Interactions: True, Heterogeneity: True
- Sample Size: 100, Covariates: 10, Error Variance: 5, Linear: True, Interactions: True, Heterogeneity: False



FIGURE B.1. Unadjusted difference-in-means comparison of reduction in absolute error

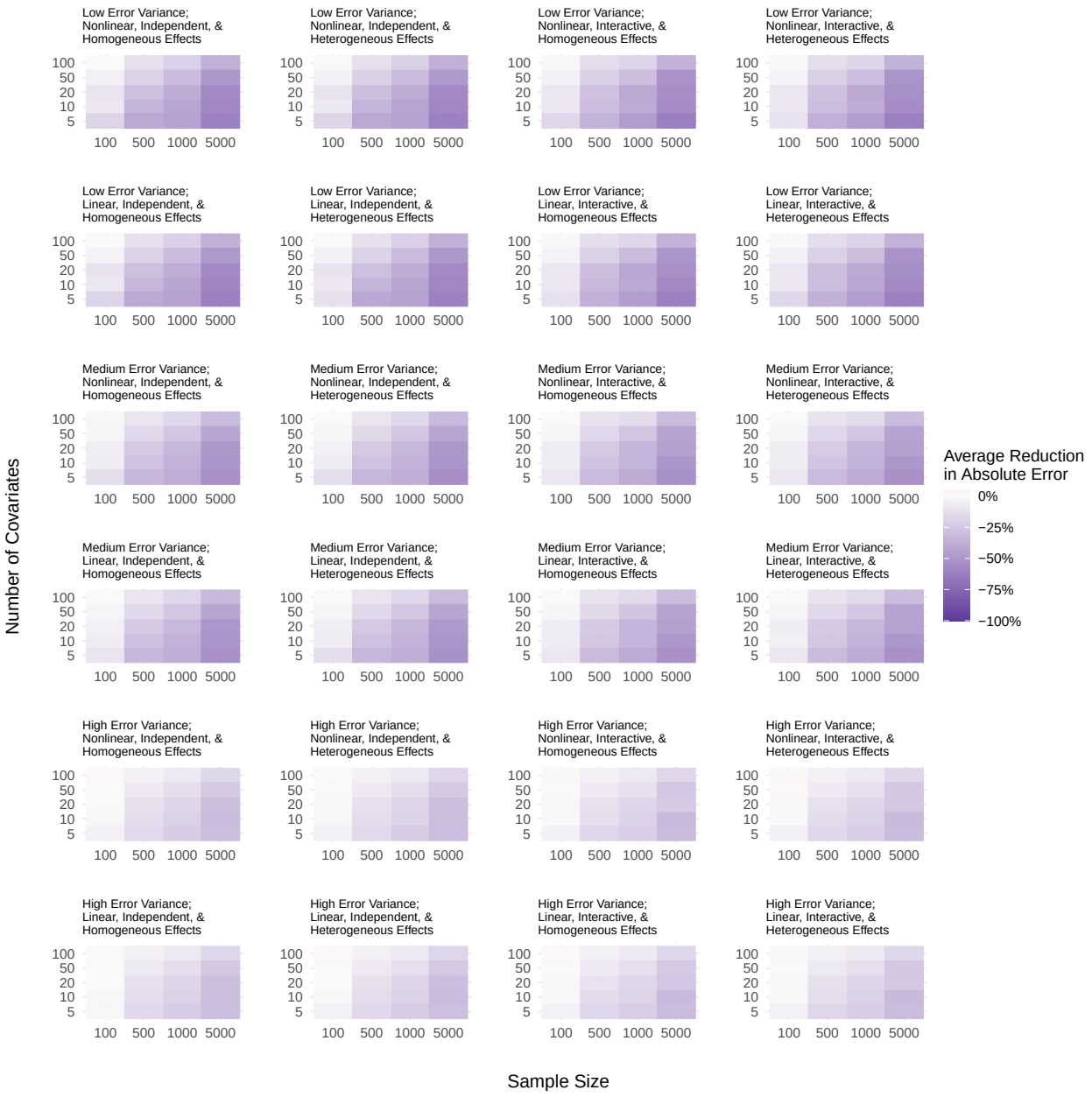


FIGURE B.2. Unadjusted difference-in-means comparison of reduction in standard error



FIGURE B.3. Simple regression adjustment comparison of reduction in absolute error

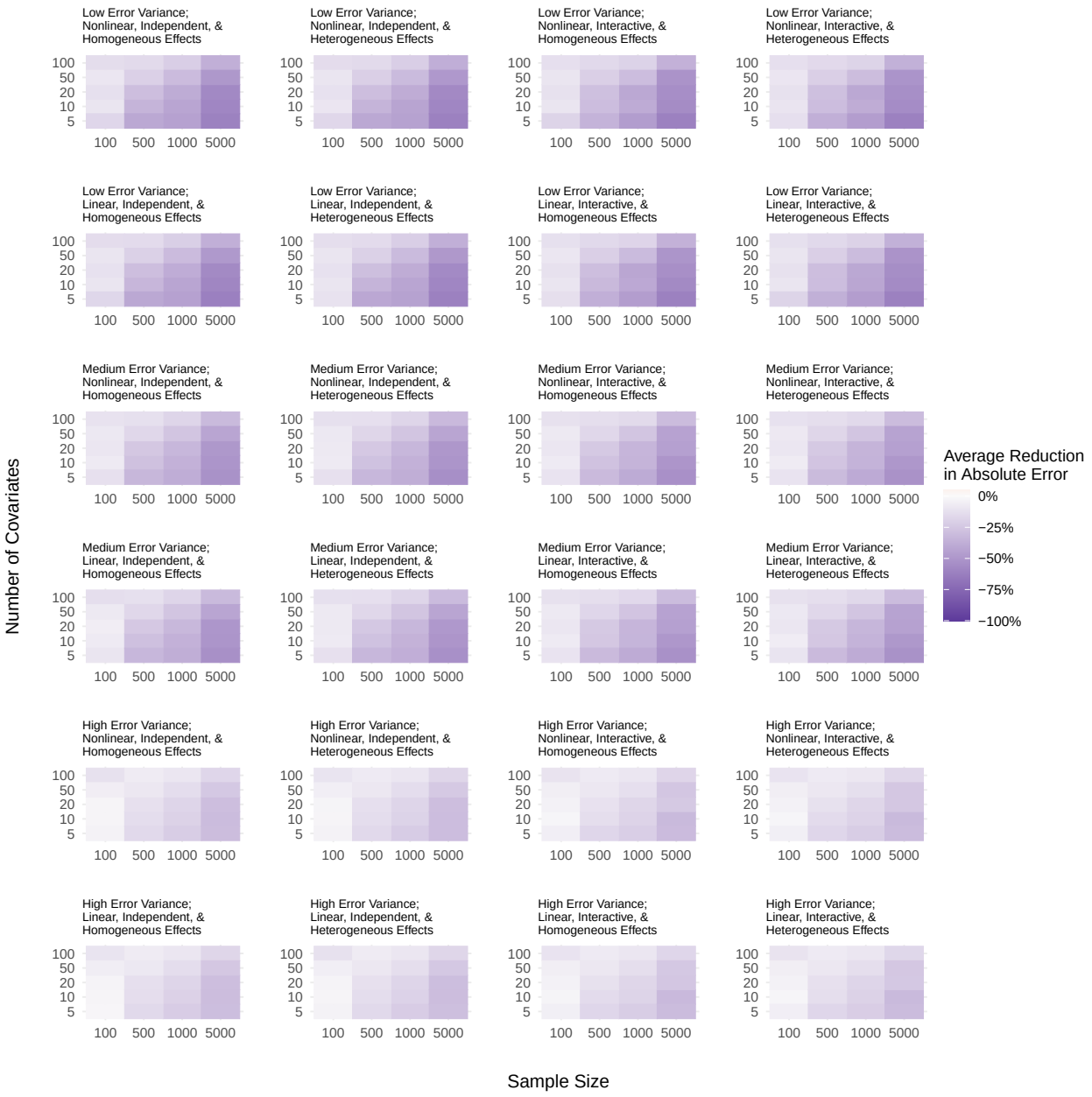


FIGURE B.4. Simple regression adjustment comparison of reduction in standard error



FIGURE B.5. Lin regression adjustment comparison of reduction in absolute error
Note: Gray boxes correspond to NA values. In these instances, the Lin estimator failed to produce estimates as a result of exhaustion of degrees of freedom. This occurs in parameter combinations where the ratio of covariates to sample size nears or exceeds 0.5, as the Lin estimator includes all two-way interactions between treatment assignment and pre-treatment covariates.

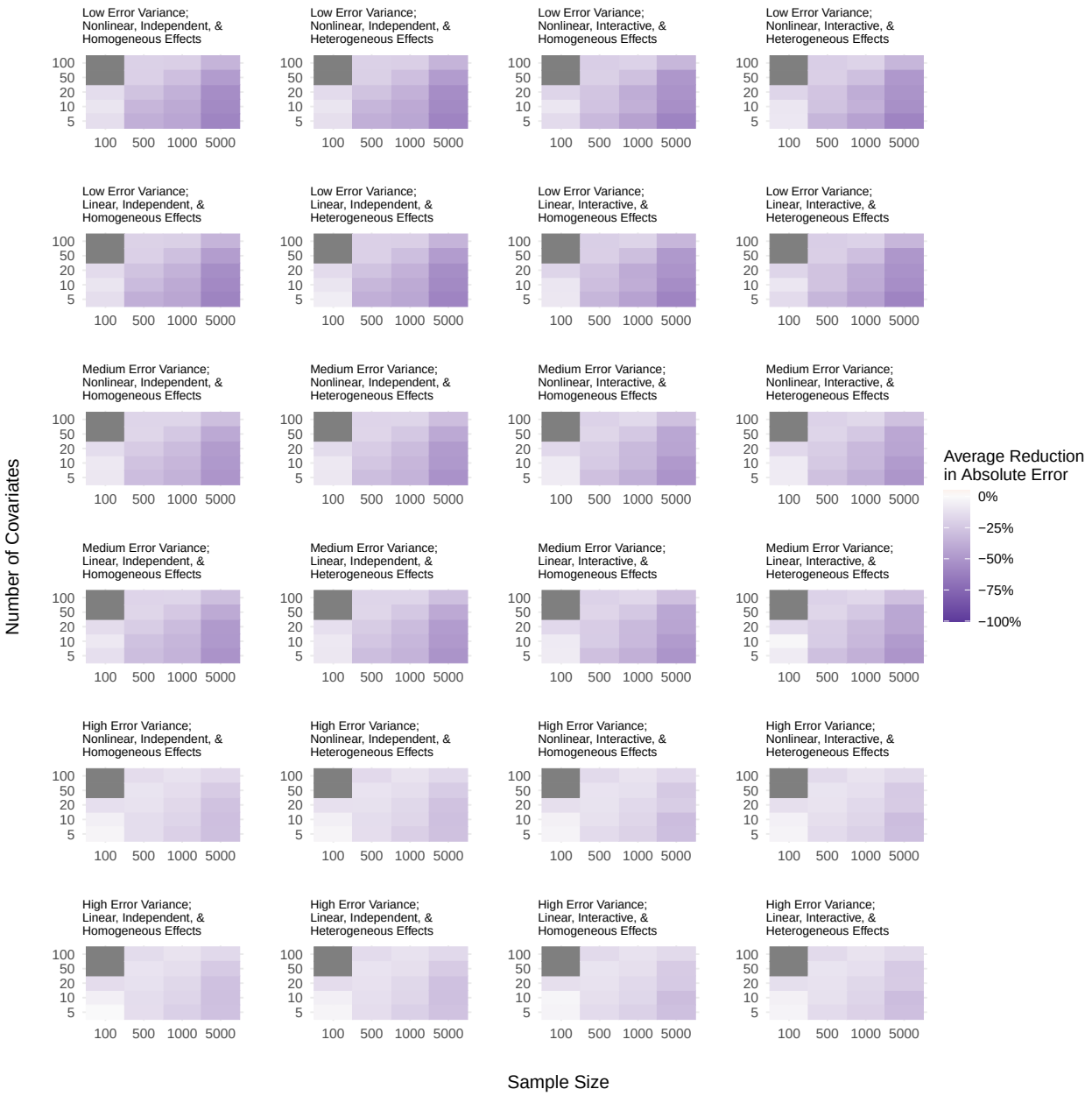


FIGURE B.6. Lin regression adjustment comparison of reduction in standard Error
Note: Gray boxes correspond to NA values. In these instances, the Lin estimator failed to produce estimates as a result of exhaustion of degrees of freedom. This occurs in parameter combinations where the ratio of covariates to sample size nears or exceeds 0.5, as the Lin estimator includes all two-way interactions between treatment assignment and pre-treatment covariates.



FIGURE B.7. Average comparison of reduction in absolute error

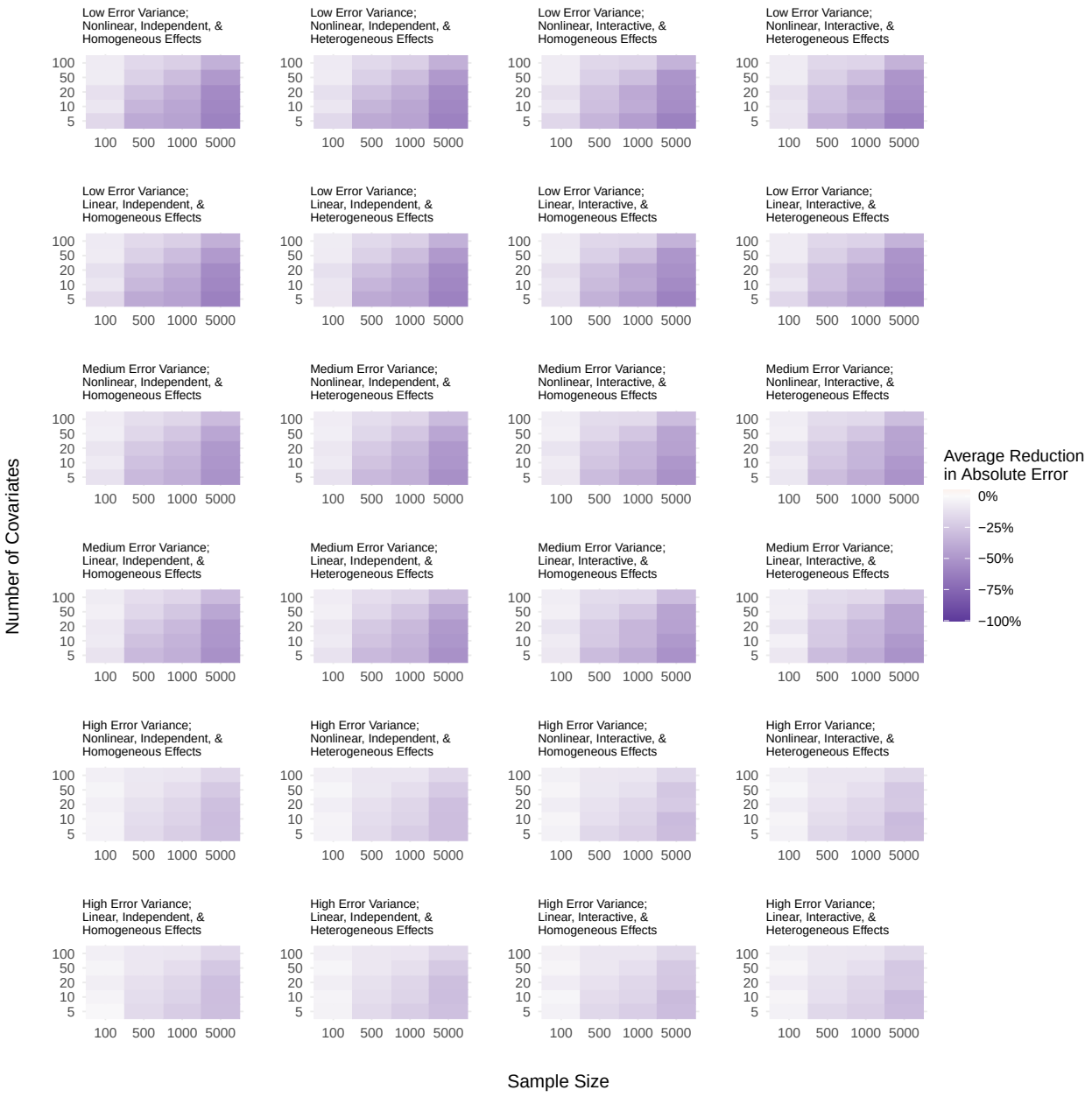


FIGURE B.8. Average comparison of reduction in standard error

Bibliography

Angelopoulos, Anastasios N, and Stephen Bates. 2021. *A gentle introduction to conformal prediction and distribution-free uncertainty quantification.* https://arxiv.org/abs/2107.07511.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Ghosal, Indrayudh, and Giles Hooker. 2020. "Boosting random forests to reduce bias; one-step boosted forest and its variance estimate." *Journal of Computational and Graphical Statistics* 30 (2): 493–502.

Lei, Lihua, and Emmanuel J Candès. 2021. "Conformal Inference of Counterfactuals and Individual Treatment Effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (5): 911–938.

Moehler, Devra C, and Jeffrey Conroy-Krutz. 2016. "Partisan media and engagement: A field experiment in a newly liberalized system." *Political Communication* 33 (3): 414–432.

Mutz, Diana C, Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1): 32–42.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108 (2): 299–319.

Tibshirani, Julie, Susan Athey, Erik Sverdrup, and Stefan Wager. 2023. *Algorithm Reference.* https://grf-labs.github.io/grf/REFERENCE.html#parameter-tuning.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.