Observational Studies () Submitted ; Published

A Falsification Framework for Investigating Observed Covariate Balance

Jeffrey J. Harden **jharden2@nd.edu**
Department of Political Science
University of Notre Dame
Notre Dame, IN 46556, USA

Abstract

Observational studies in the social and behavioral sciences often rely on unique and inventive treatment assignment mechanisms to assert causal claims. These research designs require careful empirical scrutiny of the assumption that treatment is (conditionally) ignorable. One common partial strategy for quantitatively interrogating this assumption is the examination of balance, or conditional distributional equality of observed covariates. However, conventional practice often employs a suboptimal "folk definition" of balance that prioritizes the statistical significance of covariate mean differences. Rigorously probing the observable implications of identification necessitates more and better scrutiny of the conditional covariate distributions. I propose a falsification framework for evaluating observed covariate balance that aligns with the concept's formal definition. The framework combines existing tools in a novel manner to support thorough, informative, and intuitive balance investigations in observed covariates. First, I illustrate the value of assessing balance in multiple moments of observed covariate distributions. Second, I recommend estimating balance statistic uncertainty using classical or Bayesian bootstrapping. Quantifying sampling variability permits analysts to qualify balance diagnostics with compelling inferential statements. Finally, I reanalyze a natural experiment for which treatment exogeneity merits a priori skepticism: as-if random legal delays in the implementation of voter identification laws in the American states.

Keywords: Observational studies, Covariate balance, Bootstrapping, Equivalence testing, Causal inference

1. Introduction

An enduring empirical paradigm in the social and behavioral sciences leverages clever manipulations of social contexts to identify the causal effects of phenomena that are otherwise difficult or impossible for researchers to control. Examples include haphazard allocation of land titles to unhoused persons (Galiani and Schargrodsky, 2010), the arbitrary locations of artillery strikes (Lyall, 2009), or the happenstance completion of a survey on a day of prayer rather than an ordinary day (Brooke et al., 2023). This approach to scientific inquiry has greatly advanced the discovery of knowledge and reflects the substantial value of social scientists' domain expertise. However, the seemingly far-fetched nature of the causal identification strategies described in these studies requires strong, multifaceted evidence consistent with researchers' claims of (conditional) ignorability of treatment. Such claims are not directly verifiable, but their tenability is predicated on both theoretical plau-

sibility and observable implications. Researchers must report a triangulation of qualitative arguments as well as quantitative falsification tests that subject the observed data to rigorous scrutiny. In this paper, I propose an improvement in the conduct of one particular practice for evaluating these observable implications that is commonly employed, but often not implemented rigorously. Specifically, I describe a novel falsification framework for extensively *investigating* instead of simply *checking* balance in observed covariates. While balance assessment on its own does not solve the central problem in observational studies of unobserved confounding, I show that my framework can strengthen one widely-used means of generating partial insight into causal credibility from the observed data.

The key issue I address is that conventional implementation employs a "folk definition" of balance—the difference in observed covariate means across treatment conditions, which may *indirectly* suggest balance in the potential outcome means. This long-standing approach to balance assessment provides, at best, incomplete insight into a research design's credibility. Its concentration on only a single sample-based quantity reflects a minimalist approach to evaluating the design. Given that observed covariate balance cannot, on its own, guarantee identification, I contend that such a bare-minimum assessment is not the optimal perspective. In general, researchers can implement a more holistic approach to scrutinizing their identification claims by combining observable balance with other quantitative assessments as well as qualitative context on the empirical case. However, I contend that there is also room for improvement in balance evaluations themselves. My proposed framework aligns balance assessment with the concept's formal definition: equality in the conditional distributions of observed covariates by treatment status. Doing so facilitates informative and intuitive observed covariate balance analyses that promote rigorous falsification of one type of observable implication of a research design.

I discuss the current problem and my solution with a case study of political science—a discipline that heavily relies on observational studies to advance many of its literatures. As a crossroads of the social sciences that incorporates many elements of its sister fields, insights from political science can broadly inform best practices for researchers in a variety of disciplines. The most extensive example I consider is the study of voting costs—how reforms such as state voter identification (ID) laws affect political engagement, elections, and policymaking. Causal identification is difficult with these policies because their adoption by state governments is largely determined by citizens' and politicians' preferences (Erikson and Minnite, 2009; Grimmer et al., 2018). A recent set of studies purports to mitigate this selection into treatment by leveraging as-if random variation in the timing of legal challenges to voter ID laws in the American states (Harden and Campos, 2023; Campos et al., 2024). I contend that such a design is promising, but must be empirically evaluated with strict scrutiny to interrogate its observable implications. Investigating balance in observed covariates is one tool that can support that endeavor.

Current practice for balance assessment in political science involves comparing observed covariate means by treatment group to test either the null hypothesis of equality (i.e., the treatment and control group are equal, on average) or the null of average difference between groups (Hartman and Hidalgo, 2018; Hartman, 2021). As they are implemented, neither of these alternatives can detect imbalance in higher moments of the covariate distributions. Moreover, both rely on null hypothesis significance testing (NHST) to evaluate observable implications of identification strategies. My proposed framework equips researchers to (1)

easily measure the sampling variability of multiple statistics that collectively encode balance in the full covariate distributions and (2) move beyond the NHST paradigm, avoiding the use of p-values as indicators of causal credibility. These two contributions yield a simple, but powerful framework that aligns with the formal definition of balance and is broadly applicable to common observational studies, including those employing selection-on-observables assumptions, natural experiments, and regression discontinuity designs.

More specifically, the framework I present evaluates observed covariate balance using either classical bootstrapping or my preferred approach—the Bayesian bootstrap (Rubin, 1981)—to measure balance statistic uncertainty. Further, it conducts these bootstrapping procedures for virtually any balance statistic of interest, including covariate-level measures (e.g., variance ratios) and global indicators (e.g., L1). Utilizing a wider array of metrics provides more comprehensive information on balance than the conventional focus on mean differences alone. Further, the Bayesian version of the method produces posterior distributions for the statistics of interest, which permit novel interpretations of balance with intuitive probabilistic statements. For instance, rather than reporting the results of a hypothesis test, analysts can communicate the posterior probability that an observed covariate meets an analyst-defined target for balance. To promote useful definitions of balance for such statements, I incorporate reasoning and insight from equivalence testing and related literature into the framework (e.g., Rainey, 2014; Gross, 2015; Hartman and Hidalgo, 2018; Hartman, 2021). The result is clearer, more informative, more intuitive, and more rigorous investigations of balance in observational studies. Additionally, the method is easily integrated into researchers' standard workflows with well-known software packages (e.g., Iacus et al., 2011; Greifer, 2024).[1] Thus, there is minimal cost and considerable benefit in adopting my proposed framework.

In what follows, I motivate the need for a new balance assessment framework in observational studies that offers stronger falsification of identifying assumptions in the observed data. I contend that, given the choice to compute observed covariate balance as part of a larger evaluation of identification, researchers should seek evidence *consistent with distributional equality* in the measured variables rather than simply suggest that they display the *absence of mean differences*. I demonstrate that research in top political science journals rarely uses most of the available tools for balance assessment, such as higher-moment statistics or equivalence testing. Then I show how my framework consolidates these methods to improve applied practice. Finally, I employ Campos et al.'s (2024) data on voter ID laws in American state governments to illustrate how the framework can yield a more rigorous falsification of the observable implications of a design that warrants strict scrutiny.

2. Balance Evaluations in Observational Studies

Formally, balance in an observed covariate x given binary treatment variable T implies that the conditional distribution of x is identical for the treatment and control groups, or $p(x \mid T = 1) = p(x \mid T = 0)$. However, because applied researchers have largely focused on average differences between groups (see Section 2.3 below), balance assessment carries a folk definition of tests of mean balance with conventional (NHST-based) inference. In what follows I contend that this conventional approach comprises only part of a thorough empirical

1. See the Appendix for example code.

evaluation regime that rigorously subjects a design to falsification. My proposed framework advocates for balance investigation of the full conditional distribution of x, reflecting the formal definition of balance.

In observational studies, researchers do not control the assignment mechanism and thus do not enjoy the experimentalist's "benefit of the doubt" that the treatment effect estimate is bias-free in expectation. Accordingly, the demand for scrutinizing identification assumptions increases dramatically when moving from an experiment to an observational analysis. The social world is inherently complex, strategic, and nuanced, leaving considerable doubt regarding (conditional) ignorability of treatment a priori when studying observational data. Researchers must address this complexity by reporting a constellation of distinct conceptual and empirical tasks interrogating whether a treatment was assigned by nature as-if randomly or that, given the observed covariates, confounding of the treatment effect has been sufficiently mitigated.

First, substantive evidence—which includes quantitative or qualitative description—can be used to justify the theoretical case for treatment exogeneity (Dunning, 2012). Second, falsification tests such as computing balance on held-out observed covariates, estimation on placebo outcomes, or leads and lags of treatment in panel data are useful for scrutinizing observable implications of the design (Eggers et al., 2024). If a covariate adjustment method is warranted, the analyst can assess its effectiveness by computing balance on the variables in the conditioning set.[2] Finally, sensitivity analyses are useful for assessing robustness to unobserved confounding (Cinelli and Hazlett, 2020).

The method I propose here fits into the second and/or third steps. It is not a substitute for or competitor to any of the tools noted above. Instead, it *complements* them by strengthening the evidence provided by the balance aspects of the taxonomy. The framework assists researchers in presenting a more detailed assessment of observed covariate balance. It discourages simply checking balance, which I define as computing observed covariate mean differences and relying on NHST to test nulls of equality (the most common practice in political science; see below) or average difference (Hartman and Hidalgo, 2018). Instead, the method permits analysts to truly *investigate* balance in the observed data in a compelling manner, focused on falsification.

2.1 Two Objectives for Balance Evaluations

Before considering the components of a balance investigation or the steps in my proposed framework, it is important to distinguish between two related, but distinct objectives for balance evaluations as I discuss them. These objectives correspond with two different classes of observed covariates for which researchers might be interested in computing balance statistics, as noted above. One is the set of covariates Z which the analyst argues must be conditioned upon to satisfy the ignorability assumption. These variables may display imbalance before adjustment, but should demonstrate balance in the adjusted data. Balance evaluation in this setting reflects an important "feasibility check" on the adjustment procedure itself and is likely most relevant to designs that rely heavily on a selection-on-observables

2. Several matching and weighting methods are designed with the specific purpose of optimizing (1) a balance quantity such as the mean and/or a higher moment or (2) balance and sample size simultaneously (e.g., Iacus et al., 2011; Hainmueller, 2012; Zubizarreta, 2015; King et al., 2017).

assumption. A matching algorithm, for instance, should create a subset dataset on which Z displays better distributional similarity than in the raw data. Ben-Menachem and Morris (2023) provide an example of this objective. The authors estimate the individual-level effect of police traffic stops on voter turnout. They argue that a host of individual, neighborhood, and institutional variables must be controlled to credibly isolate the effect of the traffic stop. Their table of balance results includes entries for all of those variables using the matched data (825). The authors then leverage those results to claim that "the selected control voters are very similar to the treated voters" (Ben-Menachem and Morris, 2023, 828).

The other objective targets unconditional ignorability or, if conditional ignorability is claimed, tests the validity of the conditioning set. It is most relevant to high-credibility designs in which the analyst conducts a falsification test on a held-out set of covariates (X). An observable implication of treatment exogeneity is that any covariate in X should display balance—before or after covariate adjustment and even if the covariate is not directly theorized to affect selection into treatment. Any distributional inequality is direct evidence against the identification claim. Rossiter and Harden (2026) provide an example of this perspective. They study a natural experiment generated by a ticket lottery in which participants exerted some control over their total number of entries. The authors adjust for variables predicting entries with balancing weights, then report balance results for that conditioning set (Z) as well as several other unadjusted variables not included in the weighting algorithm (X).

My framework is sufficiently general that it is useful to researchers seeking to address either of these objectives. The method can be utilized for an internal audit of the design—assessing whether the chosen statistical adjustment operated as expected. However, such an analysis on its own does not fully test treatment exogeneity. By employing the framework to investigate balance in X, analysts set up a direct falsification of causal identification. Ideally, researchers should pursue both objectives; the most robust designs should show evidence that their adjustment procedure is effective at balancing observed variables *and* that distributional equality across treatment conditions extends beyond those covariates on which the data are adjusted. Importantly, my framework can seamlessly evaluate balance on both types of observed variables.

2.2 Components of a Balance Investigation

Any causal identification strategy must begin with a substantive, theoretical justification of (conditional) treatment ignorability (see Dunning, 2012). However, beyond this first aspect of the argument for a design, researchers must also conduct rigorous empirical falsification testing. Given that an analyst has chosen to assess observed covariate balance as part of this examination, what elements comprise a rigorous investigation? I contend that they should focus on maximizing *information* and *intuition* in their assessments. A core requirement for conditional mean independence is (near) zero mean differences in the covariates between the treatment and control groups (Morgan and Winship, 2015).[3] Equal means alone may be consistent with the requirements for identification in cases where treatment was truly random (and thus ignorability holds). But in most observational settings there is rela-

3. Thresholds for determining if the absolute value of the standardized mean difference in a finite sample is sufficiently small to assert balance include 0.10, 0.20, and 0.25 (e.g., Ho et al., 2007; Harder et al., 2010).

tively high initial skepticism toward the research design. In these circumstances, analysts can strengthen their observed covariate balance assessments by providing more information than the minimalist folk definition implies and instead target the formal definition. Doing so involves demonstrating that the full observed covariate distributions look similar across treatment groups rather than just the means. Balance quantities that measure higher moments can provide consequential insight that mean differences obscure (Ho et al., 2007; Greifer, 2024).[4] For instance, imbalance in quantiles or variance could motivate a search for previously-omitted variables or suggest meaningful substantive differences between treatment groups that are otherwise similar on average (see Section 6 for an example).[5]

Indeed, demonstrating that the mean *and* variance of an observed covariate across groups are *both* similar (rather than only the mean) is a stronger empirical diagnostic of the identification claim. Of course, such an exercise can only be conducted on observables and does not yield insight into hidden confounders. But an accumulation of evidence suggesting distributional similarity of observed variables lends support to the assumptions required for causal identification. In general, an investigation of balance necessitates a more thorough evaluation of observable similarity between the treatment and control groups. Reporting multiple types of balance statistic supports that objective by increasing the "resolution" of the observed data evaluation.

I also advocate for an inferential approach to interpreting balance statistics. The need for inference may not be immediately self-evident because balance statistics are inherently tied to the sample from which they are computed; they do not represent an estimate of a population parameter. However, inference is necessary for deploying balance statistics as falsification tests of the observable implications of a research design's identifying assumptions. Bias to the treatment effect from a confounding variable is a function of the degree to which that variable is imbalanced and its effect on the outcome. Estimating the latter association is difficult, which makes imbalance a key measure of interest for scrutinizing the design. To maximize the effectiveness of this strategy, I contend that it is useful to distinguish systematic variation from stochastic noise in balance statistics. Any finite sample is likely to reveal some imbalance, even in a design deemed highly credible a priori. An inferential approach equips the analyst with a means of formally communicating their uncertainty in the chosen statistics, which strengthens their interpretation of the results.

Importantly, however, this uncertainty must be communicated intuitively and through substantively relevant comparisons. As noted above, current tools typically involve the use of NHST—either with a null of equality or difference. The shortcomings of NHST and the broader over-reliance on p-values are well-established in the scientific literature (e.g., Gill, 1999; Benjamin et al., 2018). I contend that these issues could arise when conducting inference with observed covariate balance as well; manipulation of the researcher degrees of freedom that yield statistical significance (or not) is an omnipresent risk. More importantly, however, NHST does not facilitate meaningful balance discussions. A substantively small

4. This logic is most applicable to non-binary covariates. In the case of a binary covariate, balance in the mean is sufficient.

5. Analysts can also assess the full distribution of a covariate with Kolmogorov-Smirnov (KS) or complement of overlap (OVL) statistics (Sekhon, 2011; Franklin et al., 2014). Measures proposed by Hansen and Bowers (2008), Iacus et al. (2011), and Caughey et al. (2017) permit omnibus or multidimensional balance assessments across multiple covariates.

difference between groups could be statistically significant with enough power and a large difference might yield a p-value just outside the level of the test. Equivalence testing helps mitigate these problems to some extent (Hartman and Hidalgo, 2018). However, that approach still typically relies on NHST.[6] Moving away from NHST and communicating results in natural probabilistic terms would simplify interpretation and appeal to broad audiences of the research. Consider the following three statements:

- The null hypothesis that the covariate difference in means between the treatment and control groups is zero cannot be rejected at the 0.05 level.
- The null hypothesis that the covariate difference in means between the treatment and control groups is substantively meaningful can be rejected at the 0.05 level.
- The posterior probability that the covariate means are substantively balanced between the treatment and control groups is 0.45.

The first two statements conceptualize balance as a binary decision based on a chosen statistical significance level and require an understanding of NHST. Conversely, the third statement leverages probability as an intuitive measure of uncertainty, enhancing precision and accessibility for both expert and general audiences (Crauder et al., 2018).

In brief, observational analysts faced with the need to interrogate causal identification should *investigate* observed covariate balance—rather than *check* it—as a means of providing one compelling falsification test of their designs. I contend that balance investigations necessitate fuller insight than conventional practice provides as a means of credibly searching for threats to inference and for supplying substantive experts with a multidimensional picture of the data. Additionally, the results of balance investigations must be reported in an intuitive manner. Doing so communicates more detailed evidence regarding covariate balance compared to the result of a hypothesis test. However, as I show next, a substantial gap exists between this ideal approach and reality in one discipline that heavily relies on observational studies.

2.3 Meta-Analysis: Balance Assessment Practices in Political Science

How do researchers analyzing observational data actually evaluate and discuss observed covariate balance in practice? I answer this question with a systematic description of methods employed within political science. Table 1 summarizes information from observational studies that report observed covariate balance assessments published in the discipline's top tier of journals—the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*—from January 2019 to June 2024.[7] Using the articles' main text and supplementary materials, I coded indicators for whether each study (1) reports covariate means by group and/or their differences, (2) uses mean difference p-values to argue that imbalance is absent (e.g., t-tests), (3) conducts equivalence tests of the null of difference, (4) reports any other balance statistics, and (5) offers *substantive* interpretation of the balance results.

6. Furthermore, as I demonstrate in Section 2.3, applied political scientists rarely use equivalence tests to assess balance.
7. The articles in the sample report natural experiments, regression discontinuity designs, and other studies. See the Appendix for searching, selection, and coding procedure details.

Table 1: Covariate Balance Assessments in Observational Studies in Political Science, 2019–2024

Journal	Articles	Report means	Report p-values	Equivalence test	Report other	Interpret results
APSR	18	94%	61%	0%	22%	22%
AJPS	18	94%	56%	6%	17%	28%
JOP	18	89%	56%	6%	50%	33%
Total	54	93%	57%	4%	30%	28%

Note: Cell entries report information on covariate balance assessment in the *American Political Science Review* (*APSR*), *American Journal of Political Science* (*AJPS*), and *Journal of Politics* (*JOP*). See the Appendix for full details.

This meta-analysis reveals several noteworthy patterns about balance assessment in practice.[8] First, in line with the folk definition the mean difference by group is by far the most common balance statistic used by political scientists. In about seven of 10 cases, it is the *only* statistic that researchers report. While mean differences are undoubtedly useful for balance evaluation, I contend that observational analysts need to offer more (see Section 2.2). Relying only on mean differences may preclude the possibility of learning more about the data generating process (DGP) of interest (see Section 3).

Additionally, a majority of studies report p-values from tests of statistical significance in the mean differences. This practice arises from the property that, in a given sample under random treatment assignment, a small subset of the covariates (e.g., 5%) can be expected to show significant differences due to chance alone. One perspective would indicate that such an inferential evaluation is actually not necessary because balance is a property of the sample, not the population (Ho et al., 2007; Greifer, 2024). However, a bigger problem is that even within the framework of NHST these observational studies do not employ best practices. The meta-analysis reveals that they nearly all use inferential tools to test a null hypothesis of equality. Just two articles included in the sample conduct equivalence tests of the null of difference.[9]

The meta-analysis also indicates that political scientists analyzing non-experimental data typically do not offer interpretations of or intuitive inferential statements about their balance results. Only 15 of 54 articles in the sample provide any kind of interpretation. Many of them simply report a table or figure with balance statistics and no additional context or discussion. Some provide minimal descriptive information about the results, but nothing about what they mean for the degree to which the treatment and control groups are similar. And even those that provide interpretation generally rely on statistical significance as their underlying justification for balance. Consider the following paraphrased examples from articles in the meta-analysis sample:

8. Many of these studies also employ other tools and analyses intended to scrutinize and justify identification assumptions beyond observed covariate balance assessment.
9. In some cases, equality might be the most appropriate null (Caughey et al., 2017). But in a case in which the standard for evidence of balance is relatively high, the null of difference is the more compelling test (Hartman and Hidalgo, 2018).

- Matching improves the comparability of treated and control cases; the standardized mean difference is less than 0.14 in each pretreatment year, and the maximum difference across all covariates is 0.27 (basic description, *American Journal of Political Science*, 2023).
- . . . of more than 58 pretreatment covariates, fewer than five show differences that are significant at $p < 0.05$. This level of covariate balance is what one would expect had the groups been formed by random assignment (statistical significance, *Journal of Politics*, 2024).
- After matching and refinement, the samples are well-balanced using the 0.2 standard deviation criterion (balance threshold, *American Political Science Review*, 2024).

Simply put, there is limited information here for domain experts to use in substantively evaluating the identification strategies.

To summarize, my motivation for a new, falsification-based, observed covariate balance assessment framework stems from the typical folk practice of reporting minimal information and a reliance on NHST in existing research.[10] The evidence and argumentation offered in support of balance in observational studies published in political science's most prominent outlets are generally quite weak when considered against the demands of identification implied by those designs. This pattern is especially problematic given that observed covariate balance is, at best, only circumstantial evidence consistent with identification—it does not ensure the validity of the identification assumptions (Ho et al., 2007). Observational analysts must improve discussions of how their balance results indicate meaningful similarity across groups. They can accomplish this objective by evaluating balance beyond average differences in the observed covariates and presenting and reasoning with intuitive measures of uncertainty in their balance statistics rather than relying on NHST.

3. Investigate Beyond the Mean

The first component of my framework involves computing statistics that assess balance in distributional moments other than the mean for non-binary variables. I largely focus on the following four statistics, which are easily computed by standard software packages (e.g., Greifer, 2024). But many others exist in the literature and are also available under my framework, such as Iacus et al.'s (2011) multidimensional L1 statistic (see Section 6).

10. A complementary alternative to the framework I describe here is the use of flexible predictive models—such as a machine learning (ML) algorithm—for balance assessment. This approach appears in the literature (e.g., Hübert and Copus, 2022) and holds several advantages: it is elegant, adaptable, and yields straightforward, omnibus measures of balance. It could be utilized fruitfully in tandem with my framework. However, there are some potential drawbacks as well. Its streamlined nature could lead researchers to inadvertently interpret the results as verification of causal identification rather than a falsification test with observed data. Additionally, ML-based methods tend to serve as "black boxes." They can identify complex combinations of observed variables that predict treatment better than chance. But they are often less well-suited to describing specific characteristics of the imbalance, such as its originating distributional moment or whether it is a substantively meaningful difference. I argue that statistical parsimony may not provide the most informative diagnostic information in the context of balance evaluation. There is value in the granular approach of iterating through multiple variables and statistics to inform a substantive discussion of balance.

- Standardized mean differences (SMD): The difference in group means, scaled by the pooled standard deviation. This conventional measure of the first moment remains a useful starting point for balance evaluation.
- Variance ratios (VR): The ratio of the treatment group to control group variance. This statistic measures imbalance in the second moment of the covariate distributions.
- Kolmogorov-Smirnov (KS) statistics: The largest vertical distance between the empirical cumulative distribution functions of the two groups. KS statistics measure the maximum point of imbalance across the full range of the covariate distributions.
- Complement of overlap (OVL) statistics: The total density between the distributions that is not shared. Specifically, the OVL statistic computes the complement of the area under the minimum of both curves: $1 - \int \min_x \left[\widehat{f_{\text{treated}}}(x), \widehat{f_{\text{control}}}(x)\right]$ (Franklin et al., 2014).

The first statistic encodes only locational differences in the distributions while the other three measure balance in location, spread, and/or shape. My recommendation is that applied researchers utilize my framework to compute SMDs, VRs, and either KS or OVL statistics (see Section 7 for more details).

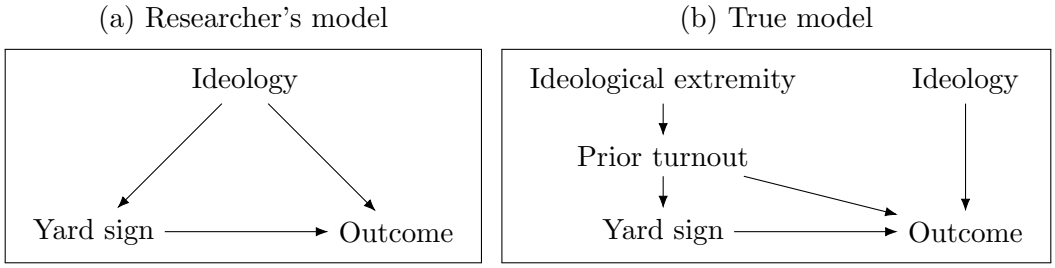
I use these statistics in a simulation exercise to demonstrate a possible consequence of limiting evaluations of observed covariate balance to average differences only (and not higher moments). In this case, I consider a situation in which a covariate is balanced across treatment conditions on the mean, but not on the variance. The example highlights one instance in which implementing the folk definition is insufficient for detecting substantively important observed covariate imbalance. Specifically, identifying imbalance in the second moment of an observed covariate falsifies the assumption of treatment ignorability. Thus, the analyst learns that they must reconsider their specification more generally.[11] I adopt a substantive context from an observational study in political science to facilitate intuition, but this logic is broadly applicable to other observational settings.

Makse et al. (2019) investigate the effects of displaying a political yard sign on the size and tone of individuals' political discussion networks using survey data from presidential election years in the United States. Importantly, a clear threat to inference emerges from several factors that systematically predict selection into treatment, such as ideological extremity, interest in politics, and others (Makse et al., 2019, 47–48). Consider a case in which a researcher posits the causal model for this process described in Figure 1, panel (a). This model identifies self-reported ideology as a confounder of the treatment effect. For example, the researcher might believe that conservative voters are more enthusiastic about an election than liberals and thus are more likely to display yard signs *and* engage in political discussion. This hypothesized process leads them to control for ideology in the estimation, but no other covariates. However, for the sake of the illustration, assume that the true process operates according to the model in panel (b) of Figure 1. Extreme ideologues on *both* sides are more likely to have voted in the previous election compared to moderates,

11. Detecting imbalance does not necessarily imply that there is an omitted variable or that the analyst's causal model is incorrect. It may simply empirically validate a known confounder that the analyst plans to mitigate with an adjustment strategy.

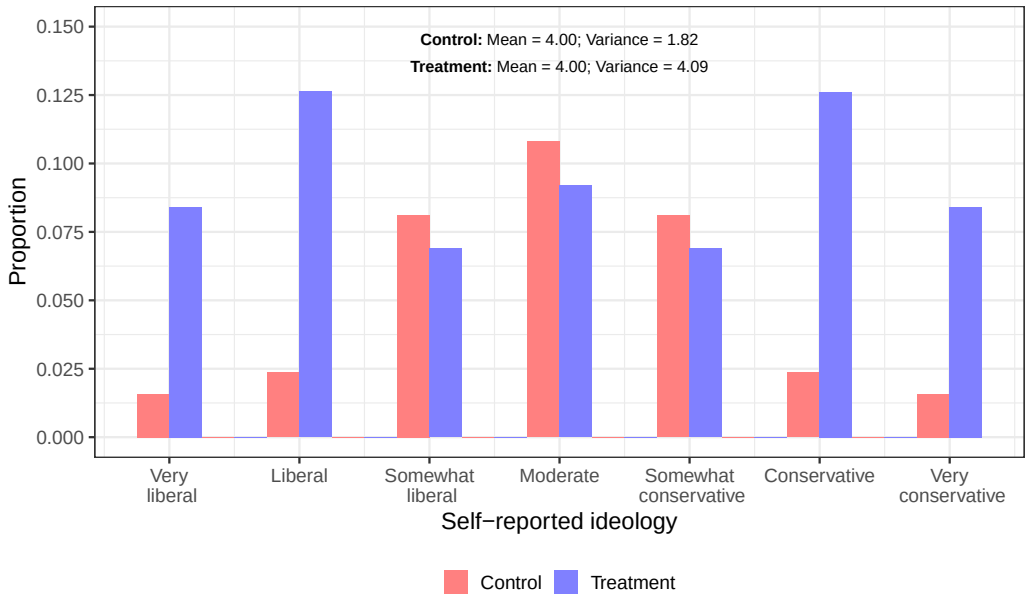
which increases their likelihood of displaying yard signs, which in turn affects the outcome of interest. Ideology affects the outcome, but is not associated with any other variables.

Figure 1: Causal Models in the Simulated Data Example



I simulate the DGP of Figure 1b to assess the consequences of the researcher's model for balance assessment and treatment effect estimation (see the Appendix for replication code). I draw 1,000 samples of 500 observations with a true treatment effect of 0.50. I estimate the effect in each sample under three specifications: (1) no covariates, (2) a control for ideology, and (3) controls for ideology and prior turnout. Figure 2 graphs the distribution of ideology by treatment status in the population. Consistent with the true model in Figure 1b, the treatment group (individuals with yard signs) includes relatively more extremists (distributed equally at each pole) whereas more moderates appear in the control group. The result is a mean of four in each group, but more than double the variance in the treatment group.

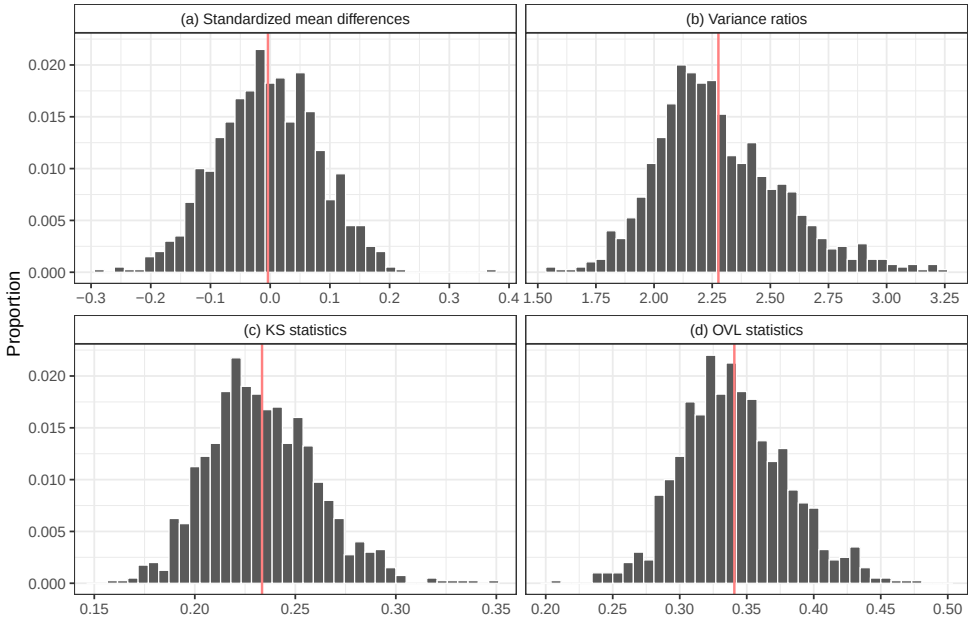
Figure 2: Population Distribution of Simulated Self-reported Ideology by Treatment Status



Note: The graph presents the population distribution of self-reported ideology by treatment status in the simulation.

Figure 3 graphs the distributions of balance statistics across the 1,000 simulated samples. Panel (a) shows that the SMD statistic is essentially zero, on average, indicating no imbalance in ideology across the two groups. This result is a product of the DGP, which sets the population means of the two group to the same value (Figure 2). And across the 1,000 simulated samples, only 23 (2.3%) generate statistically significant mean differences ($p < 0.05$). This frequency would, of course, vary with sample size and other power considerations. But in general, Figure 3a demonstrates that assessing balance by checking mean differences—the typical approach in applied work—would not give the researcher any reason for concern.

Figure 3: Distributions of Balance Statistics in the Simulated Data



Note: The graphs present the distributions of balance statistics in the simulated data. Red vertical lines indicate the mean for each statistic.

However, as Figures 1 and 2 imply, only computing mean differences would lead to erroneous conclusions. It would suggest that ideology is not a confounder, which might prevent the analyst from considering the influence of other, related variables like ideological extremity and prior turnout.[12] Importantly, ideology is a confounder—just not through its mean. The other graphs in Figure 3 correctly identify the variance-based imbalance in the ideology covariate. The VRs (panel b) indicate treatment variance of more than double the control group variance, on average; the benchmark for that statistic is one (i.e., equal variances). Panel (c) shows KS statistics near 0.20 and 0.30. Markoulidakis et al. (2023) suggest that values greater than 0.10 denote imbalance. Finally, the OVL statistics demonstrate that about 35% of the two groups' densities do not overlap, on average. In

12. Controlling for ideology only yields a biased treatment effect estimate—a mean value of 0.683—and a root mean squared error (RMSE) of 0.206. Specifying the correct model (Figure 1b) with ideology *and* prior turnout as covariates yields a mean estimate of 0.501 (RMSE = 0.115).

short, these three statistics correctly encode the clear differences between the groups in the population shown in Figure 2. Unlike mean differences, incorporating them into balance evaluation would bring this divergence to the researcher's attention. Ideally, they would then recognize the pattern that extreme ideologues are more likely to be treated and reformulate their model of the process to focus on ideological extremity and its correlates (e.g., prior turnout) as additional confounders.

This simulation example illustrates why the folk approach of checking for imbalance by determining the statistical significance of mean differences is insufficient in observational studies. Further, *the use of equivalence testing would not solve the problem*; mean differences would fail to detect the imbalance regardless of whether the analyst specified a null of equality or a null of difference. Unlike in an experimental setting, observational analysts should expect more complex relationships between treatment, covariates (observed and unobserved), and the outcome. Investigating balance beyond just the average can assist researchers in uncovering these nuanced patterns.

4. Measure Balance Statistic Uncertainty

Improving observational balance assessment in practice also requires meaningful reasoning and inference with the chosen balance statistics' sampling variability. My proposed approach for meeting this need involves evaluating observed covariate balance via bootstrapping, either from a classical[13] or—preferably—a Bayesian perspective (Rubin, 1981).[14] The goal is to compute a sampling or posterior distribution for a balance statistic of interest, then summarize it with probabilistic statements about the consistency of the data with a balanced design rather than test a null of equality.[15] This component of the framework is flexible and adaptable to researchers' specific needs. It can be used to analyze a wide range of balance statistics and is suitable to many types of observational studies, including covariate adjustment strategies, natural experiments, and regression discontinuity designs (see the Appendix).

For data of length n, classical bootstrapping can be conceptualized as repeatedly drawing probability weights $\pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ for the n data points and computing the statistic of interest weighted by π. The weights are constructed by drawing counts of each observation in a bootstrap sample from a multinomial distribution with n trials and all probabilities $p \in (p_1, p_2, \ldots, p_n)$ equal to $\frac{1}{n}$. These count weights sum to n and are normalized to create the probability weights, which sum to 1.[16] Repeatedly drawing bootstrap weights and applying them in the computation of a balance statistic of interest yields a sampling distribution for

13. Sekhon (2011, 10) recommends classical bootstrapping to compute KS statistic p-values. My method builds on this suggestion, but reduces the need to rely on NHST in balance assessments.
14. The framework I propose here does not require the analyst to posit a Bayesian model for the data generating process. Because it is only used to assess balance, it adds no new assumptions to the method for estimating treatment effects.
15. Bootstrapping can perform poorly in finite samples for some statistics without corrective procedures such as studentization (Hesterberg, 2015). In the Appendix I report results from a simulation of coverage rates using my proposed method on several balance statistics. The method performs well in general and especially as the sample size increases.
16. For example, consider a vector of four data points. One possible draw of counts from the multinomial distribution is (1, 0, 2, 1), which translates to probability weights of (0.25, 0.00, 0.50, 0.25). In other words, that sample would give no weight to observation #2 (i.e., leave it out) and twice the weight of

the statistic, which can be used to compute confidence intervals and other related quantities that support frequentist inferential statements about the observed similarity between the treatment and control groups.

Alternatively, the analyst could extend this conceptualization of classical bootstrapping to a Bayesian framework and gain even more analytic leverage to evaluate balance. The Bayesian bootstrap reflects the logic described above, but draws the probability weights on each observation from a uniform posterior Dirichlet distribution with parameter $\alpha = 1$.[17] The sampling distribution for π remains a Multinomial$(p_1, p_2, \ldots, p_n)$ as in the classical case. The posterior is the product of this sampling distribution and a limiting Dirichlet prior, which concentrates density near the vertices of an n-dimensional simplex as $\alpha \to 0$. This prior is data-driven in that it gives the most weight (but not all) to values that appear in the observed data. In fact, classical bootstrapping is a special case of the Bayesian bootstrap with a Dirichlet prior and $\alpha = 0$, which places all of the prior weight on the observed values. The posterior of the balance statistic of interest is generated by sampling π from its posterior and using each draw as a weight in the computation. See the Appendix for a graphical illustration of classical and Bayesian bootstrapping as described here.

Both versions of this process are essentially equivalent in practice, especially as n grows larger. Hastie et al. (2009, 272) refer to classical bootstrapping as generating a "poor man's Bayes posterior." My method is feasible with either approach. However, I recommend Bayesian bootstrapping for several reasons. First, in small samples the classical bootstrapping weights generated from normalizing discrete multinomial counts can be "choppy" due to the presence of very few observed values. The uniform Dirichlet distribution, by contrast, allows for smoother draws of the weights at all sample sizes. This property is particularly beneficial in applications with, for instance, a small number of treated units (e.g., Campos et al. 2024; see Section 6). The Bayesian version also eliminates "corner cases" in which classical bootstrapping fails, such as bootstrap samples with perfect collinearity between variables or no variance in a binary variable measuring a rare event (Courthard, 2022). Bayesian bootstrapping can also be vectorized and/or parallelized to decrease computation time relative to classical bootstrapping (Bååth, 2018; Courthard, 2022). Most importantly, the posterior distributions of the balance statistics facilitate more informative and intuitive inferences, as I discuss next.[18]

5. Intuitively Reason About Balance

The estimated balance statistic sampling or posterior distribution provides a measure of uncertainty that can be used to investigate the observed similarity of the treatment and control groups. The next key issue is how researchers can make informative inferential statements about balance with this quantity. My suggestions involve (1) principles from equivalence testing and related literature on the substantive significance of statistical estimates (Rainey, 2014; Gross, 2015; Hartman and Hidalgo, 2018) and (2) probabilities gen-

the original data to observation #3 (appears twice in the bootstrap sample). The other observations appear once each in the bootstrap sample.

17. Increasing α reduces variance in the weights such that as $\alpha \to \infty$ there is no weight variance, which results in the original sample (Shao and Tu, 1995; Courthard, 2022).

18. Another resampling method that could be utilized with my framework is randomization inference. I summarize this approach and discuss why I do not emphasize it in the Appendix.

erated from the posterior distribution of the balance statistic of interest. I focus on the Bayesian implementation in discussing this component of the framework because I prefer its probabilistic interpretations and obviation of NHST. Analysts who choose to employ classical bootstrapping can complete this step by adapting the procedures for evaluating substantive significance in a frequentist context described by Rainey (2014) and/or Gross (2015).

5.1 The Equivalence Range

My first recommendation for effectively reasoning about observed covariate balance is to construct a range of values around the statistic's benchmark of "perfect" balance—which I refer to as τ—that are "effectively equal" to τ and thus would not be considered meaningfully different. Hartman and Hidalgo (2018, 1003) refer to this concept as an *equivalence range* (ER). As an example, for SMDs τ is zero because that value indicates no difference between groups, on average. Non-zero SMDs indicate imbalance, but some of those values are likely inconsequential and can be considered essentially zero. For VRs, this logic holds for values above and below $\tau = 1$. The challenge for the analyst is selecting the ER values, which can be symmetric around the benchmark, asymmetric, or one-sided, depending on the statistic and substantive context.[19]

The literature on equivalence testing and substantive significance provides useful guidance for establishing the ER. Rainey (2014) uses the term m to define the range such that, in the case of a symmetric ER, values within $\tau \pm m$ would yield the same substantive conclusion. Gross (2015, 780) characterizes the range as "practically indistinguishable" from the target value (i.e., τ). Rainey (2014, 1085) and Gross (2015, 779–780) place responsibility on the researcher—as the substantive expert—to select and defend the ER (see also Hartman and Hidalgo, 2018, 1005). Indeed, engaging with this process is an important use of the analyst's substantive expertise. Wellek (2010, 15–16) explicitly argues *against* developing general rules for the process because there is considerable heterogeneity across the questions, data, and designs of individual studies. Nonetheless, this literature also includes heuristics for use in the absence of substantive information, such as standardized treatment effect sizes and default values (Bonett, 2009; Hartman and Hidalgo, 2018). For example, one recommendation for SMDs based on simulation evidence is $m = 0.36\sigma$, where σ is the pooled covariate standard deviation (Hartman and Hidalgo, 2018, 1006).

I also maintain that selecting and defending the ER is the researcher's responsibility. Despite the challenge, studying the data and past literature, consulting with other researchers, and thinking through the substantive details required to make sensible choices are all valuable practices that holistically improve the quality of balance assessments and the research in general. Moreover, the analyst can survey experts (Fitzgerald, 2024) and/or preregister the ER (Harden et al., 2019) to guard against selecting values after balance results are known. I recommend generally following the logic described by Gross (2015) to complete this process. Specifically, the researcher should structure their decisionmaking around the following questions.

19. Conventional equivalence tests include directional null hypotheses (non-inferiority and non-superiority tests; see Arel-Bundock et al., 2024). I adopt the general form here because my method accommodates both types of null.

- What values of the balance statistic of interest would indicate perfect balance in a given observed variable (τ)?
- Moving away from τ in a positive and/or negative direction, what value(s) reflect inflection points ($\pm m$) such that the treatment and control groups are substantively different enough to suggest a falsification of the identifiction assumptions?
- What value(s) of m does past literature and/or heuristics suggest are inflection points? Are these values sensible for the current application?
- Does the sample size, treatment assignment mechanism, nature of the covariate(s), or other substantive or empirical considerations make finding values outside a candidate range $[-m,\ m]$ more or less likely? If so, does the range need to be expanded or contracted to accommodate these expectations?
- Do imbalances revealed in observed covariates suggest any unobserved factors not yet considered that might confound the treatment effect?

The answers to these questions can help determine an appropriate ER within the context of the substantive question and available data. And while this exercise may seem somewhat simple and self-evident, it is important to reiterate a finding from the meta-analysis in Section 2.3: *current practice among applied researchers includes virtually no reflection on observed covariate balance of any kind*. Moreover, if there is uncertainty or disagreement among experts on these points, the values themselves can be further interrogated in a sensitivity analysis (Hartman and Hidalgo, 2018). See the Appendix for an example of this decisionmaking in practice.

5.2 Incorporating Posterior Probability

The Bayesian version of my framework permits intuitive summaries of the balance statistic posterior distribution. Probability is a powerful tool for communicating many types of quantitative information because it is defined by precise axioms and yet intuitive and familiar to a broad audience (Crauder et al., 2018). Evaluating balance with an emphasis on probability is crucial to providing meaningful informational context for balance assessment. Indeed, my central motivation for proposing bootstrapping from a Bayesian perspective is to yield balance statistic posterior distributions that can be summarized with probabilistic statements. In particular, I suggest two categories of statement that are useful for evaluating balance.[20]

One type of statement involves reasoning with the plausible values denoted by variation in the posterior. The analyst could focus their investigation solely on the posterior if they prefer to avoid defining an ER. But if they identified values for τ and m, they could also make comparisons between posterior credible intervals and those values. To incorporate probability directly, they could compute the *probability of observed covariate balance*—an intuitive quantity that is only available with the Bayesian implementation of the framework.

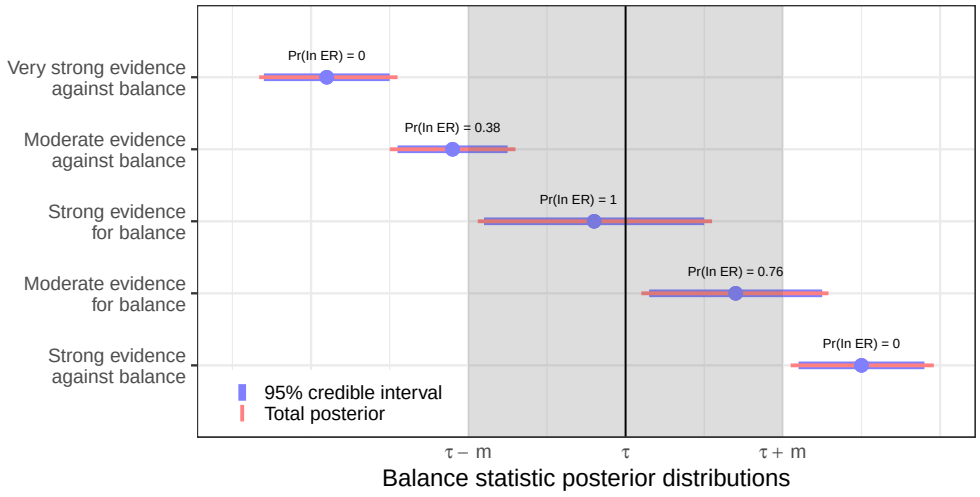
20. I propose these quantities because they are simple and intuitive, but others are available that may be preferable in a given application. My suggested metrics are not intended to optimize a specific quantity, but rather facilitate informative and natural inferential statements about a given balance statistic.

I define this metric for estimated posterior distribution F as the probability that a realization x of balance statistic X falls inside the ER:[21]

$$\Pr(\text{In ER}) = \Pr(-m \leq x \leq m) = \widehat{F_X(m)} - \widehat{F_X(-m)}. \quad (1)$$

Figure 4 illustrates this quantity with hypothetical posterior distributions for an arbitrary balance statistic. The ER is shaded and the benchmark of perfect balance in an observed covariate is located at the midpoint of that range (τ). The posteriors reflect varying degrees of evidence consistent with balance. Rather than simply evaluating whether the computed balance statistic is different from τ, the researcher could compare the posterior to the ER. For instance, the first posterior falls entirely below $\tau - m$, providing the most convincing evidence of a violation of the as-if random assignment claim. The next one suggests imbalance, but also that some values practically equivalent to τ are plausible—the posterior probability of the statistic falling inside the ER is 0.38. The strongest consistency with the standard for observed covariate balance is represented by the third posterior, in which the probability is one because the entire density is contained within the ER. The final two posteriors represent, respectively, moderate consistency with the balance requirement (posterior density inside the ER of 0.76) and strong evidence against it (posterior near the ER, but fully outside).

Figure 4: Posterior Probability of Observed Covariate Balance Using a Bootstrapped Balance Statistic and an Equivalence Range



Note: The graph illustrates the process of evaluating the bootstrapped posterior distribution of a balance statistic against a benchmark τ and its equivalence range (ER), defined by $\tau \pm m$ (shaded region). The text reports the posterior probability that the balance statistic falls inside the ER.

Another useful quantity that is only available with the Bayesian version of my framework relates to cases in which a researcher employs an adjustment method, such as matching or weighting, and seeks to assess whether the method improved balance in observed covariates

21. In cases in which τ is the lower bound of the measure, such as zero for KS or OVL statistics, $-m$ could be replaced with τ.

contained in the conditioning set. This scenario yields two balance statistic posteriors for a given covariate: one before and one after adjustment. The analyst could first choose a target improvement level, δ, defined as the smallest meaningful improvement in the balance statistic.[22] Then, using the posterior density before (f) and after (g) adjustment, they could compute the *probability of observed covariate balance improvement* according to that minimum level:

$$\Pr(\delta \text{ improvement}) = \int_{\delta} \left[\widehat{f_{(x)}} - \widehat{g_{(x)}}\right] dx. \tag{2}$$

Typically this type of comparison is problematic because covariate adjustment can reduce statistical power (Ho et al., 2007). However, Bayesian bootstrapping alleviates this issue because f and g encode the uncertainty in the statistic of interest before and after the procedure, respectively. In the absence of specific substantive guidance about δ, I suggest one-fourth of the value chosen for m as a minimum improvement level.

Figure 5 demonstrates this quantity using LaLonde's (1986) well-known data on the effect of job training programs on individuals' earnings, which have been analyzed extensively in economics, political science, and other fields. This example constitutes a "feasibility check," in which the analyst assesses balance in the conditioning set of observed covariates before and after adjustment (see Section 2.1). In this case, nearest neighbor matching is used to balance several observed covariates listed along the y-axis (for details, see Greifer, 2024). The graph presents the posterior densities of the OVL statistic (as an example) for each covariate before (red) and after (blue) matching adjustment. Lower values indicate better balance, with perfect balance achieved at zero (i.e., $\tau = 0$).[23] The points and lines represent posterior means and 95% credible intervals. The text on each line of Figure 5 reports the posterior probability of a five percentage-point improvement (decrease) in the OVL statistic (i.e., $\delta = 5$pp).[24]

For some covariates (e.g., age), the improvement probability is zero because the *unadjusted* balance is better, on average. With others, such as marital status, matching clearly improves balance. But for some covariates (e.g., 1975 earnings or no high school), the probabilities fall in the middle, indicating notably different degrees of balance improvement. These quantities provide meaningful information to communicate in using my framework to falsify a research design. The probability of observed covariate balance improvement yields an intuitive measure that can be compared across observed covariates, across adjustment procedures, and even across other datasets and applications.

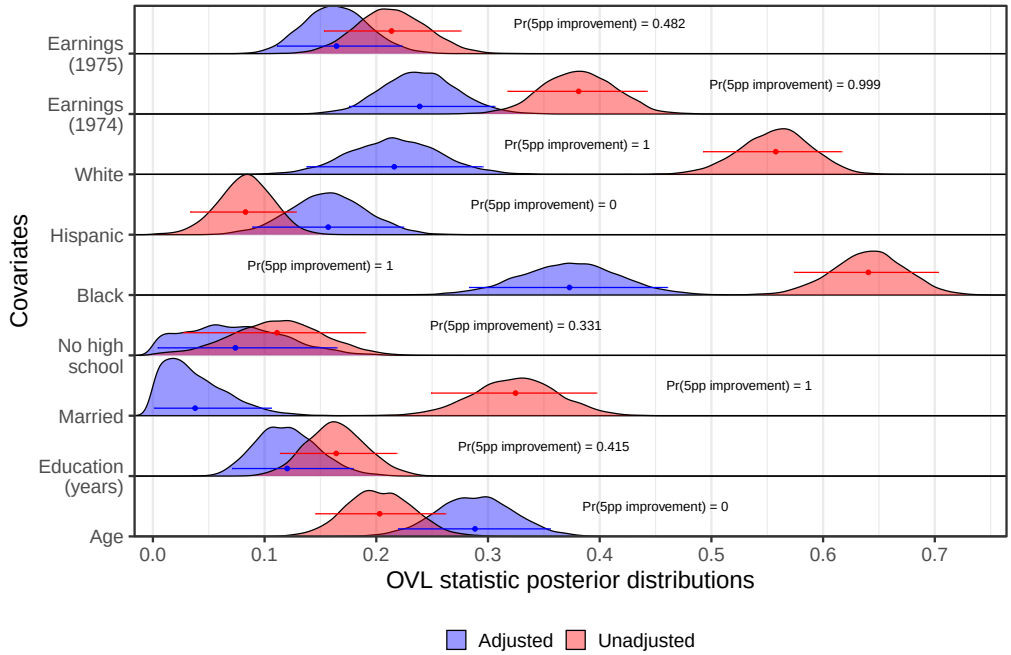
In sum, my proposed framework uses classical or (preferably) Bayesian bootstrapping to facilitate in-depth, rigorous evaluations of observed covariate balance in observational studies. Bootstrapping measures uncertainty in a statistic of interest, yielding the capacity to more precisely investigate whether the data are consistent with the assumptions of the identification strategy compared to common practice. Bootstrapping from a Bayesian perspective permits this inference to be framed in probabilistic terms, promoting intuition. Analysts can use the method to assess observed covariate balance with a variety of metrics

22. As with the ER, δ can be formed with guidance from independent experts (Fitzgerald, 2024) and should be preregistered (Harden et al., 2019).

23. A complete matching analysis would employ more than one balance metric and examine covariates other than those used for matching (Ho et al., 2007).

24. A common threshold for OVL statistics is $m = 0.20$ (Greifer, 2024). Thus, $\delta = 0.05$ (5pp) aligns with my suggested target improvement level of $\frac{m}{4}$.

Figure 5: Posterior Probability of OVL Balance Improvement After Matching Adjustment



Note: The graph presents OVL posterior distributions before (red) and after (blue) matching adjustment in the LaLonde (1986) data. Points denote posterior means and line segments denote 95% credible intervals.

that target multiple moments of the covariate distributions. They can also specify informative ranges of statistic values rather than test an unrealistic null hypothesis. Moreover, the framework is compatible with many types of observational designs, one of which I illustrate next.

6. A Natural Experiment on Voter ID Laws

I illustrate the diagnostic capacity of my balance assessment framework by applying it to a natural experiment on voter ID law implementation in the American states. In contrast to a feasibility check on an adjustment procedure, this balance assessment represents a more direct falsification test on a sample for which treatment exogeneity is plausible, but requires scrutiny (see Section 2.1). Campos et al. (2024) study the effects of state voter ID laws on ideological polarization between parties in state legislatures—a key predictor of dysfunction and policy gridlock. Their observational design must address selection into treatment because several factors systematically predict voter ID adoption, including party control of state government, ideological preferences, and racial resentment (Campos et al., 2024, 1480–1481). One element of their identification strategy involves subsetting their state-year panel data to only those state-years in which a voter ID law had previously been adopted. Many voter ID laws faced legal obstacles after adoption but before they could be implemented, which created variation in this subset with respect to when a voter ID law was in force. Campos et al. (2024) argue that, due to randomness in the timing of the

legal processes, this subsetting approach creates an exogenous source of treatment variation based on whether a law was implemented or not (see also Harden and Campos, 2023).

This application highlights the motivation for my falsification framework. The proposed identification strategy likely yields some initial skepticism from domain experts. As a natural experiment, this case requires extensive *investigation* to assess its credibility, including an examination of the history of voter ID laws and additional qualitative evidence as well as a quantitative assessment of the observable factors that may influence voter ID laws and the outcomes. It is notably different from a randomized experiment, in which the analyst can expect that observed (and unobserved) covariates are balanced. Here, the researchers must scrutinize their design as much as possible to understand the degree to which the data are consistent with their claim of treatment exogeneity.[25]

6.1 The Conventional Approach

I begin by employing common practice for balance assessment to evaluate the identification strategy using only the folk definition of covariate mean similarity. I evaluate observed covariate balance in (1) the full panel dataset of all states except Nebraska, spanning 2003–2018 ($n = 784$) and (2) the legal delay subset described above ($n = 194$; 147 treated cases). I assess 13 observed covariates contained in the authors' replication files, which measure several potential confounders of the treatment effects, as implied by the literature (Campos et al., 2024, 1485). In each dataset, I compute SMDs between treated and control state-years. I also report p-values from tests of the null hypothesis that the covariate means are equal.

Table 2 demonstrates clear imbalance in the observed full data according to conventional analysis with the folk definition. Among the 13 covariates, 11 display SMDs reaching statistical significance at the 0.05 level. In many cases these differences appear substantively large as well (e.g., governmental ideology). However, subsetting to isolate variation from legal delays noticeably mitigates the problem. Only two SMDs are distinguishable from zero—a plausible result that would occur about 11% of the time with 13 covariates purely due to random chance. One of the variables (racial resentment) is a core driver of support for voter ID laws in the states (Wilson and Brewer, 2013). But the other (term limits) is peripheral to the known process of selection into treatment and could have arisen due to chance.

Overall, according to the typical standards for balance assessment in observational studies, this analysis fails to reject the null of no difference for most of the observed covariates. However, failure to find imbalance is not the same as evidence consistent with the identification assumptions. A closer inspection suggests potential problems with this conclusion. Some SMDs in the subset data just barely miss the 0.05 threshold for statistical significance (e.g., governmental ideology and proportion nonwhite). Moreover, a few of the *nonsignificant* SMDs are also quite large in magnitude (e.g., neighbors adopting). Given the substantive context from the literature that selection into treatment is quite strong in this case, I argue that establishing the tenability of observed covariate balance requires a more extensive investigation.

25. This study is included in the meta-analysis reported in Table 1. Campos et al. (2024) report SMDs, OVL statistics, equivalence tests, and provide an extensive interpretation of their evaluation.

Table 2: Standardized Covariate Mean Differences in the Voter ID Natural Experiment

Sample	Covariate	μ_{Control}	$\mu_{\text{Treatment}}$	SMD	SMD SE	SMD p-value
Full data	Republican governor	−0.188	0.813	1.001	0.099	**0.000**
	Party competition	0.021	−0.091	−0.112	0.092	0.223
	Legislative party control	0.429	−1.857	−2.286	0.119	**0.000**
	Governmental ideology	0.305	−1.321	−1.626	0.096	**0.000**
	Citizen ideology	0.187	−0.809	−0.996	0.097	**0.000**
	Court strength	0.119	−0.516	−0.635	0.090	**0.000**
	Proportion nonwhite	0.024	−0.103	−0.127	0.099	0.200
	Racial resentment	0.039	−0.167	−0.205	0.092	**0.026**
	Term limits in effect	−0.054	0.232	0.286	0.089	**0.002**
	Legislative capacity	0.071	−0.308	−0.380	0.096	**0.000**
	Agenda size	0.054	−0.234	−0.287	0.093	**0.002**
	Sources adopting (%)	−0.149	0.644	0.792	0.089	**0.000**
	Neighbors adopting (%)	−0.035	0.152	0.187	0.092	**0.045**
Legal delay subset	Republican governor	−0.294	0.094	0.388	0.203	0.060
	Party competition	−0.141	0.045	0.187	0.133	0.164
	Legislative party control	0.126	−0.040	−0.166	0.218	0.450
	Governmental ideology	0.272	−0.087	−0.359	0.180	0.050
	Citizen ideology	0.203	−0.065	−0.267	0.181	0.144
	Court strength	−0.021	0.007	0.027	0.150	0.856
	Proportion nonwhite	0.257	−0.082	−0.339	0.171	0.051
	Racial resentment	0.446	−0.143	−0.589	0.166	**0.001**
	Term limits in effect	−0.375	0.120	0.495	0.131	**0.000**
	Legislative capacity	0.083	−0.027	−0.109	0.155	0.484
	Agenda size	0.173	−0.055	−0.229	0.176	0.199
	Sources adopting (%)	−0.086	0.028	0.114	0.174	0.513
	Neighbors adopting (%)	0.262	−0.084	−0.346	0.216	0.114

Note: Cell entries report covariate standardized means in the control and treatment groups, standardized mean differences (SMD), standard errors (SE), and p-values. The top panel summarizes the full data and the bottom panel summarizes the data after subsetting such that legal delays induce all treatment variation. Bolded p-values indicate statistical significance at the 0.05 level. See Campos et al. (2024) for covariate descriptions.

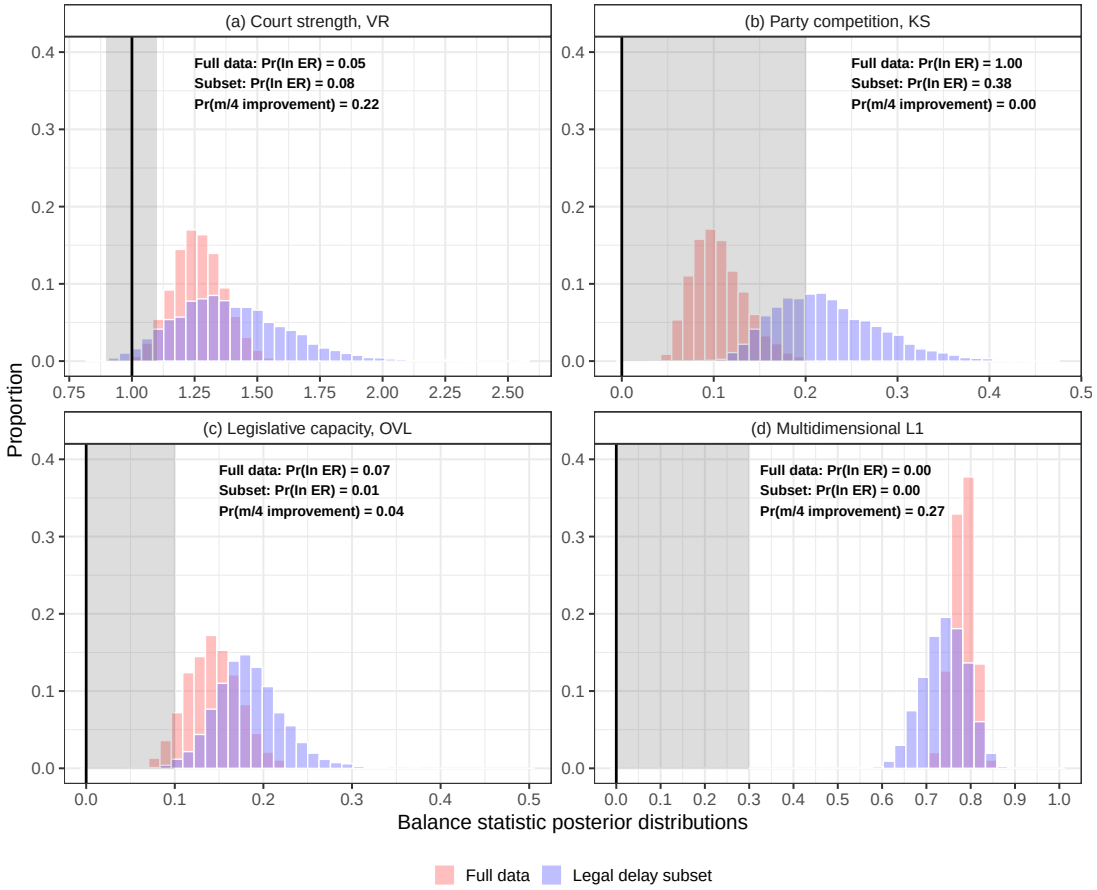
6.2 The Alternative Framework

My proposed Bayesian bootstrapping approach permits a more rigorous falsification test of this natural experiment. I use it to assess balance in the observed data with five statistics discussed previously: SMDs, VRs, KS statistics, OVL statistics, and Iacus et al.'s (2011) multidimensional L1 distance, which is an omnibus measure of balance.[26] For brevity, I focus here on covariates in which the conventional method indicates balance, but my framework suggests imbalance. See the Appendix for an expanded analysis of this application, including a substantive justification of the ERs.

Table 2 shows that, after subsetting, the absolute SMDs for court strength (0.027), party competition (0.187), and legislative capacity (0.109) do not reach statistical significance and fall below one or more of the common thresholds for balance used in the literature (e.g., 0.10

26. Specifically, the L1 metric measures overlap between the treated and control groups in a multidimensional histogram that includes coarsening the covariates under study into a reduced number of bins (Iacus et al., 2011).

Figure 6: Balance Statistic Posterior Distributions With and Without Subsetting to Isolate Legal Delays in the Voter ID Natural Experiment



Note: The graphs present balance statistic posterior distributions in the full data (red) and legal delay subset (blue) from Campos et al. (2024). The gray shaded areas denote ERs associated with $m = 0.10$ (panel a), $m = 0.20$ (panel b), $m = 0.10$ (panel c), and $m = 0.30$ (panel d).

or 0.25). Thus, an analyst employing the conventional method would likely not consider those covariates problematic. However, my framework offers a different perspective. Figure 6 presents the posterior distributions of other balance statistics and their associated ERs. The graphs show evidence of *imbalance* in the variables. The court strength VR (panel a), party competition KS statistic (panel b), and legislative capacity OVL statistic (panel c) generally become *worse* after subsetting, as shown by the low probabilities of balance and improvement probabilities. Additionally, the multidimensional L1 statistic (panel d) shows only marginal improvement after subsetting. In other words, several key variables fail the falsification test set up by my framework.

Consider the substantive context underlying the court strength results in panel (a). The bootstrapped VRs in the subset data (blue posterior) support the inference that the treatment group variance is meaningfully larger than the control variance. The posterior probability that the VR is greater than 1 is 98% and the probability that it is larger than

1.3 (i.e., a 30% larger treatment variance) is 0.64. These findings highlight the following substantive pattern in the data (Hall and Windett, 2015):

- The treatment group contains Southern states—where courts are historically weak—*and* states from other regions with traditionally stronger court systems;
- The control group is primarily composed of Southern states prior to voter ID implementation.

Recall that the *average* values of court strength are very similar across the two groups after subsetting (see Table 2). But substantive experts would likely view the above scenario as representing two qualitatively different groups. And this difference is potentially problematic because there is regional clustering in voter ID adoption (e.g., more common in the South and less common in the mid-Atlantic). The fact that the subsetting strategy sharpened this discrepancy—the VR posterior mean in panel (a) increases from 1.26 (unadjusted) to 1.40 (adjusted)—falsifies the claim of distributional equivalence that would support as-if randomness of the treatment.

This application (analyzed further in the Appendix) illustrates the value of conducting a rigorous and comprehensive evaluation of observed covariate balance as a falsification test for observational designs. The conventional approach of counting statistically significant mean differences signals substantial balance improvement from subsetting the observed data to post-voter ID adoption state-years (Table 2). An analyst using that method might simply proceed to estimation under the assumption that their identification strategy is valid. However, employing the Bayesian bootstrap on mean differences (see the Appendix) *and* other statistics that encode balance in higher moments (Figure 6) provides much more information with which to detect design violations. Some of this analysis is consistent with balance improvement in the observed data (see the Appendix). But the results shown in Figure 6 fail to rule out meaningful imbalances. This mixed evidence suggests that the subsetting strategy reduces observable bias, but also underscores the necessity of considering new covariates and additional covariate adjustment to guard against potential confounding.

7. Conclusions

Analyzing covariate balance in the observed data is one of several empirical options for scrutinizing causal identification strategies in social science. Reporting that several relevant variables are (essentially) equal across treatment and control groups provides empirical evidence consistent with assumed balance in the potential outcomes, which helps justify unbiased estimation of a causal effect. This element of identification naturally arises in experiments because randomization implies balance across observed and unobserved covariates in expectation. However, researchers studying observational data face a stricter standard. Without control over treatment assignment, known and unknown confounders are plausible and thus more qualitative and quantitative evidence is necessary to maintain the tenability of the design compared to an experiment. But in practice, my case study of one discipline shows that political scientists' balance evaluations do not typically meet this higher bar for evidence. Most published work *checks* for balance by assessing the statistical significance of observed covariate mean differences rather than conducting rigorous falsification tests of the full observed covariate distributions.

This paper proposes a novel solution to this problem by combining bootstrapping with a variety of balance statistics and principles from the literature on substantive significance. The falsification framework described here can greatly improve applied practice by providing tools for intuitively discussing observable balance across the full covariate distributions without relying on NHST. Moreover, by bootstrapping balance statistics from a Bayesian perspective, the method yields a posterior distribution that can be summarized with intuitive probabilistic statements. The result is an investigation of balance that yields more information to scrutinize the credibility of a design. In short, this new framework provides researchers studying observational data with the tools they need to rigorously audit observed covariate balance rather than cursorily check for the absence of imbalance.

How should social and behavioral scientists incorporate this novel framework into their own research? The method is inherently flexible to the needs of a particular study. But as general guidance, I suggest that researchers take the following actions to prepare for and execute an investigation of balance in their observational data.

- Select balance statistics on which to focus. I recommend SMDs, VRs, and OVLs as a set that provides a sufficiently comprehensive picture of the observed covariate distributions (i.e., the mean, variance, and distributional overlap).[27]
- Using guidance from the literature and/or heuristics discussed here, select values for τ, m, and δ to define the ER and minimum standard for balance improvement. These values could vary across covariates. If feasible, choose these values by surveying independent experts (Fitzgerald, 2024) and preregister them (Harden et al., 2019).
- Compute the posterior distributions of the selected statistics with and without the covariate adjustment method (if applicable).
- Summarize the posteriors graphically and compute the probability quantities defined above. Interpret these results with thorough discussion of the substantive context.
- If the results indicate imbalance on observables, consider additional covariates and/or conduct covariate adjustment methods to reduce the risk of potential confounding of the treatment effects. Transparently report all balance values with measures of uncertainty and acknowledge that results from observed covariates do not preclude unobserved confounders.

This falsification framework can help researchers structure balance analyses to provide transparent, thorough investigations of observed covariate similarity across treatment conditions. It yields a more intuitive and informative perspective toward balance that de-emphasizes p-values as misleading indicators of causal credibility. Critically, it yields uncertainty measures to account for sampling variability when analyzing balance statistics. Summarizing the posterior with principles from equivalence testing and substantive significance ensures that researchers are incentivized toward high-powered, information-rich designs. Overall, this method facilitates informative inferential statements about the tenability of a (conditional) ignorability assumption while remaining straightforward to implement in practice. I conclude that it holds the potential to greatly improve scrutiny of identification strategies in observational studies without placing a heavy burden on researchers.

27. The coverage rate simulation I present in the Appendix shows that bootstrapped KS statistics do not perform well relative to these other statistics.

Appendix

1 Meta-Analysis Details

This section provides further details on the meta-analysis of journal articles discussed in the main text (Table 1). In June 2024 I searched Google Scholar for articles containing the string "covariate balance" published since 2019 in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*. This process returned 20 (*APSR*), 23 (*AJPS*), and 31 (*JOP*) hits. From this list 12 articles appeared in other journals with similar names, which I omitted immediately. I downloaded the remaining 62 articles for use in the meta-analysis.

My next step was to read the articles to identify whether they reported covariate balance as I discuss here *for an observational study*. These articles report results from selection-on-observables designs, natural experiments, and regression discontinuity designs (RDD). Researchers also sometimes employed balance assessment to verify randomization in experiments. However, Mutz et al. (2019) recommend against this practice for a variety of reasons and so I do not analyze what they refer to as "clean experiments" here. This decision removed eight articles, producing a final sample of 54. Importantly, I left in the few articles that reported balance assessments on experimental data that the researchers conducted due to a suspected failure in the randomization process. Mutz et al. (2019) consider these cases to be experiments unintentionally (and unfortunately) transformed into observational studies. Under that logic, these studies are eligible for inclusion in my meta-analysis.

Finally, I coded the indicators noted in the main text by examining the articles and, when necessary, their supplementary materials available on the journal website.

- **Did the article report covariate means by group and/or differences in those means?** Across the sample there was some heterogeneity in how researchers conducted this step. Most were computed with t-tests, but a few were done with regression.
- **Did the article use mean difference p-values as evidence in favor of assumed balance in the observed covariates?** Here I looked for any formal tests of the statistical significance (or more specifically, its absence) of mean differences. Many studies simply reported p-values for each covariate. Some also counted the number of significant p-values and compared that count to the number that would be expected due to random chance, given the chosen significance level.
- **Did the article report results from equivalence tests?** I assessed whether the articles conducted equivalence tests to evaluate balance as described in Hartman and Hidalgo (2018).
- **Did the article report any other balance statistic?** I coded any other quantity (e.g., KS statistics) that could be used to assess balance beyond means or mean differences as a "yes" to this question.
- **Did the article offer any substantive interpretation of the balance results?** I searched for any discussion of the *substantive* element of balance evaluation, such as the meaning of the magnitude of difference in a covariate. I maintained a fairly

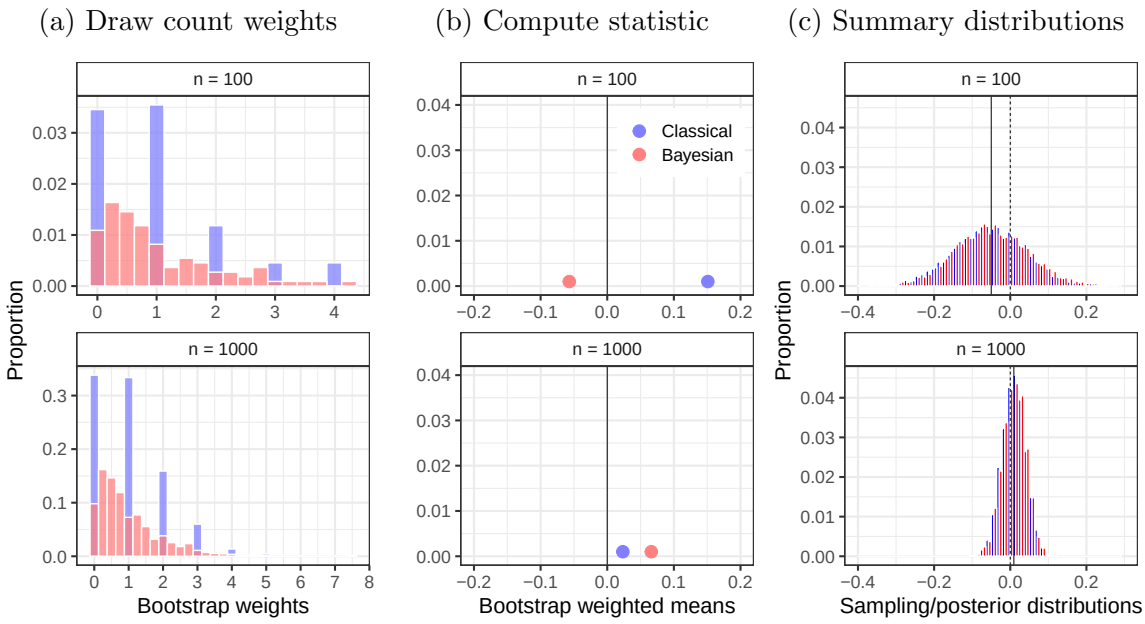
generous standard here by looking for any evidence beyond the (lack of) statistical significance of a mean difference or other metric.

The aggregated results of this set of coding procedures are reported in Table 1 of the main text.

2 An Illustration of Classical and Bayesian Bootstrapping

Figure 7 illustrates both versions of the bootstrapping process described in the main text at sample sizes of 100 and 1,000 observations. The observed variable, x, is drawn from a standard normal distribution and the inferential goal is to summarize uncertainty in the observed mean of x with a sampling (classical) or posterior (Bayesian) distribution. Panel (a) graphs example sets of "counts" generated for weight construction using classical (blue) and Bayesian (red) bootstrapping. Note that the classical counts are integers, indicating the number of times a given observation from the original data appears in the bootstrap sample. The values 0 and 1 are most common; many observations either never appear in that particular bootstrap sample or appear once. But some cases appear two or more times as well. In contrast, the "counts" implied by the Bayesian probability weights are continuous, more smoothly spanning the same range as the classical counts.[28]

Figure 7: Classical and Bayesian Bootstrapping of a Sample Mean via Weighting



Note: The graphs present the process of bootstrapping a sample mean via weighting from a classical (blue) and Bayesian (red) perspective at sample sizes of 100 (top row) and 1,000 (bottom row).

28. The Bayesian implementation never draws weight values of zero. This property captures the intuition that all of the observed data are possible and thus assigning a weight of zero to some cases may seem illogical (Courthard, 2022).

Panel (b) depicts the next step, which is computing the statistic of interest (here, the mean of x) by weighting the data with the bootstrap weights from panel (a). Those values appear as the points on the graph. Then, the final step repeats the computation in panel (b) many times with new sets of weights (here I use a total of 5,000 iterations). This step is the analog to repeatedly drawing new samples with replacement from the original data in the more familiar resampling-based description of bootstrapping.[29] Panel (c) summarizes the resulting distributions of bootstrap means. The solid and dashed vertical lines in those graphs, respectively, denote the sample mean and true population mean (zero).

3 Bootstrapping versus Randomization Inference

My framework employs classical or Bayesian bootstrapping (Rubin, 1981) to measure balance statistic uncertainty. However, bootstrapping is not the only approach that could be used to this end. It is also important to consider an alternative resampling method that is broadly similar (Carsey and Harden, 2014). Randomization inference (RI), which Hartman and Hidalgo (2018) discuss in the context of equivalence testing, is a potentially useful option that is commonly employed in the analysis of experiments and observational data. This approach repeatedly reshuffles the treatment variable to generate a "sharp null" distribution against which the quantity of interest (e.g., a balance statistic) in the original sample can be compared. It is useful to causal designs because it conditions on the data and reflects the uncertainty over which particular units were assigned treatment rather than uncertainty generated via sampling from a population (Athey and Imbens, 2017). In this case, an alternative approach for executing my framework could involve the use of test inversion to form a randomization confidence interval for the balance statistic of interest (Hartman and Hidalgo, 2018).

While acknowledging the value of the method, I argue that bootstrapping—and especially Bayesian bootstrapping—is nonetheless the preferable option for my approach to balance assessment. My reasoning is primarily practical, which I view as relevant to concerns about uptake by applied researchers (e.g., Table 1 in the main text). First, the Bayesian bootstrap is less computationally expensive. The process described in Section 4 of the main text is faster than even classical bootstrapping because all or some of it can be vectorized and/or parallelized (see Bååth, 2018; Courthard, 2022). In contrast, computing a randomization confidence interval is noticeably slower. It requires an iterative grid search for all non-zero null hypotheses that cannot be rejected. The analyst must increment the grid in small steps to ensure precise estimates and conduct the iterative randomization process at each step.

Second, moving beyond mean differences as a balance statistic is difficult with RI. Constructing a non-zero null hypothesis for a mean difference can be done by adjusting the covariate of interest: adding some chosen non-zero value to each observed value, then testing the sharp null of zero using this adjusted version of the covariate (e.g., Caughey et al., 2023). Thus, this test is performed on the scale of the variable under consideration (or a standardized version). However, this process would not necessarily be the correct ad-

29. Bootstrapping by resampling the data and bootstrapping by drawing observation-level weights as described here are equivalent. I emphasize the latter perspective because it intuitively connects classical and Bayesian bootstrapping.

justment for the null of a balance statistic that measures other moments of the covariate distribution (e.g., variance or density overlap). Conducting the test on a density function (such as OVL), for instance, would be difficult and dependent on the sample (see Hartman, 2021, 513). Put differently, RI is primarily used to conduct inference on a treatment effect. And while it can accommodate effects beyond the mean (Caughey et al., 2023), it is not well-suited to several statistics available for evaluating covariate balance. Bootstrapping can accommodate these statistics in the methodology I propose.

I also argue that the interpretation of uncertainty as arising from the sample is relevant when considering observed covariate balance because balance is inherently a property of the observed sample (Ho et al., 2007). This perspective best matches the logic of bootstrapping rather than RI. Furthermore, the main objective of my framework is to facilitate falsification of the assumption of balance in an observed covariate. Bootstrapping is superior for this goal compared to RI's focus on testing the sharp null of no effect. And my implementation of the Bayesian bootstrap does not employ strong assumptions for assessing balance and does not add *any* assumptions to the estimation of treatment effects. Indeed, an analyst could feasibly employ my framework to assess observed covariate balance, then use RI to test hypotheses about the treatment effect.

4 Expanded Analysis of the Voter ID Application

In this section I present additional balance assessment of the voter ID natural experiment described in the main text (Campos et al., 2024). I begin by demonstrating substantive reasoning for selecting values for decisionmaking associated with the balance statistics and ERs. Specifically, I base my chosen values around the following questions posed in the main text:

- What values of the balance statistic of interest would indicate perfect balance (τ)?
- Moving away from τ in a positive and/or negative direction, what value(s) reflect inflection points ($\pm m$) such that the treatment and control groups are substantively different enough to suggest a falsification of the identification assumptions?
- What value(s) of m does past literature and/or heuristics suggest are inflection points? Are these values sensible for the current application?
- Does the sample size, treatment assignment mechanism, nature of the covariate(s), or other substantive or empirical considerations make finding values outside a candidate range $[-m,\ m]$ more or less likely? If so, does the range need to be expanded or contracted to accommodate these expectations?
- Do imbalances revealed in observed covariates suggest any unobserved factors not yet considered that might confound the treatment effect?

The balance statistics I utilize and their τ values are as follows: SMD ($\tau = 0$), VR (1), KS (0), and OVL (0). The literature on voter ID laws makes clear that their adoption is a highly systematic process (e.g., Hicks et al., 2015; Campos et al., 2024). This work demonstrates that identifying causal effects of voter ID laws is fraught with empirical and methodological complications (see Erikson and Minnite, 2009; Grimmer et al., 2018). Thus,

as a baseline I contend that strict values for m are highly preferred because the threat of bias is likely high. Accordingly, past guidance from the literature on equivalence testing might suggest values such as 0.10 for the SMD statistics, indicating that a standardized mean difference within $[-0.10, 0.10]$ would be considered balanced. Strict values for the VR, KS, and OVL thresholds would include 0.10 or even 0.05 (Harder et al., 2010; Markoulidakis et al., 2023).

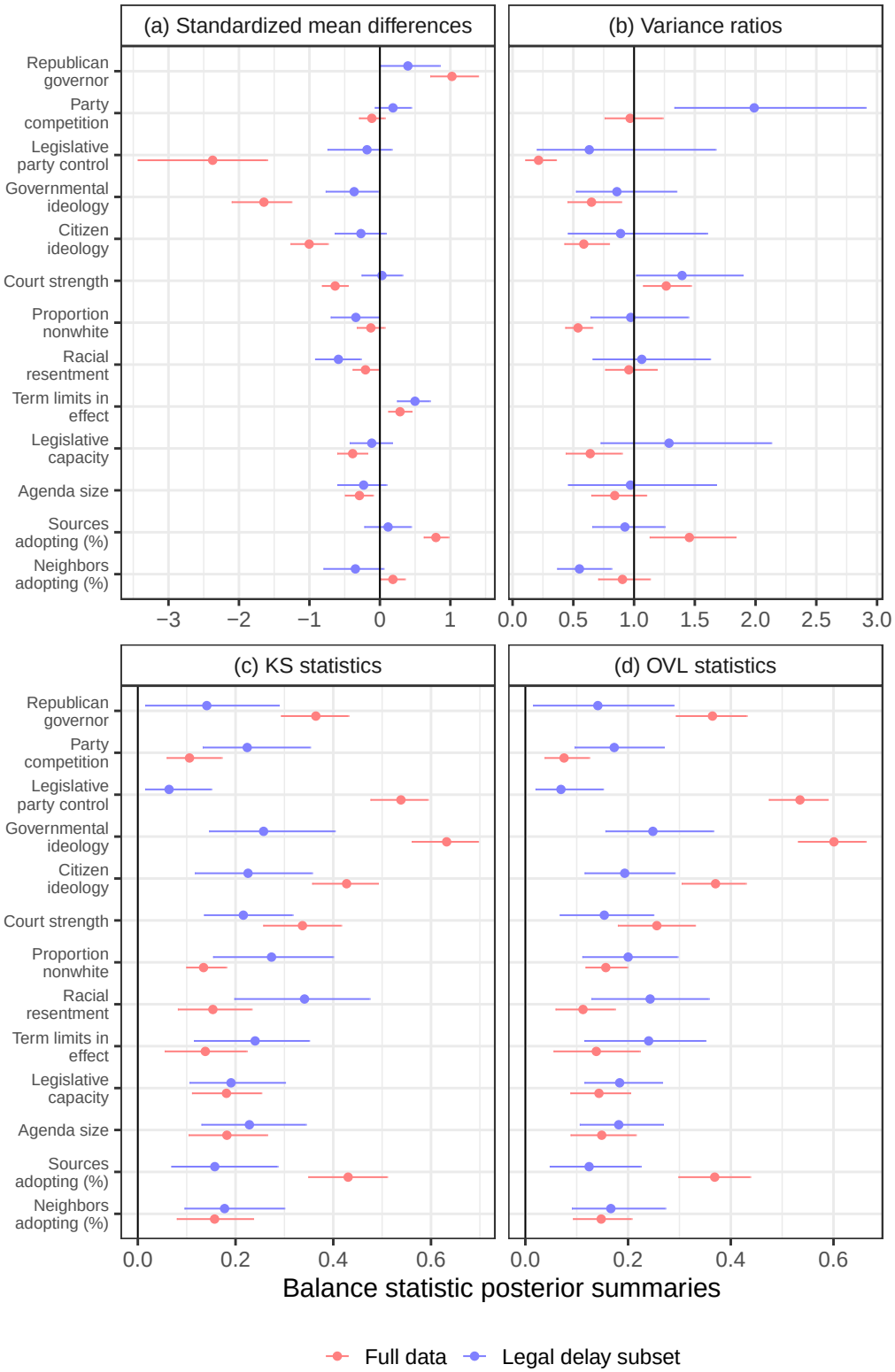
However, it is also important to note that the adjustment procedure—subsetting the data—decreases the sample size considerably ($n = 194$). This reduction in power may create challenges to producing evidence of balance. As such, I adjust the SMD and KS thresholds to $m = 0.25$ and 0.20, respectively (Ho et al., 2007, 220). This choice balances the need for rigorous scrutiny of the design against the realities of a small sample. In the case of SMD, this choice means that differences within one-fourth of a standard deviation of each variable are small enough to consider the two groups practically equivalent on that variable. A difference of that size is approximately equal to the observed difference (in 2018) between Georgia and South Dakota for the government ideology variable and Louisiana and North Carolina on proportion nonwhite (see below). I consider these comparisons substantively similar, justifying the choice of $m = 0.25$. For KS, 0.20 as a *maximum* distance separating the two groups on a given covariate still suggests a reasonably small discrepancy in my view.

I keep the values of 0.10 for VR and OVL mentioned above to ensure that some elements of my balance evaluation are done using a stricter standard, as justified above. There is limited guidance for determining a balance threshold for the multidimensional L1 statistics (Iacus et al., 2011). In that case, I select the value 0.30. Overall, this set of criteria acknowledges the need to be sensitive to possible confounders while addressing the relatively small sample size.

Full Results

Figure 8 presents the full balance statistic posterior summaries (posterior means and 95% credible intervals). Vertical solid lines in each graph indicate τ, the reference point denoting perfect balance for that statistic. Variance ratios (panel b) are omitted for binary variables (Republican governor and term limits in effect).

Figure 8: Balance Statistic Posterior Summaries in the Voter ID Natural Experiment (Campos et al., 2024)

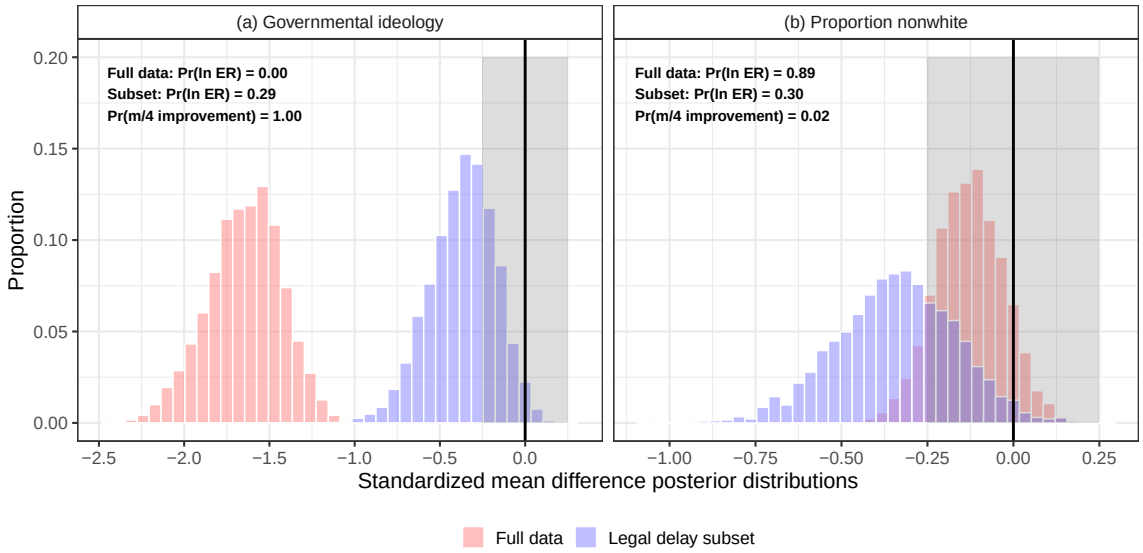


Note: The graphs present posterior means and 95% credible intervals of four Bayesian bootstrapped balance statistics in the voter ID natural experiment analyzed by Campos et al. (2024). The vertical solid lines in each graph indicate τ, the reference point denoting perfect balance for that statistic.

Detailed Results

I next examine select results depicted in Figure 8 in more detail. I begin with SMDs—the same statistic utilized with the conventional approach. Figure 9 graphs the posterior distributions of this quantity for two cases from Table 2 in the main text in which more investigation is warranted: governmental ideology and proportion nonwhite. Both variables are highly relevant to the substantive context of this natural experiment. The literature demonstrates that state government ideology and partisanship interact with state demographics to systematically predict voter ID support among elites (e.g., Hicks et al., 2015). They are also potentially associated with the study's outcome variable—party polarization in state legislatures—and thus are likely confounders of the treatment effect. The graphs show the bootstrapped SMD statistics in the full data and the legal delay subset. The shaded gray regions denote the ER of $[-0.25, 0.25]$, as discussed above.[30] Values falling inside the gray region are considered practically equivalent to zero (i.e., perfect average balance).

Figure 9: Standardized Mean Difference Posterior Distributions With and Without Subsetting to Isolate Legal Delays in the Voter ID Natural Experiment



Note: The graphs present standardized mean difference posterior distributions for governmental ideology (panel a) and proportion nonwhite (panel b) in the full data and legal delay subset from Campos et al. (2024). The gray shaded area denotes an ER associated with $m = 0.25$.

Figure 9 facilitates an in-depth examination of average balance in the covariates, including the effectiveness of the proposed source of treatment exogeneity. Panel (a) demonstrates that governmental ideology is highly imbalanced in the full data (red); the posterior mean is more than 1.5 standard deviations lower (more conservative) in the treatment group compared to the control and the tail of the distribution is below two standard deviations

30. As suggested in the main text, I use $\delta = \frac{m}{4}$ as the target improvement level for the the balance improvement probabilities.

of difference. None of the density falls above -0.25, indicating a posterior probability of observed covariate balance (as it is defined here) of zero. The adjusted distribution in blue is shifted much closer to zero; the probability of an improvement of $\frac{m}{4}$ after subsetting is one. However, the subset data posterior mean is still -0.37 and the posterior probability of observed covariate balance is just 0.29.

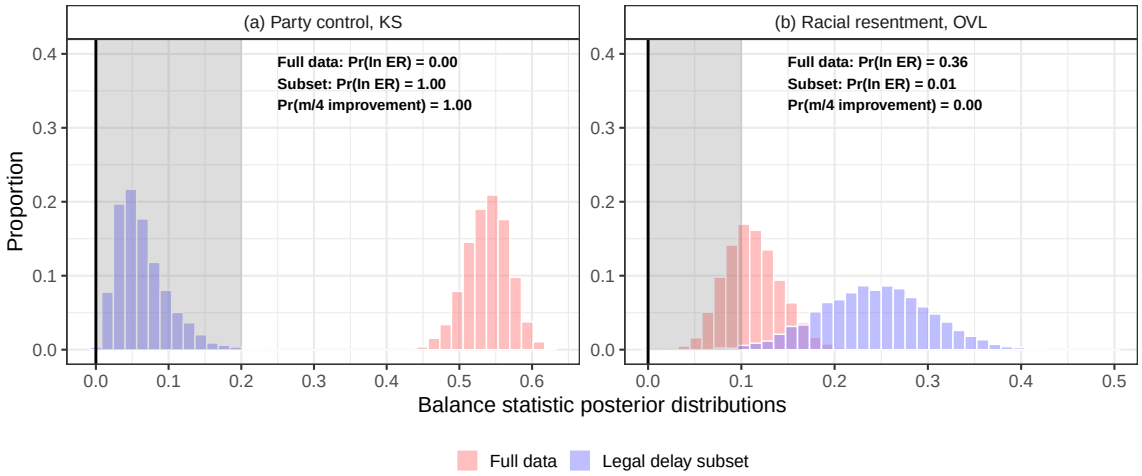
Panel (b) shows that the two posteriors are closer to each other and to zero for the proportion nonwhite SMD. Additionally, in this case the full data actually reflect better observable balance, on average, compared to the subset—the posterior probability of observed covariate balance is 0.89 in the former and 0.30 for the latter. The probability of a balance improvement of $\frac{m}{4}$ is nearly zero (0.02). Thus, the natural experiment actually leads to more imbalance in this case, although comparing more broadly reveals that it yields about the same absolute level of observable balance as seen in the adjusted balance of governmental ideology in panel (a). Overall, whereas the conventional approach in Table 2 of the main text shows that these covariates' mean differences by treatment group are not statistically distinguishable from zero after subsetting the data, my framework marshals additional information *from the same balance statistic* and reveals observed imbalance that the conventional approach fails to detect.

The utility of my approach is further highlighted by its straightforward extension to other balance measures, as shown in Figure 6 (main text) and Figure 10 below. Panels (a) and (b) plot, respectively, KS statistics of the covariate measuring legislative party control and OVL statistics for state-level racial resentment. The shading denotes ERs informed by the discussion above on each statistic: $m = 0.20$ (KS) and 0.10 (OVL). Using the conventional approach, the absolute value of the SMD for legislative party control in the subset data (Table 2) is 0.166, which meets the threshold of 0.25 for balance. The racial resentment absolute SMD is -0.589 with a p-value of 0.001. Thus, even the conventional approach to balance assessment deems this covariate as problematic.

The results demonstrate my framework's capacity to provide evidence consistent with the identification assumptions and evidence of design violations. They underscore why researchers should evaluate balance in higher moments of the covariate distributions. The KS statistics on legislative party control (panel a) show clear improvement after subsetting, including a $\delta = \frac{m}{4}$ improvement probability of 1. The full data are highly imbalanced while all of the posterior density in the subset data fall inside the ER of [0, 0.20]. The OVL statistics for racial resentment (panel b) indicate better balance in the *full data* (posterior mean of 0.11) compared to the subset (0.24). The posterior probability of observed covariate balance is higher for the full data as well (0.36 and 0.01) and the probability of improvement is zero. Thus, there is clear reason for concern that the identification strategy did not balance the observed data on this measure.

As noted in the main text, this analysis suggests that the subsetting strategy employed by Campos et al. (2024) is useful. However, additional covariate adjustment is needed to further guard against confounders. Indeed, Campos et al. (2024) employ this approach in their own analysis of these data. In addition to subsetting the data, they utilize Imai et al.'s (2023) PanelMatch estimator, which further improves balance on key covariates and estimates treatment effects in a difference-in-differences framework (see Campos et al., 2024, 1486–1487). They also provide a detailed discussion of the substantive context and summaries of the histories of each voter ID law covered in the data.

Figure 10: Additional Balance Statistic Posterior Distributions With and Without Subsetting to Isolate Legal Delays in the Voter ID Natural Experiment



Note: The graphs present balance statistic posterior distributions in the full data and legal delay subset from Campos et al. (2024). The gray shaded areas denote ERs associated with $m = 0.20$ (panel a) and $m = 0.10$ (panel b).

5 Coverage Rate Simulation

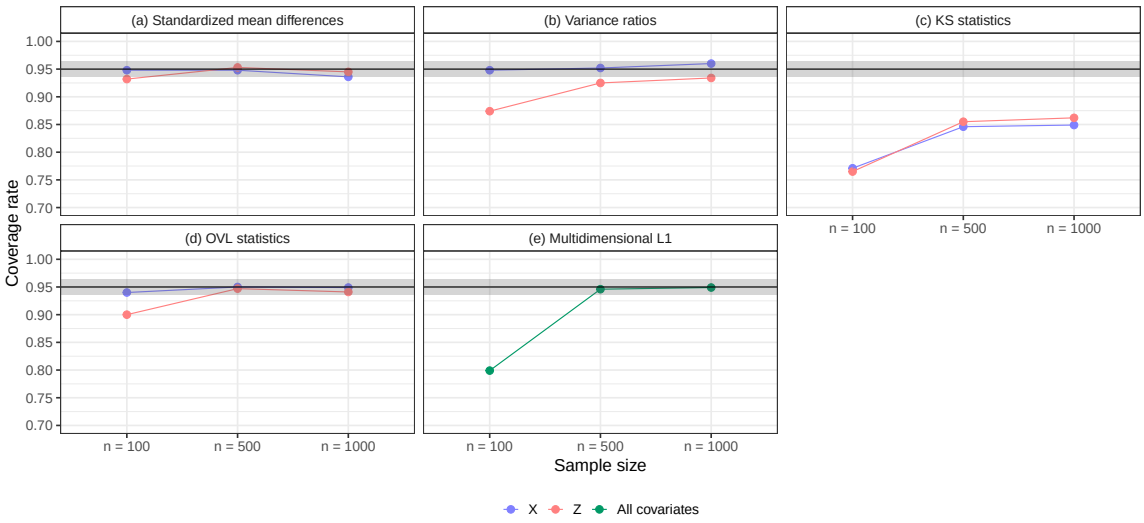
My proposed method relies on bootstrapping to measure uncertainty, which is known to perform poorly in finite samples for some statistics (Hesterberg, 2015). Accordingly, it is important to assess the performance of the uncertainty measures for the statistics discussed in this paper.[31] I do so here with a simulation of coverage rates (Carsey and Harden, 2014). This method relies on the frequentist definition of confidence intervals by computing the proportion of simulated data samples for which the true parameter value falls inside the interval. My method generates a Bayesian posterior distribution, which is somewhat at odds with this frequentist perspective (Gill, 1999, 2014). Nonetheless, assessing coverage with 95% credible intervals is still useful and informative as a means of checking that the computed posteriors accurately characterize uncertainty in the balance statistics.

To conduct the simulation I first generated a large population of data with four variables: an outcome Y, treatment D, and two covariates X and Z. I constructed both X and Z as correlated with D and Y (see the simulation code in Section 7). Thus, they are confounders that demonstrate imbalance in this population. I computed the "true" imbalance in the population on these data using five balance statistics that appear in the main text: SMD, VR, KS, OVL, and L1. I computed the first four for both covariates; L1 is an omnibus measure. Next, I randomly drew 1,000 samples each from this population using sample sizes of 100, 500, and 1,000 observations (a total of 3,000 samples). On each sample I computed the Bayesian bootstrap for each statistic and stored the 95% credible interval.[32] Then I

31. I focus on the statistics reported in the main text here. I also conducted an analogous simulation on the RDD statistics (see below) and found similar results (see the replication materials).
32. The results reported in the main text come from 5,000 posterior draws. In this case I specified 1,000 draws to save on computational expense. Results are similar with more draws.

computed, across the 1,000 samples for each sample size, the proportion of credible intervals that included the true parameter defined at the beginning. This proportion should be (close to) 0.95 to demonstrate proper coverage. Figure 11 graphs the results.

Figure 11: Coverage Rates of Bootstrapped Balance Statistics in Simulated Data



Note: The graphs present coverage rates for 95% credible intervals of the bootstrapped balance statistics. The solid black lines at 0.95 indicate correct coverage and the shading denotes simulation error bounds.

A common pattern shown here and in many other applications of bootstrapping is the increase in performance as the sample size increases. At sample sizes of 100 the coverage rates tend to fall below their nominal value, indicating undercoverage. For several statistics, the results improve moving to $n = 500$ or 1,000 (SMD, VR, OVL, and L1). However, the KS statistic coverage flattens out between 0.85 and 0.90. This result suggests that other options besides KS are a better choice for assessing balance with my method, as shown by the example in Section 6 of the main text. The OVL statistic, for instance, encodes similar information by measuring balance across the full covariate distribution. And Figure 11 demonstrates that its bootstrap coverage rate is correct or very near correct, depending on sample size. Overall, this simulation demonstrates that bootstrapping balance statistics is feasible because several commonly-used statistics can achieve correct coverage through the procedure.

6 Applying the Framework to Regression Discontinuity

Regression Discontinuity Designs (RDD) are widely used across the social sciences to identify causal effects in observational data (Cattaneo et al., 2020). Relative to many other observational strategies, RDD typically provide strong causal leverage. By comparing cases above and below a threshold (or cutoff) of a running variable, the method generally employs plausible assumptions and convincing control of unmeasured confounders.[33] However,

33. See Cattaneo et al. (2020) for a complete overview of the RDD methodology, including its strengths, weaknesses, and potential problems to address in its execution.

treatment assignment is deterministic in the RDD and so the researcher must still defend the assumptions of the design (Eggers et al., 2015). An array of placebo and falsification tests exist to detect violations of the RDD assumptions. However, many of these tests evaluate the null hypothesis that the design is valid. Hartman (2021) recently developed tools for structuring these methods to test a null of *inconsistency*. As in the equivalence testing framework more generally, Hartman's (2021) proposed methods require the analyst to establish the tenability of their design by failing to detect meaningful violations of its assumptions. The result is a more compelling evaluation of the identification strategy.

Following the logic of Hartman (2021), here I extend my proposed falsification framework to common tools for assessing the validity of a RDD. Specifically, I adapt my methodology to three common tests of the design (Cattaneo et al., 2020):

- RDD effects on pretreatment covariates;
- Density tests at the running variable cutoff;
- RDD effects on placebo cutoffs.

All of these tests yield statistics that can be evaluated effectively by bootstrapping to obtain posterior distributions and making probabilistic inferential statements with the results.[34] To illustrate this process, I use data from a RDD investigating the effect of Islamic political representation on women's education in Turkey (Meyersson, 2014; Cattaneo et al., 2020). This application employs data from 1994 mayoral elections in Turkey and specifies the margin of victory of an Islamic political party as the running variable. The outcome is the proportion of women aged 15-20 who completed high school by 2000. The data are coded such that zero on the running variable represents the cutoff—a (hypothetical) tie between an Islamic party and a non-Islamic party.

RDD Effects on Pretreatment Covariates

One test of RDD validity that is most similar to the concept of covariate balance is the estimation of treatment effects using covariates as placebo outcomes. The logic of this falsification test is that units above and below the cutoff should look similar on observed covariates; in other words, effects of zero represent the ideal result for a falsification test ($\tau = 0$). The analyst could simply test the null hypothesis that an effect is zero, but doing so would incentivize lower statistical power to "pass" the test and de-emphasize information-rich discussion of the effect (Hartman and Hidalgo, 2018). As an alternative, the analyst could bootstrap the RDD effect to form a posterior, then evaluate the probability of large, non-zero effects with the result.[35]

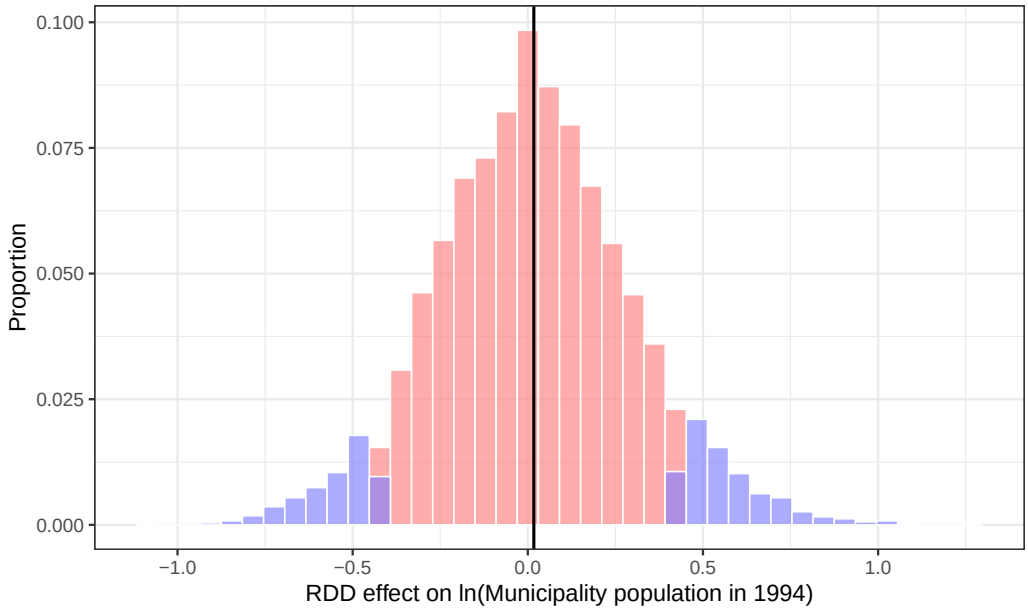
Figure 12 graphs the posterior distribution of the RDD effect on one pretreatment covariate from the data: the log of municipality population in 1994. The coloring denotes values of the posterior inside (red) and outside (blue) of an equivalence range (ER) discussed in the main text and recommended by Hartman and Hidalgo (2018, 1006): $m = \pm 0.36\sigma$, where σ is the pooled standard deviation of the covariate. By this definition, values in

34. I confirmed accurate finite sample performance of bootstrapping these statistics in a coverage rate simulation (see the replication materials).

35. Hartman (2021) provides a means of evaluating this effect via equivalence testing.

blue denote effects that are large enough to be meaningful, which weakens the design's credibility. The solid vertical line indicates the posterior mean.

Figure 12: Posterior Distribution of the RDD Effect of Islamic Political Representation on a Pretreatment Covariate (Meyersson, 2014)



Note: The graph presents the posterior distribution of the RDD effect of Islamic political representation on a pretreatment covariate (the log of municipality population in 1994). Red (blue) indicates posterior density inside (outside) the ER defined in the text. The solid vertical line denotes the posterior mean.

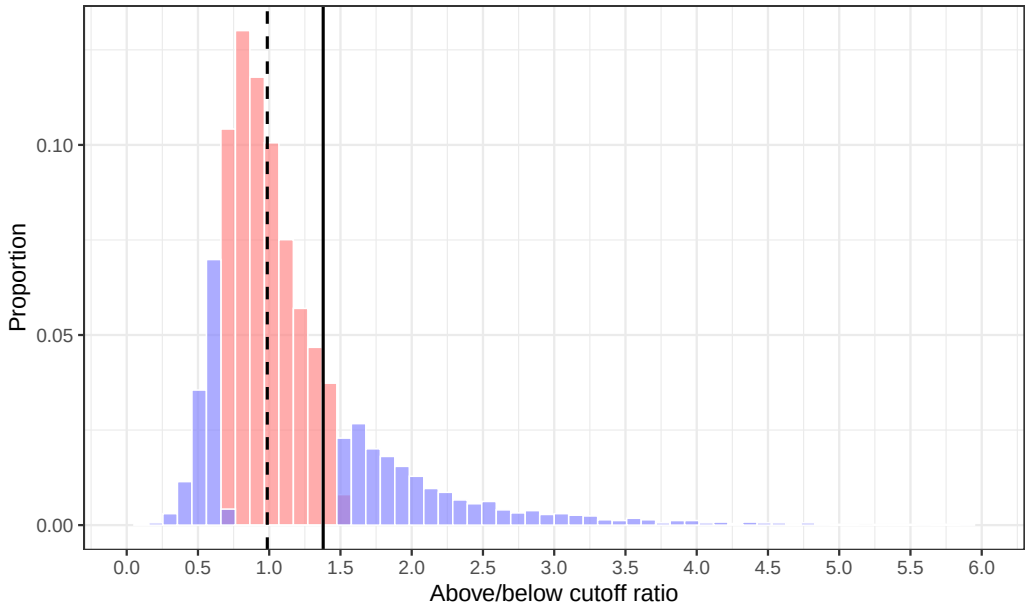
The graph shows that, on average, there is no meaningful effect of Islamic party representation on population. The posterior mean is 0.017 and its standard deviation is 0.29. However, the variability in that effect is also relevant to the evaluation. The ER as defined above is $[-0.43, 0.43]$, but the full density ranges from -1.08 to 1.27. Most notably, 87% of the posterior density falls inside the ER, reflecting a high probability that the RDD effect on population is substantively negligible according to the definition of equivalence I employ. That finding is understandable and intuitive. Moreover, it provides richer evidence with which an analyst could scrutinize the design's assumptions compared to testing the null hypothesis that the effect is zero.

Density Tests at the Cutoff

Another test of RDD validity is to evaluate the density of the running variable at the cutoff. A difference in the densities above and below that point suggests sorting around the threshold, which is problematic for the design because it creates discontinuities in the potential outcomes. Several tests exist in the literature for evaluating the null hypothesis of equality in density above and below the cutoff (see Cattaneo et al., 2020). Hartman (2021) extends the logic of these tests to an equivalence test by using the ratio of the two densities at the threshold as the statistic of interest. Under a valid design, the ratio should be one

($\tau = 1$). Values above (below) one indicate more density—and possibly sorting—above (below) the cutoff.

Figure 13: Posterior Distribution of the Density Ratio Above and Below the Cutoff (Meyersson, 2014)



Note: The graph presents the posterior distribution of the ratio of the running variable density above and below the cutoff in the Meyersson (2014) application. Red (blue) indicates posterior density inside (outside) the ER defined in the text. The solid and dashed lines denote the posterior mean and median, respectively.

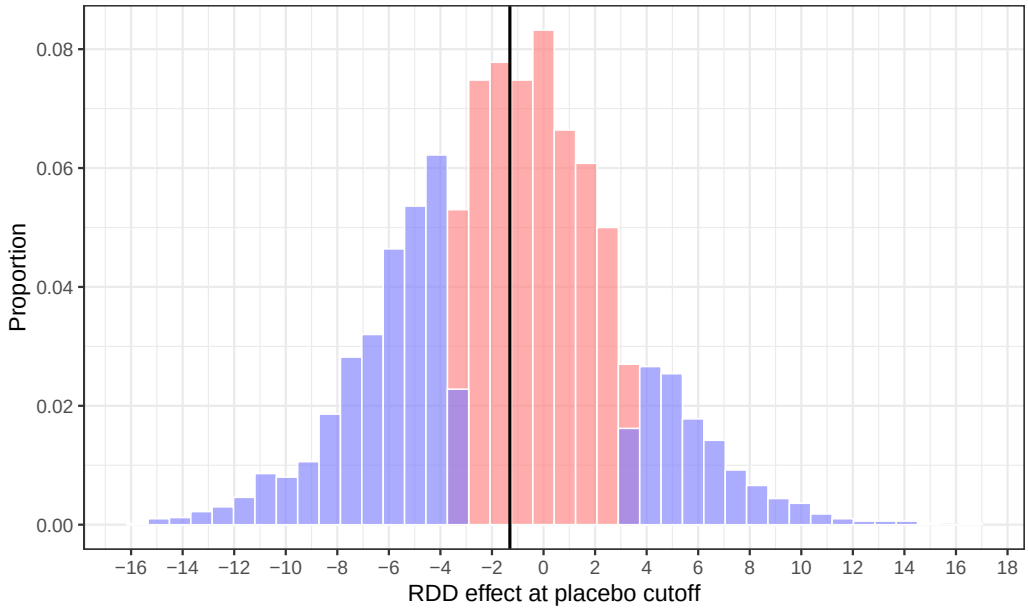
I adapt my framework to this test by bootstrapping to form the posterior distribution of this density ratio. Figure 13 graphs the results from the Meyersson (2014) data. The solid vertical line denotes the mean ratio and the dashed line indicates the posterior median. As before, red and blue denote values inside and outside an ER, respectively. I use Hartman's (2021) suggested ER for a density ratio of $\left[\frac{2}{3}, \frac{3}{2}\right]$ (515). Others, such as $\left[\frac{4}{5}, \frac{5}{4}\right]$ appear in the literature on bioequivalence (Berger and Hsu, 1996). As before, values falling inside the ER are deemed not practically different from the ideal ratio of $\frac{1}{1}$.

The graph shows a modal value near one, although the posterior evidences a long right tail of values in which the density above the cutoff is much larger than below. The median ratio, however, is very near one (0.99). About 68% of the posterior density falls within the ER—a substantially smaller amount compared to the first test. Put differently, the probability of a noteworthy imbalance in the densities on either side of the cutoff is about one in three. This finding indicates a possible design failure. Importantly, the density ratio is not statistically distinguishable from one (p = 0.17), so a conventional density test would not yield reason for concern. But investigating further with my methodology gives reason for more skepticism.

RDD Effects on Placebo Cutoffs

Finally, another test of RDD validity is the estimation of effects at placebo cutoffs. This falsification procedure tests for continuity (or lack thereof) in the regression function for treated and control units in the absence of treatment. This continuity *at the cutoff* is most important, but untestable. However, evidence of discontinuities at placebo cutoff values are still valuable because finding them would constitute falsification of the continuity assumption. The analyst can look for such discontinuities by estimating the RDD effect at other values of the running variable. Figure 14 presents this test using a placebo cutoff of one and estimating the effect with only treated cases (see Cattaneo et al., 2020, 89–90). The expected effect under a valid design is zero ($\tau = 0$), which would indicate continuity in the outcome variable comparing mayors who won by 1% or more to winners with margins less than 1%.

Figure 14: Posterior Distribution of the RDD Effect of Islamic Political Representation at a Placebo Cutoff (Meyersson, 2014)



Note: The graph presents the posterior distribution of the RDD effect of Islamic political representation on the outcome at a placebo cutoff value (1% instead of 0%). Red (blue) indicates posterior density inside (outside) the ER defined in the text. The solid line denotes the posterior mean.

The graph of the posterior distribution for the RDD effect with a placebo cutoff is slightly less than zero (posterior mean of -1.31, standard deviation of 4.29). Its ER—defined here by $m = \pm 0.36$ of an outcome standard deviation—is $[-3.45, 3.45]$. Thus, the estimate falls within the bounds and is not statistically distinguishable from zero at the 0.05 level. Nonetheless, these results still give some reason to suspect a possible discontinuity. The probability that the effect is within the ER is just 57%. When combined with the density test results, my framework provides more rigorous scrutiny and suggests reason for more skepticism toward the validity of this design compared to conventional methods.

7 R Functions

The code below demonstrates the Bayesian bootstrapping method described in the main text (Section 4). In brief, the code first creates a matching object (`m.out`) using the LaLonde (1986) data stored in the `cobalt` package. The objects `bt_args` and `bt_orig` create arguments and execute the `bal.tab()` function to compute balance on the original sample. The new function `bt_boot()` is then set to produce four balance statistics in a form that can be bootstrapped, as described in the main text (see Greifer, 2024):

- Standardized mean differences (SMD);
- Variance ratios (VR);
- Kolmogorov-Smirnov (KS) statistics;
- Complement of overlap (OVL) statistics.

Finally, the code uses the `bayesboot()` command from the `bayesboot` package (Bååth, 2018) to conduct classical bootstrapping (`use.weights = FALSE`) or Bayesian bootstrapping (`use.weights = TRUE`). Similar code can be used to compute balance statistic posteriors for quantities available in other packages, such as Iacus et al.'s (2011) L1 measure (see the replication materials).

```
### Bootstrapping balance statistics ###

## Packages ##
library(cobalt)
library(bayesboot)
library(tidyverse)
library(MatchIt)

## Matching example from cobalt documentation ##
options(scipen = 99)
data("lalonde", package = "cobalt")

m.out <- matchit(treat ~ age + educ + race + married +
                         nodegree + re74 + re75,
                 data = lalonde)

# Functions to bootstrap balance statistics #
bt_args <- list(x = m.out, un = TRUE, abs = FALSE,
                stats = c("mean.diffs", "variance.ratios",
                          "ks.statistics", "ovl.coefficients"))
bt_orig <- do.call(bal.tab, args = bt_args)

bt_boot <- function(dat, w, m.out, args){
  bt <- bal.tab(m.out, un = args$un, abs = args$abs,
                stats = args$stats, s.weights = w)

```

```
  result <- as.data.frame(cbind(t(bt$Balance[ , "Diff.Un"]), t(bt$Balance[ , "Diff.Adj"]),
                                t(bt$Balance[ , "V.Ratio.Un"]), t(bt$Balance[ , "V.Ratio.Adj"]),
                                t(bt$Balance[ , "KS.Un"]), t(bt$Balance[ , "KS.Adj"]),
                                t(bt$Balance[ , "OVL.Un"]), t(bt$Balance[ , "OVL.Adj"])))
  colnames(result) <- c(paste(rownames(bt$Balance), "Diff.Un", sep = "_"),
                        paste(rownames(bt$Balance), "Diff.Adj", sep = "_"),
                        paste(rownames(bt$Balance), "V.Ratio.Un", sep = "_"),
                        paste(rownames(bt$Balance), "V.Ratio.Adj", sep = "_"),
                        paste(rownames(bt$Balance), "KS.Un", sep = "_"),
                        paste(rownames(bt$Balance), "KS.Adj", sep = "_"),
                        paste(rownames(bt$Balance), "OVL.Un", sep = "_"),
                        paste(rownames(bt$Balance), "OVL.Adj", sep = "_"))
  result <- result %>% mutate_all(~ replace(., is.na(.), 0))
  return(result)
}

set.seed(13355)
iter <- 5000
boot.out <- bayesboot(data = lalonde, statistic = bt_boot,
                      R = iter, .progress = "text", use.weights = TRUE,
                      m.out = m.out, args = bt_args)
summary(boot.out)
```

8 Simulation R Code

Balance Example Simulation

The R code below conducts the balance example simulation from Section 3 of the main text.

```
library(cobalt)
library(tidyverse)
library(ggh4x)

### Simulate DGP ###
m <- 1000 # Iterations
N <- 500 # Sample size

bal_res <- matrix(NA, nrow = m, ncol = 4)
colnames(bal_res) <- c("mean.diffs", "variance.ratios",
                       "ks.statistics", "ovl.coefficients")

est <- matrix(NA, nrow = m, ncol = 3)
colnames(est) <- c("D", "DX", "DXZ")
tc_var <- matrix(NA, nrow = m, ncol = 2)
```

```r
colnames(tc_var) <- c("treatment", "control")
sig_count <- numeric(m)

# X = ideology
# E = ideological extremity
# Z = prior turnout
# Y = outcome

set.seed(555853)
for(i in 1:m){
  U <- rnorm(N)
  X <- sample(c(1:7), N, replace = TRUE,
              prob = c(.1, .15, .15, .2, .15, .15, .1))
  E <- abs(4 - X)
  Z <- rbinom(N, 1, prob = ifelse(E > 1, .85, .15))
  D <- rbinom(N, 1, prob = plogis(-0.5 + 3 * Z))
  Y_D_0 <- 0.4 * X + 0.3 * Z + U
  Y_D_1 <- Y_D_0 + 0.5
  Y <- ifelse(D == 1, Y_D_1, Y_D_0)

  bal_res[i, ] <- as.numeric(bal.tab(D ~ X, s.d.denom = "all",
                             stats = c("mean.diffs", "variance.ratios",
                                       "ks.statistics", "ovl.coefficients"))$Balance[2:5])

  est[i, 1] <- coef(lm(Y ~ D))[2]
  est[i, 2] <- coef(lm(Y ~ D + X))[2]
  est[i, 3] <- coef(lm(Y ~ D + X + Z))[2]

  tc_var[i, ] <- c(var(X[D == 1]), var(X[D == 0]))

  mm <- lm(D ~ X)
  sig_count[i] <- ifelse(abs(mm$coef[2]/sqrt(vcov(mm)[2, 2])) > 1.96, 1, 0)

  cat(i, " ")
  if (i %% 10 == 0) cat("\n")
}

mean(sig_count)
colMeans(est)
sqrt(mean((.5 - est[ , 1])^2))
sqrt(mean((.5 - est[ , 2])^2))
sqrt(mean((.5 - est[ , 3])^2))
```

Coverage Rate Simulation

The R code below conducts the coverage rate simulation described above.

```r
library(cobalt)
library(cem)
library(bayesboot)
library(tidyverse)

### Define true imbalance in the population ###
set.seed(7733)
N <- 1e6
U <- rnorm(N)
X <- runif(N)
Z <- rexp(N)
D <- rbinom(N, 1, prob = plogis(-0.5 + 3 * X - 2 * Z))
Y_D_0 <- 0.4 * X + 0.3 * Z + U
Y_D_1 <- Y_D_0 + 0.5
Y <- ifelse(D == 1, Y_D_1, Y_D_0)
dat <- as.data.frame(cbind(Y, D, X, Z))

true_st <- bal.tab(D ~ X + Z, data = dat, s.d.denom = "all", un = TRUE,
                   abs = FALSE, stats = c("mean.diffs", "variance.ratios",
                                          "ks.statistics", "ovl.coefficients"))$Balance

pop_l1 <- L1.meas(group = dat$D, data = dat, drop = c("D", "Y"))
brks <- pop_l1$breaks
brks$X <- quantile(brks$X, c(1/5, 2/5, 3/5, 4/5))
brks$Z <- quantile(brks$Z, c(1/5, 2/5, 3/5, 4/5))

true_l1 <- L1.meas(group = dat$D, data = dat,
                   breaks = brks, drop = c("D", "Y"))$L1

### Balance statistic functions ###
bt_boot <- function(dat, w, form){
  bt <- bal.tab(form, data = dat, s.d.denom = "all", un = FALSE,
                abs = FALSE, stats = c("mean.diffs", "variance.ratios",
                                       "ks.statistics", "ovl.coefficients"),
                s.weights = w)

  result <- as.data.frame(cbind(t(bt$Balance[ , "Diff.Un"]),
                                t(bt$Balance[ , "V.Ratio.Un"]),
                                t(bt$Balance[ , "KS.Un"]),
                                t(bt$Balance[ , "OVL.Un"])))
  colnames(result) <- c(paste(rownames(bt$Balance), "Diff.Un", sep = "_"),
                        paste(rownames(bt$Balance), "V.Ratio.Un", sep = "_"),
```

```
                        paste(rownames(bt$Balance), "KS.Un", sep = "_"),
                        paste(rownames(bt$Balance), "OVL.Un", sep = "_"))
  result <- result %>% mutate_all(~ replace(., is.na(.), 0))
  return(result)
}

l1_boot <- function(dat, w, brks, todrop){
  l1 <- L1.meas(group = dat$D, data = dat, breaks = brks,
                drop = todrop, weights = w)
  return(l1$L1)
}


### Simulation ###
# Sample from the population,
# Compute balance statistic posteriors
# Check coverage
m <- 1000
s <- 100
iter <- 1000
pe <- lo <- hi <- cover <- matrix(NA, nrow = m, ncol = 9)
cn <- c("X_Diff.Un", "Z_Diff.Un", "X_V.Ratio.Un", "Z_V.Ratio.Un",
        "X_KS.Un", "Z_KS.Un", "X_OVL.Un", "Z_OVL.Un", "L1")

colnames(pe) <- colnames(lo) <- colnames(hi) <- colnames(cover) <- cn


set.seed(2877779)
for(i in 1:m){
  samp <- dat %>% slice_sample(n = s)

  boot.out <- bayesboot(data = samp, statistic = bt_boot,
                        R = iter, .progress = "none", use.weights = TRUE,
                        form = as.formula(D ~ X + Z))

  l1.boot.out <- bayesboot(data = samp, statistic = l1_boot, brks = brks,
                           R = iter, .progress = "none", use.weights = TRUE,
                           todrop = c("D", "Y"))

  pe[i, ] <- c(apply(boot.out, 2, mean), mean(l1.boot.out$V1))
  lo[i, ] <- c(apply(boot.out, 2, quantile, prob = .025),
               quantile(l1.boot.out$V1, prob = .025))
  hi[i, ] <- c(apply(boot.out, 2, quantile, prob = .975),
               quantile(l1.boot.out$V1, prob = .975))

  cover[i, 1] <- true_st[1, 2] >= lo[i, 1] & true_st[1, 2] <= hi[i, 1]
```

```
  cover[i, 2] <- true_st[2, 2] >= lo[i, 2] & true_st[2, 2] <= hi[i, 2]
  cover[i, 3] <- true_st[1, 3] >= lo[i, 3] & true_st[1, 3] <= hi[i, 3]
  cover[i, 4] <- true_st[2, 3] >= lo[i, 4] & true_st[2, 3] <= hi[i, 4]
  cover[i, 5] <- true_st[1, 4] >= lo[i, 5] & true_st[1, 4] <= hi[i, 5]
  cover[i, 6] <- true_st[2, 4] >= lo[i, 6] & true_st[2, 4] <= hi[i, 6]
  cover[i, 7] <- true_st[1, 5] >= lo[i, 7] & true_st[1, 5] <= hi[i, 7]
  cover[i, 8] <- true_st[2, 5] >= lo[i, 8] & true_st[2, 5] <= hi[i, 8]
  cover[i, 9] <- true_l1  >= lo[i, 9] & true_l1 <= hi[i, 9]

  cat(i, " ")
  if (i %% 10 == 0) cat("\n")
}
```

References

Vincent Arel-Bundock, Noah Greifer, and Andrew Heiss. How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, 111(9):1–32, 2024.

Susan Athey and Guido W. Imbens. The econometrics of randomized experiments. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 73–140. North-Holland, Amsterdam, 2017.

Jonathan Ben-Menachem and Kevin T. Morris. Ticketing and turnout: The participatory consequences of low-level police contact. *American Political Science Review*, 117(3):822–834, 2023. doi: 10.1017/S0003055422001265.

Daniel J. Benjamin, James O. Berger, and et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Douglas G. Bonett. Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, 14(3):225–238, 2009.

Rasmus Bååth. bayesboot: An implementation of Rubin's (1981) Bayesian bootstrap. R package, version 0.2.2. `https://CRAN.R-project.org/package=bayesboot`, 2018.

Steven Brooke, Youssef Chouhoud, and Michael Hoffman. The friday effect: How communal religious practice heightens exclusionary attitudes. *British Journal of Political Science*, 53(1):122–139, 2023.

Alejandra Campos, Jeffrey J. Harden, and Austin Bussing. The legislative legacy of voter identification laws. *Journal of Politics*, 86(4):1479–1494, 2024.

Thomas M. Carsey and Jeffrey J. Harden. *Monte Carlo Simulation and Resampling Methods for Social Science*. Sage, Thousand Oaks, CA, 2014.

Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, New York, 2020.

Devin Caughey, Allan Dafoe, and Jason Seawright. Nonparametric combination (NPC): A framework for testing elaborate theories. *Journal of Politics*, 79(2):688–701, 2017.

Devin Caughey, Allan Dafoe, Xinran Li, and Luke Miratrix. Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. Working paper, Massachusetts Institute of Technology. https://arxiv.org/abs/2101.09195., 2023.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 82(1): 39–67, 12 2020.

Matteo Courthard. The Bayesian bootstrap. https://matteocourthoud.github.io/post/bayes_boot/. August 17, 2022.

Bruce C. Crauder, Benny Evans, Jerry A. Johnson, and Alan V. Noell. *Quantitative Literacy: Thinking Between the Lines*. W.H. Freeman and Company, New York, 3rd edition, 2018.

Thad Dunning. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, New York, 2012.

Andrew C. Eggers, Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Snyder Jr. On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, 59(1):259–274, 2015.

Andrew C. Eggers, Guadalupe Tuñón, and Allan Dafoe. Placebo tests for causal inference. *American Journal of Political Science*, 68(3):1106–1121, 2024.

Robert Erikson and Lorraine Minnite. Modeling problems in the voter identification-voter turnout debate. *Election Law Journal*, 8(2):85–101, 2009.

Jack Fitzgerald. Manipulation tests in regression discontinuity design: The need for equivalence testing. Institute for Replication (I4R) Discussion Paper Series. https://hdl.handle.net/10419/300277, 2024.

Jessica M. Franklin, Jeremy A. Rassen, Diana Ackermann, Dorothee B. Bartels, and Sebastian Schneeweiss. Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, 33(10):1685–1699, 2014.

Sebastian Galiani and Ernesto Schargrodsky. Property rights for the poor: Effects of land titling. *Journal of Public Economics*, 94(9):700–729, 2010.

Jeff Gill. The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3):647–674, 1999.

Jeff Gill. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC Press, Boca Raton, FL, 3rd edition, 2014.

Noah Greifer. cobalt: Covariate balance tables and plots. R package version 4.5.5. https://ngreifer.github.io/cobalt/, 2024.

Justin Grimmer, Eitan Hersh, Marc Meredith, Jonathan Mummolo, and Clayton Nall. Obstacles to estimating voter ID laws' effect on turnout. *Journal of Politics*, 80(3): 1045–1051, 2018.

Justin H. Gross. Testing what matters (if you must test at all): A context-driven approach to substantive and statistical significance. *American Journal of Political Science*, 59(3): 775–788, 2015.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Matthew E. K. Hall and Jason Harold Windett. Understanding judicial power: Divided government, institutional thickness, and high-court influence on state incarceration. *Journal of Law and Courts*, 3(1):167–191, 2015.

Ben B. Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

Jeffrey J. Harden and Alejandra Campos. Who benefits from voter identification laws? *Proceedings of the National Academy of Sciences*, 120(7):1–3, 2023.

Jeffrey J. Harden, Anand E. Sokhey, and Hannah Wilson. Replications in context: A framework for evaluating new methods in quantitative political science. *Political Analysis*, 27(1):119–125, 2019.

Valerie S. Harder, Elizabeth A. Stuart, and James C. Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3):234–249, 2010.

Erin Hartman. Equivalence testing for regression discontinuity designs. *Political Analysis*, 29(4):505–521, 2021.

Erin Hartman and F. Daniel Hidalgo. An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013, 2018.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015.

William D. Hicks, Seth C. McKee, Mitchell D. Sellers, and Daniel A. Smith. A principle or a strategy? Voter identification laws and partisan competition in the American states. *Political Research Quarterly*, 68(1):18–33, 2015.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

Ryan Hübert and Ryan Copus. Political appointments and outcomes in federal district courts. *Journal of Politics*, 84(2):908–922, 2022.

Stefano M. Iacus, Gary King, and Giuseppe Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106 (493):345–361, 2011.

Kosuke Imai, In Song Kim, and Erik H. Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.

Gary King, Christopher Lucas, and Richard A. Nielsen. The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2): 473–489, 2017.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

Jason Lyall. Does indiscriminate violence incite insurgent attacks? Evidence from Chechnya. *Journal of Conflict Resolution*, 53(3):331–362, 2009.

Todd Makse, Scott L. Minkoff, and Anand E. Sokhey. *Politics on Display: Yard Signs and the Politicization of Social Spaces*. Oxford University Press, New York, 2019.

Andreas Markoulidakis, Khadijeh Taiyari, Peter Holmans, Philip Pallmann, Monica Busse, Mark D. Godley, and Beth Ann Griffin. A tutorial comparing different covariate balancing methods with an application evaluating the causal effects of substance use treatment programs for adolescents. *Health Services & Outcomes Research Methodology*, 23(2):115–148, 2023.

Erik Meyersson. Islamic rule and the empowerment of the poor and pious. *Econometrica*, 82(1):229–269, 2014.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, New York, 2015.

Diana C. Mutz, Robin Pemantle, and Philip Pham. The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician*, 73(1):32–42, 2019.

Carlisle Rainey. Arguing for a negligible effect. *American Journal of Political Science*, 58 (4):1083–1091, 2014.

Erin Rossiter and Jeffrey J. Harden. The development of an issue public: Evidence from The Eras Tour. Forthcoming, *Journal of Politics*. https://doi.org/10.1086/737279, 2026.

Donald B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42 (7):1–52, 2011.

Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC Press, Boca Raton, FL, 2010.

David C. Wilson and Paul R. Brewer. The foundations of public opinion on voter ID laws: Political predispositions, racial resentment, and information effects. *Public Opinion Quarterly*, 77(4):962–984, 2013.

José R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.