Design-based, Bayesian Causal Inference for the Social-Sciences

Thomas Leavitt

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

© 2021

Thomas Leavitt

All Rights Reserved

Abstract

Design-based, Bayesian Causal Inference for the Social-Sciences

Thomas Leavitt

Scholars have recognized the benefits to science of Bayesian inference about the relative plausibility of competing hypotheses as opposed to, say, falsificationism in which one either rejects or fails to reject hypotheses in isolation. Yet inference about causal effects — at least as they are conceived in the potential outcomes framework (Neyman, 1923; Rubin, 1974; Holland, 1986) — has been tethered to falsificationism (Fisher, 1935; Neyman and Pearson, 1933) and difficult to integrate with Bayesian inference. One reason for this difficulty is that potential outcomes are fixed quantities that are *not* embedded in statistical models. Significance tests about causal hypotheses in either of the traditions traceable to Fisher (1935) or Neyman and Pearson (1933) conceive potential outcomes in this way; randomness in inferences about about causal effects stems entirely from a physical act of randomization, like flips of a coin or draws from an urn. Bayesian inferences, by contrast, typically depend on likelihood functions with model-based assumptions in which potential outcomes — to the extent that scholars invoke them — are conceived as outputs of a stochastic, data-generating model. In this dissertation, I develop Bayesian statistical inference for causal effects that incorporates the benefits of Bayesian scientific reasoning, but does not require probability models on potential outcomes that undermine the value of randomization as the "reasoned basis" for inference (Fisher, 1935, p. 14).

In the first paper, I derive a randomization-based likelihood function in which Bayesian infer-

ence of causal effects is justified by the experimental design. I formally show that, under weak conditions on a prior distribution, as the number of experimental subjects increases indefinitely, the resulting sequence of posterior distributions converges in probability to the true causal effect. This result, typically known as the Bernstein-von Mises theorem, has been derived in the context of parametric models. Yet randomized experiments are especially credible precisely because they do not require such assumptions. Proving this result in the context of randomized experiments enables scholars to quantify how much they learn from experiments without sacrificing the design-based properties that make inferences from experiments especially credible in the first place.

Having derived a randomization-based likelihood function in the first paper, the second paper turns to the calibration of a prior distribution for a target experiment based on past experimental results. In this paper, I show that usual methods for analyzing randomized experiments are equivalent to presuming that no prior knowledge exists, which inhibits knowledge accumulation from prior to future experiments. I therefore develop a methodology by which scholars can (1) turn results of past experiments into a prior distribution for a target experiment and (2) quantify the degree of learning in the target experiment after updating prior beliefs via a randomization-based likelihood function. I implement this methodology in an original audit experiment conducted in 2020 and show the amount of Bayesian learning that results relative to information from past experiments. Large Bayesian learning and statistical significance do not always coincide, and learning is greatest among theoretically important subgroups of legislators for which relatively less prior information exists. The accumulation of knowledge about these subgroups, specifically Black and Latino legislators, carries implications about the extent to which descriptive representation operates not only within, but also between minority groups.

In the third paper, I turn away from randomized experiments toward observational studies, specifically the Difference-in-Differences (DID) design. I show that DID's central assumption of parallel trends poses a neglected problem for causal inference: Counterfactual uncertainty, due to the inability to observe counterfactual outcomes, is hard to quantify since DID is based on parallel trends, not an as-if-randomized assumption. Hence, standard errors and p-values are too small

since they reflect only sampling uncertainty due to the inability to observe all units in a population. Recognizing this problem, scholars have recently attempted to develop inferential methods for DID under an as-if-randomized assumption. In this paper, I show that this approach is ill-suited for the most canonical DID designs and also requires conducting inference on an ill-defined estimand. I instead develop an empirical Bayes' procedure that is able to accommodate both sampling and counterfactual uncertainty under the DIDs core identification assumption. The overall method is straightforward to implement and I apply it to a study on the effect of terrorist attacks on electoral outcomes.

Table of Contents

List of Tables iv

List of Figures v

Acknowledgments vi

Dedication ix

Preface 1

Chapter 1: Bernstein-von Mises Theorem for Design-based Causal Inference 8

1.1 Introduction 8

1.2 Formal setup 9

1.3 Extant approaches to Bayesian causal inference 11

1.4 Asymptotic framework for design-based Bayesian causal inference 14

1.5 Bernstein-von Mises Theorem for Design-based Causal Inference 19

1.5.1 Weak causal hypotheses 20

1.5.2 Sharp causal hypotheses 23

1.5.3 Incorporating covariate information 27

1.5.4 Discussion and comparison of weak versus sharp causal hypotheses 29

1.6 Conclusion 34

Chapter 2: Turning Past Experiments into Priors for Design-based Bayesian Learning: Application to Audit Experiment on Racial Responsiveness 36

2.1 Introduction 36

2.2 Audit experiment 39

2.3 Constructing a prior distribution from past experimental results 43

2.3.1 Review of target validity methods 43

2.3.2 Setup 46

2.3.3 Complementing the CATE model with matching and weighting 50

2.3.4 Model of conditional average treatment effect (CATE) 55

2.4 Quantifying Bayesian learning from target experiment 57

2.5 Empirical analysis 61

2.5.1 Subgroup analysis 67

2.6 Conclusion 71

Chapter 3: Identification and Inference for Difference-in-Differences under Uncertainty in Parallel Trends 73

3.1 Introduction 73

3.2 Running Example and Formal Setup 77

3.3 Pathology of Causal Identification in DID 82

3.4 Generalized Nonparametric DID framework 85

3.5 Empirical Bayes' Identification and Inference under Uncertainty in Parallel Trends 87

3.6 Comparing and combining sampling versus counterfactual uncertainty 100

3.7 Discussion and Conclusion 105

References 128

Appendix A: Proofs 129

A.0.1 Proof of Lemma 1 129

A.0.2 Proof of Theorem 1 131

A.0.3 Proof of Theorem 2 133

A.0.4 Proof of Proposition 3 135

A.0.5 Proof of Proposition 4 136

A.0.6 Proof of Corollary 1 140

Appendix B: Simulation Results 142

List of Tables

2.1 Overall results of 2020 audit experiment 41

2.2 Maximum within-stratum absolute distance on heterogeneity score between prior and target units and maximum within-stratum ratio of target units to either treated or control prior units 64

2.3 Means and standard deviations of prior and posterior distributions of the average effect in the target experiment for each of the six contrasts. The Bayes' learning statistic is the standardized statistic given in Equation (2.16). 66

2.4 Subgroup results of audit experiment 67

2.5 Bayesian learning among Black and Latino legislators 69

2.6 Bayesian learning among GOP and Democrat legislators 71

3.1 Point estimates and standard errors under different inferential procedures in the Montalvo (2011) study 101

B.1 Simulation results 142

List of Figures

1.1 Distributions of likelihoods and posterior probabilities over repeated randomizations 22

1.2 Comparison of distribution of observable test-statistic and distribution implied by a false causal hypothesis conditional on realization of data 26

1.3 Mean Squared Errors (MSEs) of posterior distributions over repeated random assignments 32

2.1 Posterior distributions of treatment contrasts from Table 2.1 42

2.2 Heterogeneity score distributions (logit scale) between prior and target experiments 62

2.3 Heterogeneity score distributions (logit scale) after matching for all treatment contrasts 63

2.4 Prior and posterior distributions on the average effect in the target experiment for all treatment contrasts 65

2.5 Prior and posterior distributions of treatment contrasts by race of legislators 68

2.6 Prior and posterior distributions of treatment contrasts by party of legislators 70

3.1 Trends in mean Partido Popular (PP) vote shares in Spain's 52 provinces (data from Montalvo, 2011) 79

3.2 Linear time trends separately fit to pre-treatment data in treated and control groups (data from Montalvo, 2011) 98

3.3 Posterior distributions of ATT given different choices for calibrating distribution of Δ (data from Montalvo, 2011) 104

Acknowledgments

First and foremost, I would like to thank my parents and sister. Whether it was through watching my dad's reading some tome by Aristotle every morning on the subway while he went to work and I to school; my mom's vigorous dissection of the plot holes of a movie we had just watched; or my sister's impassioned arguments (for as long as I can remember) on the moral importance of vegetarianism, all of them laid the foundation for what eventually culminated in this dissertation. I feel lucky to have you all as my family.

I would also like to thank the members of my dissertation committee: Macartan Humphreys, Don Green, Jake Bowers and Naoki Egami. Perhaps what inspires me most about Macartan is his willingness to be vulnerable among his colleagues and students alike in pursuit of knowledge. I like to think that his willingness to be wrong (although I can't say I've encountered a situation in which he actually is) is why he seems to be perpetually learning. That more than anything is something I would like to take with me. I deeply admire Don for his optimism and especially for his incredible ability to think so clearly about complex topics. I often tell myself that if something doesn't make sense to Don, then it probably doesn't make sense at all. I feel so fortunate to have met Jake and learned from him for many years, well before beginning my Ph.D. Beyond his fundamental shaping of the way I think, I will appreciate him for always encouraging me to be the best version of myself. Finally, I met Naoki at the tail end of my graduate studies, right after I had decided to overcome my fear and switch to writing a methods dissertation. I don't know where I would be without Naoki's guidance and his unparalleled devotion to research and mentoring.

There are many others who I've been fortunate enough to learn from. Winston Lin walked me through so many details of causal inference and asymptotic statistics, which were genuinely fun to learn from him. Luke Miratrix (and all of the members of his C.A.R.E.S. lab), José Zubizarreta, and others provided maybe the most enriching year of my graduate studies during an exchange year at Harvard. Peter Aronow, Fredrik Sävje, Kimuli Kasara, John Marshall, Greg Wawro, Nadia Urbinati and others at Columbia, thanks so much for your guidance in big and small ways. To Iza Hussin, Bob Gooding-Williams, Lisa Wedeen, Jean and John Comaroff, John Brehm, Betsy Sinclair, Alberto Simpser, Harini Kumar, Diana Kim, and all of the other brilliant professors and students during my M.A. at U Chicago, you all gave me an intellectual awakening that changed my life. Thank you! To Birungi Solomon, Kabugho Loice, Masika, Muhindo and now Biira, as well as Manvule Peter and family, my first ever research was guided by you all and I'm grateful we're in each other's lives so many years later. Wasingya kutsibustibu.

Anna and Gosha, I consider you two my academic family and am so grateful to have begun the Ph.D. together and come out the other end 7 years later. My favorite grad school moments consisted of our arguing for hours in Nous about Bayes' and causal inference. I had always heard that, as students, we learn more from our peers than from anyone else, and with both of you, that couldn't be more true. Ben, Justin, Ira, Rick, Egor, Ash, Julia, Abhit, Jasper, all of the members of Macartan's advising group, and many others, I'm so happy to have shared our experience at Columbia together. To Antoine, Evan and Tim, we've known each other forever, and it makes me smile that we're still arguing about basketball 20 years later. That, just as much as anything, helped me see this dissertation to the finish line. To Consue, Ferdy, Ana Ceci y Rafa, experiencing your love is the best thing anyone could ask for. I'm so grateful for all the time we spent together in Puerto Rico (where I wrote most of this dissertation), your trips to visit us and all of the dinner conversations that would inevitably wind up on the question of Puerto Rican independence. You all mean the world to me.

And, most importantly, thank you to Viviana. There is no one I admire and who inspires me more. You are my home and my happy place. I can't think of any greater gift for Felipe than to

have you as his mother (y por supuesto Lila también). And to Felipe, your smiles bring me my happiest moments. Hugging you after defending my dissertation was the best moment any parent could ask for. I'm so excited to watch you grow up and to be lifelong learners together.

Dedication

For Felipe.

Introduction

At first glance, design-based and Bayesian methods for causal inference might appear to stand in tension: In the potential outcomes conception of causality (Neyman, 1923; Rubin, 1974; Holland, 1986), potential outcomes are fixed quantities that are *not* generated by an underlying probability model (Imbens and Wooldridge, 2009, p. 10). Randomness in inferences of causal effects stems solely from the physical act of randomization (Neyman, 1923; Fisher, 1935). The randomness in outcomes is inherited from the assignment process that randomly selects which of a units' multiple potential outcomes is observed. In studies without randomization, it still represents a methodological ideal whereby the quality of an observational study depends on the extent to which it is analogous to an ideal experiment (Cochran, 1965). Bayesian methods, by contrast, typically depend on likelihood functions that impose strong modeling assumptions on potential outcomes. In Bayesian causal inference, as developed in Rubin (1978) and Imbens and Rubin (2015), a stochastic, data-generating model for potential outcomes is used to stochastically predict unobserved potential outcomes. While the value of randomization in Bayesian inference has been subject to perennial debate, randomization's value in this framework stems from enabling researchers to "ignore" the assignment mechanism when using a potential outcomes model to predict counterfactual quantities (Rubin, 1976; Rubin, 1978). In light of these differing (but not incompatible) statistic traditions, what good comes out of the development of new methods in design-based, Bayesian causal inference? And how can they improve applied practice in political science and the social sciences more broadly?

Bayesian inference, at its core, consists of two central pillars. The first pillar is the prior distri-

bution. The prior distribution consists of a researcher's subjective beliefs, represented as a probability measure, about the plausibility of competing hypotheses. The second pillar is the likelihood function, which is a formal quantitative rule for updating subjective beliefs about competing hypotheses upon observing evidence. In developing Bayesian inference of causal effects, my aim is pragmatic; it is not to ground Bayesian inference in a priori, rational principles (see Ramsey, 1929; Ramsey, 1931; Savage, 1954, and related "Dutch book" arguments). Bayesian inference, I will argue, has several benefits for social-scientists that improve upon existing design-based methods of causal inference.

First, in contrast to null hypothesis significance testing (NHST), Bayesianism permits more nuanced conclusions from data. NHST is a coarse form of inference wherein a researcher picks hypotheses to test and then either rejects or fails to reject them. Among the hypotheses that a researcher fails to reject, are all equally plausible? Or, when combined with estimation, is the point estimate the most plausible value of the parameter with plausibility monotonically decreasing in distance from this estimate? The paradigm of NHST does not offer definitive answers to these questions. Bayesianism, by contrast, permits a continuous measure of the relative plausibility of competing hypotheses, not simply a binary (reject or fail to reject) claim about each hypothesis.

Second, social scientists are increasingly interested in how knowledge cumulates. Bayesian inference enables scholars to encode prior information about causal quantities in a target population and, subject to certain constraints on the prior distribution, to change beliefs in light of new evidence. In short, researchers can do more than evaluate whether a new finding, viewed in isolation, is statistically significant or not. Incorporating prior information enables scholars to assess how much a finding contributes to learning relative to what they already know going into the experiment.

Third, existing design-based methods fail to explicitly quantify the plausibility of competing identification assumptions. For example, the sensitivity analyses of Cornfield et al. (1959) played a crucial role in convincing the research community that smoking causes lung cancer. These sensitivity analyses showed that an extremely severe violation of a design's identification assumption

would have to be present in order to alter the substantive conclusion that smoking causes lung cancer. But absent the belief that such a severe violation is implausible, such a sensitivity analysis has no bearing on the causal relationship between smoking and lung cancer. A sensitivity analysis tells us how our conclusions would change if a violation of a key assumption *were* to exist, but cannot tell us whether such a violation *does* exist. Hence, drawing inferences about causal effects from sensitivity analyses, which scholars routinely (albeit informally) do, implicitly invokes Bayesian reasoning. Bayesianism therefore gives us a quantitative apparatus that we can use to formalize this reasoning.

In this dissertation, I show how design-based causal inference can be augmented by Bayesian methods, but without sacrificing the valuable properties that have drawn political scientists to experiments and design-based inference more broadly. In this introduction, I first seek to put randomization — a foundation of design-based inference — on firm footing in the context of Bayesian inference about causal effects. Doing so necessarily engages with arguments in philosophy of science that claim randomization possesses no special status for Bayesian inference. After hopefully demonstrating the valuable — if not essential — role randomization plays in Bayesian inference, I then summarize the three chapters of this dissertation. The first two chapters provide a treatment of the two pillars of Bayesian inference — the likelihood and the prior — in randomized experiments. The third and final chapter turns to observational studies and develops empirical Bayesian inference for the Difference-in-Differences design.

The role of randomization in design-based inference

The case for random assignment, a pillar of design-based methods, stands on shaky ground in the context of Bayesian inference. Indeed, there exists, as Berry and Kadane (1997, p. 813) state, "a standard result that Bayesians need not randomize." This statement refers to the result that, in the context of a single agent who wants to infer the value of a fixed, but unknown, causal parameter, the deterministic selection of an optimal assignment dominates random assignment. For

this reason, a range of scholars argues that a Bayesian agent ought to deterministically (rather than randomly) select an optimal assignment (e.g., Kasy, 2016; Bertsimas, Johnson, and Kallus, 2015; Kallus, 2018; Gibson, Caldeira, and Spence, 2002; Letham et al., 2019), which is an argument that builds on a long research tradition on the value of optimum assignment (Kiefer, 1959; Fedorov, 1972; Harville, 1975), but is especially relevant today in light of advances in computing power and the increasing use of machine-learning methods to tackle causal problems.

In light of this standard result that Bayesians do not need to randomize, Bayesian arguments for or against randomization differentiate between "experiments to learn" and "experiments to prove" (Kadane and Seidenfeld, 1990). The former pertains to a nonstrategic setting in which the researcher alone is the only actor who seeks to learn (i.e., revise prior beliefs) about a fixed causal parameter. The latter pertains to a setting with strategic interactions between a researcher, on the one hand, and either "nature" (e.g., Wu, 1981) or an adversarial audience (e.g., Banerjee, Chassang, and Snowberg, 2017; Banerjee et al., 2020), on the other. Bayesian arguments for randomization state that it can be justified in "experiments to prove," but *not* in "experiments to learn" (Savage, 1954; Savage, 1962a; Savage, 1962b; Stone, 1969; Basu, 1980; Lindley, 1982; Suppes, 1982).

In short, randomization can be justified within a Bayesian framework by changing the setting from a nonstrategic one with a single agent who wants to infer a fixed causal target to a strategic setting with a Bayesian agent pitted against either "nature" or an adversarial audience. Yet a Bayesian argument for randomization need not be based on *rational* grounds. It can be based instead on *epistemic* grounds. For some scholars, the rational basis for optimum assignment implies for it an epistemic justification (Urbach, 1985; Urbach, 1993; Howson and Urbach, 2006; Lindley and Novick, 1981). By contrast, I argue that a single Bayesian agent whose sole aim is to learn about the true value of a fixed causal parameter ought to reasonably (albeit irrationally) choose to randomize.

In this framework wherein Bayesian inference is justified due to its practical benefits, *not* axioms of rational decision theory, randomization is valuable because it severs the relationship between how the data are generated and a researcher's subjective beliefs. That is, under random

assignment, data are stochastically generated by a known, physical mechanism (e.g., independent coin flips or draws from an urn), *not* by a researcher's belief about which assignment is optimal. Randomization's severing of the relationship between the data generating process and subjective beliefs, implies that a Bayesian agent revises one's prior (subjective) beliefs *toward* the true value of the causal parameter. The first chapter of this dissertation aims to formally establish this point.

Outline of dissertation

In the context of Bayesian inference, epistemic justifications for random assignment typically refer to the *Bernstein-von Mises theorem*. A key condition for this theorem is Cromwell's rule, so named in Lindley (1971), which states that a prior probability of either 1 or 0 ought to be assigned to hypotheses that are either logically true or false.[1] Subjective probability distributions that satisfy Cromwell's rule ensure that one's prior beliefs can indeed change upon observing new evidence. Under this assumption, the Bernstein-von Mises theorem implies that a Bayesian agent's posterior distribution will converge in probability to the true value of the target parameter.

This theoretical result, however, has been derived in the context of parametric models and an assumed sampling process from an infinite superpopulation (Vaart, 1998, Section 10.2). Yet randomized experiments are especially credible precisely because they do not require such assumptions. Therefore, in the first chapter of this dissertation, I derive a randomization-based likelihood function in which Bayesian inference of causal effects is justified by the experimental design. I then show that, so long as the prior distribution obeys Cromwell's rule (i.e., does not assign prior probability 1 or 0 to a hypothesis that could logically be true or false), as the number of experimental subjects increases indefinitely, the resulting sequence of posterior distributions converges in probability to the true causal effect. In practical terms, if a researcher conducts a sufficiently large experiment, the posterior distribution will concentrate closely around the true effect regard-

[1] In explaining the rationale for Cromwell's rule, Lindley (1971, p. 110) writes, "if a decision-maker thinks something cannot be true and interprets this to mean it has zero probability, he will never be influenced by any data, which is surely absurd. So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved."

less of of the idiosyncrasies of an individual's prior distribution. Establishing this Bernstein-von Mises result enables scholars to quantify how much they learn from experiments without sacrificing the design-based properties that make inferences from experiments especially credible in the first place.

Chapter 1 establishes the value of randomization for Bayesian inference by showing that, under mild conditions on a prior distribution, belief revision will be driven by reliable data, not idiosyncratic prior beliefs. But if one's aim is to avoid undue influence of researchers' prior distributions, then why incorporate prior beliefs at all? Chapter 1 focuses on the likelihood function and its role in ensuring that the *revision* of beliefs satisfies a specific epistemic property. Chapters 2 and 3 provide two answers to the question of why scholars ought to incorporate prior beliefs in both an experimental and observational context.

In chapter 2, I turn my attention to the second pillar of Bayesian inference, the prior distribution. I show that usual methods for the analysis of experiments are equivalent to assuming an absence of any prior information. This implicit assumption inhibits knowledge accumulation from prior to future experiments, as well as assessments of how much one has learned from a new experiment relative to baseline knowledge. I therefore develop a methodology that enables scholars to (1) turn results of past experiments into a prior distribution for a target experiment and (2) quantify the degree of learning in the target experiment after updating prior beliefs via a design-based likelihood, the properties of which are established in chapter 2. I implement this methodology via an original audit experiment I conducted in 2020 on racial responsiveness among state legislators. I draw on data from prior audit experiments by Butler and Broockman (2011), Butler (2014), and Butler and Crabtree (2017) to construct a prior distribution for this experiment and then show the extent of Bayesian learning that results. Bayesian learning is greatest among subgroups, specifically Black and Latino legislators, which in turn has implications about the extent to which descriptive representation exists not only within, but also between minority groups.

The third and final chapter shifts attention to observational studies, specifically the Difference-in-Differences (DID) design. In this chapter, I focus specifically on inference, not estimation. I

show that standard errors in DID designs statistically account for only sampling uncertainty, due to the inability to sample all units from a target population, *not* counterfactual uncertainty, due to the inability to observe counterfactual potential outcomes among whichever units one samples. Existing methods of uncertainty quantification therefore yield incorrect and potentially misleading standard errors in DID applications.

Recent attempts to resolve this problem statistically represent counterfactual uncertainty via an assumed random assignment mechanism. Yet a key feature of the DID design is that its causal conclusions are based not on assumptions about an assignment mechanism, but rather about average changes in counterfactual outcomes over time (parallel trends). I show further that such design-based methods predicated on an as-if-randomized assumption are (1) often unable to generate meaningful inferences in the most canonical DID applications and (2) require interest in ill-defined causal estimands. In contrast to these methods, I decompose the causal estimand of interest into two parameters, one of which is characterized by sampling uncertainty and the other by counterfactual uncertainty. I then develop an empirical Bayes' procedure that is able to statistically represent both sampling and counterfactual uncertainty under the DID's core identification assumption of parallel trends.

Chapter 1: Bernstein-von Mises Theorem for Design-based Causal Inference

1.1 Introduction

Causal inferences from randomized experiments are especially credible — and indeed have spawned a so-called "credibility revolution" (Angrist and Pischke, 2010) — since their validity depends on only the integrity of the data collection and the adherence of statistical analysis to the stated design. Importantly, causal inferences from randomized experiments do *not* depend on probability models for the response variable, an assumed sampling process from an often vaguely defined superpopulation or other (often tenuous) assumptions (Berk and Freedman, 2003). These properties of design-based estimators and tests, ensured by random assignment, make experiments ideally suited to either estimate or test hypotheses about causal effects.

Methods of Bayesian causal inference, by contrast, appear to come at the expense of randomization as the "reasoned basis" for inference (Fisher, 1935, p. 14). Bayesian causal inference typically invokes probability models on potential outcomes and/or assumes random sampling from an infinite superpopulation (Rubin, 1978; Imbens and Rubin, 1997; Imbens and Rubin, 2015; Zhang, Rubin, and Mealli, 2009). Randomization serves as the basis for Bayesian causal inference only for specific cases in which binary or ordinal outcomes lend additional structure to potential outcomes (Ding and Miratrix, 2019; Chiba, 2018; Keele and Quinn, 2017; Humphreys and Jacobs, 2015).

The contribution of this paper is to develop Bayesian inference that is justified by the experimental design. The development of such inference has been difficult because, absent a probability model of potential outcomes, a likelihood function based on only the assignment mechanism will be unidentified for causal effects of interest— i.e., flat over multiple hypothetical values of causal effects. Hence, Bayesian belief revision about causal effects is *not* generally possible. To circum-

vent this problem, I derive a randomization-based likelihood function that conditions *not* on the full realized data, but a suitably defined function (test-statistic) of them. I then prove that, under weak conditions on a prior distribution, as the number of experimental subjects increases indefinitely, the resulting sequence of posterior distributions converges in probability to the true causal effect. This result, typically known as the Bernstein-von Mises theorem, has been derived in the context of parametric models and an assumed sampling process from an infinite superpopulation (see, e.g., Vaart, 1998, Section 10.2). Establishing this result in the context of randomized experiments opens up new possibilities for scholars to formally quantify what they learn from experiments relative to prior knowledge, but without sacrificing their desirable design-based features.

The remaining portions of the paper proceed as follows: The first section lays out the general framework for design-based causal inference. The succeeding section introduces a prior distribution on causal targets of interest and then derives a design-based likelihood function. The following section then presents the Bernstein-von Mises theorem for design-based causal inference before the final section provides a discussion and conclusion.

1.2 Formal setup

Consider a randomized experiment on a finite study population that consists of $N \geq 4$ units and let the index $i = 1, \ldots, N$ runs over these N units. Of the $N \geq 4$ units in the finite study population, $n_T \geq 2$ are assigned to the treatment condition and the remaining $N - n_T = n_C \geq 2$ are assigned to control. The binary indicator variable $Z_i = 1$ or $Z_i = 0$ denotes whether individual unit i is assigned to treatment ($Z_i = 1$) or control ($Z_i = 0$). Let $\Omega = \{\boldsymbol{z} : \sum_{i=1}^{N} z_i = n_T\}$ be the set of possible values of $\boldsymbol{Z} = \left[Z_1, \ldots, Z_N\right]^\top$ and let $|\Omega|$ denote the number of assignments in (i.e., the cardinality of) Ω.[1] In a uniform, randomized experiment $\Pr(Z_i = 1) = \dfrac{n_T}{N}$ and $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \Pr(\boldsymbol{Z} = \boldsymbol{z}) = |\Omega|^{-1}$.[2]

[1] Under N independent Bernoulli assignments, there are 2^N possible assignments. However, even if n_T is not fixed by design, n_T can be fixed by conditioning on its observed value. The randomization distribution conditional on the realized n_T yields the same randomization distribution one would obtain if n_T had been fixed ex ante by design. Hence, this general setup pertains to both simple and completely randomized assignments even though the argument by which one can regard n_T as fixed is slightly different for each type of assignment.

[2] Or, equivalently, $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \Pr(\boldsymbol{Z} = \boldsymbol{z} \mid \mathcal{Z}) = \dfrac{n_T!(N - n_T)!}{N!}$.

However, in general, one can consider arbitrary PDFs on Ω in which $0 < \Pr(\boldsymbol{Z} = \boldsymbol{z}) < 1$ for all $\boldsymbol{z} \in \Omega$.

Adopting the terminology of Freedman (2009) and later Gerber and Green (2012), define a potential outcomes schedule as a vector-valued function, $\boldsymbol{y} : \Omega \mapsto \mathbb{R}^N$, which maps the set of assignments, Ω, to an N-dimensional vector of real numbers, $\mathbb{R}^N$. More intuitively, a potential outcomes schedule is a listing of how each study participant would have responded to any $\boldsymbol{z} \in \Omega$ that a random assignment process could produce. The vectors of potential outcomes are the elements in the image of the potential outcomes schedule, $\boldsymbol{y} : \Omega \mapsto \mathbb{R}^N$, and the individual potential outcomes for unit i are the ith entries of each of the N-dimensional vectors of potential outcomes. With $|\Omega|$ possible assignments, there are in principle $|\Omega|$ vectors of potential outcomes. However, under the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980; Rubin, 1986)[3], let y_{Ti} denote the common outcome value of unit i for all $\boldsymbol{z} \in \Omega$ with $z_i = 1$. Likewise, let y_{Ci} denote the common outcome value of unit i for all $\boldsymbol{z} \in \Omega$ with $z_i = 0$. The vectors $\boldsymbol{y_C}$ and $\boldsymbol{y_T}$ denote the collection of control and treatment potential outcomes, respectively, for all N units. The observed outcome for unit i is $Y_i = Z_i y_{Ti} + (1 - Z_i) y_{Ci}$, which is either y_{Ti} or y_{Ci} depending on whether the randomly selected $\boldsymbol{z} \in \Omega$ is with $z_i = 1$ or $z_i = 0$. If baseline covariate data exist, denote the K-dimensional row vector of baseline covariates for unit i by $\boldsymbol{x}_i^\top$, where K is the number of covariates, and the $N \times K$ matrix of baseline covariates by $\boldsymbol{x}^\top$. All inferences condition on the potential outcomes and baseline covariates of all N study units, as well as on the event that the assignment vector is an element of Ω. To make the notation cleaner, I leave this conditioning implicit.

Denote the individual treatment effect on the additive scale by $\tau_i \equiv y_{Ti} - y_{Ci}$ and the collection of the N individual treatment effects by $\boldsymbol{\tau} = \begin{bmatrix} \tau_1 & \tau_2 & \dots & \tau_N \end{bmatrix}^\top$. At least since Neyman (1923), one inferential target is the average treatment effect given by $\bar{\tau} \equiv N^{-1} \sum_{i=1}^{N} \tau_i$. Denote a hypothetical value of the true average causal effect by $\bar{\tau}_h$ and the parameter space of hypothetical values of $\bar{\tau}$

[3]SUTVA implies that (1) units in the experiment respond to only the treatment condition to which each unit is individually assigned and (2) the treatment condition is actually the same treatment for all units assigned to treatment and the control condition is the same for all units assigned to control.

by $\Theta_{\bar{\tau}_h}$. Alternatively, another target of inference, often associated with Fisher (1935), is the entire N-dimensional vector of individual causal effects, $\boldsymbol{\tau}$. Denote a hypothetical vector of individual causal effects by $\boldsymbol{\tau}_h$. In principle, the space of $\boldsymbol{\tau}_h$ could be N-dimensional, reflecting hypothetical values of τ_i for all N units. However, I consider only the set of constant, additive causal hypotheses denoted by Θ_{τ_h}, the implications of which I discuss in Section 1.5.2. I refer to $\bar{\tau}_h$ as a weak causal hypothesis and τ_h as a sharp causal hypothesis, where τ_h is no longer in bold to emphasize that it is a 1-dimensional constant effect for all N units.

1.3 Extant approaches to Bayesian causal inference

In the formal setup in Section 1.2, randomness stems solely from the PDF on the set of assignments, Ω. In a randomized experiment, this distribution is known by the researcher. Bayesian inference typically proceeds by defining a likelihood function that conditions on the full realized data, which are typically conceived as i.i.d draws from a random distribution of potential outcomes. However, when potential outcomes are fixed quantities and inference is based on the assignment mechanism (as described in Section 1.2 above), an identified likelihood function that conditions on the full realized data cannnot generally be constructed.

To see this point, note that an assumed weak causal hypothesis does not imply a distribution of potential outcomes. However, an assumed sharp causal hypothesis does. Therefore, we can define an exact likelihood function that, when supplied a sharp causal hypothesis, assigns some probability to the full realized data. Once we have postulated values for missing potential outcomes according to a sharp causal hypothesis under one assignment, the vectors of outcomes we would observe for all possible assignments are known. Under random assignment, the proportion of these vectors that are equal to the vector of outcomes we actually did observe is the probability that a sharp causal hypothesis assigns to the observed data. More generally, without imposing any assumptions on potential outcomes, we can derive what Aronow and Miller (2019, p. 93) refer to

as a finite population mass function:

$$f_{FP}(\boldsymbol{y}) = \sum_{z \in \Omega} \mathbb{1}\{\boldsymbol{y} = \boldsymbol{y}_{zh}\} \Pr(\boldsymbol{Z} = z) \text{ for } \boldsymbol{y}_{zh} \in \mathcal{Y}_{zh}, \tag{1.1}$$

where $\mathcal{Y}_{zh}$ is the set of outcome vectors we would observe for each possible assignment in Ω if a sharp causal hypothesis were true and $\mathbb{1}\{\cdot\}$ is the indicator function that returns the value of 1 if its argument is true and 0 otherwise.

Such an exact likelihood function suffers from at least two pathological properties. First, under the sharp causal hypothesis of $\tau_h = 0$, the vector of observed outcomes, $\boldsymbol{y}$, is fixed over all assignments. Hence, $\mathbb{1}\{\boldsymbol{y} = \boldsymbol{y}_{zh}\}$ in Equation (1.1) will return the value of 1 for all $z \in \Omega$, which implies that the probability $\tau_h = 0$ assigns to the observed data will always be $\sum_{z \in \Omega} \Pr(\boldsymbol{Z} = z) = 1$. For other causal hypotheses, $\tau_h : \tau_h \neq 0$, the likelihood function will typically be flat over different values of τ_h and, hence, will be equally consistent with a range of sharp causal hypotheses.

These pathologies can be somewhat resolved when outcomes are binary or ordinal. In particular, Copas (1973) shows that it is possible to derive a partially identified randomization-based likelihood function when outcomes are binary. Ding and Miratrix (2019) then show how scholars can conduct model-free Bayesian inference via such a randomization-based likelihood function (see also Keele and Quinn, 2017; Humphreys and Jacobs, 2015). Chiba (2018) extends this logic for the case of binary outcomes to that of ordinal outcomes. Yet the lack of binary or ordinal outcomes in many applications makes such a randomization-based likelihood function untenable.

Given the difficulties of such model-free Bayesian inference, an alternative is to derive a likelihood function from a probability model of the joint distribution of potential outcomes. In this approach, as Imbens and Rubin (2015, p. 141) state, "potential outcomes themselves are also viewed as random variables, even in the finite sample." With stochastic potential outcomes, the essential role of randomization is that it implies that one can "ignore the assignment mechanism when making causal inferences" (Rubin, 1976, p. 233), hence the term "ignorability" (Rubin, 1976; Rubin, 1978).[4] Random assignment ensures the independence between $\boldsymbol{Y_C}, \boldsymbol{Y_T}$ and $\boldsymbol{Z}$. Therefore, we can

[4]Rubin (1976) and Rubin (1978) originally used the term "ignorability" in a Bayesian context to denote the situation

ignore the assignment process and instead need to consider only (1) the prior distribution of the potential outcome model's parameters, $\boldsymbol{\theta}$, and (2) the likelihood of the potential outcomes conditional on the model parameters. In practice, since no two potential outcomes are observable for the same unit, Imbens and Rubin (2015) transform the potential outcomes into observed and missing potential outcomes and derive the likelihood function of $f(\boldsymbol{Y}|\boldsymbol{\theta})$, which is the marginal distribution of observed potential outcomes after integrating out missing potential outcomes.

With a likelihood function derived from a model of the joint distribution of potential outcomes, we can derive a posterior distribution of $\boldsymbol{\theta}$. Upon updating on $\boldsymbol{\theta}$, we can (1) draw from the posterior distribution of $\boldsymbol{\theta}$, (2) input each draw as the parameters of the model of potential outcomes, (3) draw from the model of potential outcomes, (4) impute these draws for the missing potential outcomes and (5) directly calculate a function of the two vectors of (partially observed and partially imputed) potential outcomes, e.g., the mean causal effect. Repeating this procedure many times yields a simulation-based approximation to the posterior distribution of an estimand, such as the mean causal effect.

Yet a central concern of this methodology is that inference now depends on a stochastic model of potential outcomes, not only a known assignment process. Thus, one of the central appeals of experiments — the ability to draw inferences based on only a known assignment process — no longer obtains. Indeed, Imbens and Rubin (2015, p. 142) emphasize that "[o]ne of the practical issues in the model-based approach is the choice of a credible model for imputing the missing potential outcomes" and that "fundamentally the resulting inference may be more sensitive to the modeling assumptions." Thus, the ability to conduct Bayesian inference and quantify how much we learn from an experiment appears to come at the expense of inference that depends on only a known assignment process, *not* unverifiable probability models of potential outcomes.

As a solution to this impasse, I derive a likelihood function of causal effects that conditions *not* on the full realized data, but rather a test-statistic of them. Conditioning on a suitable test-statistic

in which a correct posterior distribution will result from a likelihood function that ignores the stochastic assignment process. Random assignment satisfies this ignorability condition. This original conception of ignorability differs from later notions of "strong ignorability" (Rosenbaum and Rubin, 1983) that pertain to Frequentist inference of superpopulation parameters.

yields a likelihood function that avoids the pathologies of a design-based likelihood function that conditions on the full data. In particular, I show that the likelihood function I derive implies the following property: Given weak conditions on a prior distribution — namely, that the true effect is in the prior distribution's support — scholars will revise their beliefs in a way that (so long as the experimental population is sufficiently large) the posterior distribution concentrates around the true causal effect. Such Bayesian inference therefore satisfies an important theoretical guarantee regarding its ability to track the true causal effect. Section 1.4 to follow lays out the asymptotic framework for this result and then Section 1.5 formally establishes this property, first for weak and then for sharp causal hypotheses.

1.4 Asymptotic framework for design-based Bayesian causal inference

The theoretical property described above is an asymptotic one that pertains to limiting statistical properties over an infinite sequence of finite populations of increasing sizes. Thus, the argument to follow embeds a finite experimental population in an imaginary sequence of finite populations of increasing sizes. In actual practice, we only ever conduct a randomized experiment on a finite number of N units. Nevertheless, following the insight of Lehmann (1999, p. 255), the purpose of embedding an experiment in this imaginary sequence of finite populations is "to obtain a simple and accurate approximation. The embedding sequence is thus an artifice and has only this purpose." In other words, the asymptotic properties of design-based Bayesian inference to follow offer a valuable approximation to the statistical properties of design-based Bayesian inference in an actual experiment with a fixed, but large N.

A common asymptotic regime in design-based causal inference (as in, e.g., Middleton and Aronow, 2015; Bowers and Leavitt, 2020) is given by Brewer (1979). In this conception of asymptotic growth, each finite population in the infinite sequence of finite populations is a specified number of copies of the original finite population. Hence, all relevant quantities, namely, n_T/N, as well as the means, variances and covariance of potential outcomes — denoted by $\bar{y}_T$, $\bar{y}_C$, $\sigma^2_{y_T}$, $\sigma^2_{y_C}$ and σ_{y_T,y_C}, respectively — are fixed constants for each N over the entire sequence of finite popu-

lations. This asymptotic regime implicitly embeds several regularity conditions that are standard in the literature (see, e.g., Lin, 2013; Freedman, 2008; Cohen and Fogarty, 2020, among others). In contrast to the asymptotic regime in Brewer (1979), I assume only these regularity conditions whereby there is no other information whatsoever between any two populations in the sequence of finite populations.[5] The mild regularity conditions on the sequence of finite populations are as follows:[6]

1. **Condition 1**: Potential outcomes have bounded fourth moments, i.e., for all $N = 4, 5, \ldots$, $\frac{1}{N}\sum_{i=1}^{N} y_{Ci}^4 < L < \infty$ and $\frac{1}{N}\sum_{i=1}^{N} y_{Ti}^4 < L < \infty$.

2. **Condition 2**: As $N \to \infty$, the proportion of treated units, $\frac{n_T}{N}$, tends to a positive value strictly greater than 0 and less than 1, i.e., $\frac{n_T}{N} \to p \in (0, 1)$ as $N \to \infty$, which, since $n_T + n_C = N$, also implies that $\frac{n_C}{N} \to 1 - p$ as $N \to \infty$.

3. **Condition 3**: The population quantities $\frac{1}{N}\sum_{i=1}^{N} y_{Ti}$, $\frac{1}{N}\sum_{i=1}^{N} y_{Ci}$, $\frac{1}{N}\sum_{i=1}^{N} y_{Ci}\, y_{Ti}$, $\frac{1}{N}\sum_{i=1}^{N} y_{Ti}^2$ and $\frac{1}{N}\sum_{i=1}^{N} y_{Ci}^2$ are Cesàro summable, i.e., tend to finite limits as $N \to \infty$, where, adopting the angle bracket notation of Freedman (2008), these finite limits are denoted by $\langle y_T \rangle$, $\langle y_C \rangle$, $\langle y_C\, y_T \rangle$, $\langle y_T^2 \rangle$ and $\langle y_C^2 \rangle$, respectively.

Regularity condition 1 ensures that the variance of an estimator of the Difference-in-Means' variance tends to 0 as $N \to \infty$. The relevance of this property for Bayesian causal inference will be explained shortly. The importance of Conditions 2 and 3 is to ensure that limits of $\bar{\tau}$ and $N \operatorname{Var}\left[\hat{\bar{\tau}}\right]$ exist over the sequence of finite populations of increasing sizes.

Under regularity conditions 1 — 3, Theorem 1 of Freedman (2008) shows that

$$\sqrt{N}(\hat{\bar{\tau}} - \mathrm{E}[\hat{\bar{\tau}}]) \xrightarrow{d} \mathcal{N}(0, \nu), \tag{1.2}$$

[5]For the value of such an asymptotic regime, see the brief but insightful discussion in Sävje, Aronow, and Hudgens (2021, Section 5). Delevoye and Sävje (2020) also discuss the value of this asymptotic regime in passing.

[6]In writing these conditions and in laying out the general asymptotic argument moving forward, one ought to index potential outcomes and other quantities in the infinite sequence of finite populations by N; however, for cleaner notation and in accordance with standard practice, I leave this indexing implicit.

where $\hat{\bar{\tau}}$ is the Difference-in-Means of the observed data and ν is the asymptotic variance of the centered Difference-in-Means scaled by $\sqrt{N}$, both of which are given by

$$\hat{\bar{\tau}} = \frac{\boldsymbol{Z}^\top \boldsymbol{Y}}{\boldsymbol{Z}^\top \mathbf{1}} - \frac{(\mathbf{1}-\boldsymbol{Z})^\top \boldsymbol{Y}}{(\mathbf{1}-\boldsymbol{Z})^\top \mathbf{1}} = \frac{1}{n_T}\sum_{i=1}^{N} Z_i Y_i - \frac{1}{n_C}\sum_{i=1}^{N}(1-Z_i)Y_i \text{ and} \tag{1.3}$$

$$\nu = \lim_{N\to\infty} \operatorname{Var}[\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau})] = \lim_{N\to\infty} N \operatorname{Var}[\hat{\bar{\tau}}] = \frac{1-p}{p}\langle y_T^2 \rangle + \frac{p}{1-p}\langle y_C^2 \rangle + 2\langle y_C\, y_T \rangle. \tag{1.4}$$

Due to Slutsky's theorem (Slutsky, 1925), Equation (1.2) can be equivalently expressed as

$$\frac{\hat{\bar{\tau}} - \mathrm{E}[\hat{\bar{\tau}}]}{\sqrt{\operatorname{Var}[\hat{\bar{\tau}}]}} \xrightarrow{d} \mathcal{N}(0,1). \tag{1.5}$$

In addition to the finite population CLT, the application of the Berry-Esseen theorem in Höglund (1978) bounds the error of the Normal approximation to the distribution of the sample sum under simple random sampling from a finite population. For this theorem to be relevant for the Difference-in-Means, one needs only to note that, when the number of treated units is fixed, the Difference-in-Means in Equation (1.3) can be expressed as a sample sum via scale and shift factors that do not vary across treatment assignments:

$$\hat{\bar{\tau}} = \sum_{i=1}^{N} Z_i q_i, \text{ where } q_i = \frac{y_{Ti}}{n_T} + \frac{y_{Ci}}{n_C} - \left(\frac{1}{n_C n_T}\right)\sum_{i=1}^{N} y_{Ci}. \tag{1.6}$$

This representation of the Difference-in-Means as a sample sum calculated on a random sample from a finite population assumes complete random assignment, which renders n_T and n_C fixed constants. However, n_T and n_C would not be fixed under a random assignment process of, e.g., N independent Bernoulli trials. Consequently, Höglund's bound on the error of the Normal approximation to the Difference-in-Means would no longer be available. To alleviate this concern, we can fix n_T by conditioning on its observed value. The randomization distribution of the Difference-in-Means conditional on the realized n_T yields the same distribution one would obtain if n_T had been fixed ex ante by design. Thus, in sum, Höglund's Berry-Esseen theorem implies that, so long as an experiment is of at least moderate size and experimental units' potential outcomes are not too

skewed or characterized by extreme outliers,

$$\frac{\hat{\tau} - \mathrm{E}[\hat{\tau}]}{\sqrt{\mathrm{Var}[\hat{\tau}]}} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1) \text{ or, equivalently, } \hat{\tau} \overset{\text{approx}}{\sim} \mathcal{N}(\mathrm{E}[\hat{\tau}], \mathrm{Var}[\hat{\tau}]). \tag{1.7}$$

The equivalence in (1.7) holds due to the Normal distribution's closure under location-scale transformations; any errors in using $\mathcal{N}(0,1)$ to approximate the distribution of the standardized Difference-in-Means estimator will be scaled by the standard error when using $\mathcal{N}(\mathrm{E}[\hat{\tau}], \mathrm{Var}[\hat{\tau}])$ to approximate the distribution of the unstandardized Difference-in-Means estimator. Taken together, the finite population CLT and Höglund's Berry-Esseen theorem justify the use of a Normal likelihood function in which the probability density of the observed data, summarized by the Difference-in-Means, is a function of two parameters, $\mathrm{E}[\hat{\tau}]$ and $\mathrm{Var}[\hat{\tau}]$.

Under random assignment and SUTVA, $\mathrm{E}[\hat{\tau}]$ is equal to $\bar{\tau}$, which is our target parameter of interest. The parameter $\mathrm{Var}[\hat{\tau}]$, however, is a nuisance parameter. The standard Bayesian approach to eliminating a nusiance parameter is by marginalizing over its distribution (Berger, Liseo, and Wolpert, 1999; Liseo, 2005). Other common approaches include conditioning on a sufficient statistic (Reid, 1995), plug-in estimation and worst case inference over all possible values of the nuisance parameter or its confidence set (Berger and Boos, 1994). This issue of eliminating a nuisance parameter arises only for a weak causal hypothesis, which implies a value of only the expectation of the Difference-in-Means, not its variance. Sharp causal hypotheses, by contrast, imply a values for both the expected value and variance parameters of the Difference-in-Means.

In the context of inference of average affects, applied scholars are unlikely to have well motivated prior beliefs about the variance of the Difference-in-Means. Hence, eliminating the variance parameter by marginalizing over its distribution is unattractive. Thus, for weak causal hypotheses, the Bayesian procedure to follow eliminates the variance nuisance parameter *not* by integrating over its distribution as is standard in Bayesian approaches, but instead via conservative plug-in estimation. This approach effectively plugs in a conservative estimator for the variance and then proceeds *as if* the variance parameter is known.

With a uniform PDF on the set of possible assignments, Ω, Neyman (1923) showed that the variance of the Difference-in-Means estimator is

$$\text{Var}\left[\hat{\bar{\tau}}\right] = \frac{S_{y_T}^2}{n_T} + \frac{S_{y_C}^2}{n_C} - \frac{S_\tau^2}{N}, \tag{1.8}$$

where

$$S_{y_T}^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_{Ti} - \bar{y}_T)^2$$
$$S_{y_C}^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_{Ci} - \bar{y}_C)^2$$
$$S_\tau^2 = (N-1)^{-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2 .$$

A natural plug-in estimator for the variance parameter is the conservative variance estimator whose expected value is always at least as great as $\text{Var}[\hat{\bar{\tau}}]$ for any fixed N (Neyman, 1923).[7] Neyman's conservative variance estimator is

$$\widehat{\text{Var}}[\hat{\bar{\tau}}] = \frac{s_{y_T}^2}{n_T} + \frac{s_{y_C}^2}{n_C}, \tag{1.9}$$

where $s_{y_T}^2$ and $s_{y_C}^2$ are unbiased estimators of $S_{y_T}^2$ and $S_{y_C}^2$ given by

$$s_{y_T}^2 = (n_T - 1)^{-1} \sum_{i=1}^{N} Z_i (y_{Ti} - \frac{1}{n_T} \sum_{i=1}^{N} Z_i y_{Ti})^2 \text{ and}$$
$$s_{y_C}^2 = (n_C - 1)^{-1} \sum_{i=1}^{N} (1 - Z_i)(y_{Ci} - \frac{1}{n_C} \sum_{i=1}^{N} (1 - Z_i) y_{Ci})^2.$$

For inference of weak causal effects, this plug-in estimation approach obviates the need for scholars to define priors over the variance parameter. Instead, scholars can simply define prior beliefs over

[7] The use of alternative, improved variance estimators is also possible, e.g., from Aronow, Green, Lee, et al. (2014) in completely randomized experiments and Imai (2008), Fogarty (2018) and Pashley and Miratrix (2021) in finely stratified experiments.

the average causal effect and subsequently update those beliefs upon observing new experimental data. Therefore, let $p(\bar{\tau}_h)$ denote the prior PDF on the hypothetical values of the mean causal effect and, in practice, consider the following standardized test-statistic:

$$\frac{\hat{\bar{\tau}} - \bar{\tau}_h}{\sqrt{\widehat{\text{Var}}[\hat{\bar{\tau}}]}}, \tag{1.10}$$

which serves as the argument to the standard Normal likelihood function. The output of the standard Normal likelihood function is a probability density of the observed data, summarized by $\hat{\bar{\tau}}$, given a hypothesis about the average causal effect, $\bar{\tau}_h$, and an estimate of the variance of the Difference-in-Means, $\widehat{\text{Var}}[\hat{\bar{\tau}}]$. With a prior distribution of $p(\bar{\tau}_h)$, the standardized test-statistic in Equation (1.10) and standard Normal likelihood function, inference proceeds via Bayes' rule.

For sharp causal hypotheses, Section 1.5.2 derives an exact likelihood function. However, Normal approximations to the Difference-in-Means distribution implied by a sharp causal hypothesis are possible. One needs only to substitute into Equation (1.10) the average effect and variance of the Difference-in-Means implied by a sharp causal hypothesis. With such a likelihood function — either exact or a Normal approximation — and a prior distribution, $p(\tau_h)$, inference of sharp causal effects also proceeds via Bayes' rule.

1.5 Bernstein-von Mises Theorem for Design-based Causal Inference

It is important not only to demonstrate that design-based Bayesian causal inference is possible, but also to show that such inference reliably recovers the true causal effect. Therefore, I now prove a version of the Bernstein-von Mises theorem for design-based Bayesian inference of the average causal effect. I first consider weak causal hypotheses and show in Theorem 1 that, under mild conditions on the prior distribution, $p(\bar{\tau}_h)$, the posterior distribution of $\bar{\tau}_h$ converges in probability (over a sequence of finite populations of increasing sizes) to the true mean causal effect. I then establish an analogous result for sharp causal hypotheses in Theorem 2. In practical terms, both results imply that, so long as one has not unduly excluded possible values of the causal effect

from the prior distribution, then, in a sufficiently large experiment, the posterior distribution will concentrate closely around the true causal effect.

1.5.1 Weak causal hypotheses

To prove a version of the Bernstein-von Mises theorem for weak causal hypotheses, I first establish Lemma 1, which states that the conservative variance estimator in Equation (1.9) converges in probability to a constant at least as great as the true asymptotic variance, ν, in Equation (1.2). The proof of Lemma 1, as well as all other proofs, are in **??**.

Lemma 1. *Under regularity conditions 1 – 3, as* $N \to \infty$, $\dfrac{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}{\nu} \xrightarrow{p} c \geq 1$, *where* $\nu = \lim_{N\to\infty} N \mathrm{Var}[\hat{\bar{\tau}}] = \dfrac{1-p}{p}\langle y_T^2\rangle + \dfrac{p}{1-p}\langle y_C^2\rangle + 2\langle y_C\, y_T\rangle$ *is the asymptotic variance of the centered Difference-in-Means test-statistic scaled by* $\sqrt{N}$.

Under conditions 1 – 3, the variance estimator, $\widehat{\mathrm{Var}}[\sqrt{N}\hat{\bar{\tau}}]$, is not consistent for the asymptotic variance, $\nu = \lim_{N\to\infty} \mathrm{Var}[\sqrt{N}\hat{\bar{\tau}}]$. Yet an application of Slutsky's theorem, an important step of the theoretical result to follow, requires only that the variance estimator converge in probability to a constant, not necessarily the target parameter. Lemma 1 shows that the random quantity $\widehat{\mathrm{Var}}[\sqrt{N}\hat{\bar{\tau}}]$ converges to a constant that is always at least as great as the true asymptotic variance denoted by ν.

Armed with Lemma 1, Theorem 1 below states that the limiting posterior probability of weak causal hypotheses arbitrarily close to true average causal effect tends to 1 as $N \to \infty$.

Theorem 1. *Define* $\Theta^*_{\bar{\tau}_h} \equiv \{\bar{\tau}_h : \bar{\tau} - \varepsilon < \bar{\tau}_h < \bar{\tau} + \varepsilon\}$, *where* ε *is an arbitrarily small constant greater than* 0, $\varepsilon > 0$, *and assume regularity conditions 1 – 3. If* $\Theta^*_{\bar{\tau}_h}$ *is in the support of the prior distribution,* $p(\bar{\tau}_h)$, *then, as the size of the experiment increases indefinitely,* $N \to \infty$, *the posterior probability of* $\Theta^*_{\bar{\tau}_h}$ *tends to* 1*:*

$$\lim_{N\to\infty} \frac{\int_{\bar{\tau}_h \in \Theta^*_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h}{\int_{\bar{\tau}_h \in \Theta_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h} = 1.$$

To provide more intuition for Theorem 1, note that we can equivalently write the test-statistic in Equation (1.10) as

$$\frac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau}) + \sqrt{N}(\bar{\tau} - \bar{\tau}_h)}{\sqrt{N}\sqrt{\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}}. \tag{1.11}$$

As Lemma 1 shows, the denominator converges in probability to a constant that is at least as great as the true asymptotic variance of the test-statistic. When the weak causal hypothesis is true, $\bar{\tau}_h = \bar{\tau}$, Slutsky's theorem and the continuous mapping theorem imply that Equation (1.11) converges in distribution to Z/c, where Z is the standard Normal distribution and c is a constant greater than or equal to 1 (as defined in Lemma 1). By contrast, when $\bar{\tau}_h \neq \bar{\tau}$, note that $\sqrt{N}(\bar{\tau} - \bar{\tau}_h)$ will increase or decrease without bound . Hence, Equation (1.11) will diverge in probability to either $-\infty$ or $+\infty$. Finally, since the standard Normal likelihood function is monotonically decreasing in distance from 0, the probability density that the standard Normal distribution assigns to the observable test-statistic when τ_h is false will tend to 0 as $N \to \infty$. As the likelihood of any false $\bar{\tau}_h$ tends to 0 as $N \to \infty$, so does the posterior probability of false hypothetical values of the average. Thus, by the law of total probability, the posterior distribution will concentrate increasingly around the true average effect.

To illustrate this point, consider the results of the following simple simulation depicted in Figure 1.1. The experimental population consists of 20 units whose true control potential outcomes are randomly drawn from the distribution $\mathcal{N}(50, 100)$ and then fixed at their realized values. Subsequent references to convergence or divergence in probability refer to the probability distribution over repeated randomizations (i.e., over Ω as described in Section 1.2). I define the treated potential outcomes as the control potential outcomes plus a constant, additive effect of $\tau = 10$, which implies that $\bar{\tau} = 10$. For expository purposes, I define a uniform prior on only two weak causal hypotheses, $\bar{\tau}_h = 5$ and $\bar{\tau}_h = \bar{\tau} = 10$. Drawing on the asymptotic regime of Brewer (1979), which satisfies regularity conditions 1 – 3, I then let the sequence of finite populations increase by simply copying this initial finite population an increasing number of times.

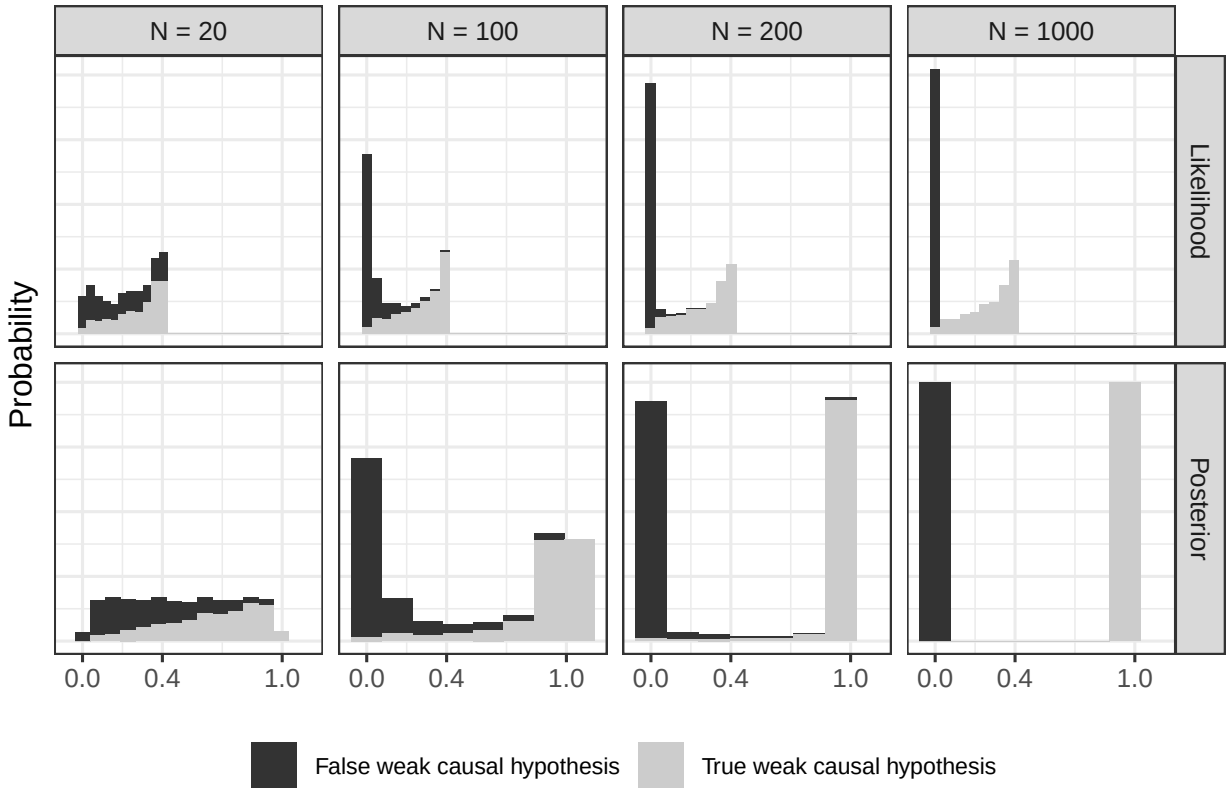

Figure 1.1: Distributions of likelihoods and posterior probabilities over repeated randomizations

Since $\bar{\tau}_h = 5$ is too small, the randomization distribution of the standardized Difference-in-Means in Equation (1.11) diverges in probability to $-\infty$. Thus, as Figure 1.1 shows, when $\bar{\tau}_h \neq \bar{\tau}$, the probability density that the standard Normal distribution assigns to a draw of the standardized test-statistic in Equation (1.11) tends to 0 as $N \to \infty$. When $\bar{\tau}_h = \bar{\tau}$, the probability density that the standard Normal distribution assigns to a draw of the standardized test-statistic takes on values between 0 and $1/\sqrt{2\pi} \approx 0.4$, which is the maximum probability density of the standard Normal distribution. Since the likelihood of the false weak causal hypothesis ($\bar{\tau}_h = 5$) tends to 0 asymptotically, so does the product of its prior and likelihood. Normalizing by the total probability of the evidence thereby implies, by the law of total probability, that the posterior distribution concentrates increasingly around the true weak causal hypothesis ($\bar{\tau}_h = 10$) as $N \to \infty$

1.5.2 Sharp causal hypotheses

The general spirit of Theorem 1 also applies to sharp causal hypotheses. Theorem 1 for weak causal hypotheses invokes a Normal likelihood function that is justified by the finite population CLT and Höglund's Berry-Esseen theorem. It is possible, however, to derive a similar result without invoking the asymptotic Normality of the Difference-in-Means so long as inferences pertain to sharp rather than weak causal effects. I now derive a result analogous to Theorem 1 for sharp causal hypotheses.

In the same way, that the likelihood function for weak causal hypotheses in Section 1.5.1 is based on the centered Difference-in-Means, define the individual outcome for unit i adjusted (or centered) by its true individual causal effect as $\tilde{y}_i \equiv y_i - \tau z_i$. Under SUTVA, $\tilde{y}_i = y_{Ci}$ regardless of whether $z_i = 0$ or $z_i = 1$ for all $i = 1, \ldots, N$. Hence, regardless of whichever data are realized, the collection of all N adjusted outcomes, $\tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{y}_1 & \ldots & \tilde{y}_N \end{bmatrix}^\top$, is fixed over all $\boldsymbol{z} \in \Omega$ and thereby satisfies the sharp causal hypothesis of no effect. (Recall that the sharp causal hypothesis of no effect implies that an outcome vector is fixed over all possible assignments, just like $\tilde{\boldsymbol{y}}$.)

We can therefore derive a likelihood function in which we adjust the observed data via a hypothetical causal effect and then assess how consistent the adjusted data are with the sharp null of no effects. That is, for a sharp causal hypothesis, τ_h, given some realization of data, $(\boldsymbol{z}, \boldsymbol{y})$, we adjust each observed outcome as $\tilde{y}_{ih} = y_i - \tau_h z_i$. If τ_h is true, then $\tilde{y}_{ih}$ is equal to the true adjusted outcome, $\tilde{y}_i$. Under the sharp null, treatment and control distributions are identical, so when the τ_h used to adjust outcomes is true, treatment and control distributions in the adjusted data should appear similar. Conversely, when τ_h is false, treatment and control distributions in the adjusted data should appear dissimilar.

Analogous to the result in Theorem 1, let the test-statistic be the Difference-in-Means given by

$$t(\boldsymbol{z}, \boldsymbol{v}) = n_T^{-1} \sum_{i=1}^{N} z_i v_i - n_C^{-1} \sum_{i=1}^{N} (1 - z_i) v_i, \text{ where } \boldsymbol{v} \in \mathbb{R}^N . \tag{1.12}$$

Conditional on some realization of data, $(\boldsymbol{z}, \boldsymbol{y})$, the Difference-in-Means test-statistic for a test of

the sharp causal hypothesis of no effect on the adjusted outcome vector is $t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)$. The outcome vector adjusted by a sharp causal hypothesis, $\tilde{\boldsymbol{y}}_h$, is in lowercase because, conditional on a realization of data, it is fixed over all $z \in \Omega$ under the sharp causal hypothesis of no effect. When referring to the test-statistic *unconditional* on a realization of data, the test-statistic is $t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)$, where $\tilde{\boldsymbol{Y}}_h$ is in uppercase because it can vary over different possible realizations of $(\boldsymbol{Z}, \boldsymbol{Y})$.[8] Regardless of the τ_h one uses to adjust the observed outcomes, $\mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)] = 0$ and $\mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] = \bar{\tau} - \tau_h$.

An exact likelihood function for a test of τ_h that conditions on the Difference-in-Means test-statistic of the realized data is given by

$$f(T \mid \tau_h) = \Pr(t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h) = T) = \sum_{z \in \Omega} \mathbb{1}\{t(z, \tilde{\boldsymbol{y}}_h) = T\} \Pr(\boldsymbol{Z} = z), \tag{1.13}$$

where T is the observed test-statistic calculated on the outcomes adjusted by a sharp causal hypothesis. Intuitively, the likelihood function in Equation (1.13) returns the probability of observing the test-statistic we did observe if the sharp causal hypothesis of τ_h were true. This likelihood function resembles the one in Equation (1.1) in Section 1.3. The key difference is that, instead of conditioning on the realized outcomes, it conditions on the realized test-statistic.

Theorem 2 below shows that, under weak conditions on the prior, Bayesian inference will recover the true causal effect asymptotically.

Theorem 2. *Define* $\Theta^*_{\tau_h} \equiv \{\tau_h : \tau - \varphi < \tau_h < \tau + \varphi\}$*, where* φ *is an arbitrarily small constant greater than* 0*,* $\varphi > 0$*. Let the test-statistic be the Difference-in-Means and assume regularity Conditions 2 – 3. If* $\Theta^*_{\tau_h}$ *is in the support of the prior distribution,* $p(\tau_h)$*, then, as the size of the experiment increases indefinitely,* $N \to \infty$*, the posterior probability of* $\Theta^*_{\tau_h}$ *tends to* 1*:*

$$\lim_{N \to \infty} \frac{\int_{\tau_h \in \Theta^*_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\, \tau_h}{\int_{\tau_h \in \Theta_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\, \tau_h} = 1.$$

[8] In particular, the test-statistic *unconditional* on a realization of data, $t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)$, will vary over assignments if and only if τ_h is false.

To provide intuition for Theorem 2, consider the same simple simulation in Section 1.5.1, except now let inference pertain to a sharp rather than weak causal effect. To reiterate, the experimental population consists of 20 units whose true control potential outcomes are randomly drawn from the distribution $\mathcal{N}(50, 100)$ and then fixed at their realized values. Treated potential outcomes are equal to the control potential outcomes plus a constant shift of $\tau = 10$. For simplicity, I consider only two sharp causal hypotheses $\tau_h = 5$ and $\tau_h = \tau = 10$ over the sequence of finite populations that grow according to the aforementioned asymptotic regime from Brewer (1979). Since the false hypothesis, $\tau_h = 5$, is too small, the distribution of the observable test-statistic, $t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)$, is stochastically larger than any distribution of $t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)$ regardless of whichever data are realized. Figure 1.2 below shows the distribution of the observable test-statistic alongside the distribution of $t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)$ that has the largest variance over all possible assignments. For the true hypothesis, $\tau_h = 10$, the adjusted outcomes are fixed over all assignments, so the distribution of the observable test-statistic and the distribution implied by $\tau_h = 10$ under any realization of data are all identical.

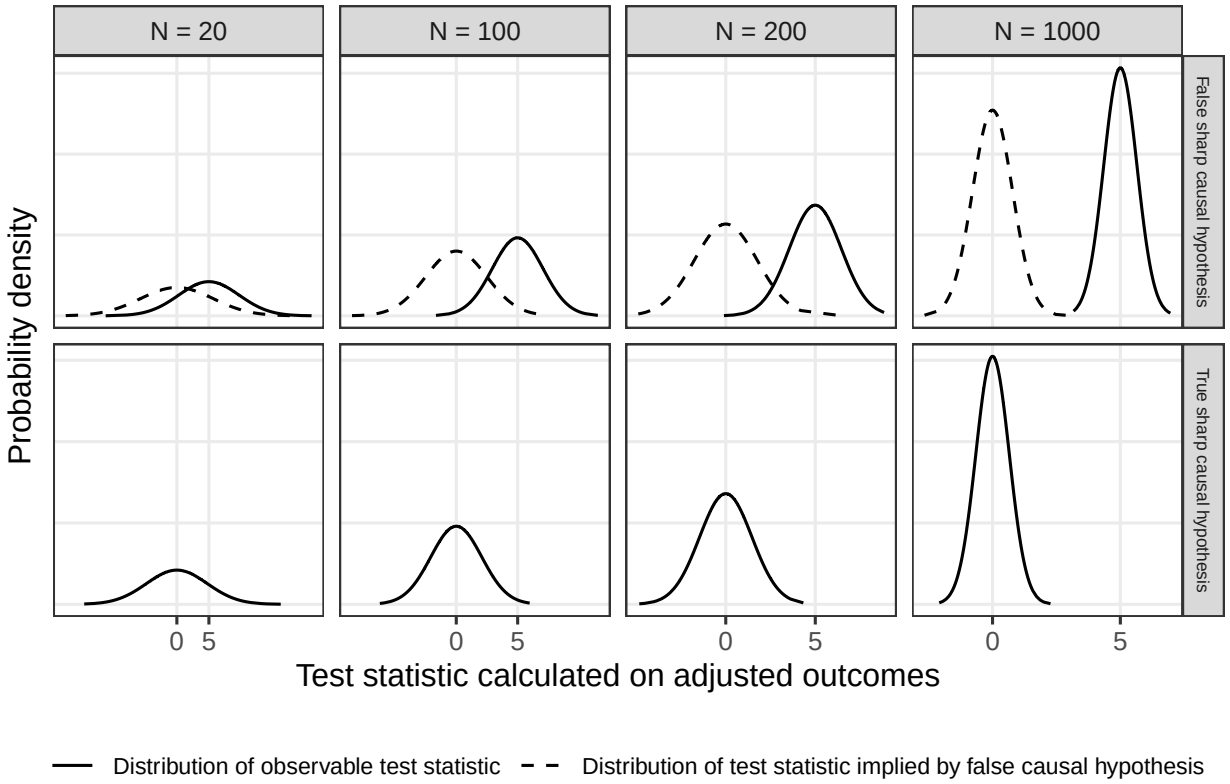

Figure 1.2: Comparison of distribution of observable test-statistic and distribution implied by a false causal hypothesis conditional on realization of data

In the first row of Figure 1.2, the solid line is the distribution of the observable test-statistic and the dashed line is a distribution conditional on a realization of data — namely, the conditional distribution of the test-statistic calculated on the adjusted outcomes implied by the false sharp causal hypothesis of $\tau_h = 5$. The expected value of the distribution in the dashed line is equal to 0 for every population in the sequence of finite populations. As Figure 1.2 shows, for a false causal hypothesis, the distribution in the solid line is stochastically larger than the distribution in the dashed line. Nevertheless, when an experiment is small, the probability that the distribution in the dashed line assigns to a draw from a test-statistic from the solid line may still be high. Yet, as the size of the experiment becomes larger, Figure 1.2 illustrates that both distributions converge in probability to their limiting expected values. Since these limiting expected values differ, the

probability that the distribution in the dashed line assigns to a draw from a test-statistic from the solid line tends to 0 as $N \to \infty$.

The implication of the property illustrated in Figure 1.2 is that the probability density that a false causal hypothesis assigns to the observed data (summarized by a test-statistic) tends to 0 as $N \to \infty$. As the likelihood of a false causal hypothesis tends to 0, so does the product of the prior and the likelihood. Then, just as with weak causal hypotheses, normalizing by the total probability of the evidence implies, by the law of total probability, that the posterior distribution concentrates increasingly around the true τ as $N \to \infty$.

1.5.3 Incorporating covariate information

Just as in the case of standard estimation and hypothesis testing, incorporating covariate information can help increase the precision of the posterior distribution without sacrificing its convergence in probability to the true average effect. In general, we can delineate two approaches to incorporating covariates. One approach, blocking, increases precision by excluding assignments that, on average, yield a Difference-in-Means (or test-statistic of choice) far from its expected value. A second approach, e.g., linear regression, increases precision by linearly rescaling potential outcomes in order to reduce their variances. Incorporating covariates via blocking requires slightly different regularity conditions than those in Section 1.4, but implies the same general result of Theorem 1. Inference on potential outcomes rescaled via covariates will also satisfy Theorem 1 so long as regularity conditions 1 and 3 also hold on the baseline covariates.

In a block (i.e., stratified) randomized experiment, each block (i.e., stratum) is a completely randomized experiment. Hence, one needs only to apply Theorem 1 to each block and then average over all blocks to obtain a result analogous to Theorem 1 under blocked random assignment. However, the appropriate asymptotic regime is slightly different under blocked random assignment. The closest analogue to the asymptotic regime described by regularity conditions 1 – 3 imagines an infinite sequence of experiments that consists of experiments in which the total number of blocks, B, is bounded, but the number of units within blocks grows large, $N_b \to \infty$. Con-

versely, an alternative asymptotic regime conceives of an infinite sequence of experiments with an increasing number of total blocks B, while the block sizes, N_b, are bounded as $B \to \infty$. Aside from these technical details, the logic of Theorem 1 carries over to block randomized designs.

An alternative approach to incorporating covariate information draws on linear regression (but without the regression model's usual assumptions), which, in the case of a single categorical covariate, can also be interpreted as a post-stratified Difference-in-Means (Miratrix, Sekhon, and Yu, 2013). Consider the regression-based test statistic proposed by Lin (2013), denoted by $\hat{\bar{\tau}}_{\text{reg}}$, which consists of a fully saturated linear regression with all interactions between treatment and covariates centered by their means. With the caveat that regularity conditions 1 and 3 for potential outcomes also apply to baseline covariates, then the use of $\hat{\bar{\tau}}_{\text{reg}}$ instead of the Difference-in-Means cannot hurt (and will likely decrease) the asymptotic mean squared error of the posterior distribution.

Formally, we can add another assumption by substituting x_{ik} for each of the $k = 1, \ldots, K$ covariates into conditions 1 and 3. Then, under conditions 1 – 3, where conditions 1 and 3 also apply to the K covariates, it follows that

$$\lim_{N\to\infty} \sqrt{N}\left(\hat{\bar{\tau}}_{\text{reg}} - \bar{\tau}\right) = \sqrt{N}\left(\frac{1}{n_T}\sum_{i=1}^{N} Z_i e_{Ti} - \frac{1}{n_C}\sum_{i=1}^{N}(1 - Z_i)e_{Ci}\right), \text{ where} \tag{1.14}$$

$$e_{Ti} = (y_{Ti} - \bar{y}_T) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \hat{\boldsymbol{\beta}}_T,$$

$$e_{Ci} = (y_{Ci} - \bar{y}_C) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \hat{\boldsymbol{\beta}}_C,$$

and $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_C$ are the coefficient vectors that minimize the sum of squared residuals separately among units with $Z_i = 1$ and among units with $Z_i = 0$. (See Cohen and Fogarty, 2020, Proposition 2.) In other words, when the regularity conditions apply to potential outcomes and covariates, then these conditions also apply to residualized outcomes, e_{Ti} and e_{Ci}. (See Cohen and Fogarty, 2020, Lemma C for Proposition 2.) Importantly, the individual effects, centered by the limiting average

effect, manifest entirely in the residuals of the regression:

$$e_{Ti} - e_{Ci} = (y_{Ti} - \bar{y}_T) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \hat{\boldsymbol{\beta}}_T - \left((y_{Ci} - \bar{y}_C) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \hat{\boldsymbol{\beta}}_C\right) = (y_{Ti} - y_{Ci}) - (\bar{y}_T - \bar{y}_C). \tag{1.15}$$

The implication, then, is that scholars can incorporate covariate information in their Bayesian inferences by substituting suitably transformed outcomes for the actual outcomes themselves.

1.5.4 Discussion and comparison of weak versus sharp causal hypotheses

Both Theorems 1 and 2 pertain to the Difference-in-Means test statistic. The Difference-in-Means is a natural choice for inference of average effects. For sharp effects, the Difference-in-Means is especially suitable for sharp hypotheses that postulate a 1-dimensional, constant effect for all N experimental units. Such a constant effect is perhaps implausible and, hence, the posterior distribution will not converge in probability to the true causal effect, $\boldsymbol{\tau}$. However, with the Difference-in-Means test-statistic, the posterior distribution will converge in probability to the value of τ_h equal to the limiting value of $N^{-1} \sum_{i=1}^{N} \tau_i$. Since the mean minimizes the sum of squared Euclidean distances, the value of τ_h to which the posterior distribution converges in probability has an interpretation as the 1-dimensional causal effect that best approximates $\boldsymbol{\tau}$. This general line of reasoning offers a justification for inference of a 1-dimensional constant effect rather than the full N-dimensional vector of effects since inferences of the latter are often "so complex, so faithful to the minute detail of reality, that they are unintelligible" (Rosenbaum, 2010, p. 45).

Nevertheless, such a justification for inference about the 1-dimensional effect that best approximates the true N-dimensional effect, $\boldsymbol{\tau}$, presumes that "best approximation" is based on a sum of squared Euclidean distances metric. But this conception of "best approximation" may gloss over different ways in which treatments exert effects. For example, if there are uncommon, but dramatic responses to treatment (Rosenbaum, 2007), then inference about the 1-dimensional τ that minimizes the sum of squared Euclidean distances from $\boldsymbol{\tau}$ may poorly summarize the causal effect. Fortunately, inference about a 1-dimensional effect also warrants an interpretation as in-

ference about the maximum individual effect (Caughey et al., 2020). Such an interpretation may be of interest when researchers suspect uncommon but dramatic effects. If so, then a test-statistic like the Stephenson rank statistic (Stephenson, 1981; Stephenson and Ghosh, 1985) may be preferable even if, unlike the Difference-in-Means, the posterior distribution may not converge on the 1-dimensional effect equal to the true average effect. Expanding the theory of design-based Bayesian inference to other such test statistics is a potentially fruitful extension.

Insofar as our target of inference is the average effect, then we can conduct Bayesian inference of it via either weak or sharp causal hypotheses. The key difference between the two approaches is how each deals with the unknown variance of the Difference-in-Means. The first approach plugs in a conservative estimator for the variance while the second tests a causal effect strong enough to imply a value for the variance. Both approaches ensure that the posterior distribution converges on the true average effect. Which approach is preferable, however, will depend in part on which likelihood function contains more information. To assess the conditions under which either of these two approaches is favorable, we can mount an asymptotic comparison. The aim is to assess which approach is more likely to yield a posterior distribution with a lower mean squared error.

To facilitate a comparison, we can draw on a standard Normal asymptotic approximation to the likelihood function of the respective standardized Difference-in-Means test-statistics given by

$$\textbf{Weak causal hypothesis}: \frac{\hat{\bar{\tau}} - \bar{\tau}_h}{\sqrt{\widehat{\text{Var}}\left[\hat{\bar{\tau}}\right]}} \qquad \textbf{Sharp causal hypothesis}: \frac{t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\tau_h) - 0}{\sqrt{\text{Var}\left[t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\tau_h)\right]}}.$$

Proposition 3 provides the exact differences in the first two moments of the respective distributions of the Difference-in-Means under each approach for any N.

Proposition 3. *Let* $\bar{\tau}_h = \tau_h$. *For* $N = 4, 5, \ldots,$

$$t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\tau_h) - 0 - \left(\hat{\bar{\tau}} - \bar{\tau}_h\right) = 0 \text{ and} \tag{1.16}$$

$$\text{Var}[t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\tau_h)] - \widehat{\text{Var}}[\hat{\bar{\tau}}] = s^2_{y_T}\alpha_T + s^2_{y_C}\alpha_C + (N-1)^{-1}\left(\bar{\tilde{Y}}_T - \bar{Y}_C\right)^2, \text{ where} \tag{1.17}$$

$$\alpha_T = \left(n_C^{-1}\frac{N}{N-1}n_T^{-1}(n_T - 1) - n_T^{-1}\right) \text{ and}$$

$$\alpha_C = \left(n_T^{-1}\frac{N}{N-1}n_C^{-1}(n_C - 1) - n_C^{-1}\right).$$

For any $\bar{\tau}_h = \tau_h$, the numerator of the two standardized test-statistics will be equal. If $\bar{\tau}_h = \tau_h \neq \bar{\tau}$, then whichever test-statistic yields a lower variance will assign a lower likelihood (and, hence, all else equal, lower posterior probability density) to false average effects. The expression for the difference in variances in Equation (1.17) mirrors the Theorem 3 of Ding (2017), but differs in that Equation (1.17) is an exact expression for any N and pertains to any $\bar{\tau}_h$ and τ_h, not only that they are equal to 0. Analogous to Ding (2017, Theorem 3), we could also write the asymptotic difference in the two variances as

$$\text{Var}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] - \widehat{\text{Var}}[\hat{\bar{\tau}}] = \left(S_T^2 - S_C^2\right)\left(n_C^{-1} - n_T^{-1}\right) + N^{-1}\left(\bar{\tau} - \tau_h\right)^2 + o_p(1/N), \text{ where} \tag{1.18}$$

$o_p(1/N)$ is a quantity that converges in probability to 0 at rate $1/N$. Equation (1.18) differs from Equation (1.17) only in that the former, first, substitutes expectations of random quantities for the random quantities themselves and, second, ignores the differences between N and $N - 1$, n_T and $n_T - 1$, and n_C and $n_C - 1$.

As Proposition 3 shows, whichever procedure yields a smaller variance depends on three key factors: (1) the relative numbers of units in treatment and control conditions, (2) the variances of treated and control potential outcomes and (3) the difference between the true average effect and 1-dimensional sharp causal hypothesis. In the simulations below, I assess the implications of this difference in variances by comparing the mean squared error of posterior distributions over repeated random assignments. In particular, I begin with a standard Normal prior distribution and

then draw 100 treated and control potential outcomes under four scenarios described in Figure 1.3 below. In both cases, I use a standard Normal approximation to the likelihood function in which either a weak or sharp hypothesis assigns some probability density to the observed test-statistic. After forming the posterior distribution, I then assess its mean squared error with respect to the true average effect. I repeat this procedure over 1000 randomizations and calculate the mean squared error of the posterior distribution in each case. Figure 1.3 below plots the results.

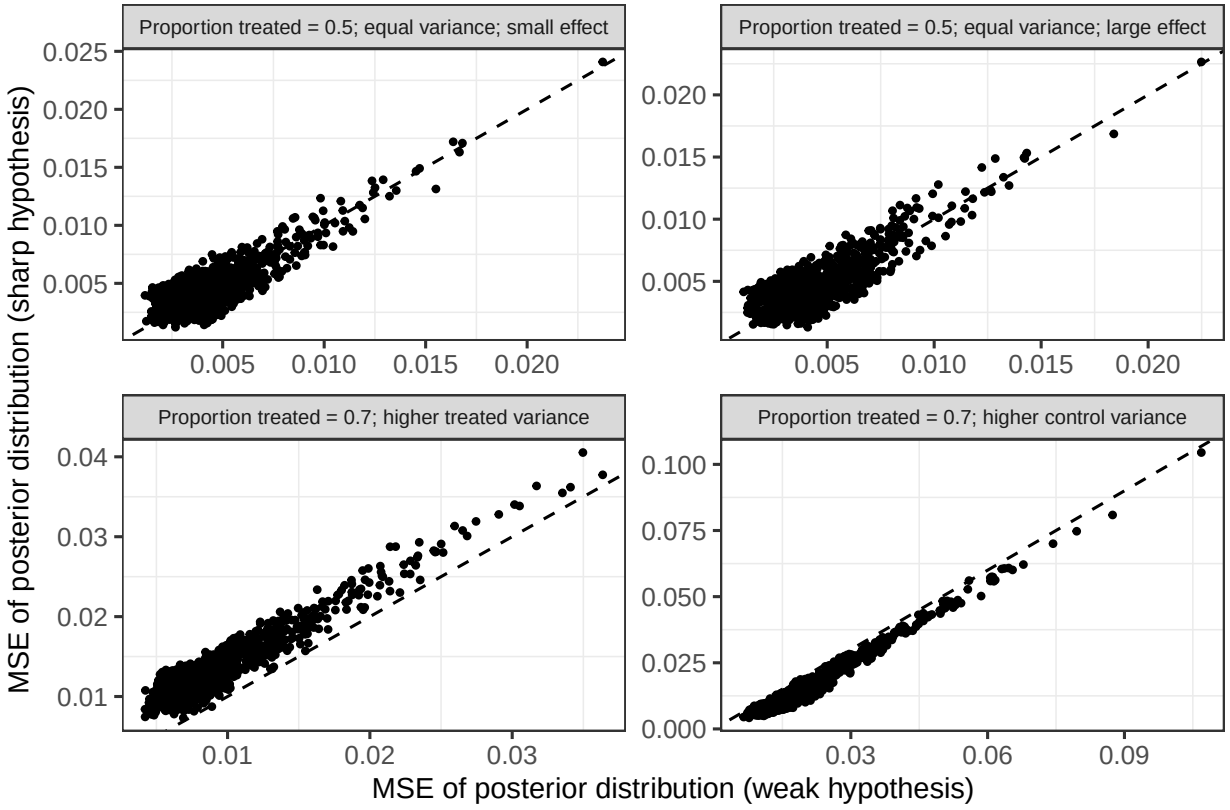

Figure 1.3: Mean Squared Errors (MSEs) of posterior distributions over repeated random assignments

Table B.1 in Chapter B provides the expected MSE under each of the four scenarios. In general, eliminating the variance nuisance parameter via plug-in estimation rather than via a sharp hypothesis performs better on average. This greater performance is especially pronounced when there is higher variance in treated potential outcomes (relative to control potential outcomes) and

more than half of the units are in the treatment condition. On the other hand, when more than half of the units are in the treatment condition, but the variance is higher among control potential outcomes, then eliminating the variance nuisance parameter via a sharp hypothesis performs better. In practice, scholars should be able to diagnose which scenario they are in by assessing the empirical variance of observed potential outcomes in an experiment. For example, an experimenter might erroneously expect higher variance in treated potential outcomes and thereby assign more units to the treatment condition. Yet, after conducting the experiment, such a scenario can be diagnosed by examining the variance of observed potential outcomes in each condition.

In addition, Figure 1.3, as well as the average MSEs in Table B.1, shows a lack of any substantial difference in posterior MSEs between each panel in the first row. That is, a larger true average effect does not, all else equal, appear to yield a worse posterior MSE. This result points to a key difference between Bayesian inference and null hypothesis significance testing.

To lay out this difference, we can refer back to Equation (1.18), given again below

$$\text{Var}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] - \widehat{\text{Var}}[\hat{\bar{\tau}}] = \left(S_T^2 - S_C^2\right)\left(n_C^{-1} - n_T^{-1}\right) + N^{-1}\left(\bar{\tau} - \tau_h\right)^2 + o_p(1/N),$$

and imagine that the true effect is constant for all units, which implies that $S_T^2 = S_C^2$, and that $n_C = n_T$. To square our focus solely on the role of the true average effect size, also assume that the size of the average effect grows proportional to N such that, asymptotically, the difference in variances for some τ_h is driven solely by the size of the average effect: $\text{Var}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] - \widehat{\text{Var}}[\hat{\bar{\tau}}] = N^{-1}\left(\bar{\tau} - \tau_h\right)^2$. In the context of null hypothesis significance testing, plug-in variance estimation will have greater relative performance in terms of rejecting the null hypothesis that $\tau_h = 0$ as the true average effect, $\bar{\tau}$, increases.

Yet this greater relative performance in terms of evidence against $\tau_h = 0$ does not necessarily translate to a Bayesian framework. When the true average effect increases, the relative performance of plug-in estimation increases for small causal hypotheses, like $\tau_h = 0$, but decreases for large causal hypotheses closer to the true average effect. Thus, the performance of each variance method

with respect to $\tau_h = 0$ may change in isolation, but not relative to other hypothetical causal effects. A posterior distribution, in effect, lines up each hypothesis and assesses their *relative* plausibility, while p-values pertain to each causal hypothesis in *isolation*. Factors, such as the size of the true average effect, that affect relative performance in terms of the latter measure do not necessarily do so in terms of the former.

Overall, eliminating the variance parameter via plug-in estimation performs better, on average, than does eliminating the variance parameter via a sharp hypothesis. Despite this difference in performance, Bayesian inference of sharp causal hypothesis might still be preferable. A prior distribution on a sharp causal hypothesis implies a prior distribution on both the mean and variance of the likelihood function. By contrast, a prior distribution on a weak causal hypothesis implies a prior distribution on only the likelihood function's mean, not its variance. In principle, one could conduct fully Bayesian inference with a marginal prior distribution on the average effect and another on the variance of the Difference-in-Means. But since the latter parameter is unlikely to be of intrinsic interest, doing so would complicate inference with little valued added. Thus, Bayesian inference of sharp as opposed to weak causal effects has the benefit of coherence in its treatment of both the mean and variance parameters of the likelihood function.

1.6 Conclusion

In their critique of randomized experiments, Deaton and Cartwright (2018, p. 3) emphasized the value of "understanding how the results from RCTs [randomized controlled trials] relate to the knowledge that you already possess about the world." The logic of null hypothesis significance testing may not map neatly on to prior knowledge that scholars possess. In some cases, null experimental results may conflict with scholars' prior expectations and contribute to learning about causal effects in a given substantive domain. Conversely, significant results might merely confirm what scholars already agree upon due to the results of prior studies. Hence, scholars may be interested not solely in whether results are statistically significant, but also in how much one actually learns from a new experimental finding.

This paper open up new possibilities for scholars to quantify the extent of Bayesian learning from randomized experiments. A common concern with extant Bayesian methods (in experiments or otherwise) is that they assume randomness in potential outcomes, which conflicts with what scholars like about experiments — namely, that randomization alone constitutes the "reasoned basis" for inference. In this paper, I have shown that, if the likelihood function conditions *not* on the full data, but a suitable function of them, then Bayesian inference of either weak (average) or sharp causal effects can be justified by the experimental design. Such Bayesian inference reliably tracks the true causal effect of interest without any further assumptions than those made in classical randomized experiments.

Chapter 2: Turning Past Experiments into Priors for Design-based Bayesian Learning: Application to Audit Experiment on Racial Responsiveness

2.1 Introduction

In this chapter, I examine the degree of Bayesian learning that results from a new audit experiment on the role of racial discrimination in the responsiveness of state legislators to putative constituents. This experiment was conducted in 2020 on the email addresses of $5,925$ state representatives in 49 US states (excluding Nebraska due to its unicameral and nonpartisan legislature). The experiment randomly assigns these $5,925$ email addresses to a request for constituency service from either a White, Black or Latino alias. The content of the constituency service email was randomly chosen from three email types: an inquiry about internship opportunities, a request for information about how to get more involved in politics, and a question about campaign work. Within each email type, the email's wording was randomly selected from three possible scripts. All emails, regardless of the type or script, signal that the putative constituent belongs to the same political party as that of the legislator. The outcome of interest is whether a legislator (or staff member who is in charge of the legislator's official email account) responds to the constituent service request.

The question of Bayesian learning from randomized experiments is especially relevant for this application. Ample evidence from past audit experiments shows that state legislators discriminate against putatively Black (Butler and Broockman, 2011; Butler, 2014) and Latino (Mendez and Grose, 2018; Mendez, 2018; Wong, Nicholson, and Lajevardi, 2017) constituents relative to putatively White constituents. This abundance of evidence has led some scholars to claim that little additional knowledge can be generated from new audit experiments by themselves (Gaddis,

2019).[1] A new randomized experiment — specifically another audit experiment — can certainly uncover additional causal facts; however, its actual contribution to learning relative to scholars' baseline knowledge may be minimal. This concern is especially salient given critiques of randomized experiments to this effect (Cartwright and Hardie, 2012; Deaton, 2009; Deaton, 2010; Deaton and Cartwright, 2018; Harrison, 2011; Harrison, 2014; Heckman, 1992; Heckman, 2020; Ravallion, 2009; Ravallion, 2020), as well as scholars' increasing emphasis on understanding both randomized and nonrandomized studies within a framework of Bayesian learning (Pritchett and Sandefur, 2015; Little and Pepinsky, 2021; Gerber, Green, and Kaplan, 2004; Vivalt, 2020; Imbens, 2021; Dunning et al., 2019).

In spite of — in fact *because* of — concerns about how much one learns from new experiments (of which audit experiments are a particularly "tough case"), I develop a methodology for Bayesian learning from randomized experiments. In particular, I develop methods through which scholars can quantify (1) the state of prior knowledge based on past experimental results and (2) the extent of Bayesian learning that occurs from before to after an experiment. While the prior distribution is constructed through the tools (but not necessarily the assumptions) of model-based inference, the likelihood function — explained in greater detail elsewhere in this dissertation — is design-based. It invokes assumptions no stronger than usual methods for analyzing randomized experiments.

An answer to the question of what we learn from a new randomized experiment depends first and foremost on prior knowledge going into the experiment. The informativeness of past experiments about effects in a new experiment depends on the experiments' distributional differences in variables that explain effect heterogeneity (Coppock, Leeper, and Mullinix, 2018; Mullinix et al., 2015). I therefore repurpose methods from the target validity literature (Westreich et al., 2019) to generate a stochastic distribution of the average effect one would expect to see in a new experiment given results of past experiments. Within the target validity literature, I draw specifically on models of effect heterogeneity (Kern et al., 2016; Nguyen et al., 2017), which I complement

[1] Gaddis (2019, p. 443) writes that "as the use of the audit method to examine racial-ethnic discrimination has increased over the last decade, the amount and diversity of knowledge that can be gained from new audits alone have decreased."

with both full matching (Hansen, 2004; Rosenbaum, 1991; Sävje, Higgins, and Sekhon, 2021) and weighting (Miratrix et al., 2018). Constructing a prior distribution is fundamentally different from generalizing or transporting an estimated effect from one study to another. Hence, in drawing on tools from target validity, I prioritize the transparent and intuitive representation of prior knowledge, *not*, say, consistent or efficient estimation of causal effects in a target study. In short, the procedures one uses to generate a prior distribution do not need to be "right," so to speak; rather, they ought to clearly, intuitively and (ideally) simply represent the state of prior knowledge based on past experimental results.

Having developed a method for constructing a prior distribution, I then turn to quantifying the extent of learning based on differences between the prior and posterior distributions of the target experiment. Drawing on value of information theory, as explicated in foundational Bayesian decision theory texts (Raiffa and Schlaifer, 1961) (see also Howard, 1966), I propose two statistics that quantify what I (and others, e.g., Little and Pepinsky, 2021) refer to as first- and second-moment Bayesian learning. I then develop a standardized test statistic between 0 and 1 for the extent of Bayesian learning from an experiment. This statistic's benchmark for an experiment's degree of Bayesian learning is the amount of learning that would result in the absence of any prior information whatsoever. Such an assumption of prior ignorance is conspicuously ill-suited for audit experiments on legislative responsiveness. Nevertheless, I show that such an assumption is implicit under a Bayesian interpretation of usual, non-Bayesian analyses of randomized experiments.

In the case of the 2020 audit experiment, I show that it contributes to substantial Bayesian learning in the context of prior information from past experiments. In particular, the experiment shows that, contrary to what one would expect from past experiments, legislators discriminate, on average, against putatively Latino constituents in favor of putatively Black constituents. The most likely percentage of legislators who would respond to a Black constituent, but not a Latino constituent, changes from 0.46% before the experiment to 3% after. In addition, after this experiment, the extent of legislators' discrimination against Black in favor of White constituents is revised downwards due to the lack of any experimental evidence for such discrimination. In an alternative

analysis, such a result might be ignored since it does not pass the threshold of statistical significance; in a Bayesian context, however, insignificant results can nevertheless yield meaningful degrees of learning. Overall, the greatest amount of Bayesian learning occurs among subgroups, particularly those that are typically too small to draw inferences about from a single experiment analyzed in isolation. The 2020 audit experiment yields greatest learning about discrimination by Black and Latino legislators between putatively Black and Latino constituents. From this experiment, we learn that Black legislators discriminate against Latino constituents (*standardized Bayesian learning statistic* of 0.22), but Latino legislators don't discriminate against Black constituents (*standardized Bayesian learning statistic* of 0.17). Learning about these subgroup effects carries implications for theories of descriptive representation — specifically, how it operates not only within, but between minority groups.

Section 2.2 to follow motivates the ensuing methodological developments by, first, providing a standard, non-Bayesian analysis of the 2020 audit experiment and, second, showing that this analysis implicitly assumes an absence of prior information. Section 2.3 then focuses on how to construct a prior distribution that represents information from past experiments. Section 2.4 subsequently turns to quantifying learning that occurs from before to after a target experiment. Section 2.5 then provides a design-based Bayesian analysis of the 2020 audit experiment before the succeeding section provides an overall discussion and conclusion.

2.2 Audit experiment

Audit experiments have become a staple method of studies that assess whether politicians racially discriminate among their constituents. In one of the seminal audit experiments, conducted among US state legislators in 2008, Butler and Broockman (2011) estimate that 5.1% of legislators would respond to a service request from a White constituent, but not a Black one. Subsequent studies have found even more pronounced patterns of discrimination. In a meta-analysis of 41 published and unpublished audit experiments between 2011 and 2016, Costa (2017) estimates that, compared to putatively White constituents, putatively Latino and Black constituents are 14.2% and

7.3% less likely to receive responses to their inquiries from politicians.

Despite this extensive information from past audit experiments, methods for the analysis of new experiments are unable to incorporate such information. For example, a standard analysis of the 2020 audit experiment described in the introduction would unbiasedly estimate an average effect and either reject or fail to reject the null hypothesis of no average effect. While such an analysis has many valuable properties, it offers little insight into how much more information a new experiment contributes to what scholars already know from past experiments. To further unpack this point, consider the following preliminary analysis of the 2020 audit experiment described in the introduction. I use this preliminary analysis as motivation for the subsequent methodological argument.

In the 2020 audit experiment, there are six initial causal targets of interest, which pertain to six different treatment contrasts. One causal target is the mean difference in state legislators' responses if they were contacted by a putatively Black constituent compared to if they were contacted by a putatively White constituent. Other targets of interest include the same average causal effect for different treatment contrasts among the Black, White and Latino treatment conditions. Table 2.1 below provides the Difference-in-Means and Neyman's conservative standard error (Neyman, 1923) estimates for all six of these contrasts.

	Outcome: Reply from legislator's email address
Treatment contrast	**Diff-in-Means (SE)**
Black (1) vs White (0)	0.0212
	(0.0154)
Latino (1) vs White (0)	−0.0221
	(0.0149)
Latino (1) vs Black (0)	−0.0433***
	(0.0149)
Black or Latino (1) vs White (0)	−0.0011
	(0.0132)
Latino (1) vs Black or White (0)	−0.0328***
	(0.0128)
Latino or White (1) vs Black (0)	−0.0327***
	(0.0132)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2.1: Overall results of 2020 audit experiment

The usual point estimates and confidence sets implied by the results in Table 2.1 fail to incorporate information from past audit experiments. In fact, the results in Table 2.1 are equivalent to the Bayesian inferences one would draw with a uniform prior distribution (representing a lack of any information whatsoever) on the average causal effect.

To see this point, note that, as Gerber, Green, and Kaplan (2004, p. 253) allude to, the finite population CLT (see Li and Ding, 2017) justifies the use of a Normal likelihood function in which possible realizations of the experimental data are summarized by the Difference-in-Means. The Difference-in-Means is essentially a random draw from a Normal distribution with unknown mean equal to the true average effect and variance that is unknown, but can be conservatively estimated (Neyman, 1923). With this Normal likelihood function and a uniform prior on the average effect, which is naturally bounded between -1 and 1, the resulting posterior distribution after observing experimental data is a truncated Normal. This posterior distribution's mean is equal to the realized

Difference-in-Means and the distribution's variance is equal to the realized conservative variance estimate. Therefore, the mean causal effect with the greatest posterior density, the so-called maximum a posteriori (MAP) estimate, will be equal to the observed Difference-in-Means. Likewise, the $(1-\alpha)100\%$ credible set with $\alpha \in (0, 1)$ will be equal to the analogous $(1-\alpha)100\%$ confidence set.

Figure 2.1 below shows the $(1-\alpha)100\%$ *confidence* sets with $\alpha = 0.05$ overlaid on top of the *posterior* distributions that result from a uniform prior for each of the six treatment contrasts in Table 2.1.

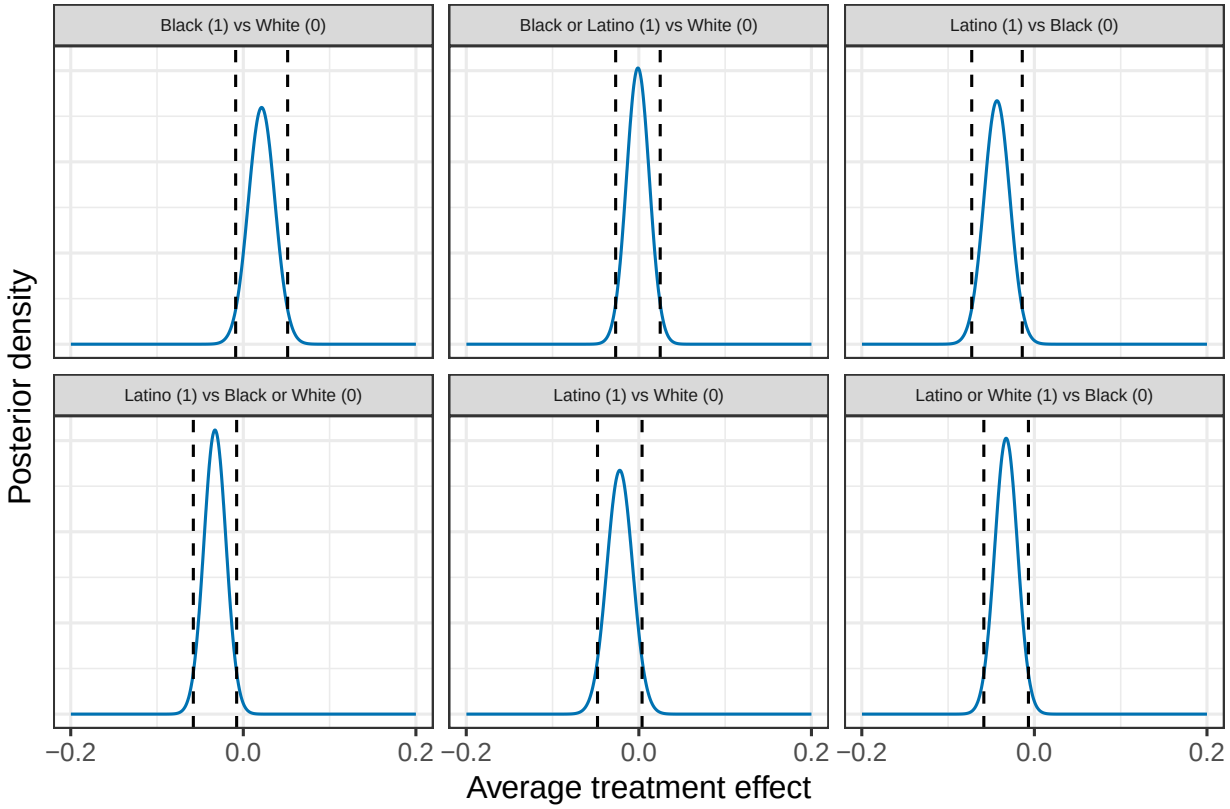

Figure 2.1: Posterior distributions of treatment contrasts from Table 2.1

The blue lines in the panels of Figure 2.1 are the posterior distributions of the average causal effect in the 2020 audit experiment. The dashed lines in each of the panels are the espective $(1-\alpha)100\%$ confidence sets with $\alpha = 0.05$. We can see that these two-sided 95% confidence intervals are

located exactly at the 2.5 and 97.5 percentiles of the posterior distribution. In addition, the MAP estimates in each of the panels are identical to the respective Difference-in-Means estimates.

This preliminary analysis underscores that the usual methods for analyzing randomized experiments assume that we have no prior information. At least for this particular audit experiment, such an assumption is erroneous given the extensive number of past audit experiments about public officials' responsiveness to constituents of different ethnoracial identities. Insofar as we possess prior knowledge coming into an experiment, then the usual practice of estimating average causal effects and Normal approximation-based confidence sets will *not* capture what we learn from randomized experiments. If we are to quantify what we learn from a new experiment relative to baseline knowledge from past experiments, then we need a methodology that turns information from past experiments into prior distributions for target experiments.

2.3 Constructing a prior distribution from past experimental results

2.3.1 Review of target validity methods

To construct a prior distribution from past experimental results, I draw on tools from the target validity literature. Target validity, as described in Westreich et al. (2019), pertains to two types of inference: One is generalizability, which refers to when a set of units under study is a proper subset of a target population of interest; the other is transportability (Bareinboim and Pearl, 2013; Pearl and Bareinboim, 2014), which refers to when units under study and the target population are disjoint sets, i.e., study participants are not themselves members of the target population.

For both generalizability and transportability, scholars have drawn upon weighting (Cole and Stuart, 2010; Stuart et al., 2011; Olsen et al., 2013), stratification (Tipton, 2013; Tipton et al., 2014; O'Muircheartaigh and Hedges, 2014), outcome modeling (Kern et al., 2016; Nguyen et al., 2017) or combinations of weighting and outcome modeling via "doubly robust" (Egami and Hartman, 2021b; Dahabreh, Robertson, and Hernán, 2019) or targeted maximum likelihood (Rudolph et al., 2014; Rudolph and Laan, 2017) estimation. I briefly review weighting, stratification and outcome modeling for target validity since the methodology I propose draws most heavily from these three

approaches.

Broadly speaking, weighting estimators for target validity generalize or transport experimental results via adjustment for units' study participation probabilities. Cole and Stuart (2010) propose an inverse probability weighted (IPW) estimator for generalizing effects from a study population to a target population from which the study's subjects were presumably sampled. Westreich et al. (2017) show that this estimator extends to the context of transportability in which one need not assume that the units in the study are a subset of the target population.[2] Tipton (2013) and Tipton et al. (2014) also estimate units' study participation probabilities, but generalize to a target population by constructing strata of study and target units that have similar estimated study participation probabilities. Occupying a liminal space between weighting and outcome modeling approaches, Josey et al. (2021) draw on entropy balancing (Hainmueller, 2012) and Lu et al. (2021) draw on approximate balancing (Zubizarreta, 2015) to construct weights that balance covariates between a study and target population. Such weights aim to minimize covariate imbalance and need not reflect units' study participation probabilities.

Outcome modeling approaches to target validity aim to directly model and adjust for variables that explain effect heterogeneity (Kern et al., 2016; Nguyen et al., 2017). In principle, if one could exactly stratify units in a study and target population on the values of all causal moderators, then it would be straightforward to transport effects between units within the same stratum. Alternatively, if one knew or could reliably predict how average potential outcomes vary across different values of baseline covariates, then one could transport causal effects by predicting them in a target population via a model of effect heterogeneity. Kern et al. (2016) employ this modeling approach using linear regression, as well as Bayesian additive regression trees (BART) (Hill, 2011; Green and Kern, 2012). Other models are also possible, such as those from the machine learning literature that have been used to estimate heterogeneous effects (Wager and Athey, 2018; Imai and Ratkovic, 2013; Grimmer, Messing, and Westwood, 2017).

[2]The IPW estimator is an extension of the classic Horvitz-Thompson estimator from survey sampling (Horvitz and Thompson, 1952). Other weighting estimators from the observational studies and survey sampling literatures are also possible (and typically preferable), e.g., the Hájek (Hájek, 1960) and Des Raj (Raj, 1965) estimators among others.

Assessments of the target validity methods thus described largely focus on the conditions under which estimators are unbiased or consistent for causal quantities in a target population. If an outcome model of average potential outcomes given baseline covariates is correctly specified, then an outcome modeling estimator is consistent for the average causal effect in the target experiment. Alternatively, if the model of study participation probabilities is correctly specified, then the IPW estimator is consistent for the target average causal effect.[3] By contrast, estimators that weight for covariate balance typically depend on an assumption that the true, stochastic outcome model lies in a particular class of models. Scholars then typically construct weights that minimize some function of the chosen estimator's error, such as the maximum squared error, under this assumption. Such assumptions are sensible if one is to establish desirable properties (e.g., consistency) of effect generalization estimators. However, the aim of constructing a prior distribution for a target experiment based on past experimental results is fundamentally different.

Importantly, the assumptions of the methods I draw upon — i.e., weighting, stratification and outcome modeling — need not be "right." Going forward, potential outcomes are not assumed to be embedded in an underlying probability model (or class of models). Outcome models in the methodology to follow serve as predictive algorithms of fixed potential outcomes (**rosenbaum2002**; see Sales, Hansen, and Rowan, 2018). The primary concern of these models is not whether their assumptions are correct; instead, such models ought to transparently and intuitively predict the average causal effect one would expect to see in a target experiment given the results of past experiments. Such predictions, which, in turn, constitute a prior distribution for the target experiment's average effect, might be wrong. Yet such erroneous beliefs will be revised upon actually conducting the target experiment. A comparison of posterior and prior beliefs in a target experiment therefore reflects how much one learns from a new experimental result relative to what one would expect the effect to be given past experiments.

[3]The doubly robust estimator (Robins, Rotnitzky, and Zhao, 1994; Kang and Schafer, 2007) will be consistent when the model of either the conditional average treatment effect or the study participation probabilities are correctly specified.

2.3.2 Setup

To more rigorously lay out how one constructs such a prior distribution, consider a setting with potentially multiple past experiments and one target experiment of interest. Concatenate the prior and target experiments and let the index $i = 1, \ldots, N$ run over the N total units. Denote the collection of indices for prior experimental units by $\mathcal{P}$ and for target experimental units by $\mathcal{T}$. Let the sets $\mathcal{P}_1 \subset \mathcal{P}$ and $\mathcal{P}_0 \subset \mathcal{P}$ contain the indices for treated and control prior units, respectively, where $\mathcal{P}_1 \cup \mathcal{P}_0 = \mathcal{P}$. The cardinality (i.e., the number of elements in) each respective set are denoted by $|\mathcal{P}_1| = N_{\mathcal{P}_1}$, $|\mathcal{P}_0| = N_{\mathcal{P}_0}$, $|\mathcal{P}| = N_{\mathcal{P}}$ and $|\mathcal{T}| = N_{\mathcal{T}}$.

Under the usual stable unit treatment value assumption (SUTVA), each of the $i = 1, \ldots, N$ units has one potential outcome, y_{zi}, for each treatment condition, $Z_i = z$, to which the ith unit could be assigned. The observed outcome is equal to the potential outcome corresponding to the condition to which the ith unit is assigned: $Y_i = \sum_z \mathbb{1}\{Z_i = z\} y_{zi}$, where $\mathbb{1}\{\cdot\}$ is the indicator function that is equal to 1 when its argument is true and 0 otherwise. The treatment and outcome variables are observed for units in prior experiments, but not for units in the target experiment. The specific causal quantity of interest is the average treatment effect in the target experiment, i.e., the target average treatment effect (TATE). For a specific treatment contrast, e.g., z versus z', the TATE is formally defined as

$$\tau_{\mathcal{T}} \coloneqq N_{\mathcal{T}}^{-1} \sum_{i \in \mathcal{T}} (y_{zi} - y_{z'i}) \,. \tag{2.1}$$

Going forward, I refer to a binary treatment variable in which $z = 1$ or $z = 0$. In the 2020 audit experiment, there are 3 experimental conditions and 6 different binary contrasts of these 3 conditions. It should be clear from context the specific contrast to which the binary treatment variable refers.

Denote the Difference-in-Means estimator of the TATE in the target experiment by $\hat{\tau}_{\mathcal{T}}$. The finite population CLT and associated theory imply that, in experiments of at least moderate size, the Difference-in-Means is approximately Normal, i.e., $\hat{\tau}_{\mathcal{T}} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mathrm{E}\left[\hat{\tau}_{\mathcal{T}}\right], \mathrm{Var}\left[\hat{\tau}_{\mathcal{T}}\right]\right)$. Under random assignment, $\mathrm{E}\left[\hat{\tau}_{\mathcal{T}}\right] = \tau_{\mathcal{T}}$, which justifies the use of a Normal likelihood function in which

a hypothetical value of $\tau_{\mathcal{T}}$ (along with a conservative plug-in estimator of the variance) assigns probability density to the observed experimental data summarized by the Difference-in-Means. Also define a Normal prior distribution on the TATE, $\tau_{\mathcal{T}} \sim \mathcal{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$, where μ_{prior} and σ^2_{prior} are the hyperparameters.[4] With this Normal prior distribution on the TATE, as well as the Normal likelihood function, inference can proceed via Bayes' rule. Scholars are unlikely to have well motivated prior beliefs about the nuisance parameter $\text{Var}\,[\hat{\tau}_{\mathcal{T}}]$; hence, use of a Normal likelihood function that substitutes for $\text{Var}\,[\hat{\tau}_{\mathcal{T}}]$ its conservative estimator enables scholars to define a prior distribution and to conduct inference on only the TATE.

The aim in the immediately succeeding sections is to calibrate the parameters of the Normal prior distribution. In particular, the prior distribution ought to encode information about the average effect in the target experiment based on results from past experiments. To do so, let $\boldsymbol{x}_i$ denote a K-dimensional vector of values for K baseline variables measured in both prior and target experiments. Each of the K baseline variables are thought to predict effect heterogeneity; hence, I refer to them as *heterogeneity variables* (see also Egami and Hartman, 2021a). To lay out what exactly it means for a baseline variable to predict effect heterogeneity, let $\hat{y}_z(\cdot)$ denote an algorithmic model that, when supplied with heterogeneity variables, $\boldsymbol{x}_i$, generates predictions of the ith unit's potential outcome, y_{zi}. Then denote the conditional average treatment effect (CATE) predictive function for the ith unit as

$$\textbf{CATE predictive function}: \ \hat{\tau}_{\mathcal{T}}(\boldsymbol{x}_i) \coloneqq \hat{y}_z(\boldsymbol{x}_i) - \hat{y}_{z'}(\boldsymbol{x}_i). \tag{2.2}$$

The CATE function in Equation (2.2) returns a prediction of the causal effect for unit i as a function of the vector of variables thought to explain effect heterogeneity, $\boldsymbol{x}_i$.

The predictive algorithm in Equation (2.2) can have high or low accuracy. Without embedding potential outcomes in a stochastic model or class of models, the squared predictive accuracy of the

[4]This distributional choice is justified as a conservative one due to the Normal distribution's property of maximum entropy among distributions with finite variance (see Cover and Thomas, 1991). For a formal proof of this property, see McElreath (2020, pp. 306–307), which is based on the treatment in Conrad (2005).

CATE function for the TATE is given by

$$\textbf{Predictive accuracy}: \; N_{\mathcal{T}}^{-1} \sum_{i \in \mathcal{T}} \left[y_{zi} - \hat{y}_z(\boldsymbol{x}_i) - (y_{z'i} - \hat{y}_{z'}(\boldsymbol{x}_i)) \right]^2, \tag{2.3}$$

which is unknowable due to its dependence on unobservable potential outcomes in the target population. As Equation (2.3) shows, predictive accuracy depends not on a correct outcome model, but rather boils down to the average squared distance between the true (but unobservable) potential outcomes and their predictions.

The CATE predictive function as explained thus far generates only deterministic predictions of potential outcomes given units' fixed values of heterogeneity variables. To calibrate the prior distribution of the TATE to results from past experiments, I use the following procedure:

1. Fit a CATE predictive model to prior experimental data and (assuming the model matrix is of full rank) set the parameters of the model to the values that minimize the prediction error of observed potential outcomes in the prior experimental data.

2. Having fit the model to prior experimental data, feed baseline data from the target experiment to the fitted model in order to generate a prediction of the individual effect for each unit in the target experiment.

3. Set the hyperparameters of the Normal prior distribution of the TATE as follows:

   (a) Set μ_{prior} to the average of target units' predicted effects:

   $$\mu_{\text{prior}} \leftarrow N_{\mathcal{T}}^{-1} \sum_{i \in \mathcal{T}} \hat{\tau}_{\mathcal{T}}(\boldsymbol{x}_i) \text{ and} \tag{2.4}$$

   (b) Set σ^2_{prior} to the variance of the average predictive effect implied by the estimated variance-covariance matrix of the predictive model fit to prior data:

   $$\sigma^2_{\text{prior}} \leftarrow \frac{1}{N_{\mathcal{P}}^2} \sum_{i \in \mathcal{P}} v_i^2 + 2 \sum_{i \in \mathcal{P}} \sum_{i<j} v_{ij}, \tag{2.5}$$

where ν_i and ν_{ij} are the diagonal and off-diagonal entries of the the variance-covariance matrix for the predicted values, where variance-covariance matrix estimated on prior data with robust (heteroskedasticity consistent) standard errors (Eicker, 1963; Eicker, 1967; White, 1980).

Note that in this procedure, both prior and target experimental data are fixed. The randomness stems from the Normal prior distribution on the parameter vector of the CATE in which the hyperparameters of the prior distribution are set to the empirical means and the empirical variance-covariance matrix. The result of this procedure is a Normal prior distribution of the TATE in which prior uncertainty is based on how well heterogeneity variables predict observed potential outcomes in past experiments. Intuitively, this empirical Bayes' procedure is sensible because it sets the most likely predicted average effect to that which best fits the prior data. Nevertheless, uncertainty in this predicated average effect remains. Overall uncertainty in predictions of the TATE depends on the accuracy of predictions from the CATE model, on average, in the prior data. The most plausible predicted TATE will be based on which parameter values of the CATE model best fit the prior data and the next most plausible values of the TATE will be based on the CATE model's parameters that do not minimize the sum of squared residuals, but yield the next best fit. Therefore, this procedure intuitively calibrates uncertainty about the TATE based on predictive accuracy of the CATE model in prior experiments.

Implementing this procedure in practice requires researchers to make specific choices about the functional form of the CATE predictive algorithm. In general, researchers ought to have flexibility in the specific algorithm they choose, which will likely depend on domain knowledge of the application at hand. Nevertheless, several common principles related to model dependence and the reliability of each prior observation are important regardless of the predictive algorithm one chooses. I now turn to how scholars can satisfy these principles by both stratification (via matching) and weighting in the context of a baseline linear projection model.

2.3.3 Complementing the CATE model with matching and weighting

Given that the CATE function is a predictive algorithm, not a specification of a true, underlying probability model, one would want to avoid an overdependence on the (often arbitrary) functional form assumptions of the CATE predictive algorithm (Ho et al., 2007; King and Zeng, 2006; King and Zeng, 2007). In other words, one would want to avoid generating predictions that depend upon either interpolation or extrapolation of the CATE algorithm's functional form to covariate regions in which the data are sparse. To alleviate this concern, I propose, first, matching prior and target units based on their similarity in heterogeneity variables before fitting a weighted model that minimizes the sum of within stratum squared prediction errors.

Matching for common support

The CATE model is fit to only prior data; hence, the first desirable feature is *internal common support* within the prior data, which refers to covariate regions of the prior data that have a sufficiently large number of observations and variation in the treatment variable. With high internal common support, the parameters of the CATE function based on the model fit to prior data will be more credible. Nevertheless, predictions of the TATE based on the CATE model fit to prior data inevitably involve extrapolation to the target data. Therefore, a second desirable feature is *external common support* between the prior and target experiments, which refers to covariate regions across both prior and target experiments with a large number of observations and variation in whether they belong to the prior or target experiments. Matching units to optimize balance between treatment conditions in prior experiments and between prior and target units addresses both concerns of internal and external common support.

Following Sävje, Higgins, and Sekhon (2021), let a matched set, $\boldsymbol{m}$, be a nonempty set of unit indices and let a matching, $\boldsymbol{M}$, be a set of disjoint matched sets, $\boldsymbol{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots\}$. Sävje, Higgins, and Sekhon (2021) and Higgins, Sävje, and Sekhon (2016) provide a near-optimal algorithm for constructing a matching that, subject to user specified constraints, minimizes the maximum value of a distance metric between any two units in the same matched set. Although driven in

part by computational tractability, this choice of objective function (minimizing the maximum within-set distance) has intuitive appeal. It allows a user to specify a maximum distance whereby any greater distance between any two units would render them incomparable due to lack of common support and related concerns of model dependence. Subject to this user-specified constraint, the algorithm of Sävje, Higgins, and Sekhon (2021) and Higgins, Sävje, and Sekhon (2016) can then find the matching that minimizes the maximum distance between any two units in the same matched set. To make this algorithm relevant for turning past experiments into a prior distribution, we need to define a relevant distance metric and a set of constraints that a matching, $\boldsymbol{M}$, ought to satisfy.

For the former task, denote the *heterogeneity score* of unit i by $h_i = \varphi(\boldsymbol{x}_i)$, where $\varphi(\cdot)$ is a function that predicts membership in the prior or target experiments as a function of heterogeneity variables. A natural choice for $\varphi(\cdot)$ is a logistic regression of an indicator of membership in the prior or target experiments on the set of heterogeneity variables. Other link functions — e.g., probit, cauchy and complementary log-log among others — are also possible. Importantly, the heterogeneity score is *not* meant to reflect the predicted probabilities of participation in either the prior or target experiments. Instead, $\varphi(\cdot)$ serves primarily to collapse the K-dimensional vector of baseline covariates of each unit into a single scalar value, $\varphi : \mathbb{R}^K \mapsto \mathbb{R}$. Once we've constructed heterogeneity scores for all N units, we can then calculate a unidimensional distance metric between any two units given by $d_{ij} = \sqrt{(h_i - h_j)^2}$, which is the absolute distance in heterogeneity scores between units i and j.

The set of constraints that a matching, $\boldsymbol{M}$, out to satisfy in this context are as follows. First, in order to be as transparent as possible, each unit can belong to at most one matched set. Second, each matched set ought to contain at least one unit from the treatment and control conditions in the prior experiments and at least one unit from the target experiment. Third, since the effect of interest is the TATE, the matching should not exclude any target units. Finally, as mentioned above, two units in the same matched set ought not to exceed a user-specified distance in heterogeneity

scores beyond which the two units are incomparable.[5] The optimal matching problem can thus be represented as

$$\underset{\boldsymbol{M}}{\text{minimize}} \quad \max_{\boldsymbol{m} \in \boldsymbol{M}} \max \left\{ d_{ij} : i, j \in \boldsymbol{m} \right\} \tag{2.6a}$$

subject to

$$\text{For all } \boldsymbol{m}, \boldsymbol{m}' \in \boldsymbol{M}, \text{ if } \boldsymbol{m} \neq \boldsymbol{m}', \text{ then } \boldsymbol{m} \cap \boldsymbol{m}' = \emptyset, \tag{2.6b}$$

$$\text{For all } \boldsymbol{m} \in \boldsymbol{M} \text{ and } z \in \{0, 1\}, \; |\boldsymbol{m} \cap \mathcal{P}_z| \geq 1, \tag{2.6c}$$

$$\text{For all } \boldsymbol{m} \in \boldsymbol{M}, \; |\boldsymbol{m} \cap \mathcal{T}| \geq 1, \tag{2.6d}$$

$$\sum_{\boldsymbol{m} \in \boldsymbol{M}} |\boldsymbol{m} \cap \mathcal{T}| = |\mathcal{T}|, \tag{2.6e}$$

$$\max_{\boldsymbol{m} \in \boldsymbol{M}} \max \left\{ d_{ij} : i, j \in \boldsymbol{m} \right\} \leq c, \tag{2.6f}$$

$$0 \leq c < \infty. \tag{2.6g}$$

Rosenbaum (1991) and Hansen (2004) show that in the context of bipartite matching, the optimal solution that reduces the sum of all treatment-control, within-stratum distances will consist of one treated unit matched to potentially multiple control units, or vice versa. The near-optimal algorithm for the different objective function in Equation (2.6a) from Sävje, Higgins, and Sekhon (2021) and Higgins, Sävje, and Sekhon (2016) will not necessarily result in a matching with this structure. However, in more complex setups beyond a single binary treatment variable and with a large number of units, this near-optimal algorithm is preferable. Upon constructing matched sets of units with similar in heterogeneity variables, it is then straightforward to fit the CATE model within these matched sets before aggregating over all of them, e.g., by including fixed effects for matched strata in the CATE model.

[5] If calipers on other distance measures, such as the Euclidean distance on a single important heterogeneity variable, are crucial in substantive applications, such additional constraints can be incorporated.

Weighting for the reliability and importance of prior observations

Even after matching, differences within matched sets on the heterogeneity score are likely to remain. Prior units that are more similar to target units presumably generate more reliable predictions of target units' effects. To reflect this greater reliability, we can upweight prior units that exhibit greater similarity to target units in the same stratum. Furthermore, in addition to the reliability of prior observations, the importance of each prior observation depends on the number of target units to which each prior unit can be compared. That is, if one prior unit is similar to one target unit, but another prior unit is similar to five target units, then the latter prior unit is more important than the former. In what follows, I propose weights for the CATE model that aims to intuitively and transparently account for the twin aims of representing the reliability and importance of prior observations.

To lay out this weighting scheme, first let the index $m = 1, \ldots, M$ run over the M matched sets and let $s_i = m$ indicate the matched set to which unit i belongs. Then write the number of prior control, prior treated and target units in matched set m as

$$N_{m1} = \sum_{i \in \mathcal{P}_1} \mathbb{1}\{s_i = m\},$$
$$N_{m0} = \sum_{i \in \mathcal{P}_0} \mathbb{1}\{s_i = m\},$$
$$N_{m\mathcal{T}} = \sum_{i \in \mathcal{T}} \mathbb{1}\{s_i = m\},$$

and let ω_i for all $i \in \mathcal{P}$ be

$$\omega_i = \mathbb{1}\{i \in \mathcal{P}_0\}\,\mathbb{1}\{s_i = m\}\,\frac{n_{m\mathcal{T}}}{N_{m0}} + \mathbb{1}\{i \in \mathcal{P}_1\}\,\mathbb{1}\{s_i = m\}\,\frac{n_{m\mathcal{T}}}{N_{m1}}. \tag{2.7}$$

If prior unit i is treated, then ω_i is simply the ratio of the number of target units to prior treated units in i's matched stratum. If prior unit i is in the control condition, then ω_i is the ratio of the number of target units to prior control units in i's matched stratum. This quantity ω_i is relative to a

baseline of 1 in which $\omega_i = 1$ implies that one treated or control prior unit stands for the treated or control potential outcome of one target unit. When ω_i is greater than 1, then one prior unit stands for a greater number of target units; hence, this prior unit is especially important since it counts for a greater number of target units. Conversely, when ω_i is less than 1, then this prior unit is one of multiple prior units that stand for 1 target unit; hence, the importance of this one prior unit by itself is downweighted. In general, if unit i belongs to a set in which there are more target units than units in the condition to which unit i was assigned, then ω_i will be less than 1. Otherwise, ω_i will be greater than or equal to 1. The quantity ω_i represents the importance of each prior observation. I now turn to each prior observation's reliability.

Optimally matching on the heterogeneity score constructs sets such that prior treated, prior control and target units are similar on the heterogeneity score, which is interpreted as a 1-dimensional summary of the K-dimensional vector of heterogeneity variables. Yet, as mentioned above, imbalances on the heterogeneity score (and the heterogeneity variables themselves) may remain after matching. Prior units with heterogeneity scores closer to the average heterogeneity score among within-stratum target units yield more trustworthy predictions of these target units' counterfactual outcomes. To assign greater weight to such prior observations, define the distance between a prior unit's heterogeneity score and the average heterogeneity score of target units in i's stratum as

$$\text{For all } i \in \mathcal{P},\ \delta_i = \sqrt{\left(h_i - \frac{1}{n_{m\mathcal{T}} \mathbb{1}\left\{s_i = m\right\}} \sum_{j \in \mathcal{T}} \mathbb{1}\left\{s_j = s_i\right\} h_j \right)^2}. \tag{2.8}$$

This quantity δ_i provides a measure of the imbalances on the heterogeneity score that remain after matching. Note that all units' heterogeneity scores take on values between 0 and 1, which implies that δ_i is also bounded between 0 and 1. Values of δ_i closer to 1 indicate higher imbalance on heterogeneity variables and values closer to 0 indicate lower imbalance.

Within a matched set, the degree of balance on heterogeneity scores gives us a sense of the reliability of predictions based on these prior units. That is, a prior unit's heterogeneity score may be equal to the average of within-stratum target units' heterogeneity scores; if so, that prior unit

ought to be assigned its full weight implied by ω_i in Equation (2.7). If a prior unit is dissimilar from the average within-stratum target unit, then that prior unit ought to count for fewer target units compared to more similar within-stratum prior units.

To quantify this overall intuition, we can bring together Equations 2.7 and 2.8 to define the normalized weight of prior unit i as

$$w_i = \mathbb{1}\{i \in \mathcal{P}_1\} \frac{(1-\delta_i)\omega_i}{\sum_{i \in \mathcal{P}_1} (1-\delta_i)\omega_i} + \mathbb{1}\{i \in \mathcal{P}_0\} \frac{(1-\delta_i)\omega_i}{\sum_{i \in \mathcal{P}_0} (1-\delta_i)\omega_i}. \tag{2.9}$$

The weight in Equation (2.9) can be incorporated into the CATE algorithm such that prior units with higher weights exert greater leverage in the fitting of the CATE model. Prior units with lesser weights exert less influence in the fitting of the CATE model. The weight in Equation (2.9) implies that prior units comparable to a greater number of target units (i.e., greater ratio of target to prior treated or control units in the same matched set) and more similar in heterogeneity scores will have the greatest weights. Conversely, prior units' weights are decreasing in the within-stratum ratio of target to prior treated or control units and in the distance between within-stratum prior and target units' heterogeneity scores.

Armed with both a matching and weighting scheme, Section 2.3.4 now provides a linear projection model for the CATE. The model is fit only to prior units, accounting for their weights, stratum memberships and heterogeneity variables. The fitted model is then used to generate predictions of potential outcomes for target units, which, in turn, are used to calibrate the target experiment's prior distribution as described in Section 2.3.2.

2.3.4 Model of conditional average treatment effect (CATE)

Having described the matching procedure and the weights reflecting the reliability and importance of prior observations, we can now state the CATE model. As a baseline model, I propose a linear, multiplicative interaction model (as in Kern et al., 2016) with fixed effects for matched strata and weights given in Equation (2.9). Formally, we can write the CATE predictive model for

target experimental units as

$$\hat{y}_{zi} = \hat{\alpha}_i + \hat{\alpha}_m + \hat{\gamma}_0 z_{im} + \sum_{k=1}^{K} \hat{\beta}_k x_{kim} + \sum_{k=1}^{K} \hat{\gamma}_k x_{kim} z_{im} \text{ for all } i \in \mathcal{T}. \tag{2.10}$$

Importantly, the coefficients in Equation (2.10) are solutions to the least squares problem in the prior data:

$$\begin{aligned} &\left(\hat{\alpha}_i, \hat{\alpha}_m, \hat{\gamma}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K, \hat{\gamma}_1, \ldots, \hat{\gamma}_K\right) = \\ &\arg\min \sum_{i \in \mathcal{P}} \sum_{m=1}^{M} \mathbb{1}\{s_i = m\} w_i \left[y_{im} - \left(\alpha_i + \alpha_m + \gamma_0 z_{im} + \sum_{k=1}^{K} \beta_k x_{kim} + \sum_{k=1}^{K} \gamma_k x_{kim} z_{im} \right) \right]^2 . \end{aligned} \tag{2.11}$$

The actual relationship between heterogeneity variables and potential outcomes is *not* assumed to be linear. The linear model in Equation (2.10) serves only as a linear approximation that represents how potential outcomes change, on average, across different values of heterogeneity variables. In the context of representing baseline knowledge from past experiments, not, e.g., consistent estimation, such a linear approximation is sensible, particularly as a baseline model. Other models, specifically those from the machine-learning literature on estimating heterogeneous effects (Wager and Athey, 2018; Hill, 2011; Green and Kern, 2012; Imai and Ratkovic, 2013; Grimmer, Messing, and Westwood, 2017), are also potentially fruitful choices. However, in the closely related context of target validity, simple linear regression often performs nearly as well as machine-learning models in a range of simulations with different data generating processes (see e.g., Kern et al., 2016).[6]

Referring back to Equation (2.4), we can set the mean of the target experiment's prior distribution to the difference in means of predicted potential outcomes in Equation (2.10). Likewise, as in Equation (2.5), we can set the variance of the target experiment's prior distribution to the uncertainty of these predictions implied by the estimated variance-covariance matrix. Heteroskedasticity-consistent (robust) standard errors (Eicker, 1963; Eicker, 1967; White, 1980)

[6]Kern et al. (2016, p. 124) write "[o]ne surprising result from our simulations is that, overall, linear regression performed better than expected; this leads one to wonder if it is worthwhile for researchers to adopt more sophisticated methodological strategies."

are an intuitive choice in that the diagonal entries of the variance-covariance matrix are the error variances of each prior observation; hence, the prior variance is intuitively calibrated to the predictive accuracy of the CATE model in past experiments. I use HC3 standard errors (MacKinnon and White, 1985; Andrews, 1991) since they are conservative (i.e., larger) compared to alternative HC variance-covariance matrix estimators (see Hansen, 2021, p. 112).[7]

2.4 Quantifying Bayesian learning from target experiment

Thus far, I have focused on how scholars can construct a target experiment's prior distribution based on results from past experiments. I now turn to how scholars can characterize how much they have learned, relative to this prior distribution, after carrying out the target experiment. As mentioned in the introduction (Section 2.1), I draw on value of information theory, as explicated in Bayesian decision theory (Raiffa and Schlaifer, 1961; Howard, 1966), in order to quantify how much one has learned from a randomized experiment

Prior to observing experimental data, the best guess for the TATE, $\tau_{\mathcal{T}}$, is whichever value maximizes its prior probability density. In addition to this best guess, we can also define a loss function that characterizes the penalty for making an incorrect guess about the value of $\tau_{\mathcal{T}}$. This loss function depends on both the best guess of $\tau_{\mathcal{T}}$ and its true value over which one has prior or posterior beliefs.

To formalize this framework, first let $\tau_{\mathcal{T}}^{d}$ denote the best guess (decision) about the value of $\tau_{\mathcal{T}}$. Then consider a squared error loss function given by $\mathcal{L}\left(\tau_{\mathcal{T}}, \tau_{\mathcal{T}}^{d}\right) = \left(\tau_{\mathcal{T}} - \tau_{\mathcal{T}}^{d}\right)^2$. The optimal guess is that which minimizes expected loss. Hence, the Bayes' solution to the decision problem before

[7]The notion of conservativeness may take on a different meaning in the context of Bayesian inference compared to other modes of inference. For example, it might be anti-conservative to use larger standard errors for the CATE model since the subsequent prior distribution will be more diffuse and, consequently, learning from the target experiment will be greater than it otherwise would be. Alternatively, one could conceive of larger standard errors for the CATE model as conservative since it aims to mitigate the power of the prior distribution to shape posterior inferences. Under trade-offs such as these two, I err on the side of letting target experimental data speak for themselves, which is conservative in terms of how one quantifies the amount of prior information.

and after observing data, respectively, is

$$\arg\min_{\tau_{\mathcal{T}}^d} \int_{\tau_{\mathcal{T}}} \mathcal{L}\left(\tau_{\mathcal{T}}, \tau_{\mathcal{T}}^d\right) p\left(\tau_{\mathcal{T}}\right) d\,\tau_{\mathcal{T}} \tag{2.12}$$

$$\arg\min_{\tau_{\mathcal{T}}^d} \int_{\tau_{\mathcal{T}}} \mathcal{L}\left(\tau_{\mathcal{T}}, \tau_{\mathcal{T}}^d\right) p\left(\tau_{\mathcal{T}} \mid \hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) d\,\tau_{\mathcal{T}}, \tag{2.13}$$

where, as mentioned above, $p\left(\tau_{\mathcal{T}}\right)$ is the prior distribution of $\tau_{\mathcal{T}}$ and $p\left(\tau_{\mathcal{T}} \mid \hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)$ is the posterior distribution after observing experimental evidence summarized by $\hat{\tau}_{\mathcal{T}}$ and plugging in Neyman's conservative variance estimator for the variance nuisance parameter. With a squared error loss function, the optimal decision rules given in Equations (2.12) and (2.13) are equal to the expectations of the prior and posterior distributions, respectively (see Jaynes, 2003, pp. 172–175). Since the optimal decision is the expected value of either the prior or posterior distribution, the expected loss of this decision with a squared error loss function is the variance of the respective prior or posterior distribution.

Proposition 4 below formally characterizes the change in the optimal guess about the TATE and the change in expected loss from the prior to posterior distribution.

Proposition 4. *Suppose a Normal prior distribution* $p\left(\tau_{\mathcal{T}}\right) \sim \mathcal{N}\left(\mu_{prior}, \sigma^2_{prior}\right)$ *and a Normal likelihood function,* $f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}\left[\hat{\tau}_{\mathcal{T}}\right]\right)$*. The change in the optimal decision from the prior to the posterior distribution is*

$$\mu_{post} - \mu_{prior} = \frac{\sigma^2_{prior}}{\left(\sigma^2_{prior} + \widehat{\text{Var}}\left[\hat{\tau}_{\mathcal{T}}\right]\right)}\left(\hat{\tau}_{\mathcal{T}} - \mu_{prior}\right) \tag{2.14}$$

and the change in the expected loss from the prior to the posterior distribution is

$$\sigma^2_{post} - \sigma^2_{prior} = -\frac{\sigma^4_{prior}}{\sigma^2_{prior} + \widehat{\text{Var}}\left[\hat{\tau}_{\mathcal{T}}\right]}. \tag{2.15}$$

As one might expect, the smaller is the variance in the target experiment, the greater is the decrease

in expected loss upon observing data from a new experiment. Moreover, since $\widehat{\text{Var}}\left[\hat{\tau}_{\mathcal{T}}\right] < \infty$, it follows that an additional randomized experiment always yields at least some degree of second-moment learning (decrease in expected loss). The change in the optimal decision about the value of $\tau_{\mathcal{T}}$ (first-moment learning) depends on the relative information that the experimental data and the prior distribution provide. In particular, the greater is prior uncertainty, then the more the optimal decision shifts in the direction of the estimated average treatment effect. Relatedly, the smaller is the estimated variance, then the less weight is assigned to the prior mean.

Both expressions in Proposition 4 provide two useful statistics that represent how much one has learned from a new randomized experiment. I refer to Equation (2.14) as the *first-moment Bayesian learning statistic* and to Equation (2.15) as the *second-moment Bayesian learning statistic*. But in addition to these first- and second-moment Bayesian learning statistics, it is helpful to have a single standardized statistic taking on values between 0 and 1. As mentioned in the motivating example in Section 2.2, when potential outcomes are naturally bounded, the usual point and standard error estimates are equivalent to a uniform prior distribution on the possible values of the TATE. This situation corresponds to the maximal amount of learning. Hence, a useful statistic compares the actual amount of learning to the maximal amount of learning without any prior information. Formally, we can write this standardized Bayesian learning statistic as

$$\textbf{Standardized Bayesian learning statistic} : \frac{\sqrt{\left(\sigma^2_{\text{post}} - \sigma^2_{\text{prior}}\right)^2}}{\sqrt{\left(\sigma^2_{\text{unif}} - \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^2}}, \tag{2.16}$$

where $\sigma^2_{\text{unif}} = \frac{1}{12}\left(U_{\tau_{\mathcal{T}}} - L_{\tau_{\mathcal{T}}}\right)^2$ with $U_{\tau_{\mathcal{T}}}$ and $L_{\tau_{\mathcal{T}}}$ representing the lower and upper bounds of the TATE. In the case of the legislative audit experiment on state legislators, the TATE is naturally bounded between -1 and 1. When the TATE is unbounded, the implicit prior distribution can no longer be uniform and the maximal amount of learning is difficult to characterize. Although an improper Jeffreys (Jeffreys, 1939; Jeffreys, 1946) or reference (Bernardo, 1979) prior can yield a proper posterior, they do not facilitate comparisons of the change in expected loss of the best guess from the prior to posterior distributions. Thus, when potential outcomes are unbounded, I propose

to compare the change in expected loss from prior to posterior distributions to what the change would have been with a uniform prior on the TATE with bounds given by

$$\max_{1\leq i\leq N} z_i y_i - \min_{1\leq i\leq N} (1-z_i) y_i \text{ and } \min_{1\leq i\leq N} z_i y_i - \max_{1\leq i\leq N} (1-z_i) y_i. \tag{2.17}$$

Equation (2.17) sets the upper bound to the maximum observed treated potential outcome minus the minimum observed control potential outcome. The lower bound is the minimum observed treated potential outcome minus the maximum observed control potential outcome. Note that the minimum and maximum outcomes under treated and control are taken over all prior experiments and the target experiment.

Finally, before turning to an empirical analysis of the 2020 audit experiment in Section 2.5, I briefly offer two other insights that can be gleaned from Proposition 4. First, it shows that some degree of learning always results from a single replication study. As a thought exercise, if a replication study were conducted on the exact same units as a prior study and the estimated effect were identical, a decrease in expected loss (second-moment learning) would nonetheless result. As Equation (2.15) shows, the posterior minus the prior variance is always negative, so second moment learning results. Continuing the thought experiment of an exact replication, imagine that σ^2_{prior} and the estimated variance in the replication experiment, $\widehat{\text{Var}}\left[\hat{\tau}_{\mathcal{T}}\right]$, are identical. The extent of learning will depend on the value of σ^2_{prior}. In this case of an exact replication, the amount of learning can be represented simply as $-\sigma^2_{\text{prior}}/2$. Hence, we always learn from a single, exact replication, but how much we learn depends on the state of prior knowledge. With a very precise state of prior knowledge, learning is minimal, but increases with greater prior uncertainty.

Second, Proposition 4 offers insight more broadly on a form of Bayesian power analysis whereby scholars can assess how large an experiment would need to be to achieve a certain degree of learning in expectation. For example, consider the difference in variances between prior and posterior distributions in Equation (2.15). Under assumptions about the means and variances of treated and control potential outcomes, as well as the total number of units and proportion of treated

units in the target experiment, it is straightforward to assess the expected amount of learning. The key insight is that $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$, whose expectation and probability limit can be analytically derived, is the only random quantity in Equation (2.15). The prior variance is fixed over possible random assignments. Hence, it would be straightforward to assess how large the N of a target experiment would need to be to achieve some expected level of Bayesian learning. Given budget constraints that researchers invariably face, they can assess how much learning would result relative to prior knowledge from a small experiment (even if that experiment were unlikely to reach the threshold of statistical significance).

2.5 Empirical analysis

I now apply the overall methodology to the 2020 audit experiment described in Section 2.2. I construct the prior distribution for this target experiment based on data from three published papers (Butler and Broockman, 2011; Butler, 2014; Butler and Crabtree, 2017) and one unpublished paper (Janusz and Lajevardi, 2016). Given the private nature of the data from audit experiments on public officials, I was able to obtain data with sufficient background information from only these four papers under the condition that I not share the data.

The first step in implementing the methods thus described is to define the set of heterogeneity variables. I define this set as

$$\mathcal{X} = \{\texttt{Gender}, \texttt{Race}, \texttt{Party}, \texttt{Level of government}, \texttt{District percent Black}, \texttt{Professionalism index}, \texttt{Southern state}\}. \tag{2.18}$$

I then construct heterogeneity scores, as described in Section 2.3.2 via logistic regression. The distributions of heterogeneity scores between prior and target observations are given in Figure 2.2 below.

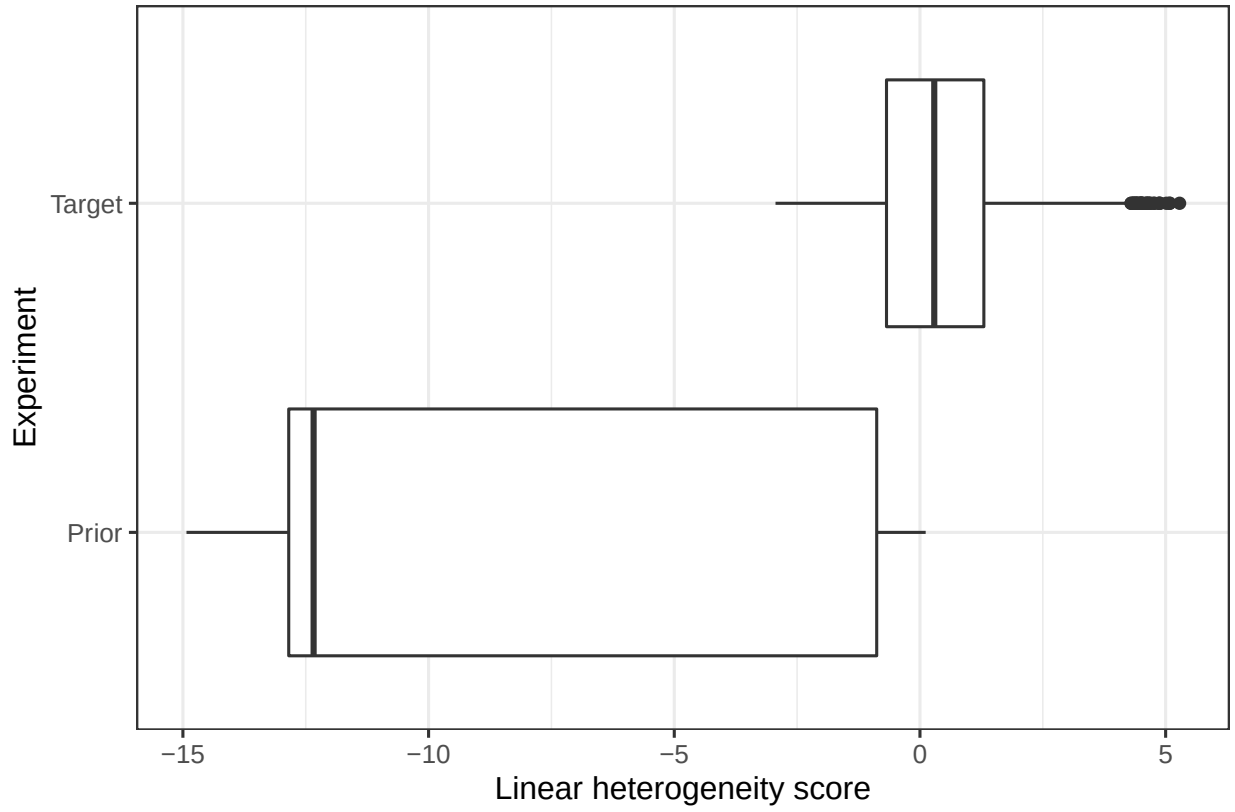

Figure 2.2: Heterogeneity score distributions (logit scale) between prior and target experiments

The heterogeneity score distributions suggest a poor degree of overlap on variables that plausibly explain effect heterogeneity. Hence, a large proportion of prior observations will be poorly suited to predict potential outcomes in the target experiment. To construct sets of prior control, prior treated and target units similar on the heterogeneity score, I implement generalized full matching via `R`'s `quickmatch` package (Sävje, Higgins, and Sekhon, 2021). I construct a matching for each treatment contrast separately. The matching for all contrasts satisfies the constraints given in Equations 2.6b – 2.6f with a caliper on the heterogeneity score set to 0.2.[8] Figure 2.3 below presents the differences in heterogeneity score distributions between prior and target experiments after matching.

[8]This caliper on the heterogeneity score also happens to be the recommended caliper on the propensity score in the closely related context of matched observational studies with a binary treatment variable (Cochran and Rubin, 1973; Rosenbaum and Rubin, 1984).

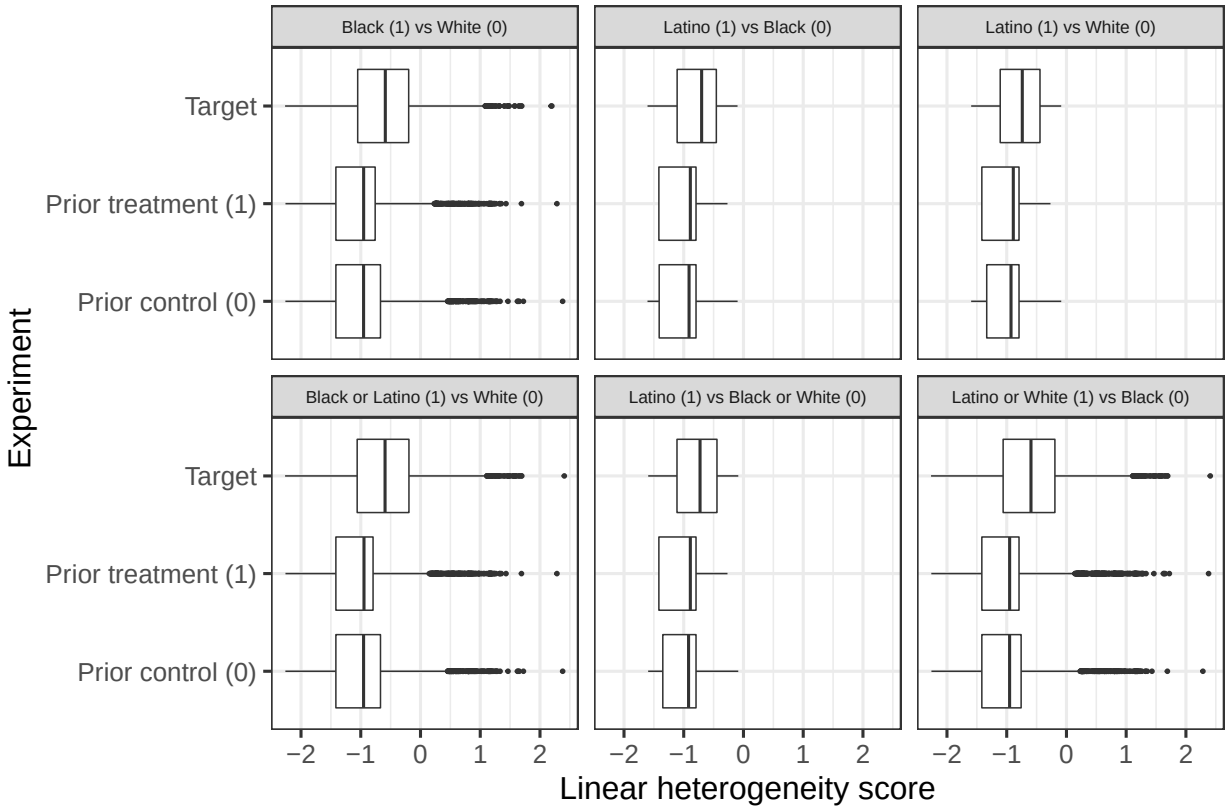

Figure 2.3: Heterogeneity score distributions (logit scale) after matching for all treatment contrasts

Overlap in heterogeneity scores is substantially improved after matching, but at the expense of fewer prior units. Moreover, discrepancies in heterogeneity scores (and the heterogeneity variables themselves) between prior control, prior treated and target units may still exist (and can be as large as 0.2 or approximately -1.39 on the logit scale). Table 2.2 below provides the maximum absolute distance in heterogeneity scores within a matched set for all treatment contrasts, along with the average within-stratum ratio of target to either treated or control prior units.

Treatment contrast	Max heterogeneity distance	Avg target-prior ratio
Black (1) vs White (0)	0.0201	0.0002
Latino (1) vs White (0)	0.0445	0.0003
Latino (1) vs Black (0)	0.0416	0.0003
Black or Latino (1) vs White (0)	0.0251	0.0002
Latino (1) vs Black or White (0)	0.0448	0.0002
Latino or White (1) vs Black (0)	0.0251	0.0002

Table 2.2: Maximum within-stratum absolute distance on heterogeneity score between prior and target units and maximum within-stratum ratio of target units to either treated or control prior units

In many cases, as is shown in Table 2.2, the within-stratum ratio of target to prior units will be small. When multiple prior experiments exist, there will likely be many more prior units than target units, thereby deflating the within-stratum ratios of target to prior treated or control units.

With the weights in Equation (2.9) and fixed effects for matched strata, I fit the linear interaction model in Equation (2.11) with the heterogeneity variables in Equation (2.18) to the prior data. I then use the fitted coefficients from this model to construct the mean and variance parameters of the truncated Normal prior distribution as described in Equations (2.4) and (2.5). I subsequently update the respective prior distributions via a Normal likelihood function in which a hypothetical value of the average effect, along with a conservative plug-in estimator of the variance, assigns probability density to the target experiment's post-treatment data summarized by the Difference-in-Means. Figure 2.4 and Table 2.3 below provide both the prior and posterior distributions for all six treatment contrasts.

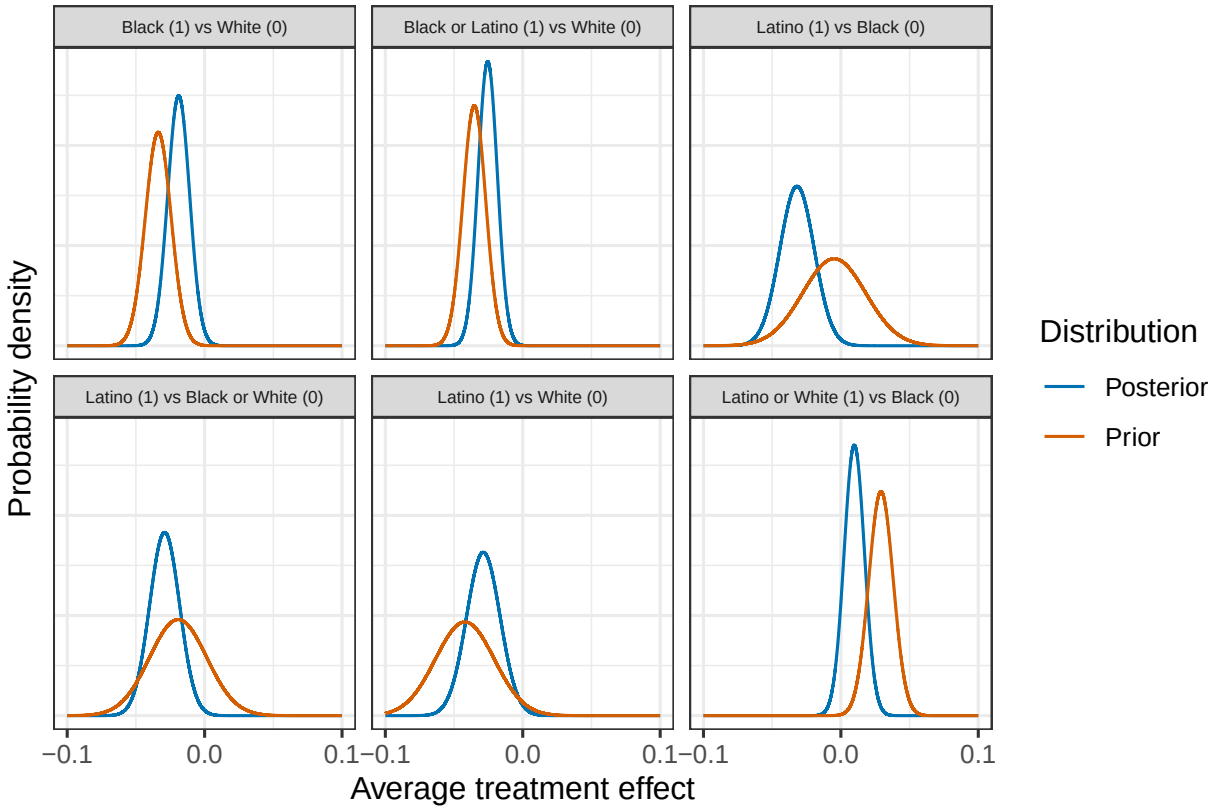

Figure 2.4: Prior and posterior distributions on the average effect in the target experiment for all treatment contrasts

	Outcome: Reply from legislator's email address		
Treatment contrast	**Prior mean (SD)**	**Posterior mean (SD)**	**Bayes' learning**
Black (1) vs White (0)	−0.034 (0.0094)	−0.0191 (0.008)	0.0084
Latino (1) vs White (0)	−0.0419 (0.0218)	−0.0284 (0.0123)	0.0312
Latino (1) vs Black (0)	−0.0046 (0.023)	−0.0318 (0.0125)	0.0334
Black or Latino (1) vs White (0)	−0.0355 (0.0086)	−0.0252 (0.0072)	0.0082
Latino (1) vs Black or White (0)	−0.0205 (0.0204)	−0.0293 (0.0108)	0.03
Latino or White (1) vs Black (0)	0.0283 (0.0092)	0.0083 (0.0076)	0.0091

Table 2.3: Means and standard deviations of prior and posterior distributions of the average effect in the target experiment for each of the six contrasts. The Bayes' learning statistic is the standardized statistic given in Equation (2.16).

As Figure 2.4 and Table 2.3 show, this audit experiment results in substantively meaningful first-moment learning, but leads to relatively little changes in expected loss. Based on results of prior audit experiments, one would expect 3.4% of state legislators to respond to a White constituent, but not a Black constituent. Yet in this experiment, referring back to Table 2.1, the average difference in legislators' response rates to putatively White and Black constituents is low. Therefore, the posterior mean suggests a substantially smaller level of discrimination, 1.91%, that legislators exhibit, on average, against Black constituents in favor of White constituents.

Unsurprisingly, the standardized Bayesian learning statistics are all relatively small. With a prior distribution informed by several past audit experiments, the change in expected loss from the prior to posterior distribution will be much smaller than what that change would have been with a uniform prior. That is what is reflected in Table 2.3, even though we do see meaningful levels of first-moment learning. Learning is likely to be greater among subgroups for which less prior

information exists. I now turn to an analysis of learning from theoretically important subgroups.

2.5.1 Subgroup analysis

One of the primary benefits of Bayesian analysis of experiments is that it enables assessments of how much information from an experiment one gains about theoretically important subgroups. I focus on subgroups defined by both party and ethnorace. Before assessing Bayesian learning in these subgroups, I present the Difference-in-Means and conservatively estimated standard errors in each subgroup.

	Outcome: Reply from legislator's email address			
Treatment contrast	**Diff-in-Means (SE)**			
	GOP legs	*Dem legs*	*Black legs*	*Latino legs*
Black (1) vs White (0)	0.0128	0.0304	0.0452	0.1343**
	(0.0216)	(0.0218)	(0.0429)	(0.0667)
Latino (1) vs White (0)	−0.0454**	0.0032	−0.0054	0.1018
	(0.0206)	(0.0216)	(0.0414)	(0.0663)
Latino (1) vs Black (0)	−0.0582***	−0.0272	−0.0505	−0.0325
	(0.0206)	(0.0217)	(0.0412)	(0.0686)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2.4: Subgroup results of audit experiment

Table 2.4 suggests that Latino legislators strongly favor putatively Black to White constituents, but still favor Latino constituents over Black constituents. These results suggest that the benefits of descriptive representation, at least among Latino legislators, accrue to both in- and out-minority groups. By contrast, among Black legislators, the benefits of descriptive representation appear to apply only to putatively Black constituents. These estimated effects among Black legislators are not statistically significant, although at least one of the results among Latino legislators is significant. Regardless of their significance, we can still assess how much we learn from these results.

Examining Table 2.5 below, we can see that this experiment yields a very different understanding of both Black and Latino legislatorsrelative to what one would have predicted going into the

experiment. Figure 2.5 below illustrates the changes from prior to posterior distributions.

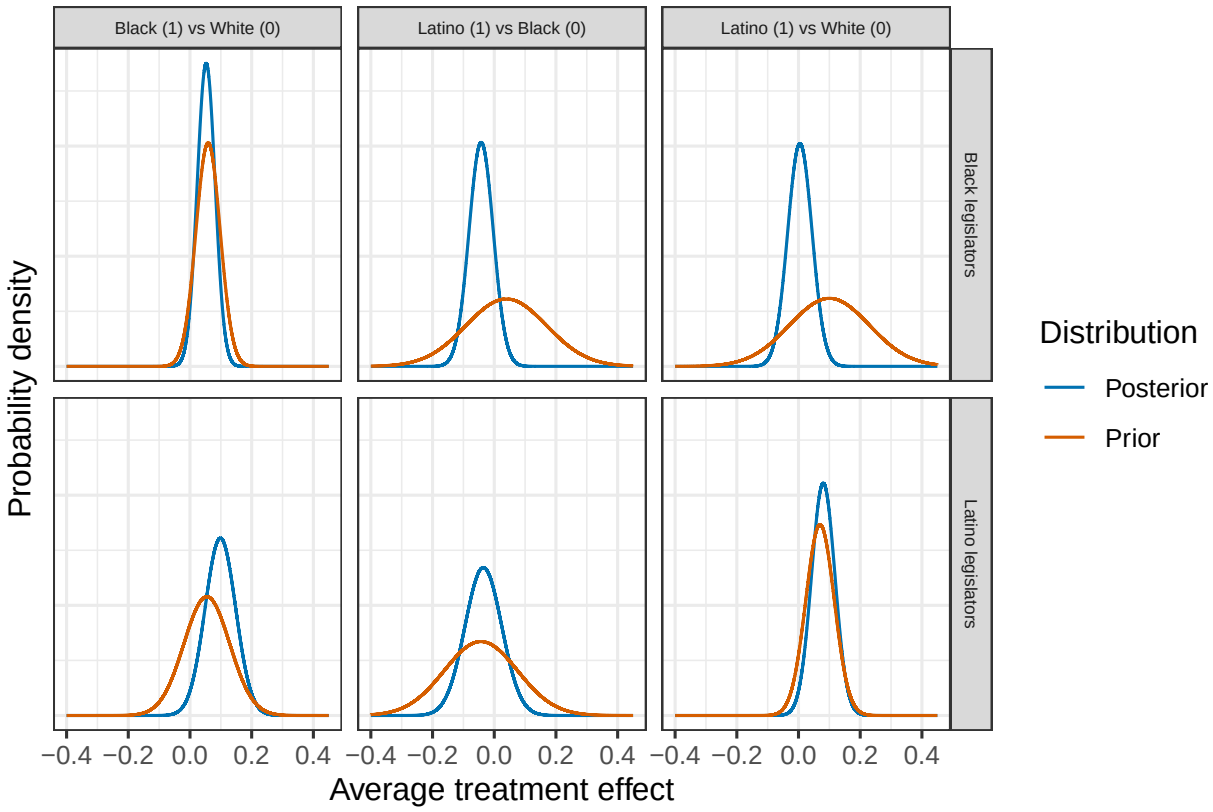

Figure 2.5: Prior and posterior distributions of treatment contrasts by race of legislators

Of particular interest is Black legislators' responsiveness to Latino vs Black constituents. The first-moment change from the prior to posterior distribution is large, 0.04 to −0.043. But more importantly, the prior to posterior standard deviation decreases from 0.133 to .039. Among Latino legislators, there is relatively large amount of learning overall about the extent to which they favor Black over White constituents. Yet the posterior distribution itself is still relatively imprecise. An even greater amount of learning among Latino legislators obtains for the Latino vs Black contrast. Note again, however, that this large amount of learning still results in a relatively imprecise posterior standard deviation. The high learning is driven by an extremely diffuse prior distribution due to the scarcity of Latino legislators in prior experiments. We can see this pattern in Table 2.5 below.

	Outcome: Reply from legislator's email address		
Treatment contrast	**Prior mean (SD)**	**Posterior mean (SD)**	**Bayes learning**
Subgroup: Black legislators			
Black (1) vs White (0)	0.057 (0.0361)	0.0521 (0.0276)	0.0404
Latino (1) vs White (0)	0.106 (0.132)	0.0046 (0.0395)	0.2187
Latino (1) vs Black (0)	0.0404 (0.1329)	−0.0426 (0.0393)	0.2204
Subgroup: Latino legislators			
Black (1) vs White (0)	0.0527 (0.0755)	0.0985 (0.05)	0.0986
Latino (1) vs White (0)	0.0664 (0.0447)	0.0774 (0.0371)	0.0435
Latino (1) vs Black (0)	−0.0385 (0.1162)	−0.0341 (0.0591)	0.1745
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2.5: Bayesian learning among Black and Latino legislators

Now turning to subgroup effects by party, recall from Table 2.4 suggests that GOP legislators, on average, discriminate quite strongly against putatively Latino constituents in favor of not only White, but also putatively Black constituents. Table 2.4 also shows that GOP and Democratic legislators exhibit little discrimination between putatively Black and White constituents. Figure 2.6 and Table 2.6 show the changes from prior to posterior distributions upon observing the results in Table 2.4.

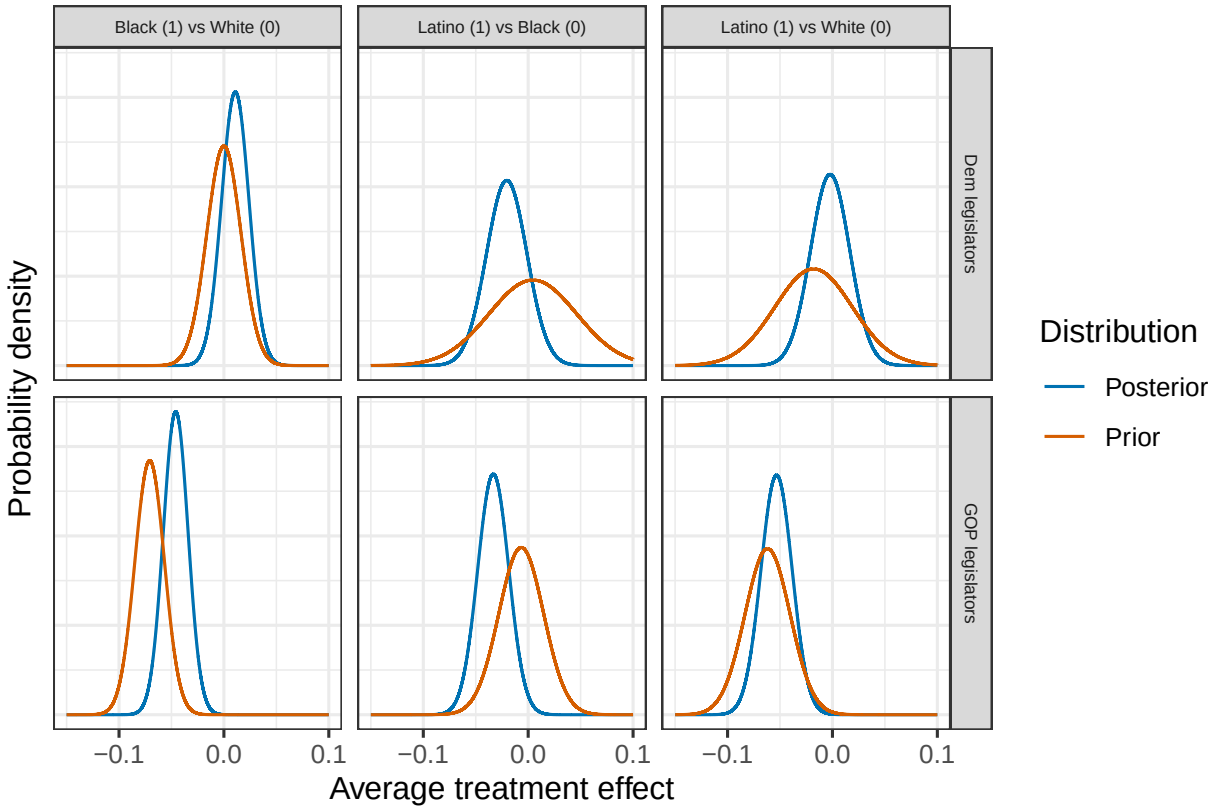

Figure 2.6: Prior and posterior distributions of treatment contrasts by party of legislators

	Outcome: Reply from legislator's email address		
Treatment contrast	**Prior mean (SD)**	**Posterior mean (SD)**	**Bayes learning**
Subgroup: GOP legislators			
Black (1) vs White (0)	−0.0711 (0.0143)	−0.0454 (0.0119)	0.0137
Latino (1) vs White (0)	−0.0602 (0.0223)	−0.0522 (0.0151)	0.0283
Latino (1) vs Black (0)	−0.0063 (0.021)	−0.0328 (0.0147)	0.0261
Subgroup: Dem legislators			
Black (1) vs White (0)	0.0004 (0.0164)	0.0113 (0.0131)	0.0171
Latino (1) vs White (0)	−0.016 (0.0366)	−0.0017 (0.0186)	0.0547
Latino (1) vs Black (0)	0.0053 (0.042)	−0.0203 (0.0193)	0.0646

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2.6: Bayesian learning among GOP and Democrat legislators

Figure 2.6 and Table 2.6 show that meaningful first-moment learning obtains within party subgroups. The standardized Bayesian learning statistics, however, are all relatively small. However, the Bayesian learning for the Latino vs White and Latino vs Black contrasts among Democratic legislators is twice as large as all measures of overall learning in Table 2.3.

2.6 Conclusion

In this paper, I have demonstrated how scholars can generate a prior distribution, empirically justified from past experiments, for a new target experiment. I have also shown how scholars can provide meaningful quantitative summaries of how much one has learned from a target experiment relative to the prior. Both the prior distribution and likelihood function are the two central pillars

of Bayesian inference. This chapter has focused on the prior distribution and how it changes when updated by new experimental data. Chapter 2 focused on the likelihood function, which showed that Bayesians do not simply revise their beliefs, but do so in a way that tracks the true causal target of interest. This likelihood function and the prior distribution, taken together, enables scholars to quantify what they learn from randomized experiments and to assess how knowledge cumulates from one experiment to the next. Having laid out these properties in the context of an ideal randomized experiment, the next chapter shifts focus to observational studies and the Difference-in-Differences design in particular.

Chapter 3: Identification and Inference for Difference-in-Differences under Uncertainty in Parallel Trends

3.1 Introduction

At least since the seminal studies of Ashenfelter and Card (1985), Card (1990) and Card and Krueger (1994),[1] Difference-in-Differences (DID)[2] has emerged as one of the most popular non-parametric (model-free) methods for inference of causal effects in studies with outcomes measured over time. Scholars use the canonical DID design in contexts where one (treated) subpopulation is subject to an intervention but another (control) subpopulation is not. In such contexts, the average treatment effect in the treated subpopulation (the ATT) is difficult to infer. First, baseline factors often differ between treated and control groups. Hence, a treated-vs-control comparison of post-treatment outcomes may reflect *not* a causal effect, but only this baseline difference between groups. Second, factors other than treatment often vary over time in the treated population. Hence, a comparison of pre- and post-treatment outcomes in the treated group may reflect these factors, *not* a causal effect. The DID design offers a solution to both problems. So long as there would have been no difference between treated and control populations' average changes in outcomes over time had the treatment not occurred,[3] the DID design identifies the ATT.[4] This identification

[1] The actual genesis of the DID design can be traced at least as far back as Snow (1854) and Obenauer and Nienburg (1915). For more on the role that DID reasoning played in John Snow's analysis of London's 1854 Broad Street cholera outbreak, see Coleman (2019).

[2] For some of the core literature on DID designs (albeit sometimes under different names), see Abadie (2005), Angrist and Pischke (2008, chapter 5), Athey and Imbens (2006), Lechner (2011), Rosenbaum (1989b) and Rosenbaum (2001), Sofer et al. (2016) and Tchetgen Tchetgen (2014) among others.

[3] In other words, as Abadie (2005, pp. 1–2) states, "the conventional DID estimator requires that in absence of the treatment, the average outcomes for treated and controls would have followed parallel paths over time." One could equivalently state that the DID design adjusts for time-invariant confounders, but not time-varying confounders. Hence, the crucial identification assumption of the DID design is that of no time-varying confounders.

[4] Following Koopmans and Reiersøl (1950) (see also Hurwicz, 1950; Koopmans, 1949; Koopmans, Rubin, and Leipnik, 1950), I refer to an unobservable causal parameter as *identified* when, given perfect knowledge of the probability distributions of observable variables, that parameter is consistent with only one value, not (perhaps infinitely)

assumption, known as parallel trends, provides a design-based justification for common estimators, like two-way fixed effects. However, despite its benefits for estimation, the tethering of the DID design to the identification assumption of parallel trends poses problems for inference.

Drawing upon the typology of Abadie et al. (2020), uncertainty is typically conceived in two ways. One is *sampling* uncertainty that arises from the inability to sample all units from a target population. Another is *counterfactual* uncertainty that arises from the inability to observe counterfactual potential outcomes among whichever units one samples. When making a statistical inference from observed data in a sample to a counterfactual parameter in a population, the statistical distributions on which our conclusions about this parameter are based ought to reflect both types of uncertainty. Consider, for example, a randomized experiment conducted on a random sample of units from a target population. Statistical uncertainty in conclusions about the average causal effect in this target population reflect variation in (1) the average causal effect across possible samples and (2) causal estimates across possible assignments conditional on each possible sample (Neyman, 1923; Aronow, Green, Lee, et al., 2014; Freedman, 2008; Imbens and Rubin, 2015). By contrast, statistical uncertainty in the canonical DID design reflects only sampling uncertainty that arises from the inability to sample all units from a target population. The quantification of uncertainty in inferences from observed to counterfactual outcomes is difficult in the DID design since its causal conclusions are based on assumptions about the *fixed* target parameter itself, not assumptions about a *random* assignment mechanism (see, e.g., Manski and Pepper, 2018). Hence, counterfactual uncertainty is absent from the standard errors, p-values and confidence intervals of the canonical DID design.

This failure to account for counterfactual uncertainty in the DID design has two immediate consequences for empirical practice: First, in cases where both sampling and counterfactual uncertainty exist, scholars often produce uncertainty intervals that are too narrow and hypothesis tests that too frequently detect an average causal effect when none exists. Second, in many cases, units cannot be naturally conceived as a sample from a target population, e.g., when units consist of all

many values.

50 states in the United States. When units can indeed be conceived as a sample from a population, the standard assumption of independent and identically distributed (IID) sampling may be difficult to justify (Abadie et al., 2020; Manski and Pepper, 2018). Thus, existing methods that assume an IID random sampling process yield incorrect and potentially misleading standard errors in DID applications.

One solution to these two problems is to statistically capture counterfactual uncertainty due to a random treatment assignment mechanism with unknown and potentially heterogeneous assignment probabilities (Rambachan and Roth, 2020) or the potentially random timing of treatment in the DID design (Athey and Imbens, 2021). Yet a key feature of the DID design is that, unlike other observational designs, its causal conclusions are *not* based on an analogy with randomized experiments (see Keele, 2020). Therefore, statistical uncertainty predicated on an alternative assumption about the assignment mechanism does not represent uncertainty in ones causal conclusions from the DID design.[5]

In this paper, I offer a resolution to the problem of inference via the following methodological contribution. I first generalize the DID design to the full set of counterfactual trend assumptions, which includes *not* only parallel trends, but also each possible way in which parallel trends could be violated. I use this framework to decompose the DID design's overall uncertainty into sampling and counterfactual uncertainty. To statistically capture counterfactual uncertainty, I derive an empirical Bayes' procedure that calibrates a prior distribution on the set of counterfactual trend assumptions to (a) pre-treatment outcome trends and (b) the control group's post-treatment deviation from pre-treatment trends. This empirical Bayes' procedure can be easily integrated with the usual nonparametric bootstrap in DID designs (Bertrand, Duflo, and Mullainathan, 2004) when sampling uncertainty is also present. I show that this empirical Bayes' procedure resolves the pathological inferential properties of the DID design and also show the conditions under which this procedure's

[5]This difficulty of quantifying uncertainty leads some scholars in some DID designs to eschew standard errors altogether. For example, in their DID design to assess the effect of Right-to-Carry laws on crime in US states, Manski and Pepper (2018) make clear that their DID design's identification assumption is based on *deterministic* constraints on the mean of counterfactual potential outcomes. Therefore, Manski and Pepper (2018, p. 234) write that they "do not provide measures of statistical precision" because the US states in the study are not realizations of a random sampling process nor are the design's causal conclusions based on an assumed random assignment mechanism.

uncertainty intervals have correct coverage.

This paper's contribution pertains to inference, *not* estimation. Standard approaches to estimation in the DID design typically proceed by choosing one from a set of possible identification assumptions and then estimating the ATT conditional on this assumption. The use of model-based estimators are often justified based on their equivalence with DID estimators defined in terms of potential outcomes rather than structural parameters of a model (Egami and Yamauchi, 2019; Lee, 2016; Mora and Reggio, 2012; Mora and Reggio, 2019). In this paper, I happen to draw on the machinery (but not the assumptions) of linear regression only as a way to calibrate the prior distribution on counterfactual trends. The empirical Bayes' procedure for inference is not wedded to linear regression for effecting such calibration. Thus, as an extension to empirical Bayes' inferential procedure, I show how scholars can assess the sensitivity of their inferences to alternative choices for calibrating the prior distribution of counterfactual trends.

In the rest of the paper, I proceed by introducing, first, a running example of a canonical DID design from Montalvo (2011), which seeks to infer the effect of the 2004 Madrid train bombings on subsequent outcomes in Spain's 2004 general election. I choose this application because the empirical Bayes' (maximum a posteriori) point estimate that results from the procedure I propose is roughly identical to that which would result from the canonical DID estimator, but with very different levels of associated uncertainty. Hence, the Montalvo (2011) study helps square the focus of this paper on inference rather than estimation. The succeeding sections provide relevant notation and lay out the pathological properties for inference that follow from the canonical DID design. The succeeding sections derive the generalized DID design and the empirical Bayes' inferential framework. Given the extensive literature on sampling-based inference (see, e.g., Bertrand, Duflo, and Mullainathan, 2004; Beck and Katz, 1995; Conley and Taber, 2011; Donald and Lang, 2007; Driscoll and Kraay, 1998; Ferman and Pinto, 2019; Imbens and Wooldridge, 2009; Rokicki et al., 2018; Wooldridge, 2003), these sections focus primarily on counterfactual inference in a finite sample, but nevertheless show how one can easily draw upon the nonparametric bootstrap to incorporate sampling uncertainty when it is also present. The final section offers concluding

remarks and reflects on potentially fruitful extensions of the general method.

3.2 Running Example and Formal Setup

On March 11, 2004, only three days before Spain's general election, several coordinated bombings of Madrid's commuter trains left nearly 200 people dead and thousands more injured. Many commentators perceived the Madrid train bombings to be in response to Spain's support of US military involvement in Iraq. José-Maria Aznar of the ruling *Partido Popular (PP)* (People's Party) staunchly supported US military involvement in Iraq despite widespread opposition from the Spanish public. Polls predicted a victory by the PP prior to the bombings on March 11. However, the opposition *Partido Socialista Obrero Español (PSOE)* (Spanish Socialist Workers' Party), which opposed Spanish military involvement in Iraq, ended up winning the general elections three days after the Madrid bombings. Some scholars argue that the Madrid bombings caused the unexpected PSOE victory (Bali, 2007), while others argue that the electoral outcome would have remained the same in the absence of the bombings (Torcal and Rico, 2004; Lago and Montero, 2005).

Montalvo (2011) intervenes in this debate by seeking to infer the effect of the Madrid bombings on electoral support for the PP relative to the PSOE via the DID design. As Montalvo (2011) explains, the Madrid bombings occurred three days before resident Spanish voters cast their votes. But nonresident Spanish voters (i.e., Spanish nationals living outside of Spain) voted either in person at the relevant Spanish consulate or by mail between March 2 and March 7, four days before the bombings. Resident voters therefore make up a treated group whose members did know about the attacks at the time of voting and nonresident voters make up a control group whose members did *not* know about the attacks at the time of voting. Technically, the votes of nonresidents are realized before treatment onset, but one can nevertheless regard them as post-treatment under the mild assumption that nonresidents' votes would have been the same had nonresidents voted roughly one week later in the absence of the bombings.

In the dataset from Montalvo (2011), each election — in 1989, 1993, 1996, 2000 and 2004 — consists of 104 observations, which reflect the two groups of resident and nonresident voters in

each of Spain's 52 provinces. The primary outcome of interest is the PP vote share at the province-group level, which is naturally bounded between 0 and 1. The central identification problem is that, in the absence of the bombings, resident voters' PP vote share in 2004 is unobserved.

A standard resolution to this identification problem is to assume that, as Montalvo (2011, p. 1149) states, "in the absence of treatment, the average outcome for the treated and untreated would have followed parallel trends." This assumption states that the counterfactual mean of resident groups' PP vote shares in 2004 is equal to an observable quantity — the mean of resident groups' PP vote shares in 2000 plus the change in means of nonresident groups' PP vote shares from 2000 to 2004. Although, the parallel trends assumption is fundamentally untestable without further assumptions, Montalvo (2011) draws on additional elections in 1989, 1993, 1996 and 2000 as placebos to assess the plausibility of parallel trends. As Figure 3.1 shows, the means of PP vote shares for the provinces' resident and nonresident groups follow roughly parallel trends prior to the 2004 bombings, but then come closer together in 2004 after the bombings.[6]

[6]Focusing specifically on the outcome of the mean ratio of PP to PSOE votes, Montalvo (2011, p. 1149) states, "[b]efore 2004, the lines are basically parallel; in 2004, they converge." This application using the data from Montalvo (2011) uses PP vote share as the outcome since the ratio of PP to PSOE vote share is not defined for all possible values that PSOE vote share could take on (namely, the value of 0).

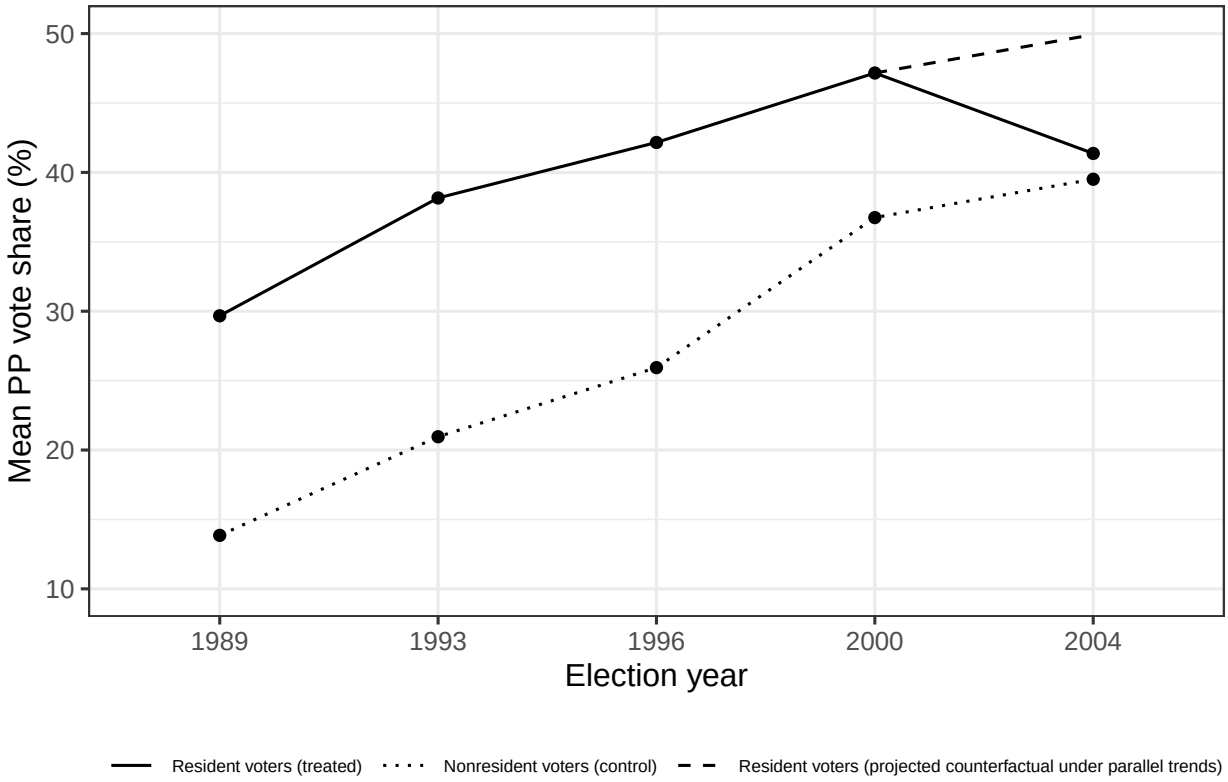

Figure 3.1: Trends in mean Partido Popular (PP) vote shares in Spain's 52 provinces (data from Montalvo, 2011)

Figure 3.1 suggests that the assumption of parallel trends is indeed plausible — i.e., that resident and nonresident groups' changes in average PP vote share from 2000 to 2004 would have been equal had the bombings not occurred. However, given that resident groups' counterfactual PP vote shares are fundamentally unobservable, a range of violations of parallel trends are also plausible. In the canonical DID design, the parallel trends assumption permits identification of the average effect of the bombings on resident groups' vote shares for the PP in 2004.

More formally, the canonical DID setup consists of two groups and only one post-treatment period.[7] Let $\mathcal{P}_N$ be a population of N units that belong to one of two groups: a control group,

[7]Recent literature on DID considers deviations from this canonical setup (see, e.g., Strezhnev, 2018; Chaisemartin and D'Haultfœuille, 2021; Goodman-Bacon, 2018; Callaway and Sant'Anna, 2018; Hudson, Hull, and Liebersohn, 2017; Yamauchi, 2020).

$Z = 0$, and a treated group, $Z = 1$, which are of sizes N_0 and N_1, respectively. From the population, $\mathcal{P}_N$, let $\mathcal{S}_n$ be a random sample of size n, which is stratified by Z, where n_0 and n_1 are the fixed numbers of sampled units from $Z = 0$ and $Z = 1$, respectively, and $n = n_0 + n_1$. Without loss of generality, assume that the first $1, \ldots, n_0$ units are sampled from group $Z = 0$ and the $n_0 + 1, \ldots, n$ units are sampled from group $Z = 1$. All units in the population, $\mathcal{P}_N$, bear measurements over $T + 1$ time periods, where the index $t \in \{0, \ldots, T\}$ runs over the $T + 1$ periods. The baseline period is $t = 0$ and T is the only post-treatment period. Throughout this paper, I write individual sample quantities with the i subscript, where the index $i \in \{1, \ldots, n\}$ runs over the n individual units in $\mathcal{S}_n$, and in either uppercase or lowercase depending on whether the sample quantity is random (upper) or not (lower). Distributions defined on all units in the population, $\mathcal{P}_N$, are written in uppercase without the i subscript.

All units in the population, $\mathcal{P}_N$, are unexposed to treatment in time periods $t \in \{0, \ldots, T - 1\}$. However, in time period T, the group $Z = 1$ is treated, while the group $Z = 0$ is not. Under the stratified sampling process described above, the groups $Z = 0$ and $Z = 1$ can be treated as two subpopulations from which units are sampled, which in turn implies that z_i is constant over all time periods conditional on the value of i.[8] The time period subscript, t, is omitted for Z and z_i since both quantities are fixed across all time periods.

In contrast to z_i, sample units' outcomes in all periods are random. Denote the individual sample outcomes as Y_{it}. For units in which $z_i = 1$, assume that each unit has two potential outcomes in period T denoted by $Y_{iT}(1)$ and $Y_{iT}(0)$, which reflect if all treated units had been treated or untreated.[9] Only the treated potential outcome, $Y_{iT}(1)$, is observed in period T. Crucially, this representation of potential outcomes in only time T, not in any prior periods, implies no anticipation of treatment for all units in which $z_i = 1$, which is distinct from the assumption of parallel trends (Malani and Reif, 2015). Potential outcomes in period T are defined only among treated units since

[8] That is, the value of z_i for a sampled unit could *not* have been different had some other unit from $\mathcal{P}_N$ been sampled instead.

[9] By defining exposure to treatment such that all treated units are either treated or untreated as a group, this assumption does not preclude interference among treated units. This assumption does, however, stipulate that there are no hidden levels of the treatment variable, which implies that all treated units are comparable to each other (Rubin, 1986).

the estimand of interest (the ATT) pertains only to them and the DID design need not invoke the assumption that control units had some treatment assignment probability greater than 0 and less than 1.[10] For control units, one can simply denote the observed outcomes in period T as Y_{iT}.

We can denote the population ATT as

$$\text{ATT} = \text{E}\left[Y_{iT}(1) - Y_{iT}(0) \mid z_i = 1\right] \tag{3.1}$$

and the population parallel trends assumption, defined on periods $T-1$ and T, as

$$\text{Parallel trends} \coloneqq \text{E}\left[Y_{iT}(0) - Y_{iT-1}(0) \mid z_i = 1\right] = \text{E}\left[Y_{iT} - Y_{iT-1} \mid z_i = 0\right], \tag{3.2}$$

where the expected value operator, $\text{E}[\cdot]$, is taken over the set of possible samples from $\mathcal{P}_N$ via the stratified, simple random sampling process described above.

The DID estimator is defined on only periods T and $T-1$, but all pre-treatment periods can be harnessed for placebo studies under additional assumptions (see, e.g., Egami and Yamauchi, 2019). The moment-based estimators of $\left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n}\left(Y_{iT}(1) - Y_{iT-1}\right)$ and $\left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0}\left(Y_{iT} - Y_{iT-1}\right)$ are unbiased for $\text{E}\left[Y_{iT}(1) - Y_{iT-1} \mid z_i = 1\right]$ and $\text{E}\left[Y_{iT} - Y_{iT-1} \mid z_i = 0\right]$, respectively. Hence, the DID estimator,

$$\widehat{\text{DID}} = \left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n}\left(Y_{iT}(1) - Y_{iT-1}\right) - \left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0}\left(Y_{iT} - Y_{iT-1}\right), \tag{3.3}$$

is unbiased for the descriptive difference of $\text{E}\left[Y_{iT}(1) - Y_{iT-1} \mid z_i = 1\right] - \text{E}\left[Y_{iT} - Y_{iT-1} \mid z_i = 0\right]$.[11]

The population parallel trends assumption in Equation (3.2) is the key identification assumption for the DID design. Equation (3.2) implies that the descriptive difference between treated and

[10] Rosenbaum (2017, chapter 8) prefers the term "counterparts" instead of "controls" for this reason.

[11] This DID estimator in Equation (3.3) is equivalent to the two-way fixed effects (FE) estimator (i.e., linear regression with unit and time fixed effects) (Imai and Kim, 2019; Sobel, 2012; Angrist and Pischke, 2008; Wooldridge, 2005; Kropko and Kubinec, 2018; Egami and Yamauchi, 2019), although this equivalence does not necessarily hold when the DID estimator is defined on more than two groups or time periods (Imai and Kim, 2021; Chaisemartin and D'Haultfœuille, 2021; Goodman-Bacon, 2021).

control populations is equal to a causal difference in the treated population — namely, the ATT in Equation (3.1). Therefore, $\mathrm{E}\left[\widehat{\mathrm{DID}}\right]$ is equal to not only the descriptive quantity of $\mathrm{E}\left[Y_{iT}(1) - Y_{iT-1} \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT} - Y_{iT-1} \mid z_i = 0\right]$, but also for the causal quantity of ATT when the assumption of parallel trends holds. However, as I will argue, this causal identification assumption leads to a pathological property of inference in the DID design.

3.3 Pathology of Causal Identification in DID

To illustrate the pathology of the parallel trends assumption in DID, it is helpful to contrast it with the archetypal model of an observational study (Cochran, 1965). In a randomized experiment, units' individual assignment probabilities are known by design. In an observational study, units' realized assignments are usually assumed to be the result of independent, individual assignment mechanisms with unknown and potentially heterogeneous probabilities. A standard observational design like optimal matching (Rosenbaum, 1989a; Rosenbaum, 1991; Hansen, 2004) aims to resolve this problem by constructing matched strata that can (at least provisionally) justify the claim that individual assignment probabilities are uniform within strata. Under this assumption, conditioning on the observed number of treated units in each matched stratum implies a uniform probability distribution on the set of assignment vectors within each stratum (Rosenbaum, 1984; Rosenbaum, 2002, Sections 3.2 and 3.4), which thereby renders methods for the analysis of complete, block randomized experiments appropriate. A sensitivity analysis can then assess whether any found impact persists over different values of a sensitivity parameter that represent increasingly severe violations of the complete random assignment within matched strata assumption (see Rosenbaum and Krieger, 1990; Gastwirth, Krieger, and Rosenbaum, 2000; Fogarty, 2020).

In this archetypal model of an observational study, identification assumptions are about the assignment mechanism. For example, consider a test of the null hypothesis of no average effect relative to the alternative of a positive average effect: Different assumptions about the assignment mechanism imply different probability distributions of the test statistic, which, in turn, determine whether or not we reject the null hypothesis (see Fogarty, 2020). Since our causal conclusions

(e.g., rejection or not of null in favor of alternative) are based on the statistical distribution of our test statistic, our causal claims are characterized by statistical uncertainty, e.g., type I and type II error probabilities. In DID, by contrast, the key identification assumption of parallel trends is about the causal parameter itself. But because the causal parameter is a fixed quantity of which there is no statistical distribution, assumptions about it do not imply measures of statistical uncertainty.

To see more clearly the problem that parallel trends poses for inference, imagine that one samples the entire population of interest ($\mathcal{S}_n = \mathcal{P}_N$) such that the DID estimator is without sampling uncertainty. In the absence of sampling uncertainty, the parallel trends assumption states that

$$\underbrace{\left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n} y_{iT}(0)}_{\text{unobserved}} - \underbrace{\left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n} y_{iT-1}}_{\text{observed}} = \underbrace{\left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0} y_{iT}}_{\text{observed}} - \underbrace{\left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0} y_{iT-1}}_{\text{observed}}, \tag{3.4}$$

where only the first term, $\left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n} y_{iT}(0)$, is unobserved in a finite sample. Thus, parallel trends boils down to an assumption about the mean of treated units' post-treatment outcomes in the absence of treatment, i.e., $\left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n} y_{iT}(0)$. Since the mean of treated units' treated potential outcomes is observed, parallel trends logically implies a value of the ATT in a finite sample. Hence, the parallel trends assumption is about a fixed causal parameter, *not* the distribution of a random quantity on which our inferences of this fixed parameter are based. Parallel trends therefore implies an absence of statistical uncertainty in our causal conclusions.

This problem cannot be solved by a sensitivity analysis to violations of parallel trends. In a sensitivity analysis to violations of assumptions about the assignment mechanism (Rosenbaum and Krieger, 1990; Gastwirth, Krieger, and Rosenbaum, 2000; Fogarty, 2020), scholars can assess how their conclusions about a *fixed* causal parameter would change over different values of a sensitivity parameter that characterizes the assignment mechanism. Yet each assumption about the assignment mechanism, which justifies one's subsequent causal conclusions, implies some measure of statistical uncertainty in one's causal conclusions. By contrast, in the DID design, an assessment of sensitivity to some violation of parallel trends boils down to another assumption about the mean

of treated units' counterfactual outcomes, a component of the *fixed* causal parameter itself. Hence, assessments of sensitivity to violations of parallel trends fail to statistically capture counterfactual uncertainty.

One solution to this pathology of the parallel trends identification assumption is to invoke an alternative identification assumption altogether. Athey and Imbens (2021), for example, invoke an assumption about the random timing of treatment adoption for all units in a finite sample. Similarly, Rambachan and Roth (2020) consider a finite sample DID setting in which the assignment mechanism is stochastic, but treatment assignment probabilities potentially vary across units. In this setting, the key identification assumption is that the covariance of assignment probabilities and the after-minus-before change in control potential outcomes is equal to 0. I propose an alternative, empirical Bayes' approach that statistically captures uncertainty over the possible violations of parallel trends. Like the canonical DID design, this approach conditions on the observed assignments and does *not* invoke assumptions about units' assignment probabilities or their covariance with potential outcomes. There are at least three reasons why scholars might prefer such an empirical Bayes' approach as opposed to an approach based on assumptions about a random assignment mechanism:

(1) A key feature of the DID design is that its core assumption is not based on analogy with a randomized experiment (Keele, 2020). Indeed, the parallel trends assumption in Equation (3.2) makes no reference to an assignment mechanism. Moreover, such an assumption about the changes in potential outcomes conditional on units' realized assignments might, in many applications, be more convincing than one about units' assignment probabilities.

(2) Even with complete knowledge of either assignment probabilities or the random timing of treatment adoption, inference on such bases is often unfeasible. For example, consider the canonical Card and Krueger (1994) DID study, which draws on employment outcomes measured over time in New Jersey (NJ) and Pennsylvania (PA) to assess the effects of NJ's minimum wage increase on employment in NJ's fast-food restaurants. Exposure of restaurants to treatment (minimum wage increase) occurs at the state level, which would imply a clustered random assignment

process for restaurants in NJ and PA. With only two possible assignments (conditioning on the event that restaurants in only one of the two states are treated), inference on the basis of an assignment mechanism is impractical even if the probability distribution on this set of two assignments were known with certainty. The same problem obtains under inferences based on the random timing of treatment when, as is the case in many applications, there are few time periods.

(3) In a setting with a stochastic assignment mechanism, but potentially heterogeneous assignment probabilities, the ATT estimand will vary over possible assignments depending on which units happen to be treated (Sekhon and Shem-Tov, 2021). The variance of the DID estimator would capture statistical uncertainty surrounding *not* a fixed ATT, but rather the expected value of a random ATT. With unknown and potentially heterogeneous assignment probabilities, the expected ATT is a weighted sum of individual treatment effects, with weights equal to the inverse of units' treatment assignment probabilities, divided by the number of treated units (see Rambachan and Roth, 2020). With unknown weights, which might be unintuitive even if they were known, such an estimand is difficult to interpret and relate to scientific quantities of interest in DID applications.

In the sections to follow, I derive an empirical Bayes' procedure that captures uncertainty over the possible violations of parallel trends. To do so, I first generalize the DID design beyond the parallel trends assumption to the full space of possible trend assumptions. I then decompose the population ATT into two parameters, one of which is characterized by only sampling uncertainty and the other by counterfactual uncertainty. I then show how scholars can calibrate a prior distribution on the causal parameter based on pre-treatment outcome trends and the control groups post-treatment deviation from pre-treatment trends, both of which inform the plausibility of different possible trend assumptions. Finally, I derive the conditions under which the empirical Bayes' uncertainty intervals I propose have correct coverage.

3.4 Generalized Nonparametric DID framework

To generalize the canonical DID design, I introduce the parameter Δ, which quantifies the difference in (1) the pre- to post-treatment change in treated units' mean outcomes had the treatment

not occurred and (2) the pre- to post-treatment change in control units' mean outcomes. Parallel trends is a special case of this general representation in which $\Delta = 0$; however, the space of Δ also captures all possible ways in which parallel trends could be false. Under stratified, simple random sampling from a target population, the parameter Δ is defined as

$$\Delta = \underbrace{\mathrm{E}\left[Y_{iT}(0) \mid z_i = 1\right]}_{\text{Inestimable}} - \underbrace{\mathrm{E}\left[Y_{iT-1} \mid z_i = 1\right]}_{\text{Estimable}} - \left[\underbrace{\mathrm{E}\left[Y_{iT} \mid z_i = 0\right]}_{\text{Estimable}} - \underbrace{\mathrm{E}\left[Y_{iT-1} \mid z_i = 0\right]}_{\text{Estimable}}\right], \tag{3.5}$$

where only the first term, $\mathrm{E}\left[Y_{iT}(0) \mid z_i = 1\right]$, cannot be estimated from sample data. This general representation of the difference in counterfactual trends between treated and control units leads to Proposition 5 below. The proof of this proposition and all others are in the supplementary materials.

Proposition 5. *Under stratified, simple random sampling from a target population, $\mathcal{P}_N$, the difference between the expected value of the DID estimator, $\mathrm{E}\left[\widehat{DID}\right]$, and the ATT estimand is equal to Δ, i.e., $\mathrm{E}\left[\widehat{DID}\right] - \Delta = ATT$. For the special case in which parallel trends is true, $\Delta = 0$ and $\mathrm{E}\left[\widehat{DID}\right] = ATT$.*

When the parallel trends assumption holds, i.e., $\Delta = 0$, then the DID estimator's expected value is equal to the average treatment effect in the treated population. However, if the parallel trends assumption does not hold, i.e., $\Delta \neq 0$, then the expected value of the DID estimator differs from the estimand by Δ. Decomposing identification of ATT in the DID design into $\text{ATT} = \mathrm{E}\left[\widehat{\text{DID}}\right] - \Delta$ enables identification of ATT under any assumption about the value of Δ, whether $\Delta = 0$ (parallel trends) or otherwise.

In addition, the decomposition of ATT in Proposition 5 clarifies that ATT is a function of two quantities, inferences of which are subject to two different types of uncertainty under random sampling from a population. Inference of one quantity, $\mathrm{E}\left[\widehat{\text{DID}}\right]$, is subject to only sampling uncertainty due to the inability to observe all units in a population. Inference of the other quantity, Δ, is subject to counterfactual uncertainty due to the inability to observe counterfactual outcomes

of whichever units are sampled. That is, the ATT can be written as

$$\text{ATT} = \underbrace{\text{E}\left[\widehat{\text{DID}}\right]}_{\substack{\text{Sampling} \\ \text{uncertainty}}} - \underbrace{\Delta.}_{\substack{\text{Counterfactual} \\ \text{uncertainty}}} \tag{3.6}$$

The standard parallel trends assumption asserts not only that $\Delta = 0$, but also that there is no uncertainty in the value of Δ. Hence, in the canonical DID design, overall uncertainty in one's inference of ATT is characterized solely by sampling uncertainty with respect to $\text{E}\left[\widehat{\text{DID}}\right]$, which is equal to a descriptive difference between treated and control populations.

In the argument to follow, the statistical representation of counterfactual uncertainty in the value of Δ is induced *not* by a classical mechanism, such as random sampling or random assignment. Instead, randomness reflects a subjective, prior distribution due to uncertainty about treated units' unobserved counterfactual outcomes in the post-treatment period.

3.5 Empirical Bayes' Identification and Inference under Uncertainty in Parallel Trends

The decomposition of the ATT into the expected value of the DID estimator, subject to only sampling uncertainty, and Δ, subject to counterfactual uncertainty, results in one linear equation with one known (or estimable) parameter — $\text{E}[\text{DID}]$ — and two unknown parameters — ATT and Δ. (See Equation 3.6.) Hence, without imposing a restriction on the value of the nuisance parameter Δ, the ATT is unidentified (and, indeed, has infinitely many solutions). We can achieve identification by imposing an assumption on the value of Δ such that only one value of ATT is consistent with a given value of $\text{E}\left[\widehat{\text{DID}}\right]$ — which is known in a finite sample and unknown, but estimable, under random sampling from a superpopulation. Yet referring back to Equation (3.5), an assumption about Δ boils down to an assumption about a component of the target parameter itself — namely, the mean of treated units' counterfactual outcomes. Thus, a restriction on the value of Δ elides any statistical representation of the fundamental uncertainty that arises from the inability to observe treated units' counterfactual outcomes. In this section, I draw on the Bayesian notion

of *identification in probability* (Drèze, 1972; Drèze, 1975; Drèze, 1976)[12] to derive an empirical Bayes' procedure that does not require exact restrictions on Δ to identify the ATT and is able to statistically capture counterfactual uncertainty.

Let $p(\Delta)$ denote a prior probability distribution function (PDF) on Δ, which, when individual outcomes are unbounded, can take on any real number. However, some values of Δ will typically be more plausible than others, which is reflected by their differing prior probability densities. In this context, we can refer to identification in probability (Drèze, 1972; Drèze, 1975; Drèze, 1976) of the ATT as follows: Instead of an exact, deterministic restriction on the value of Δ, we can instead impose such a restriction on the value of $\mathrm{E}[\Delta]$ such that only one value is consistent with $\mathrm{E}[\mathrm{ATT}]$ given the value of $\mathrm{E}\left[\widehat{\mathrm{DID}}\right]$.[13] By imposing a distribution on Δ and a concomitant assumption on $\mathrm{E}[\Delta]$ rather than Δ itself, we statistically account for counterfactual uncertainty that arises from the inability to observe treated units' counterfactual outcomes. Such counterfactual uncertainty implies that even if the expected difference in trends were equal to 0 (analogous to the canonical parallel trends assumption), other values of Δ are plausible and induce uncertainty in our inferences of ATT.

One concern, however, is that an idiosyncratic prior distribution will overly drive our inferences of ATT. Therefore, to maximize the quality of information that drives our inferences of ATT, I propose an empirical Bayes' approach (so named in Robbins, 1956; Robbins, 1964) wherein the prior distribution is not fixed before the realization of data, but is informed by pre-treatment data instead. In the developments to follow, I propose a procedure that uses the machinery of regression to calibrate the parameters of a Normal prior distribution on Δ to information contained in pre-treatment data and the control group's change in outcomes from the pre- to post-treatment period. As alluded to above, this approach is conservative, especially in contrast to the canonical DID

[12]For valuable discussions on this topic, see Aldrich (2002, especially pp. 87–96) Berger (1985), Dawid (1979), Gustafson (2005) and Gustafson (2009), Hsiao (1983), Kadane (1975), Leamer (1978, especially Chapter 7), Neath and Samaniego (1997), Poirier (1998), Richard (1973), Rothenberg (1971) and Zellner (1971, especially pp. 253–258).

[13]In this framework, I avoid the debate about whether identification of a parameter ought to be conceived as a property of a likelihood function Kadane (1975) or a prior distribution (**lindley1972**). Since identification of $\mathrm{E}[\mathrm{ATT}]$ is defined with reference to its posterior distribution, given a value or estimate of $\mathrm{E}[\mathrm{DID}]$, identification can be satisfied by conditions on either the likelihood function or prior distribution.

design, which is characterized by a degenerate prior probability distribution concentrated on the assumption of parallel trends.

Before laying out this procedure, first consider (for simplicity) finite sample inference. Then define a Normal prior distribution on the nuisance parameter Δ, which is a conservative choice due to the Normal distribution's property of maximum entropy among distributions with finite variance (Cover and Thomas, 1991; Harte, 2011).[14] Based on the definition of Δ in a finite sample, a Normal distribution on Δ implies a normal distribution on the mean of treated units' counterfactual outcomes:

$$\Delta \sim \mathcal{N}\left(\mu, \sigma^2\right), \tag{3.7}$$

which, by the definition of Δ, implies that

$$\left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \left(y_{iT}(0) - y_{iT-1}\right) - \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} \left(y_{iT} - y_{iT-1}\right) \sim \mathcal{N}\left(\mu, \sigma^2\right). \tag{3.8}$$

Because of the closure of the Normal distributions under linear combinations and the fact that $\bar{y}_{1T-1}$ and $\bar{y}_{0T} - \bar{y}_{0T-1}$ are known constants in a finite sample, Equation (3.8) implies that

$$\left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT}(0) \sim \mathcal{N}\left(\mu + \bar{y}_{1T-1} + (\bar{y}_{0T} - \bar{y}_{0T-1}), \sigma^2\right), \tag{3.9}$$

where $\bar{y}_{1T-1} = \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT-1}$, $\bar{y}_{0T} - \bar{y}_{0T-1} = \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} \left(y_{iT} - y_{iT-1}\right)$. Thus, to calibrate the prior distribution on Δ, we need to calibrate only the mean (which we can then shift by the observed values of $\bar{y}_{1T-1}$ and $\bar{y}_{0T} - \bar{y}_{0T-1}$) and variance of the mean of treated units' counterfactual outcomes.

To do this calibration, first denote a predictive (machine-learning) model by $\hat{y}_z(\cdot)$, where $z \in \{0, 1\}$ indicates whether the model is fit in the control ($z = 0$) or treated ($z = 1$) group. Once $\hat{y}_z(\cdot)$ has generated predictions for all control units and for all treated units, we can stochastically

[14] For a formal proof of this property, see McElreath (2020, pp. 306–307), which is based on the treatment in Conrad (2005).

impute the mean of treated units' counterfactual outcomes in a way that accounts for pre-treatment outcome trends and the control groups post-treatment deviation from pre-treatment trends:

Equal Expected Deviation from Trends (EEDT) imputation

$$\hat{\bar{Y}}_T(0) \equiv \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} (y_{iT} - \hat{y}_{iT}) + \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \hat{Y}_{iT}(0), \tag{3.10}$$

where $\hat{\bar{Y}}_T(0)$ is the imputed mean of treated units' control potential outcomes and the prediction of a treated unit's counterfactual outcome, $\hat{Y}_{iT}(0)$, is written in uppercase, which indicates that it is a random quantity.

In principle, any machine-learning model can be used to generate the predictive distribution of $\hat{\bar{Y}}_T(0)$. In the Montalvo (2011) example and the algorithm below, I use a linear time trend regression with individual fixed effects, although the inferential procedure to follow is by no means wedded to this particular choice. Inference of the ATT may of course be sensitive to the choice of specifically how to calibrate the prior distribution on Δ. To lay out the regression procedure for predicting $\hat{\bar{Y}}_T(0)$ (and thus implying a distribution on Δ), consider, first, how to generate the predictive distribution of the mean of treated units' counterfactual outcomes and, second, how to adjust this distribution by the expected prediction error in the control group.

First, let $\boldsymbol{x}_{1,\text{pre}}$ be the model matrix of explanatory variables for the treated group in the pre-treatment period. Denote the treated group's pre-treatment outcomes as $\boldsymbol{y}_{1,\text{pre}}$. Then let $\hat{\boldsymbol{y}}_{1,\text{pre}} = \boldsymbol{x}_{1,\text{pre}}\hat{\boldsymbol{\beta}}_1$ be the linear projection of $\boldsymbol{y}_{1,\text{pre}}$ onto the space of $\boldsymbol{x}_{1,\text{pre}}$, where $\hat{\boldsymbol{\beta}}_1$ is understood not as an estimator for the parameter of a probability model, but instead simply as $\hat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}_1 \in \mathbb{R}^K}{\arg\min} \left\| \boldsymbol{y}_{1,\text{pre}} - \boldsymbol{x}_{1,\text{pre}}\boldsymbol{\beta}_1 \right\|_2$. With the vector $\hat{\boldsymbol{\beta}}_1$, it is straightforward to generate predictions of treated units' post-treatment outcomes had pre-treatment trends continued uninterrupted: $\hat{\boldsymbol{y}}_{1T}(0) = \boldsymbol{x}_{1,T}\hat{\boldsymbol{\beta}}_1$. The vector $\hat{\boldsymbol{y}}_{1T}(0)$ contains the post-treatment projections that are most consistent with pre-treatment trends (i.e., based on the $\boldsymbol{\beta}_1$ that minimizes the sum of squared residuals in the pre-treatment period). Yet even if a particular projection is the most plausible, there are still other projections that remain plausible, albeit slightly less so. Such other projections are captured by the

other values of $\boldsymbol{\beta}_1$ that fit the pre-treatment data, although not as well as the vector $\hat{\boldsymbol{\beta}}_1$ that minimizes the sum of squared residuals. To capture this uncertainty over the set of plausible projections of post-treatment outcomes, we can express the variance-covariance matrix of the post-treatment predictions as $\boldsymbol{x}_{1,T} S^2_{1,\text{pre}} \left(\boldsymbol{x}'_{1,\text{pre}} \boldsymbol{x}_{1,\text{pre}} \right)^{-1} \boldsymbol{x}'_{1,T}$, where $S^2_{1,\text{pre}} = \frac{\boldsymbol{\epsilon}'_1 \boldsymbol{\epsilon}_1}{n_1 - K}$ and $\boldsymbol{\epsilon}_1 = \boldsymbol{y}_{1,\text{pre}} - \boldsymbol{x}_{1,\text{pre}} \hat{\boldsymbol{\beta}}_1$. This expression for $S^2_{1,\text{pre}}$ implies that, all else equal, the more accurate one's projections are in the pre-treatment period and the more pre-treatment data one has, then the more certain will be one's extrapolations to the post-treatment period.

Second, beyond the existence of pre-treatment data in the treated group, we can also directly assess the prediction error of the same linear projection in the control group, $\boldsymbol{y}_{0T} - \hat{\boldsymbol{y}}_{0T}$, where $\hat{\boldsymbol{y}}_{0T} = \hat{\boldsymbol{\beta}}_0 \boldsymbol{x}_{0,T}$, and $\hat{\boldsymbol{\beta}}_0$, $\boldsymbol{x}_{0,T}$, $\boldsymbol{y}_{0,\text{pre}}$ and $\boldsymbol{x}_{0,\text{pre}}$ are defined analogously to how they were in the treated group. The control group's post-treatment deviation from its pre-treatment trend can inform what the same deviation from trend would have been in the treated group in the absence of treatment. Hence, in addition to information from pre-treatment data, we can also draw upon a measure of exactly how informative pre-treatment data are for post-treatment outcomes.

Thus, to calibrate the Normal prior distribution on the mean of treated units' counterfactual outcomes, which implies a Normal prior distribution on Δ after shifting the distribution by the observed values of $\bar{y}_{1T-1}$ and $\bar{y}_{0T} - \bar{y}_{0T-1}$, we can set the hyperparameters μ and σ^2 to

$$\mu = \left(\frac{1}{n_0} \right) \sum_{i=1}^{n_0} \left(\boldsymbol{y}_{0,T} - \boldsymbol{x}_{0,T} \hat{\boldsymbol{\beta}}_0 \right) + \left(\frac{1}{n_1} \right) \sum_{i=n_0+1}^{n} \boldsymbol{x}_{1,T} \hat{\boldsymbol{\beta}}_1 \text{ and} \tag{3.11}$$

$$\sigma^2 = \left(\frac{1}{n_1^2} \right) \sum_{i=n_0+1}^{n} s_i^2 + 2 \sum_{i=n_0+1}^{n} \sum_{i<j} s_{ij}, \tag{3.12}$$

where s_i^2 are the diagonal and s_{ij} are the off-diagonal elements of $\boldsymbol{x}_{1,T} S^2_{1,\text{pre}} \left(\boldsymbol{x}'_{1,\text{pre}} \boldsymbol{x}_{1,\text{pre}} \right)^{-1} \boldsymbol{x}'_{1,T}$.

The reliability of this procedure in terms of finite sample coverage depends on how much the true mean of treated units' counterfactual outcomes differs from the most plausible value of that mean counterfactual, relative to the variance of the prior distribution. To be more explicit, let $\hat{\bar{Y}}_T(0) \equiv \left(\frac{1}{n_0} \right) \sum_{i=1}^{n_0} (y_{iT} - \hat{y}_{iT}) + \left(\frac{1}{n_1} \right) \sum_{i=n_0+1}^{n} \hat{Y}_{iT}(0)$ as given in Equation (3.10). Consider finite sample

inference with a fixed level of $\alpha \in (0, 1)$ and a Normal prior distribution on Δ. A two-sided $(1 - \alpha)$ 100% uncertainty interval brackets ATT with probability at least as great as $1 - \alpha$ for any Var $[\Delta]$ if and only if

$$\left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT}(0) - \mathrm{E}\left[\left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \hat{Y}_{iT}(0)\right] = \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT} - \mathrm{E}\left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} \hat{Y}_{iT}\right]. \tag{3.13}$$

By contrast, when Equation (3.13) is false, a two-sided $(1 - \alpha)$ 100% uncertainty interval brackets the ATT with probability at least as great as $1 - \alpha$ if and only if

$$\left| \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} (\hat{y}_{iT}(0) - y_{iT}(0)) - \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} (\hat{y}_{iT} - y_{iT}) \right| \leq \Phi^{-1}_{1-\alpha/2} \, \mathrm{Var}\left[\Delta\right]^{1/2}, \tag{3.14}$$

where $\Phi^{-1}_{1-\alpha/2}$ is the quantile function of the standard Normal distribution evaluated at $1 - \alpha/2$.

The coverage properties of the empirical Bayes' inferential procedure depend fundamentally on the prior distribution of Δ. If the predictions in treated and control groups have equal expected deviations from their true values, then the left-hand side of the inequality in (3.14) is equal to 0 and the $(1 - \alpha)$ 100% uncertainty interval will always possess proper coverage no matter the variance of Δ. By contrast, in the (likely) event that Equation (3.13) does not hold exactly, correct coverage depends on the magnitude of the expected difference in prediction errors relative to the level of uncertainty in Δ. For example, a violation of equal expected deviation from trends in Equation (3.13) would have to be roughly twice the standard deviation of Δ in order for the 95% uncertainty interval to have coverage probability less than 0.95. Therefore, even if counterfactual patterns like Ashenfelter's dip (Ashenfelter, 1978) — common in, e.g., job training programs — and regression to the mean more broadly (Daw and Hatfield, 2018) exist, the proposed uncertainty intervals will maintain proper coverage so long as such patterns do not deviate too extremely from the data used to calibrate the variance of Δ.

To implement this empirical Bayes' procedure in practice, we can draw on what Gelman and Hill (2006, chapter 7) term an "informal Bayesian approach." In this procedure, given in Algorithm 1, we set the hyperparameters of a multivariate Normal prior distribution, $\mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on the treated

group model's K dimensional vector, $\boldsymbol{\beta}_1$, to $\boldsymbol{\mu} = \hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}_1$. We then repeatedly draw from this multivariate Normal prior distribution and generate plausible predictions of the counterfactual mean outcome of treated units, which imply plausible values of Δ. With this simulation-based approximation to the distribution of Δ, we can then generate a simulation-based approximation to the posterior distribution of ATT by taking $\widehat{\text{DID}} - \Delta$.

When sampling uncertainty exists in addition to counterfactual uncertainty, pre-treatment outcome data and the control group's change in outcomes from the pre- to post-treatment period cannot be directly observed. Nevertheless, the procedure for calibrating the distribution of Δ remains informative about Δ under the assumption that the realized sample approximates the population. This assumption is equivalent to the standard assumption in the nonparametric bootstrap (Efron and Tibshirani, 1994) and, hence, it can be used to combine both counterfactual and sampling uncertainty. As Algorithm 1 shows, each treated unit has a random distribution of imputations for its missing potential outcome, $\hat{Y}_{iT}(0)$ for all $i \in \{n_0 + 1, \ldots, n\}$, and all control units have an expected individual prediction for their post-treatment outcomes, $\hat{y}_{iT}$ for all $i \in \{1, \ldots, n_0\}$. These quantities imply not only a distribution on Δ in the superpopulation (based on the realized sample as an approximating distribution), but also a distribution on the finite sample Δ for whichever units happen to have been sampled (with replacement) via the bootstrap. With treated units' random imputations of their missing potential outcomes and control units' individual predictions of post-treatment outcomes, it is straightforward to generate the distributions on ATT for a given bootstrapped sample and the overall distribution of ATT over all samples.

When the analogous population version of Equation (3.13) is true, the $(1 - \alpha)\,100\%$ uncertainty intervals will always have correct coverage regardless of the prior variance of Δ. However, when Equation (3.13) is false, coverage depends on the magnitude of the violation of Equation (3.13) vis-á-vis the overall variance of $\widehat{\text{DID}} - \Delta$, which consists of the variances of $\widehat{\text{DID}}$ and Δ, as well as their covariance, where $\text{Var}\left[\Delta_{\mathcal{P}_N}\right] = \text{Var}\left[\text{E}\left[\Delta_{\mathcal{S}_n}\right]\right] + \text{E}\left[\text{Var}\left[\Delta_{\mathcal{S}_n}\right]\right]$ and $\text{Cov}\left[\widehat{\text{DID}}_{\mathcal{P}}, \Delta_{\mathcal{P}_N}\right] = \text{E}\left[\left(\widehat{\text{DID}}_{\mathcal{P}} - \text{E}\left[\widehat{\text{DID}}_{\mathcal{P}}\right]\right)\left(\text{E}\left[\Delta_{\mathcal{S}_n}\right] - \text{E}\left[\text{E}\left[\Delta_{\mathcal{S}_n}\right]\right]\right)\right]$. The inner expectations and variances, defined conditional on a realized sample, reflect counterfactual uncertainty due to the inability to observe

sampled units' counterfactual outcomes. The outer expectations and variances, defined over the set of bootstrapped samples, reflect how causal quantities vary across samples.

Algorithm 1: Inference of ATT

Data: Long-format dataframe with variables measured over all time periods

Result: Vector of ATT estimates of length S if target is $\text{ATT}_{\mathcal{S}_n}$ or length RS if target is $\text{ATT}_{\mathcal{P}_N}$

Set $\widehat{\text{DID}} \leftarrow \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT}(1) - y_{iT-1} - \left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT} - y_{iT-1}\right]$

Set $\hat{\boldsymbol{\beta}}_z \leftarrow \underset{\boldsymbol{\beta}_z \in \mathbb{R}^K}{\arg\min} \left\|\boldsymbol{y}_{z,\text{pre}} - \boldsymbol{x}_{z,\text{pre}}\boldsymbol{\beta}_z\right\|_2$ for $z \in \{0, 1\}$ and $\widehat{\boldsymbol{\Sigma}}_1 \leftarrow \frac{\boldsymbol{\epsilon}_1'\boldsymbol{\epsilon}_1}{n_1 - K}\left(\boldsymbol{x}_{1,\text{pre}}'\boldsymbol{x}_{1,\text{pre}}\right)^{-1}$

Generate predictions for control units' post-treatment outcomes, $\hat{\boldsymbol{y}}_{0,T} \leftarrow \boldsymbol{x}_{0,T}\hat{\boldsymbol{\beta}}_0$

Set hyperparameters of multivariate Normal prior distribution, $\mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on treated group model's K dimensional parameter vector, $\boldsymbol{\beta}_1$, to $\boldsymbol{\mu} \leftarrow \hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\Sigma} \leftarrow \widehat{\boldsymbol{\Sigma}}_1$

for $s \in \{1, \ldots, S\}$, where S is the total number of draws, **do**

  Randomly draw coefficient vector from $\mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and call this vector $\tilde{\boldsymbol{\beta}}^s$

  Generate predictions for treated units' counterfactual post-treatment outcomes $\hat{\boldsymbol{y}}^s_{1,T}(0) \leftarrow \boldsymbol{x}_{1,T}\tilde{\boldsymbol{\beta}}^s$

end

if target is $\text{ATT}_{\mathcal{S}_n}$ **then**

  for $s \in \{1, \ldots, S\}$, **do**

    Set mean of treated units' counterfactual potential outcomes $\bar{y}_T^{\text{imp},s}(0) \leftarrow \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} (y_{iT} - \hat{y}_{iT}) + \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \hat{y}^s_{iT}(0)$

    Set difference in counterfactual trends $\Delta^{\text{imp},s} \leftarrow \bar{y}_T^{\text{imp},s}(0) - \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT-1} - \left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT} - \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT-1}\right]$

    Estimate effect $\widehat{\text{ATT}}^s \leftarrow \widehat{\text{DID}} - \Delta^{\text{imp},s}$

  end

  return $\left\{\widehat{\text{ATT}}^1, \ldots, \widehat{\text{ATT}}^S\right\}$

else if target is $\text{ATT}_{\mathcal{P}_N}$ **then**

  for $r \in \{1, \ldots, R\}$, where R is the total number of draws, **do**

    Randomly draw with replacement samples of size n_0 and n_1 from $\{y_{i0}, \ldots, y_{iT}, \hat{y}_{iT}\}_{i=1}^{n_0}$ and $\left\{y_{i0}, \ldots, y_{iT-1}, y_{iT}(1), \hat{y}^1_{iT}(0), \ldots, \hat{y}^S_{iT}(0)\right\}_{i=n_0+1}^{n}$

    Set $\widehat{\text{DID}}^r \leftarrow \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y^r_{iT}(1) - y^r_{iT-1} - \left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y^r_{iT} - y^r_{iT-1}\right]$

    for $s \in \{1, \ldots, S\}$ **do**

      Set $\bar{y}_T^{\text{imp},r,s}(0) \leftarrow \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} \left(y^r_{iT} - \hat{y}^r_{iT}\right) + \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \hat{y}^{r,s}_{iT}(0)$

      Set $\Delta^{\text{imp},r,s} \leftarrow \bar{y}_T^{\text{imp},r,s}(0) - \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y^r_{iT-1} - \left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y^r_{iT} - \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y^r_{iT-1}\right]$

      Estimate effect $\widehat{\text{ATT}}^{r,s} \leftarrow \widehat{\text{DID}}^r - \Delta^{\text{imp},r,s}$

    end

  end

  return $\left\{\widehat{\text{ATT}}^{1,1}, \ldots, \widehat{\text{ATT}}^{1,S}, \ldots, \widehat{\text{ATT}}^{R,1}, \ldots, \widehat{\text{ATT}}^{R,S}\right\}$

The output of Algorithm 1 consists of a simulation-based approximation to the posterior distribution of the ATT, which captures both sampling (if it exists) and counterfactual uncertainty. Since the posterior distribution is Normal, the most probable value of the ATT will be the mean of this distribution (see Jaynes, 2003, pp. 172–175). Hence, the maximum a posteriori (MAP) point estimate is the mean of the output returned by Algorithm 1. Scholars can combine this point estimate with $(1-\alpha)100\%$ uncertainty intervals for $\alpha \in (0,1)$ by evaluating the lower $(\alpha/2)$ and upper $(1-\alpha/2)$ quantiles of the simulation-based approximation to the posterior distribution of ATT.

To demonstrate this algorithm, consider again the example from Montalvo (2011). In this study, a reasonable choice for calibrating the distribution of Δ is a linear time trend with individual fixed effects:

$$\hat{y}_{it} = \beta_0 + \beta_1 t + \alpha_i \text{ for } i \in \{1, \ldots, n_0\} \text{ and } t \in \{0, \ldots, T-1\} \tag{3.15}$$

$$\hat{y}_{it} = \beta_0 + \beta_1 t + \alpha_i \text{ for } i \in \{n_0+1, \ldots, n\} \text{ and } t \in \{0, \ldots, T-1\}, \tag{3.16}$$

where $\hat{y}_{it}$ is the predicted value of the outcome for unit i at time t, β_0 is an intercept term, β_1 is the coefficient of t, which is a counter that is equal to 0 in the first baseline period and increases by 1 for each succeeding period up to the last pre-treatment period, and α_i is the individual fixed effect for the ith unit. Separately fitting the two trends to control and treated groups in Equations 3.15 and 3.16 is equivalent to

$$\hat{y}_{it} = \beta_0 + \beta_1 t + \beta_2 z_i + \theta z_i t + \alpha_i \text{ for } i \in \{1, \ldots, n\} \text{ and } t \in \{0, \ldots, T-1\}, \tag{3.17}$$

where β_2 is the coefficient of the treatment variable, z_i, and θ is the coefficient of the interaction between treatment and the time index. The inclusion of the interaction term enables differential pre-treatment trends between treated and control groups.

For inference in a finite sample, the data — y_{it}, z_i and t — are fixed for all units, while the vector of coefficients is random. (See Steps 4–6 in Algorithm 1.) Since the distribution on the vector of coefficients $\boldsymbol{\beta}_1$ is multivariate Normal, the distribution on ϵ_{it} is univariate Normal. The expected

value of ϵ_{it} is $\mathrm{E}[\epsilon_{it}] = 0$ since the predictive models are fit with individual intercept terms for each unit. The variance of ϵ_{it}, which is derived from the expression for the variance of a linear combination of correlated random variables, is $\mathrm{Var}[\epsilon_{it}] = \sum_{k=1}^{K} x_{itk}^2 \mathrm{Var}(\beta_k) + 2 \sum_{k=1}^{K} \sum_{k<l} x_{itk} x_{itl} \mathrm{Cov}(\beta_k, \beta_l)$, where x_{itk} and x_{itl} are the respective values (including the intercepts) for the k^{th} and l^{th} independent variables of the i^{th} unit in the t^{th} period. The distribution on ϵ_{it} is derived not from a probability model on the outcome, y_{it}, but from a random distribution on $\boldsymbol{\beta}_1$ that is calibrated to pre-treatment data.

The predictions in Equations (3.15) and (3.16) can then be used to generate the expected predictions and corresponding uncertainty intervals of the mean post-treatment outcome in treated and control groups, respectively. Step 4 in **??** above implies that the expected coefficient vector in the treated group is $\mathrm{E}[\boldsymbol{\beta}_1] = \hat{\boldsymbol{\beta}}_1$. Hence, in the treated and control groups, the expected predictions can be written as $\boldsymbol{x}_{1,T} \hat{\boldsymbol{\beta}}_1$. Figure 3.2 illustrates the trends fit separately to pre-treatment data in treated and control groups and then used to predict the mean outcome in the post-treatment period for each group.

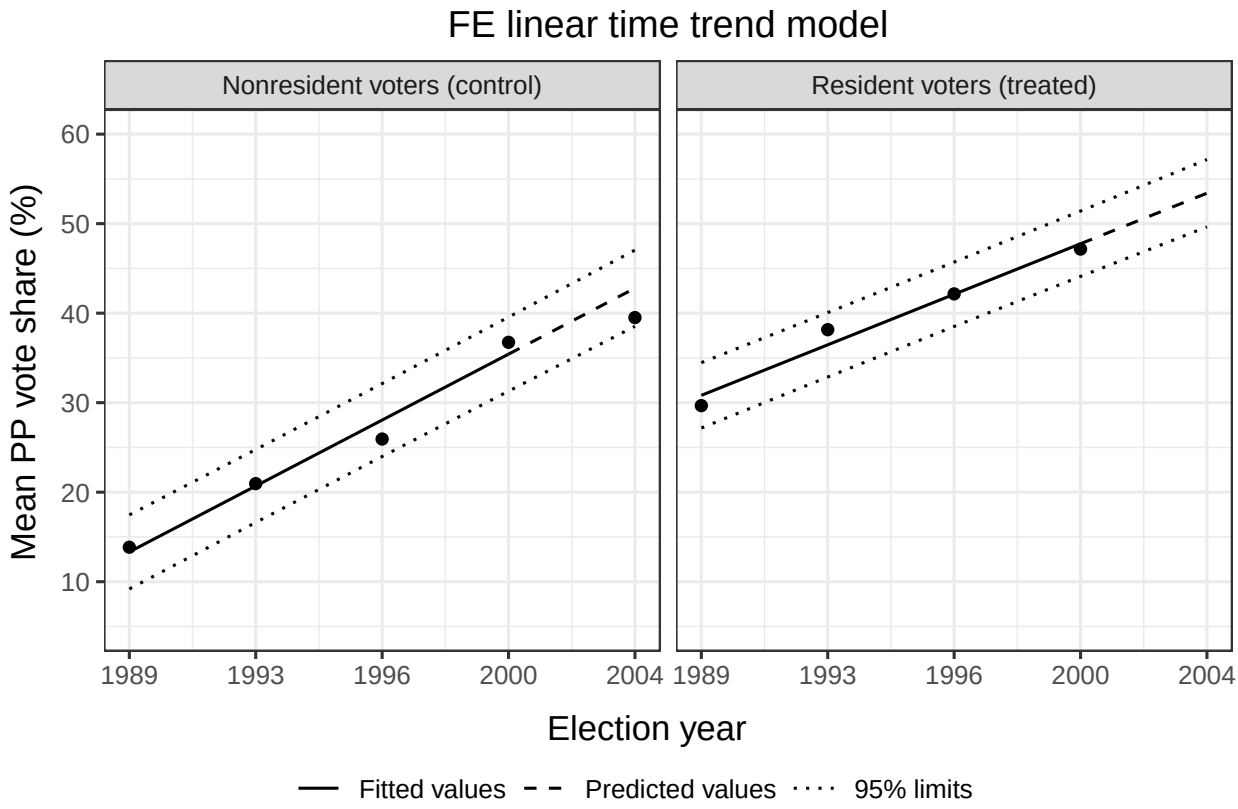

Figure 3.2: Linear time trends separately fit to pre-treatment data in treated and control groups (data from Montalvo, 2011)

Once we have generated predictions of the the post-treatment mean outcomes in treated and control groups based on pre-treatment trends, we can then impute the mean of treated units' post-treatment outcomes in the absence of treatment according to the imputation scheme given in Step 11. Following this imputation, we can then estimate ATT_{S_n} according to Steps 12 – 13 in **??**.

The output of Algorithm 1 is an entire simulation-based approximation to the posterior distribution of ATT. From this entire distribution, we can calculate a MAP point estimate, a $(1-\alpha)$ 100 uncertainty interval, or various other features of the distribution. Given the use of regression tools to calibrate the prior distribution on Δ, which in turn informs the MAP point estimate and $(1-\alpha)$ 100 uncertainty interval, the procedure thus far might seem only to be proposing a new parametric model for DID designs (much like recent literature that incorporates group-specific time trends

into regression specifications, e.g., Dobkin et al., 2018; Goodman-Bacon, 2021; Bhuller et al., 2013). Yet, as mentioned at the outset, this paper's focus is on inference, not estimation. Hence, Corollary 1 below shows that it is possible to represent the MAP point estimator of the posterior distribution returned by Algorithm 1 as a standard model-based estimator. Yet despite this equivalence in terms of estimation, the empirical Bayes' procedure above has different implications for inference.

Corollary 1. *The MAP estimator,* $\arg\max_{ATT} f\left(ATT \mid \widehat{DID}\right)$*, is numerically equivalent to the model-based estimator given by*

$$\frac{1}{n_1}\sum_{i=n_0+1}^{n}\left(y_{iT}(1)-\hat{Y}_{iT}(0)\right)-\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\hat{y}_{iT}-y_{iT}\right). \tag{3.18}$$

In terms of inference, the procedure in Algorithm 1 implies that Equation (3.18) has at least two key differences from a model-based estimator: First, according to Algorithm 1, the expression in Equation (3.18) represents a random distribution on the parameter ATT conditional on a realized sample. Hence, in contrast to a model-based estimator, Equation (3.18) does *not* represent a single point estimate that varies across possible samples from a population. Second, model-based estimators, such as two-way fixed effects regression, are often justified based on their equivalence with a model-free estimator that recovers the ATT under an identification assumption about potential outcomes. Equation (3.18), by contrast, represents a random distribution on the ATT that is induced by a prior distribution over the space of possible identification assumption (i.e., the space of possible values of the mean of treated units' counterfactual outcomes) in a fully generalized DID design.

Insofar as this model extracts information from pre-treatment data to inform the Normal prior distribution on Δ, the resulting inferences of ATT will be sensitive to the choice of prior distribution. Fortunately, this approach permits straightforward assessments of how one's inferences of ATT would change under different prior distributions on Δ. In general, there are many ways to calibrate the distribution on Δ, such as by incorporating various weights for each time period and

autoregressive structures, such as those in Bloom, Riccio, and Verma (2005), Miratrix (2019) and others. Section 3.6 assesses the robustness of results to alternative model choices to calibrate the prior distribution. Despite the sensitivity of this approach, it will be less than that of the canonical DID design wherein the identifying restriction that $\Delta = 0$ is equivalent to a degenerate prior probability distribution on Δ that is concentrated on the assumption of parallel trends ($\Delta = 0$).

3.6 Comparing and combining sampling versus counterfactual uncertainty

Thus far, I have proceeded by decomposing the ATT into the expected value of the DID estimator, which is equal to the difference in after-minus-before means between treated and control populations, and the difference in counterfactual trends. When one samples the entire population of interest, there is no sampling uncertainty for inference of the first quantity, but counterfactual uncertainty exists for the second. This section now compares the practical stakes in the Montalvo (2011) study of (1) erroneously using sampling-based standard errors for inference of the ATT in a finite sample versus (2) statistically representing counterfactual uncertainty in a finite sample as laid out in this paper or (3) capturing both sampling and counterfactual uncertainty when each of which exists. I then show how scholars can assess the sensitivity of their inferences to the choice for calibrating the Normal prior distribution on Δ.

Perfunctory variance calculations in the DID design are based on elementary theory of survey sampling (e.g., Cochran, 1977; Lohr, 2010; Kish, 1965). This theory from survey sampling yields analytic expressions for the variance of the DID estimator when the assumed treated and control superpopulations are finite, $N_0, N_1 < \infty$, and when they are infinite, $N_0, N_1 = \infty$. In the latter case, the variance of the DID estimator, $\widehat{\text{DID}}$, is

$$\text{Var}\left[\widehat{\text{DID}}\right] = \left(\frac{1}{n_0}\right) \text{Var}\left[(Y_T - Y_{T-1}) \mid Z_T = 0\right] + \left(\frac{1}{n_1}\right) \text{Var}\left[(Y_T(1) - Y_{T-1}) \mid Z_T = 1\right]. \quad (3.19)$$

The two unknown superpopulation variances in Equation (3.19) can be unbiasedly and consistently estimated by

$$\widehat{\text{Var}}\left[(Y_T - Y_{T-1}) \mid Z_T = 0\right] = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} \left((Y_T - Y_{T-1}) - \left(\frac{1}{n_0} \sum_{i=1}^{n_0} (Y_T - Y_{T-1}) \right) \right)^2 \text{ and} \tag{3.20}$$

$$\widehat{\text{Var}}\left[(Y_T(1) - Y_{T-1}) \mid Z_T = 1\right] = \frac{1}{n_1 - 1} \sum_{i=n_0+1}^{n_1} \left((Y_T(1) - Y_{T-1}) - \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n_1} (Y_T(1) - Y_{T-1}) \right) \right)^2, \tag{3.21}$$

assuming that $n_0, n_1 > 1$. Interval estimation and hypothesis tests can then proceed via Normal approximations that appeal to central limit theorems or the nonparametric bootstrap (as proposed for the DID design by Bertrand, Duflo, and Mullainathan, 2004).

The perfunctory use of such survey sampling-based standard errors when the target is a causal quantity in a finite sample often leaves information on the table. In the context of the Montalvo (2011) study, Table 3.1 below shows the differences between point estimates and standard errors calculated via the standard sampling-based approach to inference, the method described in this paper and the combination of both types of uncertainty.

	Type of uncertainty:		
	Sampling	Counterfactual	Both
Point estimate of ATT	-8.56*	-8.76*	-8.76*
	(0.88)	(0.62)	(1.14)
Note:		*p<0.05	

Table 3.1: Point estimates and standard errors under different inferential procedures in the Montalvo (2011) study

The point estimate in the first column of ≈ -8.56 is the estimate from the usual DID estimator of ATT under the assumption of parallel trends. The MAP estimate of ≈ -8.76 in the second column is the expected value of the posterior distribution of ATT in which the prior distribution on Δ is calibrated via the FE linear time trends in Equations (3.15) and (3.16). This point estimate of ≈ -8.76 differs slightly from the estimate of ≈ -8.56 calculated under the assumption of parallel trends. This slight difference is because the most probable difference in counterfactual trends, i.e., $\mathrm{E}[\Delta]$, is set to that which is most consistent with the pre-treatment data given the trends in

Equations (3.15) and (3.16). The most probable difference in counterfactual trends is ≈ 0.2, which in turn yields the MAP estimate for the ATT of ≈ -8.76. The point estimate of ≈ -8.56 under the parallel trends assumption assumes a difference in counterfactual trends slightly less consistent with pre-treatment trends.

As Table 3.1 also shows, the standard errors under counterfactual uncertainty are roughly 30% smaller than the standard errors under sampling uncertainty. Note that, for finite sample inference, the nonparametric bootstrap implicitly encodes a prior distribution on Δ in which $\mathrm{E}[\Delta] = 0$ and $\mathrm{Var}[\Delta]$ is equal to the sum of Equations (3.20) and (3.21). This approach clearly leaves information on the table given the strong trends in the pre-treatment period that continue (in the control group) into the post-treatment period. However, the standard error under the combination of counterfactual and sampling uncertainty via the nonparametric bootstrap is roughly 0.94. Hence, the corresponding 95% interval is wider than either of the intervals under only one source of uncertainty. All of the p-values are significant — i.e., have uncertainty intervals that exclude 0 — at the $\alpha = 0.05$ level. Nevertheless, the substantial differences in uncertainty point to the practical implications of clarifying exactly what the target estimand is and which forms of statistical uncertainty can be interpreted as counterfactual uncertainty surrounding that estimand.

Given that inference of the ATT will be driven by the choice of how to calibrate the prior distribution on Δ, it is important as a matter of practice to assess the sensitivity of inferences to this choice. The burgeoning literature on more flexible DID estimators that directly impute counterfactual potential outcomes can be helpful in this regard. For example, two-way linear fixed effects (Imai and Kim, 2019; Imai and Kim, 2021), interactive fixed-effects, (Gobillon and Magnac, 2016; Bai, 2009) synthetic control (Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2010; Abadie, Diamond, and Hainmueller, 2012), as well as recent generalizations of it (Amjad, Shah, and Shen, 2018; Xu, 2017), and matrix completion (Athey et al., 2020) methods can all be helpful for calibrating the prior distribution on Δ, even though these methods were originally developed in a different context focused on direct estimation of the ATT.

One issue is that the synthetic control and matrix completion methods require a large number

of pre-treatment periods and control observations. For example, Xu (2017, p. 73) recommends at least 10 pre-treatment periods and 40 control units. With only 4 pre-treatment periods, the `gsynth` and `MCPanel` packages in `R` are unable to fit a model to the Montalvo (2011) data and return error messages due to the insufficient number of pre-treatment periods. Instead, I assess the sensitivity of the FE linear time trends in Equations (3.15) and (3.16) to a linear lagged dependent variable (DV) and time trend with unit fixed effects (as in Doudchenko and Imbens, 2017) given in Equations (3.22a) and (3.22b) below:

$$\hat{y}_{it} = \alpha_i + \beta_1 y_{it-1} + \beta_2 t \text{ for } i \in \{1, \ldots, n_0\} \text{ and } t \in \{1, \ldots, T-1\} \tag{3.22a}$$

$$\hat{y}_{it} = \alpha_i + \beta_1 y_{it-1} + \beta_2 t \text{ for } i \in \{n_0+1, \ldots, n\} \text{ and } t \in \{1, \ldots, T-1\}, \tag{3.22b}$$

where the baseline time period, $t = 0$, is excluded since the lagged mean PP vote share percentage does not exist for the first time period.

The posterior distributions of the ATT under when calibrating the distribution of Δ to either Equations (3.15) and (3.16) or Equations (3.22a) and (3.22b) are given in Figure 3.3 below.

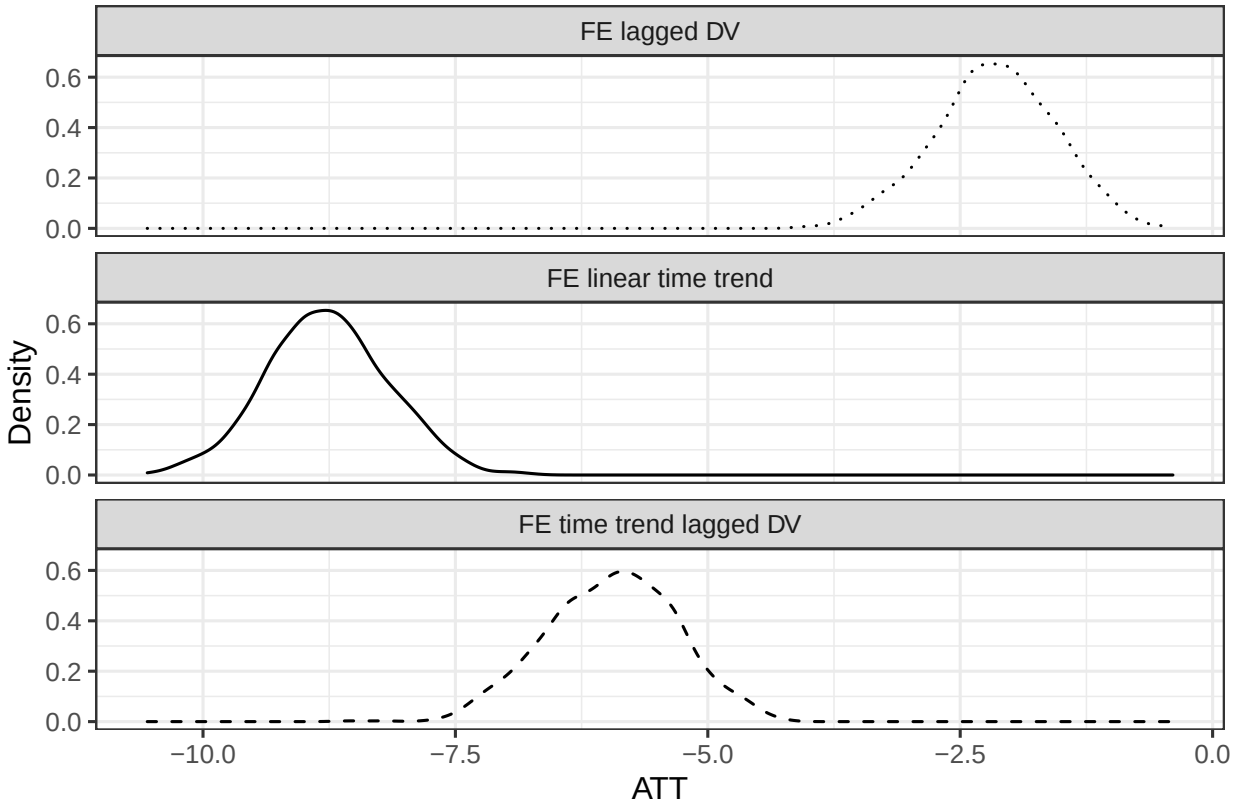

Figure 3.3: Posterior distributions of ATT given different choices for calibrating distribution of Δ (data from Montalvo, 2011)

Under both distributions on Δ, the ATT is negative. However, the effect under the FE linear time trend calibration suggests an expected ATT with slightly greater magnitude. Like any sensitivity analysis, the results in Figure 3.3 cannot tell us which among the differing approaches for calibrating Δ is most plausible. Therefore, interpreting the sensitivity analysis requires careful assessment of which approach is indeed most plausible. Nevertheless, inferences are relatively robust under these two choices and it is straightforward to assess robustness to other choices as well.

3.7 Discussion and Conclusion

I have argued that uncertainty in the ATT can be decomposed into sampling uncertainty of the DID estimator and counterfactual uncertainty of the difference in counterfactual trends. In contrast to the canonical DID design, statistical uncertainty depends on the difference in counterfactual trends distribution, which is calibrated to pre-treatment data via a predictive model fit separately to treated and control groups. Uncertainty in such inferences will accurately represent the inference from observed to counterfactual outcomes, which is the causal inference scholars are often most interested in. Standard design-based methods in the DID design reflect only sampling uncertainty, not counterfactual uncertainty.

The method I propose in the paper is just one way to express counterfactual uncertainty in the DID and other related designs. Inference under uncertainty in the difference in counterfactual trends is also possible via tools developed for the synthetic control design. Inference in this context proceeds by generating a distribution of either "in-time placebos" or "in-space placebos" (Abadie, Diamond, and Hainmueller, 2012, pp. 499–500) and then assessing the proportion of placebos that are at least as extreme as the estimated effect.[15] Recently, Hasegawa, Webster, and Small (2019) and Keele, Hasegawa, and Small (2019) have applied such placebo-based inference to the DID design. However, this inferential procedure will be uninformative for a given null hypothesis about the ATT when there are few pre-treatment periods, which is why Abadie, Diamond, and Hainmueller (2012, p. 500) "do not recommend using this method when . . . the number of pretreatment periods is small." Many DID designs do indeed have too few pre-treatment periods for placebo-based inference to be informative, which makes inference proposed in this paper attractive in such cases. In the Montalvo (2011) study, for example, there are 4 pre-treatment periods and, if one were to calculate placebo DID estimates on each pair of adjacent pre-treatment periods, the

[15]This form of placebo-based inference has been used in many applied studies on a range of topics. Examples include studies on the effect of direct democracy on naturalization decisions for immigrants (Hainmueller and Hangartner, 2019), the personal connections of former U.S. Treasury Secretary (2009 – 2013) Timothy Geithner on stock market prices (Acemoğlu et al., 2016), minimum wage laws on subsequent employment and wages (Dube and Zipperer, 2015), immigration employment laws on the size of immigrant populations (Bohn, Lofstrom, and Raphael, 2014), and employer health insurance laws on health insurance coverage and labor demand (Buchmueller, DiNardo, and Valletta, 2011) among others.

minimum possible placebo-based p-value would be 0.25. As I have attempted to illustrate, more informative inferences of $\text{ATT}_{\mathcal{S}_n}$ are possible despite the existence of only 4 pre-treatment periods. Hence, the method in this paper offers a valuable complement to other methods of causal inference when identification is not based on claims about the assignment mechanism.

The arguments and methods proposed in this paper also relate to some of the foundational literature on causal inference in observational studies (Cochran, 1965; Cochran, 1983), which states that the quality of an observational design can be evaluated based on the extent to which it is analogous with an ideal randomized experiment (see also Bind and Rubin, 2019; Hansen and Bowers, 2008; Rubin, 2007; Rubin, 2008). However, the DID design's core identification assumption is akin to the "scientific" as opposed to the "statistical solution" to causal inference (Holland, 1986, p. 947).[16] The nature of uncertainty under the "scientific solution" is clearly explicated by Abadie, Diamond, and Hainmueller (2010, pp. 496–497) who write that statistical uncertainty represents "ignorance about the ability of the control group to reproduce the counterfactual of how the treated units would have evolved in the absence of treatment." One potentially fruitful extension would be to compare the inferential properties via the method proposed in this paper with others, such as placebo-based p-values, in a range of contexts.

Other possible extensions include applying the inferential method in this paper to closely related designs. Examples of relevant designs include the aforementioned synthetic control, as well as others that draw upon measurements of the same units over time as a source of causal leverage (Blackwell and Glynn, 2018; De Boef and Keele, 2008; Wooldridge, 2010). More broadly, inference in this paper can potentially be extended to designs that solve the "fundamental problem of causal inference" (Holland, 1986, p. 947) via invariance assumptions on potential outcomes instead of assumptions about a stochastic assignment mechanism. For example, another common invariance assumption is continuity of potential outcomes, as in the regression discontinuity design (Hahn, Todd, and Klaauw, 2001) and, more specifically, the regression discontinuity in time

[16] By the "scientific solution" to causal inference, as opposed to the "statistical solution," Holland (1986) refers to designs based on "invariance" (also known as "homogeneity") assumptions. These assumptions stipulate, in some form or another, that an observed quantity is a proxy for an unobservable counterfactual quantity.

design (Hausman and Rapson, 2018).[17] Overall, designs that invoke some form of an invariance assumption for causal identification are ubiquitous in a range of applications. This paper offers one way in which expressions of statistical uncertainty can be made to reflect uncertainty in these designs' causal conclusions.

[17] For an approach to inference in the regression discontinuity design based on continuity arguments instead of local randomization, see Eckles et al. (2020).

References

Abadie, Alberto (2005). "Semiparametric Difference-in-Differences Estimators". In: *The Review of Economic Studies* 72.1, pp. 1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505.

— (2012). "Comparative Politics and the Synthetic Control Method". In: *American Journal of Political Science* 59.2, pp. 495–510.

Abadie, Alberto and Javier Gardeazabal (2003). "The Economic Costs of Conflict: A Case Study of the Basque Country". In: *The American Economic Review* 93.1, pp. 113–132.

Abadie, Alberto et al. (2020). "Sampling-Based versus Design-Based Uncertainty in Regression Analysis". In: *Econometrica* 88.1, pp. 265–296.

Acemoğlu, Daron et al. (2016). "The Value of Connections in Turbulent Times: Evidence from the United States". In: *Journal of Financial Economics* 121.2, pp. 368–391.

Aldrich, John (2002). "How Likelihood and Identification went Bayesian". In: *International Statistical Review / Revue Internationale de Statistique* 70.1, pp. 79–98.

Amjad, Muhammad, Devavrat Shah, and Dennis Shen (2018). "Robust Synthetic Control". In: *Journal of Machine Learning Research* 19.22, pp. 1–51.

Andrews, Donald W K (1991). "Asymptotic Normality of Series Estimators for Nonparameric and Semiparametric Regression Models". In: *Econometrica* 59.2, pp. 307–345.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

— (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics". In: *The Journal of Economic Perspectives* 24.2, pp. 3–30.

Aronow, Peter M, Donald P Green, Donald KK Lee, et al. (2014). "Sharp Bounds on the Variance in Randomized Experiments". In: *The Annals of Statistics* 42.3, pp. 850–871.

Aronow, Peter M and Benjamin T Miller (2019). *Foundations of Agnostic Statistics*. New York, NY: Cambridge University Press.

Ashenfelter, Orley (1978). "Estimating the Effect of Training Programs on Earnings". In: *The Review of Economics and Statistics* 60.1, pp. 47–57.

Ashenfelter, Orley and David Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs". In: *The Review of Economics and Statistics* 67.4, pp. 648–660.

Athey, Susan and Guido W Imbens (2006). "Identification and Inference in Nonlinear Difference-in-Differences Models". In: *Econometrica* 74.2, pp. 431–497.

— (2021). "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption". In: *Journal of Econometrics*.

Athey, Susan et al. (2020). "Matrix Completion Methods for Causal Panel Data Models". Preprint, `https://arxiv.org/pdf/1710.10251.pdf`.

Bai, Jushan (2009). "Panel Data Models with Interactive Fixed Effects". In: *Econometrica* 77.4, pp. 1229–1279.

Bali, Valentina A (2007). "Terror and elections: Lessons from Spain". In: *Electoral Studies* 26.3, pp. 669–687.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg (2017). "Decision Theoretic Approaches to Experiment Design and External Validity". In: *Handbook of Field Experiments*. Ed. by Esther Duflo and Abhijit Banerjee. Vol. 1. Amsterdam, NL: North-Holland. Chap. 4, pp. 141–174.

Banerjee, Abhijit et al. (2020). "A Theory of Experimenters: Robustness, Randomization, and Balance". In: *American Economic Review* 110.4, pp. 1206–1230.

Bareinboim, Elias and Judea Pearl (2013). "A General Algorithm for Deciding Transportability of Experimental Results". In: *Journal of Causal Inference* 1.1, pp. 107–134.

Basu, Debabrata (1980). "Rejoinder". In: *Journal of the American Statistical Association* 75.371, pp. 593–595.

Beck, Nathaniel and Jonathan N Katz (1995). "What to Do (and Not to Do) with Time-Series Cross-Section Data". In: *The American Political Science Review* 89.3, pp. 634–647.

Berger, James O (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd. New York, NY: Springer-Verlag.

Berger, James O, Brunero Liseo, and Robert L Wolpert (1999). "Integrated Likelihood Methods for Eliminating Nuisance Parameters". In: *Statistical Science* 14.1, pp. 1–22.

Berger, Roger L and Dennis D Boos (1994). "P Values Maximized Over a Confidence Set for the Nuisance Parameter". In: *Journal of the American Statistical Association* 89.427, pp. 1012–1016.

Berk, Richard A and David A Freedman (2003). "Statistical Assumptions as Empirical Commitments". In: *Punishment and Social Control: Essays in Honor of Sheldon L. Messinger*. Ed. by Thomas G Blomberg and Stanley Cohen. 2nd. New York, NY: Aldine de Gruyter. Chap. 10, pp. 235–254.

Bernardo, José M (1979). "Reference Posterior Distributions for Bayesian Inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2, pp. 113–128.

Berry, Scott M and Joseph B Kadane (1997). "Optimal Bayesian Randomization". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4, pp. 813–819.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" In: *The Quarterly Journal of Economics* 119.1, pp. 249–275.

Bertsimas, Dimitris, Mac Johnson, and Nathan Kallus (2015). "The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples". In: *Operations Research* 63.4, pp. 868–876.

Bhuller, Manudeep et al. (2013). "Broadband Internet: An Information Superhighway to Sex Crime?" In: *The Review of Economic Studies* 80.4, pp. 1237–1266.

Bind, Marie-Abele C and Donald B Rubin (2019). "Bridging Observational Studies and Randomized Experiments by Embedding the Former in the Latter". In: *Statistical Methods in Medical Research* 28.7, pp. 1958–1978.

Blackwell, Matthew and Adam N Glynn (2018). "How to Make Causal Inferences with Time-Series Cross- Sectional Data under Selection on Observables". In: *The American Political Science Review* 112.4, pp. 1067–1082.

Bloom, Howard S, James A Riccio, and Nandita Verma (2005). *Promoting Work in Public Housing: The Effectiveness of Jobs-Plus*. Report. New York, NY: Manpower Demonstration Research Corporation.

Bohn, Sarah, Magnus Lofstrom, and Steven Raphael (2014). "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" In: *The Review of Economics and Statistics* 96.2, pp. 258–269.

Bowers, Jake and Thomas Leavitt (2020). "Causality and Design-Based Inference". In: *The SAGE Handbook of Research Methods in Political Science and International Relations*. Ed. by Luigi

Curini and Robert Franzese. Vol. 2. Thousand Oaks, CA: SAGE Publications. Chap. 41, pp. 769–804.

Brewer, K.R.W. (1979). "A Class of Robust Sampling Designs for Large-Scale Surveys". In: *Journal of the American Statistical Association* 74.368, pp. 911–915.

Buchmueller, Thomas C, John DiNardo, and Robert G Valletta (2011). "The Effect of an Employer Health Insurance Mandate on Health Insurance Coverage and the Demand for Labor: Evidence from Hawaii". In: *American Economic Journal: Economic Policy* 3.4, pp. 25–51.

Butler, Daniel M (2014). *Representing the Advantaged: How Politicians Reinforce Inequality*. New York, NY: Cambridge University Press.

Butler, Daniel M and David E Broockman (2011). "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators". In: *American Journal of Political Science* 55.3, pp. 463–477.

Butler, Daniel M and Charles Crabtree (2017). "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions". In: *Journal of Experimental Political Science* 4.1, pp. 57–67.

Callaway, Brantly and Pedro H C Sant'Anna (2018). "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment". Preprint, `https://arxiv.org/pdf/1803.09015.pdf`.

Card, David (1990). "The Impact of the Mariel Boatlift on the Miami Labor Market". In: *Industrial and Labor Relations Review* 43.2, pp. 245–257.

Card, David and Alan B Krueger (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". In: *The American Economic Review* 84.4, pp. 772–793.

Cartwright, Nancy and Jeremy Hardie (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York, NY: Oxford University Press.

Caughey, Devin et al. (2020). "Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects". Working Paper.

Chaisemartin, Clément de and Xavier D'Haultfœuille (2021). "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110.9, pp. 2964–2996.

Chiba, Yasutaka (2018). "Bayesian Inference of Causal Effects for an Ordinal Outcome in Randomized Trials". In: *Journal of Causal Inference* 6.2.

Cochran, William G (1965). “The Planning of Observational Studies of Human Populations”. In: *Journal of the Royal Statistical Society. Series A (General)* 128.2, pp. 234–266.

— (1977). *Sampling Techniques*. 3rd. Hoboken, NJ: John Wiley & Sons.

Cochran, William G and Donald B Rubin (1973). “Controlling Bias in Observational Studies: A Review”. In: *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* 35.4, pp. 417–446.

Cochran, William Gemmell (1983). *Planning and Analysis of Observational Studies*. Ed. by Lincoln E Moses and Frederick Mosteller. New York, NY: John Wiley & Sons, Inc.

Cohen, Peter L and Colin B Fogarty (2020). “Gaussian Prepivoting for Finite Population Causal Inference”. Working paper, `https://arxiv.org/pdf/2002.06654.pdf`.

Cole, Stephen R and Elizabeth A Stuart (2010). “Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial”. In: *American Journal of Epidemiology* 172.1, pp. 107–115.

Coleman, Thomas S (2019). “Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference”. Working Paper, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262234`.

Conley, Timothy G and Christopher R Taber (2011). “Inference with ‘Difference in Differences’ with a Small Number of Policy Changes”. In: *The Review of Economics and Statistics* 93.1, pp. 113–125.

Conrad, Keith (2005). “Probability Distributions and Maximum Entropy”. Unpublished Manuscript, `https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf`.

Copas, J B (1973). “Randomization Models for the Matched and Unmatched 2 x 2 Tables”. In: *Biometrika* 60.3, pp. 467–476.

Coppock, Alexander, Thomas J Leeper, and Kevin J Mullinix (2018). “Generalizability of Heterogeneous Treatment Effect Estimates across Samples”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.49, pp. 12441–12446.

Cornfield, Jerome et al. (1959). “Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions”. In: *JNCI: Journal of the National Cancer Institute* 22.1, pp. 173–203.

Costa, Mia (2017). “How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials”. In: *Journal of Experimental Political Science* 4.3, pp. 241–254.

Cover, Thomas M and Joy A Thomas (1991). *Elements of Information Theory*. New York, NY: John Wiley & Sons, Inc.

Cox, David Roxbee (1958). *Planning of Experiments*. New York, NY: Wiley.

Dahabreh, Issa J, Sarah E Robertson, and Miguel A Hernán (2019). "On the Relation Between G-formula and Inverse Probability Weighting Estimators for Generalizing Trial Results". In: *Epidemiology* 30.6.

Daw, Jamie R and Laura A Hatfield (2018). "Matching in Difference-in-Differences: Between a Rock and a Hard Place". In: *Health Services Research* 53.6, pp. 4111–4117.

Dawid, A Philip (1979). "Conditional Independence in Statistical Theory". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 41.1, pp. 1–31.

De Boef, Suzanna and Luke Keele (2008). "Taking Time Seriously". In: *American Journal of Political Science* 52.1, pp. 184–200.

Deaton, Angus (2009). "Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development". In: *Proceedings of the British Academy, Volume 162, 2008 Lectures*. Ed. by Ron Johnston. Vol. 162. Oxford, UK: Oxford University Press, pp. 123–160.

— (2010). "Instruments, Randomization, and Learning about Development". In: *Journal of Economic Literature* 48.2, pp. 424–455.

Deaton, Angus and Nancy Cartwright (2018). "Understanding and Misunderstanding Randomized Controlled Trials". In: *Social Science & Medicine* 210, pp. 2–21.

Delevoye, Angèle and Fredrik Sävje (2020). "Consistency of the Horvitz-Thompson Estimator under General Sampling and Experimental Designs". In: *Journal of Statistical Planning and Inference* 207, pp. 190–197.

Ding, Peng (2017). "A Paradox from Randomization-Based Causal Inference". In: *Statistical Science* 32.3, pp. 331–345.

Ding, Peng and Luke Miratrix (2019). "Model-Free Causal Inference of Binary Experimental Data". In: *Scandinavian Journal of Statistics* 46.1, pp. 200–214.

Dobkin, Carlos et al. (2018). "The Economic Consequences of Hospital Admissions". In: *The American Economic Review* 108.2, pp. 308–352.

Donald, Stephen G and Kevin Lang (2007). "Inference with Difference-in-Differences and Other Panel Data". In: *The Review of Economics and Statistics* 89.2, pp. 221–233.

Doudchenko, Nikolay and Guido W Imbens (2017). "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis". Preprint, `https://arxiv.org/pdf/1610.07748.pdf`.

Drèze, Jacques H (1972). "Econometrics and Decision Theory". In: *Econometrica* 40.1, pp. 1–18.

— (1975). "Bayesian Theory of Identification in Simultaneous Equations Models". In: *Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage*. Ed. by Stephen E Fienberg and Arnold Zellner. Amsterdam, NL: North-Holland. Chap. 5.1, pp. 159–174.

— (1976). "Bayesian Limited Information Analysis of the Simultaneous Equations Model". In: *Econometrica* 44.5, pp. 1045–1075.

Driscoll, John C and Aart C Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *The Review of Economics and Statistics* 80.4, pp. 549–560.

Dube, Arindrajit and Ben Zipperer (2015). *Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies*. IZA Discussion Paper 8944. Bonn, Germany: IZA — Institute of Labor Economics.

Dunning, Thad et al., eds. (2019). *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York, NY: Cambridge University Press.

Eckles, Dean et al. (2020). "Noise-Induced Randomization in Regression Discontinuity Designs". Working Paper, `https://arxiv.org/abs/2004.09458`.

Efron, Bradley and Robert J Tibshirani (1994). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Egami, Naoki and Erin Hartman (2021a). "Covariate Selection for Generalizing Experimental Results: Application to Large-Scale Development Program in Uganda". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

— (2021b). "Elements of External Validity: Framework, Design, and Analysis". Working Paper, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3775158`.

Egami, Naoki and Soichiro Yamauchi (2019). "How to Improve the Difference-in-Differences Design with Multiple Pre-treatment Periods". Preprint, `https://naokiegami.com/paper/double_did.pdf`.

Eicker, Friedhelm (1963). "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions". In: *Annals of Mathematical Statistics* 34.2, pp. 447–456.

— (1967). "Limit Theorems for Regressions with Unequal and Dependent Errors". In: *Berkeley Symposium on Mathematical Statistics and Probability* 5.1, pp. 59–82.

Fedorov, V V (1972). *Theory of Optimal Experiments*. New York, NY: Academic Press.

Ferman, Bruno and Cristine Pinto (2019). "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity". In: *The Review of Economics and Statistics* 101.3, pp. 452–467.

Fisher, Ronald Aylmer (1935). *The Design of Experiments*. Edinburgh, SCT: Oliver and Boyd.

Fogarty, Colin B (2018). "On Mitigating the Analytical Limitations of Finely Stratified Experiments". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.5, pp. 1035–1056.

— (2020). "Testing Weak Nulls in Matched Observational Studies". Working Paper.

Freedman, David A (2008). "On Regression Adjustments in Experiments with Several Treatments". In: *The Annals of Applied Statistics* 2.1, pp. 176–196.

— (2009). *Statistical Models: Theory and Practice*. New York, NY: Cambridge University Press.

Gaddis, S Michael (2019). "Understanding the 'How' and 'Why' Aspects of Racial-Ethnic Discrimination: A Multimethod Approach to Audit Studies". In: *Sociology of Race and Ethnicity* 5.4, pp. 443–455.

Gastwirth, Joseph L, Abba M Krieger, and Paul R Rosenbaum (2000). "Asymptotic separability in sensitivity analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 545–555.

Gelman, Andrew and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gerber, Alan S and Donald P Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.

Gerber, Alan S, Donald P Green, and Edward H Kaplan (2004). "The Illusion of Learning from Observational Research". In: *Problems and Methods in the Study of Politics*. Ed. by Ian Shapiro, Rogers M Smith, and Tarek E Massoud. New York, NY: Cambridge University Press. Chap. 12, pp. 251–273.

Gibson, James L, Gregory A Caldeira, and Lester Kenyatta Spence (2002). "The Role of Theory in Experimental Design: Experiments Without Randomization". In: *Political Analysis* 10.4, pp. 362–375.

Gobillon, Laurent and Thierry Magnac (2016). "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls". In: *The Review of Economics and Statistics* 9.3, pp. 535–551.

Goodman-Bacon, Andrew (2018). "Public Insurance and Mortality: Evidence from Medicaid Implementation". In: *Journal of Political Economy* 126.1, pp. 216–262.

Goodman-Bacon, Andrew (2021). "Difference-in-Differences with Variation in Treatment Timing". In: *Journal of Econometrics*.

Green, Donald P and Holger L Kern (2012). "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees". In: *Public Opinion Quarterly* 76.3, pp. 491–511.

Grimmer, Justin, Solomon Messing, and Sean J Westwood (2017). "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods". In: *Political Analysis* 25.4, pp. 413–434.

Gustafson, Paul (2005). "On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables". In: *Statistical Science* 20.2, pp. 111–140.

— (2009). "What Are the Limits of Posterior Distributions Arising From Nonidentified Models, and Why Should We Care?" In: *Journal of the American Statistical Association* 104.488, pp. 1682–1695.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design". In: *Econometrica* 69.1, pp. 201–209.

Hainmueller, Jens (2012). "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." In: *Political Analysis* 20.1, pp. 25–46.

Hainmueller, Jens and Dominik Hangartner (2019). "Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland". In: *American Journal of Political Science* 63.3, pp. 530–547.

Hájek, Jaroslav (1960). "Limiting Distributions in Simple Random Sampling from a Finite Population". In: *Publications of the Mathematics Institute of the Hungarian Academy of Science* 5, pp. 361–374.

Hansen, Ben B (2004). "Full Matching in an Observational Study of Coaching for the SAT". In: *Journal of the American Statistical Association* 99.467, pp. 609–618.

Hansen, Ben B and Jake Bowers (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies". In: *Statistical Science* 23.2, pp. 219–236.

Hansen, Bruce E (2021). *Econometrics*. Princeton, NJ: Princeton University Press.

Harrison, Glenn W (2011). "Randomisation and Its Discontents". In: *Journal of African Economies* 20.4, pp. 626–652.

Harrison, Glenn W (2014). "Cautionary Notes on the Use of Field Experiments to Address Policy Issues". In: *Oxford Review of Economic Policy* 30.4, pp. 753–763.

Harte, John (2011). *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. New York, NY: Oxford University Press.

Harville, David A (1975). "Experimental Randomization: Who Needs It?" In: *The American Statistician* 29.1, pp. 27–31.

Hasegawa, Raiden B, Daniel W Webster, and Dylan S Small (2019). "Evaluating Missouri's Handgun Purchaser Law: A Bracketing Method for Addressing Concerns About History Interacting with Group". In: *Epidemiology* 30.3, pp. 371–379.

Hausman, Catherine and David S Rapson (2018). "Regression Discontinuity in Time: Considerations for Empirical Applications". In: *Annual Review of Resource Economics* 10.1, pp. 533–552.

Heckman, James J (1992). "Randomization and Social Policy Evaluation". In: *Evaluating Welfare and Training Programs*. Ed. by Charles F Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press. Chap. 5, pp. 201–230.

— (2020). "Epilogue: Randomization and Social Policy Evaluation Revisited". In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Ed. by Florent Bédécarrats, Isabelle Guérin, and François Roubaud. New York, NY: Oxford University Press. Chap. 12, pp. 304–330.

Higgins, Michael J, Fredrik Sävje, and Jasjeet Singh Sekhon (2016). "Improving massive experiments with threshold blocking". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7369–7376.

Hill, Jennifer L (2011). "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Ho, Daniel E et al. (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". In: *Political Analysis* 15.3, pp. 199–236.

Höglund, Thomas (1978). "Sampling from a Finite Population. A Remainder Term Estimate". In: *Scandinavian Journal of Statistics* 5.1, pp. 69–71.

Holland, Paul W (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Horvitz, Daniel G and Donovan J Thompson (1952). "A Generalization of Sampling without Replacement from a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685.

Howard, Ronald A (1966). "Information Value Theory". In: *IEEE Transactions on Systems Science and Cybernetics* 2.1, pp. 22–26.

Howson, Colin and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. 3rd. Chicago, IL: Open Court Publishing.

Hsiao, Cheng (1983). "Identification". In: *Handbook of Econometrics*. Ed. by Zvi Griliches and Michael D Intriligator. Vol. 1. Amsterdam, NL: North-Holland. Chap. 4, pp. 223–283.

Hudson, Sally, Peter Hull, and Jack Liebersohn (2017). "Interpreting Instrumented Difference-in-Differences". Econometrics Note, `http://www.mit.edu/~liebers/DDIV.pdf`.

Humphreys, Macartan and Alan M Jacobs (2015). "Mixing Methods: A Bayesian Approach". In: *The American Political Science Review* 109.04, pp. 653–673.

Hurwicz, Leonid (1950). "Generalization of the Concept of Identification". In: *Statistical Inference in Dynamic Economic Models*. Ed. by Tjalling C Koopmans. Cowles Foundation for Research in Economics 10. New York, NY: John Wiley & Sons. Chap. 4, pp. 245–257.

Imai, Kosuke (2008). "Variance Identification and Efficiency Analysis in Randomized Experiments under the Matched-Pair Design". In: *Statistics in Medicine* 27.24, pp. 4857–4873.

Imai, Kosuke and In Song Kim (2019). "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" In: *American Journal of Political Science* 63.2, pp. 467–490.

— (2021). "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data". In: *Political Analysis* 29.3, pp. 405–415.

Imai, Kosuke and Marc Ratkovic (2013). "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation". In: *Annals of Applied Statistics* 7.1, pp. 443–470.

Imbens, Guido W (2021). "Statistical Significance, p-Values, and the Reporting of Uncertainty". In: *Journal of Economic Perspectives* 35.3, pp. 157–174.

Imbens, Guido W and Donald B Rubin (1997). "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance". In: *The Annals of Statistics* 25.1, pp. 305–327.

— (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Imbens, Guido W and Jeffrey M Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation". In: *Journal of Economic Literature* 47.1, pp. 5–86.

Janusz, Andrew and Nazita Lajevardi (2016). "The Political Marginalization of Latinos: Evidence from Three Field Experiments". Unpublished Manuscript, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2799043.

Jaynes, Edwin T (2003). *Probability Theory: The Logic of Science*. Ed. by G Larry Bretthorst. New York, NY: Cambridge University Press.

Jeffreys, Harold (1939). *Theory of Probability*. Oxford, UK: Oxford University Press.

— (1946). "An Invariant Form for the Prior Probability in Estimation Problems". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186.1007, pp. 453–461.

Josey, Kevin P et al. (2021). "Transporting Experimental Results with Entropy Balancing". To appear in Statistics in Medicine.

Kadane, Joseph B (1975). "The Role of Identification in Bayesian Theory". In: *Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage*. Ed. by Stephen E Fienberg and Arnold Zellner. Amsterdam, NL: North-Holland. Chap. 5.2, pp. 175–191.

Kadane, Joseph B and Teddy Seidenfeld (1990). "Randomization in a Bayesian Perspective". In: *Journal of Statistical Planning and Inference* 25.3, pp. 329–345.

Kallus, Nathan (2018). "Optimal A Priori Balance in the Design of Controlled Experiments". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 85–112.

Kang, Joseph DY and Joseph L Schafer (2007). "Demystifying Double Robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical Science* 22.4, pp. 523–539.

Kasy, Maximilian (2016). "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead". In: *Political Analysis* 24.3, pp. 324–338.

Keele, Luke J (2020). "Differences-in-Differences: Neither Natural nor an Experiment". In: *The SAGE Handbook of Research Methods in Political Science and International Relations*. Ed. by Luigi Curini and Robert Franzese. Vol. 1. Thousand Oaks, CA: SAGE Publications. Chap. 43, pp. 822–834.

Keele, Luke J, Raiden Hasegawa, and Dylan Small (2019). "Bracketing Bounds for Differences-in-Differences with an Application to Voter ID Laws". Working paper, https://polmeth.mit.edu/sites/default/files/documents/Keele_Paper.pdf.

Keele, Luke and Kevin M Quinn (2017). "Bayesian Sensitivity Analysis for Causal Effects from 2×2 Tables in the Presence of Unmeasured Confounding with Application to Presidential Campaign Visits". In: *The Annals of Applied Statistics* 11.4, pp. 1974–1997.

Kern, Holger L et al. (2016). "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations". In: *Journal of Research on Educational Effectiveness* 9.1, pp. 103–127.

Kiefer, Jack (1959). "Optimum Experimental Designs". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 21.2, pp. 272–319.

King, Gary and Langche Zeng (2006). "The Dangers of Extreme Counterfactuals". In: *Political Analysis* 14.2, pp. 131–159.

— (2007). "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference". In: *International Studies Quarterly* 51.1, pp. 183–210.

Kish, Leslie (1965). *Survey Sampling*. New York, NY: Wiley & Sons.

Koopmans, Tjalling C (1949). "Identification Problems in Economic Model Construction". In: *Econometrica* 17.2, pp. 125–144.

Koopmans, Tjalling C and Olav Reiersøl (1950). "The Identification of Structural Characteristics". In: *The Annals of Mathematical Statistics* 21.2, pp. 165–181.

Koopmans, Tjalling C, Herman Rubin, and Roy B Leipnik (1950). "Measuring the Equation Systems of Dynamic Economics". In: *Statistical Inference in Dynamic Economic Models*. Ed. by Tjalling C Koopmans. Cowles Foundation for Research in Economics 10. New York, NY: John Wiley & Sons. Chap. 2, pp. 53–237.

Kropko, Jonathan and Robert Kubinec (2018). "Why the Two-Way Fixed Effects Model Is Difficult to Interpret, and What to Do About It". Preprint, `https://ssrn.com/abstract=3062619`.

Lago, Ignacio and José Ramón Montero (2005). "Los mecanismos del cambio electoral". In: *Claves de la Razón Práctica* 149, pp. 36–45.

Leamer, Edward E (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York, NY: Wiley.

Lechner, Michael (2011). "The Estimation of Causal Effects by Difference-in-Difference Methods". In: *Foundations and Trends in Econometrics* 4.3, pp. 165–224.

Lee, Myoung-jae (2016). "Generalized Difference in Differences With Panel Data and Least Squares Estimator". In: *Sociological Methods & Research* 45.1, pp. 134–157.

Lehmann, Erich Leo (1999). *Elements of Large-Sample Theory*. Springer Texts in Statistics. New York, NY: Springer.

Letham, Benjamin et al. (2019). "Constrained Bayesian Optimization with Noisy Experiments". In: *Bayesian Analysis* 14.2, pp. 495–519.

Li, Xinran and Peng Ding (2017). "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference". In: *Journal of the American Statistical Association* 112.520, pp. 1759–1769.

Lin, Winston (2013). "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique". In: *The Annals of Applied Statistics* 7.1, pp. 295–318.

Lindley, Dennis V (1971). *Making Decisions*. 1st. Hoboken, NJ: John Wiley & Sons.

— (1982). "The Role of Randomization in Inference". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 2: Symposia and Invited Papers. Philosophy of Science Association. Chicago, IL: University of Chicago Press, pp. 431–446.

Lindley, Dennis V and Melvin R Novick (1981). "The Role of Exchangeability in Inference". In: *The Annals of Statistics* 9.1, pp. 45–58.

Liseo, Brunero (2005). "The Elimination of Nuisance Parameters". In: *Bayesian Thinking: Modeling and Computation*. Ed. by Dipak Kumar Dey and Calyampudi Radhakrishna Rao. Vol. 25. Handbook of Statistics. Amsterdam, NL: Elsevier. Chap. 7, pp. 193–219.

Little, Andrew T and Thomas B Pepinsky (2021). "Learning from Biased Research Designs". In: *The Journal of Politics* 83.2, pp. 602–616.

Lohr, Sharon L (2010). *Sampling: Design and Analysis*. Second. Boston, MA: Brooks/Cole.

Lu, Benjamin et al. (2021). "Is it Who You Are or Where You Are? Accounting for Compositional Differences in Cross-site Treatment Variation". Working Paper, `https://arxiv.org/pdf/2103.14765.pdf`.

MacKinnon, James G and Halbert White (1985). "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties". In: *Journal of Econometrics* 29.3, pp. 305–325.

Malani, Anup and Julian Reif (2015). "Interpreting Pre-trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform". In: *Journal of Public Economics* 124, pp. 1–17.

Manski, Charles F and John V Pepper (2018). "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions". In: *The Review of Economics and Statistics* 100.2, pp. 232–244.

McElreath, Richard (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd. Boca Raton, FL: Chapman & Hall/CRC.

Mendez, Matthew S (2018). “Towards an Ethical Representation of Undocumented Latinos”. In: *PS: Political Science & Politics* 51.2, pp. 335–339.

Mendez, Matthew S and Christian R Grose (2018). “Doubling Down: Inequality in Responsiveness and the Policy Preferences of Elected Officials”. In: *Legislative Studies Quarterly* 43.3, pp. 457–491.

Middleton, Joel A and Peter M Aronow (2015). “Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments”. In: *Statistics, Politics and Policy* 6.1-2, pp. 39–75.

Miratrix, Luke W (2019). “Simulating for Uncertainty with Interrupted Time Series Designs”. Working Paper.

Miratrix, Luke W, Jasjeet S Sekhon, and Bin Yu (2013). “Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.2, pp. 369–396.

Miratrix, Luke W et al. (2018). “Worth Weighting? How to Think About and Use Weights in Survey Experiments”. In: *Political Analysis* 26.3, pp. 275–291.

Montalvo, José G (2011). “Voting after the Bombings: A Natural Experiment on the Effect of Terrorist Attacks on Democratic Elections”. In: *The Review of Economics and Statistics* 93.4, pp. 1146–1154.

Mora, Ricardo and Iliana Reggio (2012). *Treatment Effect Identification Using Alternative Parallel Assumptions*. Working Paper, Economic Series (48) 12–33. Getafe, Spain: Universidad Carlos III.

— (2019). “Alternative Diff-in-Diffs Estimators with Several Pretreatment Periods”. In: *Econometric Reviews* 38.5, pp. 465–486.

Mullinix, Kevin J et al. (2015). “The Generalizability of Survey Experiments”. In: *Journal of Experimental Political Science* 2.2, pp. 109–138.

Neath, Andrew A and Francisco J Samaniego (1997). “On the Efficacy of Bayesian Inference for Nonidentifiable Models”. In: *The American Statistician* 51.3, pp. 225–232.

Neyman, Jersey (1923). “Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes”. In: *Roczniki Nauk Rolniczych* 10, pp. 1–51.

Neyman, Jerzy and Egon Sharpe Pearson (1933). “On the Problem of the Most Efficient Tests of Statistical Hypotheses”. In: *Philosophical Transactions of the Royal Society of London* 231.694–706, pp. 289–337.

Nguyen, Trang Quynh et al. (2017). "Sensitivity Analysis for an Unobserved Moderator in RCT-to-Target-Population Generalization of Treatment Effects". In: *The Annals of Applied Statistics* 11.1, pp. 225–247.

O'Muircheartaigh, Colm and Larry V Hedges (2014). "Generalizing from unrepresentative experiments: A stratified propensity score approach". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.2, pp. 195–210.

Obenauer, Marie Louise and Bertha Marie von der Nienburg (1915). *Effect of Minimum-Wage Determinations in Oregon: July, 1915*. Washington, D.C.: U.S. Government Printing Office.

Olsen, Robert B et al. (2013). "External Validity in Policy Evaluations That Choose Sites Purposively". In: *Journal of Policy Analysis and Management* 32.1, pp. 107–121.

Pashley, Nicole E and Luke W Miratrix (2021). "Insights on Variance Estimation for Blocked and Matched Pairs Designs". In: *Journal of Educational and Behavioral Statistics* 46.3, pp. 271–296.

Pearl, Judea and Elias Bareinboim (2014). "External Validity: From Do-Calculus to Transportability Across Populations". In: *Statistical Science* 29.4, pp. 579–595.

Poirier, Dale J (1998). "Revising Beliefs in Nonidentified Models". In: *Econometric Theory* 14.4, pp. 483–509.

Pritchett, Lant and Justin Sandefur (2015). "Learning from Experiments When Context Matters". In: *American Economic Review* 105.5, pp. 471–475.

Raiffa, Howard and Robert Schlaifer (1961). *Applied Statistical Decision Theory*. Boston, MA: Division of Research, Graduate School of Business Adminitration, Harvard University.

Raj, Des (1965). "On a Method of Using Multi-Auxiliary Information in Sample Surveys". In: *Journal of the American Statistical Association* 60.309, pp. 270–277.

Rambachan, Ashesh and Jonathan Roth (2020). "Design-Based Uncertainty for Quasi-Experiments". Working paper, `https://scholar.harvard.edu/jroth/publications/design-based-uncertainty-quasi-experiments`.

Ramsey, Frank Plumpton (1929). "Knowledge". In: *F.P. Ramsey: Philosophical Papers*. Ed. by David Hugh Mellor. New York, NY: Cambridge University Press. Chap. 5, pp. 110–111.

— (1931). "Truth and Probability". In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. London, UK: Kegan, Paul, Trench, Trubner & Co. Chap. VII, pp. 156–198.

Ravallion, Martin (2009). "Should the Randomistas Rule?" In: *The Economists' Voice* 6.2, pp. 1–5.

Ravallion, Martin (2020). "Should the Randomistas (Continue to) Rule?" In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Ed. by Florent Bédécarrats, Isabelle Guérin, and François Roubaud. New York, NY: Oxford University Press. Chap. 1, pp. 47–78.

Reid, Nancy (1995). "The Roles of Conditioning in Inference". In: *Statistical Science* 10.2, pp. 138–157.

Richard, Jean-François (1973). *Posterior and Predictive Densities for Simultaneous Equation Models*. New York, NY: Springer-Verlag.

Robbins, Herbert (1956). "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by Jerzy Neyman. Vol. 1. Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press, pp. 157–163.

— (1964). "The Empirical Bayes Approach to Statistical Decision Problems". In: *Annals of Mathematical Statistics* 35.1, pp. 1–20.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Rokicki, Slawa et al. (2018). "Inference with Difference-in-Differences with a Small Number of Groups: A Review, Simulation Study and Empirical Application Using SHARE Data". In: *Medical Care* 56.1, pp. 97–105.

Rosenbaum, Paul R (1984). "Conditional Permutation Tests and the Propensity Score in Observational Studies". In: *Journal of the American Statistical Association* 79.387, pp. 565–574.

— (1989a). "Optimal Matching for Observational Studies". In: *Journal of the American Statistical Association* 84.408, pp. 1024–1032.

— (1989b). "The Role of Known Effects in Observational Studies". In: *Biometrics* 45.2, pp. 557–569.

— (1991). "A Characterization of Optimal Designs for Observational Studies". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, pp. 597–610.

— (2001). "Stability in the Absence of Treatment". In: *Journal of the American Statistical Association* 96.453, pp. 210–219.

— (2002). *Observational Studies*. 2nd. New York, NY: Springer.

— (2007). "Confidence intervals for uncommon but dramatic responses to treatment". In: *Biometrics* 63.4, pp. 1164–1171.

Rosenbaum, Paul R (2010). *Design of Observational Studies*. New York, NY: Springer.

— (2017). *Observation and Experiment: An Introduction to Causal Inference*. Cambridge, MA: Harvard University Press.

Rosenbaum, Paul R and Abba M Krieger (1990). "Sensitivity of Two-Sample Permutation Inferences in Observational Studies". In: *Journal of the American Statistical Association* 85.410, pp. 493–498.

Rosenbaum, Paul R and Donald B Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1, pp. 41–55.

— (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score". In: *Journal of the American Statistical Association* 79.387, pp. 516–524.

Rothenberg, Thomas J (1971). "Identification in Parametric Models". In: *Econometrica* 39.3, pp. 577–591.

Rubin, Donald B (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66.5, p. 688.

— (1976). "Bayesian Inference for Causality: The Importance of Randomization". In: *American Statistical Association: 1975 Proceedings of the Social Statistics Section*. Ed. by Edwin D Goldfield. Washington, D. C.: American Statistical Association, pp. 233–239.

— (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". In: *The Annals of Statistics* 6.1, pp. 34–58.

— (1980). "Comment on 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by Basu, D." In: *Journal of the American Statistical Association* 75.371, pp. 591–593.

— (1986). "Which Ifs Have Causal Answers? (Comment on 'Statistics and Causal Inference' by Paul W. Holland)." In: *Journal of the American Statistical Association* 81.396, pp. 961–962.

— (2007). "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials". In: *Statistics in Medicine* 26.1, pp. 20–36.

— (2008). "For Objective Causal Inference, Design Trumps Analysis". In: *The Annals of Applied Statistics* 2.3, pp. 808–840.

Rudolph, Kara E and Mark J van der Laan (2017). "Robust Estimation of Encouragement Design Intervention Effects transported across Sites". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 79.5, pp. 1509–1525.

Rudolph, Kara E et al. (2014). "Estimating Population Treatment Effects From a Survey Subsample". In: *American Journal of Epidemiology* 180.7, pp. 737–748.

Sales, Adam C, Ben B Hansen, and Brian Rowan (2018). "Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates". In: *Journal of Educational and Behavioral Statistics* 43.1, pp. 3–31.

Savage, Leonard J (1954). *The Foundations of Statistics*. Hoboken, NJ: John Wiley & Sons.

— (1962a). "On the Foundations of Statistical Inference: Discussion". In: *Journal of the American Statistical Association* 57.298, pp. 307–308.

— (1962b). "Subjective Probability and Statistical Practice". In: *The Foundations of Statistical Inference: A Discussion*. Ed. by Maurice Stevenson Bartlett. New York, NY: John Wiley & Sons, pp. 9–35.

Sävje, Fredrik, Peter M Aronow, and Michael G Hudgens (2021). "Average Treatment Effects in the Presence of Unknown Interference". In: *The Annals of Statistics* 49.2, pp. 673–701.

Sävje, Fredrik, Michael J Higgins, and Jasjeet S Sekhon (2021). "Generalized Full Matching". In: *Political Analysis*.

Sekhon, Jasjeet S and Yotam Shem-Tov (2021). "Inference on a New Class of Sample Average Treatment Effects". In: *Journal of the American Statistical Association* 116.534, pp. 798–804.

Slutsky, Eugen (1925). "Über stochastische Asymptoten und Grenzwerte". In: *Metron* 5.3, pp. 3–89.

Snow, John (1854). "The Cholera near Golden-square, and at Deptford". In: *Medical Times and Gazette* 9, pp. 321–322.

Sobel, Michael E (2012). "Does Marriage Boost Men's Wages?: Identification of Treatment Effects in Fixed Effects Regression Models for Panel Data". In: *Journal of the American Statistical Association* 107.498, pp. 521–529.

Sofer, Tamar et al. (2016). "On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences". In: *Statistical Science* 31.3, pp. 348–361.

Stephenson, W Robert (1981). "A General Class of One-Sample Nonparametric Test Statistics Based on Subsamples". In: *Journal of the American Statistical Association* 76.376, pp. 960–966.

Stephenson, W Robert and Malay Ghosh (1985). "Two Sample Nonparametric Tests based on Subsamples". In: *Communications in Statistics – Theory and Methods* 14.7, pp. 1669–1684.

Stone, M (1969). "The Role of Experimental Randomization in Bayesian Statistics: Finite Sampling and Two Bayesians". In: *Biometrika* 56.3, pp. 681–683.

Strezhnev, Anton (2018). "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs". Working Paper.

Stuart, Elizabeth A et al. (2011). "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 174.2, pp. 369–386.

Suppes, Patrick (1982). "Arguments for Randomizing". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982.2, pp. 464–475.

Tchetgen Tchetgen, Eric J (2014). "The Control Outcome Calibration Approach for Causal Inference With Unobserved Confounding". In: *American Journal of Epidemiology* 179.5, pp. 633–640.

Tipton, Elizabeth (2013). "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts". In: *Journal of Educational and Behavioral Statistics* 38.3, pp. 239–266.

Tipton, Elizabeth et al. (2014). "Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling". In: *Journal of Research on Educational Effectiveness* 7.1, pp. 114–135.

Torcal, Mariano and Guillem Rico (2004). "The 2004 Spanish General Election: In the Shadow of Al Quaeda". In: *South European Society and Politics* 9.3, pp. 107–121.

Urbach, Peter (1985). "Randomization and the Design of Experiments". In: *Philosophy of Science* 52.2, pp. 256–273.

— (1993). "The Value of Randomization and Control in Clinical Trials". In: *Statistics in Medicine* 12.15-16, pp. 1421–1431.

Vaart, Aad van der (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY: Cambridge University Press.

Vivalt, Eva (2020). "Using Priors in Experimental Design: How Much Are We Leaving on the Table?" In: *Randomized Controlled Trials in the Field of Development: A Critical Perspective*. Ed. by Florent Isabelle Guérin Bédécarrats and François Roubaud. Oxford, UK: Oxford University Press. Chap. 11, pp. 293–303.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Westreich, Daniel et al. (2017). "Transportability of Trial Results Using Inverse Odds of Sampling Weights". In: *American Journal of Epidemiology* 186.8, pp. 1010–1014.

Westreich, Daniel et al. (2019). "Target Validity and the Hierarchy of Study Designs". In: *American Journal of Epidemiology* 188.2, pp. 438–443.

White, Halbert (1980). "Using Least Squares to Approximate Unknown Regression Functions". In: *International Economic Review* 21.1, pp. 149–170.

Wong, Tom K, Michael Nicholson, and Nazita Lajevardi (2017). "Immigrants, Citizens, and (Un)Equal Representation". In: *The Politics of Immigration: Partisanship, Demographic Change, and American National Identity*. Ed. by Tom K Wong. New York, NY: Oxford University Press. Chap. 4, pp. 192–208.

Wooldridge, Jeffrey M (2003). "Cluster-Sample Methods in Applied Econometrics". In: *The American Economic Review* 93.2, pp. 133–138.

— (2005). "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models". In: *The Review of Economics and Statistics* 87.2, pp. 385–390.

— (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd. Cambridge, MA: The MIT Press.

Wu, Chien-Fu (1981). "On the Robustness and Efficiency of Some Randomized Designs". In: *The Annals of Statistics* 9.6, pp. 1168–1177.

Xu, Yiqing (2017). "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models". In: *Political Analysis* 25.1, pp. 57–76.

Yamauchi, Soichiro (2020). "Difference-in-Differences for Ordinal Outcomes: Application to the Effect of Mass Shootings on Attitudes towards Gun Control". Working paper, `https://soichiroy.github.io/files/papers/ordinal_did.pdf`.

Zellner, Arnold (1971). *An Introduction to Bayesian Inference in Econometrics*. New York, NY: Wiley.

Zhang, Junni L, Donald B Rubin, and Fabrizia Mealli (2009). "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification". In: *Journal of the American Statistical Association* 104.485, pp. 166–176.

Zubizarreta, José R (2015). "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data". In: *Journal of the American Statistical Association* 110.511, pp. 910–922.

Appendix A: Proofs

A.0.1 Proof of Lemma 1

Proof. First note that $(\hat{\bar{\tau}} - \bar{\tau})$ is equivalent to $\hat{\bar{\tau}}$ when potential outcomes are centered to have mean 0. Thus, without loss of generality (and without amending notation), assume that the potential outcomes are centered such that $\frac{1}{N}\sum_{i=1}^{N} y_{Ti} = 0$ and $\frac{1}{N}\sum_{i=1}^{N} y_{Ci} = 0$ and consider the variance of $\sqrt{N}\hat{\bar{\tau}}$.

From the rules of variance and the derivation of the Difference-in-Means estimator's variance in Neyman (1923), it follows that for any $N = 4, 5, \ldots$,

$$N \operatorname{Var}[\hat{\bar{\tau}}] = \frac{N}{N-1}\left(\frac{1 - n_T/N}{n_T/N}\sigma^2_{y_T} + \frac{n_T/N}{1 - n_T/N}\sigma^2_{y_C} + 2\sigma_{y_T, y_C}\right). \tag{A.1}$$

Since potential outcomes are centered, it follows that

$$\begin{aligned}
\sigma^2_{y_T} &= \frac{1}{N}\sum_{i=1}^{N} y^2_{Ti}, \\
\sigma^2_{y_C} &= \frac{1}{N}\sum_{i=1}^{N} y^2_{Ci} \text{ and} \\
\sigma_{y_T, y_C} &= \frac{1}{N}\sum_{i=1}^{N} y_{Ti}\, y_{Ci}.
\end{aligned}$$

Hence, we can write the variance in Equation (A.1) as

$$N \operatorname{Var}[\hat{\bar{\tau}}] = \frac{N}{N-1}\left(\frac{1 - n_T/N}{n_T/N}\left(\frac{1}{N}\sum_{i=1}^{N} y^2_{Ti}\right) + \frac{n_T/N}{1 - n_T/N}\left(\frac{1}{N}\sum_{i=1}^{N} y^2_{Ci}\right) + 2\left(\frac{1}{N}\sum_{i=1}^{N} y_{Ti}\, y_{Ci}\right)\right).$$

Then, from Conditions 2 and 3, it follows that

$$\lim_{N\to\infty} N \operatorname{Var}[\hat{\bar{\tau}}] = \frac{1-p}{p}\langle y^2_T\rangle + \frac{p}{1-p}\langle y^2_C\rangle + 2\langle y_C\, y_T\rangle = \nu \tag{A.2}$$

as in Theorem 1 of Freedman (2008) described in Equation (1.2) in the main text.

Now note that with centered potential outcomes we can write $N\widehat{\text{Var}}[\hat{\bar{\tau}}]$, where $\widehat{\text{Var}}[\hat{\bar{\tau}}]$ is Neyman's conservative variance estimator in Equation (1.9), as follows:

$$N\widehat{\text{Var}}[\hat{\bar{\tau}}] = \frac{N}{N-1}\left(\frac{1-n_T/N}{n_T/N}\hat{\mu}_{y_T^2} + \frac{n_T/N}{1-n_T/N}\hat{\mu}_{y_C^2} + \hat{\mu}_{y_T^2} + \hat{\mu}_{y_C^2}\right), \tag{A.3}$$

where

$$\hat{\mu}_{y_T^2} = \frac{1}{n_T}\sum_{i=1}^{N} Z_i y_{Ti}^2 \text{ and}$$
$$\hat{\mu}_{y_C^2} = \frac{1}{n_C}\sum_{i=1}^{N}(1-Z_i) y_{Ci}^2.$$

Since $\text{E}[\hat{\mu}_{y_T^2}] = \frac{1}{N}\sum_{i-1}^{N} y_{iT}^2$ and $\text{E}[\hat{\mu}_{y_C^2}] = \frac{1}{N}\sum_{i-1}^{N} y_{iC}^2$, it follows from conditions 2 and 3 that

$$\lim_{N\to\infty} \text{E}[N\widehat{\text{Var}}[\hat{\bar{\tau}}]] = \frac{1-p}{p}\langle y_T^2\rangle + \frac{p}{1-p}\langle y_C^2\rangle + \langle y_T^2\rangle + \langle y_C^2\rangle, \tag{A.4}$$

which, by the Cauchy-Schwarz inequality, is greater than or equal to ν in Equation (A.2).

Finally, to establish that $N\widehat{\text{Var}}[\hat{\bar{\tau}}]$ converges in probability to this constant in Equation (A.4) that is greater than or equal to ν, we need only to demonstrate that both $\text{Var}[\hat{\mu}_{y_T^2}]$ and $\text{Var}[\hat{\mu}_{y_C^2}]$ tend to 0 as $N \to \infty$. From elementary theory of survey sampling, it follows that, for any $N = 4, 5, \ldots$

$$\text{Var}[\hat{\mu}_{y_T^2}] = \frac{(N-n_T)}{N-1}\frac{\sigma^2_{y_T^2}}{n_T}$$
$$\text{Var}[\hat{\mu}_{y_C^2}] = \frac{n_T}{N-1}\frac{\sigma^2_{y_C^2}}{(N-n_T)},$$

where $\sigma^2_{y_T^2}$ and $\sigma^2_{y_C^2}$ can be equivalently written as

$$\sigma^2_{y_T^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{Ti}^2 - \frac{1}{N}\sum_{i=1}^{N} y_{Ti}^2\right)^2 = \frac{1}{N}\sum_{i=1}^{N} y_{Ti}^4 - \left(\frac{1}{N}\sum_{i=1}^{N} y_{Ti}^2\right)^2$$
$$\sigma^2_{y_C^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{Ci}^2 - \frac{1}{N}\sum_{i=1}^{N} y_{Ci}^2\right)^2 = \frac{1}{N}\sum_{i=1}^{N} y_{Ci}^4 - \left(\frac{1}{N}\sum_{i=1}^{N} y_{Ci}^2\right)^2.$$

By Condition 2, $\frac{(N-n_T)}{N-1} \to 1-p$ and $\frac{n_T}{N-1} \to p$ as $N \to \infty$. In addition, by Condition 3, $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} y_{Ti}^2\right)^2 = \langle y_T^2\rangle^2$ and $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} y_{Ci}^2\right)^2 = \langle y_C^2\rangle^2$, both of which are finite limits. Then, by Condition 1, $\frac{1}{N}\sum_{i=1}^{N} y_{Ti}^4 < L < \infty$ for each $N = 4, 5, \ldots$. Therefore, $\mathrm{Var}[\hat{\mu}_{y_T^2}] \to 0$ and $\mathrm{Var}[\hat{\mu}_{y_C^2}] \to 0$ as $N \to \infty$, which completes the proof.

□

A.0.2 Proof of Theorem 1

Proof. Let $\Theta_{\bar{\tau}_h}$ be the set of hypothetical mean causal effects in the support of the prior distribution. Also define $\Theta^*_{\bar{\tau}_h} \equiv \{\bar{\tau}_h : \bar{\tau} - \varepsilon < \bar{\tau}_h < \bar{\tau} + \varepsilon\}$, $\Theta^-_{\bar{\tau}_h} \equiv \{\bar{\tau}_h : \bar{\tau} - \bar{\tau}_h \geq \varepsilon\}$ and $\Theta^+_{\bar{\tau}_h} \equiv \{\bar{\tau}_h : \bar{\tau}_h - \bar{\tau} \geq \varepsilon\}$, where ε is an arbitrarily small constant in which $\varepsilon > 0$. The sets $\Theta^-_{\bar{\tau}}$ and $\Theta^+_{\bar{\tau}}$ contain the hypothetical average causal effects that are either too small or too large to be within a distance of ε from the true average causal effect, $\bar{\tau}$.

The test-statistic in Equation (1.10) is

$$\frac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau}_h)}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}} = \frac{Z}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}/\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}},$$

where $Z = \dfrac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau}_h)}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}$.

We can equivalently write this test-statistic as

$$\left(\frac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau})}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}} + \frac{\sqrt{N}(\bar{\tau} - \bar{\tau}_h)}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}\right)\frac{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}}$$
$$= \left(\frac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau})}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}\right)\frac{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}} + \left(\frac{\sqrt{N}(\bar{\tau} - \bar{\tau}_h)}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}\right)\frac{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}}.$$

Then, by the finite population CLT, Slutsky's theorem, the continuous mapping theorem and Lemma 1, the first term

$$\left(\frac{\sqrt{N}(\hat{\bar{\tau}} - \bar{\tau})}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}\right)\frac{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}}{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}} \xrightarrow{d} Z/\sqrt{c}.$$

Now analyzing the denominator of the second term, Lemma 1 and the continuous mapping theorem imply that $\dfrac{\sqrt{N\widehat{\mathrm{Var}}[\hat{\bar{\tau}}]}}{\sqrt{N\,\mathrm{Var}[\hat{\bar{\tau}}]}} \xrightarrow{p} \sqrt{c}$. When $\bar{\tau}_h \in \Theta^-_{\bar{\tau}_h}$ or $\bar{\tau}_h \in \Theta^+_{\bar{\tau}_h}$, the numerator of the second term, $\sqrt{N}(\bar{\tau} - \bar{\tau}_h)$, diverges to either $+\infty$ or $-\infty$. Therefore, the test-statistic diverges in probability, i.e., the probability that the absolute value of the test-statistic in Equation (1.11) is greater than m tends to 1 as $N \to \infty$, where m is any positive, real number. Given that the standard Normal distribution is strictly monotonically decreasing in distance from its mean, the probability density of the standard Normal likelihood tends to 0 as $N \to \infty$.

Therefore, it follows that

$$\lim_{N\to\infty}\left[\frac{\int_{\bar{\tau}_h\in\Theta^-_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}])p(\bar{\tau}_h)d\,\bar{\tau}_h + \int_{\bar{\tau}_h\in\Theta^+_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}])p(\bar{\tau}_h)d\,\bar{\tau}_h}{\int_{\bar{\tau}_h\in\Theta_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}])p(\bar{\tau}_h)d\,\bar{\tau}_h}\right]$$
$$= \frac{\int_{\bar{\tau}_h\in\Theta^-_{\bar{\tau}_h}} \overbrace{(\lim_{N\to\infty} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]))}^{=0} p(\bar{\tau}_h)d\,\bar{\tau}_h + \int_{\bar{\tau}_h\in\Theta^+_{\bar{\tau}_h}} \overbrace{(\lim_{N\to\infty} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]))}^{=0} p(\bar{\tau}_h)d\,\bar{\tau}_h}{\int_{\bar{\tau}_h\in\Theta} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}])p(\bar{\tau}_h)d\,\bar{\tau}_h} = 0,$$

and, by the Law of Total Probability,

$$
\begin{aligned}
&\lim_{N\to\infty} \int_{\bar{\tau}_h \in \Theta^*_{\bar{\tau}_h}} f(\bar{\tau}_h \mid \hat{\bar{\tau}}, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) d\,\bar{\tau}_h \\
&= 1 - \lim_{N\to\infty} \left[\frac{\int_{\bar{\tau}_h \in \Theta^-_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h + \int_{\bar{\tau}_h \in \Theta^+_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h}{\int_{\bar{\tau}_h \in \Theta_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h} \right] \\
&= \lim_{N\to\infty} \left[1 - \frac{\int_{\bar{\tau}_h \in \Theta^-_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h + \int_{\bar{\tau}_h \in \Theta^+_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h}{\int_{\bar{\tau}_h \in \Theta_{\bar{\tau}_h}} f(\hat{\bar{\tau}} \mid \bar{\tau}_h, \widehat{\mathrm{Var}}[\hat{\bar{\tau}}]) p(\bar{\tau}_h) d\,\bar{\tau}_h} \right] \\
&= \left(\lim_{N\to\infty} 1 \right) - 0 = 1,
\end{aligned}
$$

which proves the theorem. □

A.0.3 Proof of Theorem 2

Proof. First, define the following subsets

$$
\begin{aligned}
\Theta^*_{\tau_h} &\equiv \{\tau_h : \tau - \varphi < \tau_h < \tau + \varphi\} \\
\Theta^-_{\tau_h} &\equiv \{\tau_h : \tau - \tau_h \geq \varphi\} \\
\Theta^+_{\tau} &\equiv \{\tau_h : \tau - \tau_h \leq -\varphi\},
\end{aligned}
$$

all of which are subsets of Θ_{τ_h}, the set of hypothetical constant effects in the support of the prior distribution, $p(\tau_h)$.

The finite population weak law of large numbers (Lin, 2013, Lemma 1) implies that

$$
\begin{cases}
\lim_{N\to\infty} \Pr(|t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h) - \mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)]| < \vartheta) = 1, & \text{(A.5)} \\
\lim_{N\to\infty} \Pr(|t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h) - \mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)]| < \varrho) = 1, & \text{(A.6)}
\end{cases}
$$

where ϑ and ϱ are arbitrarily small constants such that $\varphi, \varrho > 0$. Since the Difference-in-Means is

unbiased, note that $\mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] = \tau - \tau_h$ and $\mathrm{E}[t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h)] = 0$, so Equation (A.5) and Equation (A.6) can be written as

$$\begin{cases} \lim_{N\to\infty} \Pr(|t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h) - (\tau - \tau_h)| < \vartheta) = 1, & \text{(A.7)} \\ \lim_{N\to\infty} \Pr(|t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h) - 0| < \varrho) = 1 & \text{(A.8)} \end{cases}$$

We can further recast Equation (A.5) as $t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h) \in (\tau - \tau_h - \vartheta, \tau - \tau_h + \vartheta)$ and Equation (A.6) as $t(\boldsymbol{Z}, \tilde{\boldsymbol{y}}_h) \in (0 - \varrho, 0 + \varrho)$, and then note that if

$$t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h) \in (\tau - \tau_h - \vartheta, \tau - \tau_h + \vartheta),$$

then

$$t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h) \notin (0 - \varrho, 0 + \varrho),$$

whenever

$$\begin{cases} \tau_h - \tau \geq \vartheta + \varrho & \text{if } \tau > \tau_h, \\ \tau_h - \tau \leq -\vartheta - \varrho & \text{if } \tau < \tau_h. \end{cases}$$

To complete the proof, simply let $\varphi = \vartheta + \varrho$ and then, analogous to the proof of Theorem 1 above, it follows that

$$\lim_{N\to\infty} \left[\frac{\int_{\tau_h \in \Theta^-_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h + \int_{\tau_h \in \Theta^+_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h}{\int_{\tau_h \in \Theta_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h} \right]$$

$$= \frac{\int_{\tau_h \in \Theta^-_{\tau_h}} \overbrace{(\lim_{N\to\infty} f(T \mid \tau_h))}^{=0} p(\tau_h) d\,\tau_h + \int_{\tau_h \in \Theta^+_{\tau_h}} \overbrace{(\lim_{N\to\infty} f(T \mid \tau_h))}^{=0} p(\tau_h) d\,\tau_h}{\int_{\tau_h \in \Theta} f(T \mid \tau_h) p(\tau_h) d\,\tau_h} = 0,$$

and, by the Law of Total Probability,

$$\lim_{N\to\infty} \int_{\tau_h \in \Theta^*_{\tau_h}} f(\tau_h \mid T) d\,\tau_h$$

$$= 1 - \lim_{N\to\infty} \left[\frac{\int_{\tau_h \in \Theta^-_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h + \int_{\tau_h \in \Theta^+_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h}{\int_{\tau_h \in \Theta_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h} \right]$$

$$= \lim_{N\to\infty} \left[1 - \frac{\int_{\tau_h \in \Theta^-_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h + \int_{\tau_h \in \Theta^+_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h}{\int_{\tau_h \in \Theta_{\tau_h}} f(T \mid \tau_h) p(\tau_h) d\,\tau_h} \right]$$

$$= (\lim_{N\to\infty} 1) - 0 = 1,$$

which proves the theorem. □

A.0.4 Proof of Proposition 3

Denoting $\boldsymbol{Y} - \boldsymbol{Z}\tau_h$ by $\tilde{\boldsymbol{Y}}_h$, note that for $N = 4, 5, \ldots,$

$$\text{Var}[t(\boldsymbol{Z}, \tilde{\boldsymbol{Y}}_h)] = \frac{N}{n_T n_C (N-1)} \sum_{i=1}^{N} \left(\tilde{Y}_i - \bar{\tilde{Y}} \right)^2$$

and that

$$\sum_{i=1}^{N} \left(\tilde{Y}_i - \bar{\tilde{Y}} \right)^2 = \sum_{i=1}^{N} \left(Z_i \tilde{Y}_{hi} - \frac{1}{n_T} \sum_{i=1}^{N} Z_i \tilde{Y}_{hi} \right)^2 + n_T \left(\frac{1}{n_T} \sum_{i=1}^{N} Z_i \tilde{Y}_{hi} - \frac{1}{N} \sum_{i=1}^{N} \tilde{Y}_{hi} \right)^2 + \tag{A.9}$$

$$\sum_{i=1}^{N} \left((1 - Z_i) \tilde{Y}_{hi} - \frac{1}{n_T} \sum_{i=1}^{N} (1 - Z_i) \tilde{Y}_{hi} \right)^2 + n_C \left(\frac{1}{n_C} \sum_{i=1}^{N} (1 - Z_i) \tilde{Y}_{hi} - \frac{1}{N} \sum_{i=1}^{N} \tilde{Y}_{hi} \right)^2. \tag{A.10}$$

Noting that

$$\widehat{\text{Var}} \left[\hat{\bar{\tau}} \right] = \frac{s^2_{y_T}}{n_T} + \frac{s^2_{y_C}}{n_C} \tag{A.11}$$

and taking the difference between Equation (A.9) and $\widehat{\text{Var}} \left[\hat{\bar{\tau}} \right]$ yields the desired expression.

A.0.5 Proof of Proposition 4

First, since $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$ is known conditional on the realized data, it follows from the definition of joint probability that

$$f\left(\hat{\tau}_{\mathcal{T}}, \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) = f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) p(\tau_{\mathcal{T}}). \tag{A.12}$$

Note that $f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)$ is the Normal likelihood function in which a value of the TATE, $\tau_{\mathcal{T}}$, and a conservative plug-in variance estimator, $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$, assign probability density to the realized Difference-in-Means in the target experiment, $\hat{\tau}_{\mathcal{T}}$. The Normal prior distribution of the TATE, $\tau_{\mathcal{T}}$, is $p(\tau_{\mathcal{T}})$.

With a Normal likelihood and a Normal prior distribution, it follows that their product is

$$\begin{aligned} & f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) p(\tau_{\mathcal{T}}) \\ &= \left(2\pi\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^{-1/2} \exp\left(-\frac{1}{2\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]}\left(\hat{\tau}_{\mathcal{T}} - \tau_{\mathcal{T}}\right)^2\right)\left(2\pi\sigma^2_{\text{prior}}\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2_{\text{prior}}}\left(\tau_{\mathcal{T}} - \mu_{\text{prior}}\right)^2\right), \end{aligned} \tag{A.13}$$

which algebraically simplifies to

$$\left(2\pi\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^{-1/2}\left(2\pi\sigma^2_{\text{prior}}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]}\left(\hat{\tau}_{\mathcal{T}} - \tau_{\mathcal{T}}\right)^2 + \frac{1}{\sigma^2_{\text{prior}}}\left(\tau_{\mathcal{T}} - \mu_{\text{prior}}\right)^2\right)\right). \tag{A.14}$$

Then note that, after some algebra, it follows that

$$\begin{aligned} & \frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]}\left(\hat{\tau}_{\mathcal{T}} - \tau_{\mathcal{T}}\right)^2 + \frac{1}{\sigma^2_{\text{prior}}}\left(\tau_{\mathcal{T}} - \mu_{\text{prior}}\right)^2 \\ &= \frac{\hat{\tau}^2_{\mathcal{T}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{\mu^2_{\text{prior}}}{\sigma^2_{\text{prior}}} - \left(\frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{1}{\sigma^2_{\text{prior}}}\right)\left(\frac{\mu_{\text{prior}}\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \hat{\tau}_{\mathcal{T}}\sigma^2_{\text{prior}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \sigma^2_{\text{prior}}}\right)^2 \\ &\quad + \left(\frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{1}{\sigma^2_{\text{prior}}}\right)\left(\tau_{\mathcal{T}} - \left(\frac{\mu_{\text{prior}}\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \hat{\tau}_{\mathcal{T}}\sigma^2_{\text{prior}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \sigma^2_{\text{prior}}}\right)\right)^2. \end{aligned} \tag{A.15}$$

To simplify the notation, now let

$$\eta^2 = \left(\frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{1}{\sigma^2_{\text{prior}}} \right) \text{ and} \tag{A.16}$$

$$\zeta = \frac{\mu_{\text{prior}} \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \hat{\tau}_{\mathcal{T}} \sigma^2_{\text{prior}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \sigma^2_{\text{prior}}} \tag{A.17}$$

and then plug Equation (A.15) back into Equation (A.14), which yields

$$\left(2\pi \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^{-1/2} \left(2\pi\sigma^2_{\text{prior}}\right)^{-1/2} \exp\left(-\frac{1}{2} \left(\frac{\hat{\tau}^2_{\mathcal{T}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{\mu^2_{\text{prior}}}{\sigma^2_{\text{prior}}} - \eta^2\zeta^2 \right) \right) \exp\left(-\frac{1}{2}\eta^2 \left(\tau_{\mathcal{T}} - \zeta\right)^2 \right). \tag{A.18}$$

Equation (A.18) further simplifies to

$$\begin{aligned} \left(2\pi \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^{-1/2} \left(2\pi\sigma^2_{\text{prior}}\right)^{-1/2} \left(2\frac{\pi}{\eta^2}\right)^{1/2} \exp\left(-\frac{1}{2\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} \left(\hat{\tau}^2_{\mathcal{T}} + \frac{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\mu^2_{\text{prior}}}{\sigma^2_{\text{prior}}} - \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\eta^2\zeta^2 \right) \right) \\ \times \left(2\frac{\pi}{\eta^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\eta^2 \left(\tau_{\mathcal{T}} - \zeta\right)^2 \right). \end{aligned} \tag{A.19}$$

In Equation (A.19) and from the expressions for η^2 and ζ in Equations (A.16) and (A.31), we can see that

$$\left(2\pi \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)^{-1/2} \left(2\pi\sigma^2_{\text{prior}}\right)^{-1/2} \left(2\frac{\pi}{\eta^2}\right)^{1/2} \exp\left(-\frac{1}{2\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} \left(\hat{\tau}^2_{\mathcal{T}} + \frac{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\mu^2_{\text{prior}}}{\sigma^2_{\text{prior}}} - \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\eta^2\zeta^2 \right) \right) \tag{A.20}$$

does not depends on the value of the TATE, $\tau_{\mathcal{T}}$, but does depend on $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$ and $\hat{\tau}_{\mathcal{T}}$ through ζ. We can also see that

$$\left(2\frac{\pi}{\eta^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\eta^2 \left(\tau_{\mathcal{T}} - \zeta\right)^2 \right) \tag{A.21}$$

is a function of $\tau_{\mathcal{T}}$, $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$ and $\hat{\tau}_{\mathcal{T}}$, and that it is the Normal density function with mean (location) parameter equal to ζ and variance (squared scale) parameter equal to η^2.

Denote the function of $\hat{\tau}_{\mathcal{T}}$ and $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$ in Equation (A.20) by $g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)$ and the function of $\tau_{\mathcal{T}}$, $\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]$ and $\hat{\tau}_{\mathcal{T}}$ in Equation (A.21) by $g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right)$, which implies that Equation (A.19) can be written as $g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right)$ and that, referring back to Equation (A.12),

$$f\left(\hat{\tau}_{\mathcal{T}}, \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) = g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) \tag{A.22}$$

Then note that

$$\begin{aligned}
f(\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}) &= \int_{\tau_{\mathcal{T}}} f\left(\hat{\tau}_{\mathcal{T}}, \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) d\,\tau_{\mathcal{T}} && \text{(A.23)}\\
&= \int_{\tau_{\mathcal{T}}} g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) d\,\tau_{\mathcal{T}} && \text{(A.24)}\\
&= g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) \int_{\tau_{\mathcal{T}}} g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) d\,\tau_{\mathcal{T}} && \text{(A.25)}\\
&= g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right). && \text{(A.26)}
\end{aligned}$$

where lines (A.25) and (A.26) follow from the fact that $g\left(\hat{\tau}_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right)$ does not depend on $\tau_{\mathcal{T}}$ and that $g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right)$ is a density function and, hence, it must integrate to 1.

Therefore,

$$\begin{aligned}
f\left(\hat{\tau}_{\mathcal{T}}, \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) &= g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) f(\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}) && \text{(A.27)}\\
f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) p(\tau_{\mathcal{T}}) &= g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) f(\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}) && \text{(A.28)}\\
\frac{f\left(\hat{\tau}_{\mathcal{T}} \mid \tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]\right) p(\tau_{\mathcal{T}})}{f(\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}})} &= g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right), && \text{(A.29)}
\end{aligned}$$

which implies that

$$g\left(\tau_{\mathcal{T}}, \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}], \hat{\tau}_{\mathcal{T}}\right) = \left(2\frac{\pi}{\eta^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\eta^2\,(\tau_{\mathcal{T}} - \zeta)^2\right) \tag{A.30}$$

is the Normal posterior density function with parameters

$$\sigma^2_{\text{post}} = \left(\frac{1}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}]} + \frac{1}{\sigma^2_{\text{prior}}} \right)^{-1} \text{ and}$$

$$\mu_{\text{post}} = \frac{\mu_{\text{prior}} \widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \hat{\tau}_{\mathcal{T}} \sigma^2_{\text{prior}}}{\widehat{\text{Var}}[\hat{\tau}_{\mathcal{T}}] + \sigma^2_{\text{prior}}}.$$

Then taking $\sigma^2_{\text{post}} - \sigma^2_{\text{prior}}$ and $\mu_{\text{post}} - \mu_{\text{prior}}$ yields the desired expressions.

Proof of Proposition 5

Proof. By the definition of $\Delta_{\mathcal{P}_N}$ in Equation (3.5) and the unbiasedness of the DID estimator in Equation (3.3) for the population difference in differences, $\text{E}\left[\widehat{\text{DID}}_{\mathcal{P}}\right] = \text{E}\left[Y_{iT}(1) \mid z_i = 1\right] - \text{E}\left[Y_{iT-1} \mid z_i = 1\right] - (\text{E}\left[Y_{iT} \mid z_i = 0\right] - \text{E}\left[Y_{iT-1} \mid z_i = 0\right])$, where $\text{E}[\cdot]$ is taken over the set of possible random samples from population $\mathcal{P}_N$, it follows that

$$\begin{aligned}
&\text{E}\left[\widehat{\text{DID}}_{\mathcal{P}}\right] - \Delta_{\mathcal{P}_N} \\
&= \left[\text{E}\left[Y_{iT}(1) \mid z_i = 1\right] - \text{E}\left[Y_{iT-1} \mid z_i = 1\right] - (\text{E}\left[Y_{iT} \mid z_i = 0\right] - \text{E}\left[Y_{iT-1} \mid z_i = 0\right])\right] \\
&- \left[\text{E}\left[Y_{iT}(0) \mid z_i = 1\right] - \text{E}\left[Y_{iT-1} \mid z_i = 1\right] - (\text{E}\left[Y_{iT} \mid z_i = 0\right] - \text{E}\left[Y_{iT-1} \mid z_i = 0\right])\right].
\end{aligned}$$

Then, by simple algebra and the linearity of expectations, it follows that

$$
\begin{aligned}
&\mathrm{E}\left[\widehat{\mathrm{DID}}_{\mathcal{P}}\right] - \Delta_{\mathcal{P}_N} \\
&= (\mathrm{E}\left[Y_{iT}(1) \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT-1} \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT} \mid z_i = 0\right] + \mathrm{E}\left[Y_{iT-1} \mid z_i = 0\right]) \\
&- (\mathrm{E}\left[Y_{iT}(0) \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT-1} \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT} \mid z_i = 0\right] + \mathrm{E}\left[Y_{iT-1} \mid z_i = 0\right]) \\
&= \mathrm{E}\left[Y_{iT}(1) \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT-1} \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT} \mid z_i = 0\right] + \mathrm{E}\left[Y_{iT-1} \mid z_i = 0\right] \\
&- \mathrm{E}\left[Y_{iT}(0) \mid z_i = 1\right] + \mathrm{E}\left[Y_{iT-1} \mid z_i = 1\right] + \mathrm{E}\left[Y_{iT} \mid z_i = 0\right] - \mathrm{E}\left[Y_{iT-1} \mid z_i = 0\right] \\
&= \mathrm{E}\left[Y_{iT}(1) \mid z_i = 1\right] - \mathrm{E}\left[Y_{iT}(0) \mid z_i = 1\right] \\
&= \underbrace{\mathrm{E}\left[Y_{iT}(1) - Y_{iT}(0) \mid z_i = 1\right]}_{\mathrm{ATT}_{\mathcal{P}}},
\end{aligned}
$$

which is equal to $\mathrm{ATT}_{\mathcal{P}_N}$ as defined in Equation (3.1), thereby proving the proposition. One can analogously prove the same result in a finite sample, $\mathcal{S}_n$, albeit with fewer assumptions since $\mathrm{ATT}_{\mathcal{S}_n}$, $\widehat{\mathrm{DID}}_{\mathcal{S}_n}$ and $\Delta_{\mathcal{S}_n}$ are all fixed quantities in a finite sample. □

A.0.6 Proof of Corollary 1

Proof. Steps 11 and 12 of Algorithm 1 state that

$$
\bar{y}_T^{\mathrm{imp},s}(0) = \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} (y_{iT} - \hat{y}_{iT}) + \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} \hat{y}_{iT}^{s}(0) \text{ and} \tag{A.31}
$$

$$
\Delta^{\mathrm{imp},s} = \bar{y}_T^{\mathrm{imp},s}(0) - \left(\frac{1}{n_1}\right) \sum_{i=n_0+1}^{n} y_{iT-1} - \left[\left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT} - \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} y_{iT-1}\right]. \tag{A.32}
$$

Using Step 11 of Algorithm 1 to substitute for $\bar{y}_T^{\text{imp},s}(0)$ in Step 12 yields

$$\begin{aligned}
\Delta^{\text{imp},s} &= \left[\left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0}(y_{iT}-\hat{y}_{iT}) + \left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n}\hat{y}_{iT}^{s}(0)\right] \\
&\quad - \left(\frac{1}{n_1}\right)\sum_{i=n_0+1}^{n} y_{iT-1} - \left[\left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0} y_{iT} - \left(\frac{1}{n_0}\right)\sum_{i=1}^{n_0} y_{iT-1}\right] \\
&= \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left(\hat{y}_{iT}^{s}(0) - y_{iT-1}\right) - \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\hat{y}_{iT} - y_{iT-1}\right) \qquad \text{(A.33)}
\end{aligned}$$

With the expression for $\Delta^{\text{imp},s}$ in Equation (A.33), it follows that $\widehat{\text{DID}} - \Delta^{\text{imp},s}$ is

$$\begin{aligned}
\widehat{\text{DID}} - \Delta^{\text{imp},s} &= \frac{1}{n_1}\sum_{i=n_0+1}^{n}(y_{iT}(1) - y_{iT-1}) - \frac{1}{n_0}\sum_{i=1}^{n_0}(y_{iT} - y_{iT-1}) \\
&\quad - \left\{\frac{1}{n_1}\sum_{i=n_0+1}^{n}\left(\hat{y}_{iT}^{s}(0) - y_{iT-1}\right) + \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\hat{y}_{iT} - y_{iT-1}\right)\right\} \\
&= \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left(y_{iT} - \hat{y}_{iT}^{s}(0)\right) - \underbrace{\frac{1}{n_0}\sum_{i=1}^{n_0}(\hat{y}_{iT} - y_{iT})}_{\substack{\text{Average prediction} \\ \text{error in control group}}}
\end{aligned}$$

Then in simulation s, note that $\hat{y}_{iT}^{s}(0) = \boldsymbol{x}_{1,T}\tilde{\boldsymbol{\beta}}^{s}$, where

$$\tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\Sigma}}_1\right)$$

in which $\hat{\boldsymbol{\beta}}_1$ is a vector of regression coefficients from a pre-period regression of the treated group. If we then average over many simulation draws,

$$\lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S}\hat{y}_{iT}^{s}(0) = \boldsymbol{x}_{1,T}\hat{\boldsymbol{\beta}}_1$$

by the law of large numbers. □

Appendix B: Simulation Results

Case 1 Simulation parameters:	$n_T/N = 0.5$, $\bar{y}_T = 0.1$, $\bar{y}_C = -0.02$, $S_{y_T} = 0.06$, $S_{y_C} = 0.07$, $S_\tau = 0.14$
Average MSE (weak hypotheses)	0.0044
Average MSE (sharp hypotheses)	0.0046
Case 2 Simulation parameters:	$n_T/N = 0.5$, $\bar{y}_T = 0.31$, $\bar{y}_C = -0.02$, $S_{y_T} = 0.06$, $S_{y_C} = 0.07$, $S_\tau = 0.14$
Average MSE (weak hypotheses)	0.0045
Average MSE (sharp hypotheses)	0.0046
Case 3 Simulation parameters:	$n_T/N = 0.7$, $\bar{y}_T = 0.11$, $\bar{y}_C = -0.02$, $S_{y_T} = 0.23$, $S_{y_C} = 0.07$, $S_\tau = 0.14$
Average MSE (weak hypotheses)	0.0096
Average MSE (sharp hypotheses)	0.0136
Case 4 Simulation parameters:	$n_T/N = 0.7$, $\bar{y}_T = 0.1$, $\bar{y}_C = -0.05$, $S_{y_T} = 0.06$, $S_{y_C} = 0.27$, $S_\tau = 0.36$
Average MSE (weak hypotheses)	0.0201
Average MSE (sharp hypotheses)	0.0154

Table B.1: Simulation results