Placebo Tests for Causal Inference

Andrew C. Eggers[1], Guadalupe Tuñón[3], and Allan Dafoe[2,4]

[1] Department of Political Science, University of Chicago

[2] Department of Politics and International Relations, University of Oxford

[3] Department of Politics and Woodrow Wilson School, Princeton University

[4] Centre for the Governance of AI, Future of Humanity Institute, University of Oxford

This draft: February 18, 2021

Abstract

Placebo tests allow researchers to probe the soundness of a research design by checking for an association that should be present if the design is flawed but not otherwise. Despite the growing popularity of placebo tests, the principles for designing and interpreting them have remained obscure. Drawing on a comprehensive survey of recent empirical work in political science, we define placebo tests, introduce a typology of tests, and analyze what makes them informative. We consider examples of each type of test and discuss how to design and evaluate tests for specific research designs. In sum, we offer a guide to understanding and using placebo tests to improve causal inference.

Authors can be reached at allan.dafoe@politics.ox.ac.uk, andrew.eggers@nuffield.ox.ac.uk, tunon@princeton.edu. For helpful input, we thank Devin Caughey, Thad Dunning, Anthony Fowler, Don Green, Sophia Hatz, Luke Miratrix, Jas Sekhon, Nils Weidmann, Baobao Zhang, and the Yale and Uppsala students of *Advanced Quantitative Methods*. Hayley Pring, Rhys Dubin, and Jonne Kamphorst provided excellent research assistance. Audiences at ISA, Oxford, NYU-Abu Dhabi, Columbia, and the International Methods Colloquium provided useful feedback.

1 Introduction

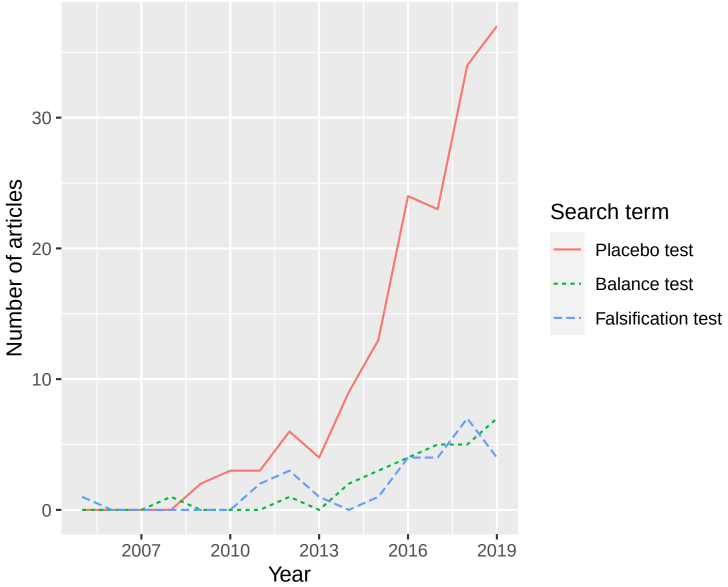
In an observational study measuring the effect of a treatment on an outcome, a researcher's job is only partly done once she estimates the treatment effect. In addition to assessing whether a similarly strong association could have arisen by chance (usually via null-hypothesis significance testing), researchers often conduct robustness checks to assess how conclusions depend on modeling choices (Neumayer and Plümper 2017), subgroup analyses[1] to check whether the treatment effect varies across units in a way that corresponds with the author's causal theory (Cochran and Chambers 1965; Rosenbaum 2002), and sensitivity analyses to assess bias due to remaining confounders (Rosenbaum and Rubin 1983; Cinelli and Hazlett 2020). These auxiliary analyses help the reader judge whether the estimated treatment effect reliably measures the treatment effect or instead reflects random error, misspecification, confounding, or something else.

In this paper, we study placebo tests, another form of auxiliary analysis for observational studies. Like the other types just mentioned, placebo tests are designed to assess possible shortcomings in observational studies: a placebo test looks for an association that may be present if the main analysis is flawed but should not be present otherwise. The term "placebo test" has its origins in medicine, where a "placebo" originally referred to an ineffective medicine prescribed to reassure a worried patient through deception (De Craen et al. 1999) and later came to refer to a pharmacologically inert passive treatment in drug trials. In observational studies in epidemiology, economics, and other social sciences, "placebo test" now refers to auxiliary analyses where the treatment (like a placebo in a drug trial) should not or cannot have an effect, and finding an apparent effect could indicate important flaws in the study.

The use of placebo tests in political science has grown rapidly in recent years. Figure 1 shows the annual number of papers including "placebo test" and closely related terms that were published in seven top political science journals (*APSR*, *AJPS*, *JOP*, *IO*, *BJPS*, *QJPS*,

[1] Ideally preregistered.

Figure 1: The number of articles mentioning "placebo test" and related terms in seven top political science journals, 2005-2019



CPS) between 2005 and 2019.[2] We found no papers mentioning "placebo test" before 2009, but by 2019 over thirty-five such articles appeared (which is over 5% of all articles mentioning the word "test"). The growing popularity of placebo tests likely reflects the diffusion of more rigorous standards for causal inference in the discipline (including exhortations by Sekhon (2009), Dunning (2012), and others to conduct placebo tests), rising expectations for robustness tests and other auxiliary analyses (Neumayer and Plümper 2017), and the considerable intuitive appeal of some of the best applications.

Despite the increasingly widespread use of placebo tests, it can be difficult to understand what makes placebo tests work, both in specific cases and in general, and how to design them. Insights about placebo tests are scattered across empirical applications and in methodological articles in several disciplines where the same basic practice is referred to by different names (e.g. refutability tests, falsification tests, balance tests, tests for known effects, tests of unconfoundedness, tests with negative controls). Many discussions address only one type of what we consider a more general class of placebo tests, and thus the un-

[2] We counted hits on Google Scholar.

derlying links between related tests remain unclear. Perhaps as a result, many studies that could include informative placebo tests still fail to do so; there is also (as we show below) apparent disagreement about how some placebo tests should be implemented, and many published papers include “placebo tests” that we argue have little evidential value.

This paper aims to improve the use and evaluation of placebo tests in social science by cutting through the existing thicket of conflicting terminology and notation to clarify what a placebo test is, what makes placebo tests informative, and how they should be designed and interpreted. After defining placebo tests, we offer a new typology of tests based on how the test alters the research design in the main analysis. For each type of test, we illustrate the basic logic using directed acyclic graphs (DAGs) and a single running example (Peisakhin and Rozenas 2018); we also clarify what properties make the test more or less informative and discuss these properties with reference to additional examples drawn mostly from recent political science research. We apply these insights in discussing how to design placebo tests for specific research designs (IV, RDD, and diff-in-diff). In the course of explicating and analyzing the various types of test, we raise and address several thorny questions: Why would we use a possible confounder in a placebo test rather than just controlling for it? Should a placebo test that uses an altered version of the treatment control for the actual treatment? What do we learn from placebo tests for RDDs that use “fake cutoffs”? We conclude by noting that “null-hacking” (i.e. p-hacking with the goal of producing insignificant results) is a particular threat for placebo tests, but that the research community can address it both by encouraging pre-registration of placebo tests and by developing clearer standards for the design and implementation of placebo tests. A final notable contribution is Appendix A, which summarizes and categorizes over one hundred placebo tests that appear in recent political science research. This library of placebo tests provides an empirical foundation for the paper, informing the definition, typology, and theoretical approach we adopt; it is also a resource scholars can use in devising informative placebo tests in the future.

2 What is a placebo test? A definition and typology

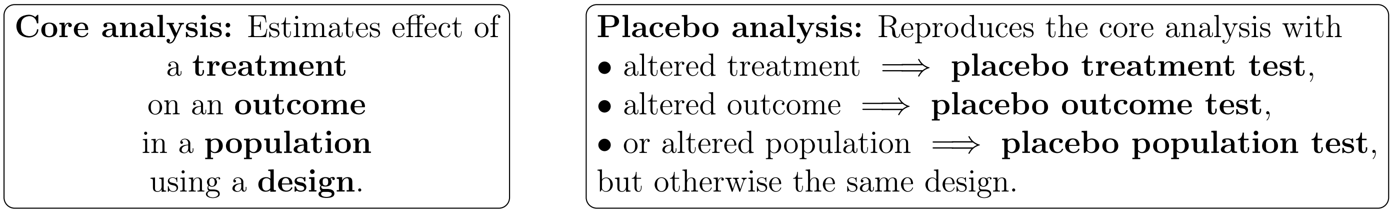
Placebo tests diagnose problems with research designs in observational studies. When a researcher estimates a treatment effect based on observational data, the estimator may be biased by confounders, model misspecification, differential measurement error, or other flaws; the researcher may also have constructed confidence intervals incorrectly, such that we would reject the null hypothesis too frequently (or infrequently) under the null. A placebo test checks for an association that should be absent if the research design is sound but not otherwise. Placebo tests can thus be seen as a strategy for checking the soundness of a research finding and, more broadly, improving causal inference.

Although in principle there may be many ways to design a placebo test, our survey of the political science literature indicates that almost all placebo tests measure the effect of the treatment on the outcome in a minimally altered version of the original research design. (We examined every observational study mentioning a "placebo test," "balance test," or "falsification test" in the *APSR*, *AJPS*, *JOP*, and *IO* between 2009 and 2018; the resulting summary of 110 placebo tests appears in Appendix A.[3]) We will call the original research design the *core analysis* and the altered version the *placebo analysis*. We observe three ways in which the core analysis is most commonly altered.[4] This suggests a typology we use throughout the paper. We use the term *placebo population test* to refer to a test that replicates the core analysis in a different population, *placebo outcome test* to refer to a test that replicates the core analysis with a different outcome variable, and *placebo treatment test* to refer to a test that replicates the core analysis with a different treatment variable. We use the terms *placebo population*, *placebo outcome*, and *placebo treatment* to refer to the

[3] This is a nearly exhaustive list of placebo tests appearing in these journals during these years; the main exception is that we include only a sample of the simplest types of placebo tests (balance tests and fake-cutoff tests from RDD studies); we also include only one test of each type per paper, and we omit the two tests we could not categorize.

[4] The exceptions are Model 6 in Thachil (2014), which puts the treatment on the LHS, and Gilardi (2015), which changes both the treatment and the outcome (in a manner similar to Cohen-Cole and Fletcher (2008)).

Figure 2: Schematic illustrating typology and key terms



Core analysis: Estimates effect of
a **treatment**
on an **outcome**
in a **population**
using a **design**.

Placebo analysis: Reproduces the core analysis with
- altered treatment $\Longrightarrow$ **placebo treatment test**,
- altered outcome $\Longrightarrow$ **placebo outcome test**,
- or altered population $\Longrightarrow$ **placebo population test**,

but otherwise the same design.

component that has been altered in each case. Figure 2 summarizes the typology and related key terms.

What we refer to as a placebo test has been referred to by a variety of other terms. "Falsification test" is used in the same way we use "placebo test" (e.g. Pizer 2016; Laitin and Ramachandran 2016; Healy and Lenz 2017). "Balance test" is also widely used to refer to what we call a placebo outcome test in the case where the placebo outcome is a pre-treatment variable. Rosenbaum (1984, 1989, 2010) describes as "tests of known effects" and "tests of strongly ignorable treatment assignment" procedures that we would call placebo tests, though our definition of placebo tests also encompasses tests where the treatment's effect is not known or where strongly ignorable treatment assignment is not assumed. Epidemiologists use the term "negative control outcome" and "negative control exposure" to refer to what we call a placebo outcome and placebo treatment, respectively (Lipsitch, Tchetgen Tchetgen and Cohen 2010). As we describe below, a common logic unites these disparate practices, and we aim to highlight that common logic by discussing them in a common framework with common terminology.[5]

[5] There is a close parallel between randomization inference (e.g. as implemented for the synthetic control method by Abadie, Diamond and Hainmueller 2010, 2015) and placebo tests that randomly permute the treatment. The purpose of randomization inference is to generate a null distribution and p-value for a test statistic (as in Abadie, Diamond and Hainmueller (2010) or Berry and Fowler (2021)); a similar procedure might be called a "placebo test" when the purpose is to check the rejection rate for a particular procedure of statistical inference (as in Bertrand, Duflo and Mullainathan (2004)).

3 What makes a placebo test informative?

Generally, a placebo test is *informative* to the extent that it is capable of influencing our beliefs about flaws in the core analysis. To see what kind of test has that property, focus (to simplify the exposition) on placebo tests that yield a binary result: either we find a statistically significant (conditional) association between the treatment and the outcome in the placebo test or we do not. (To some extent this reflects the common practice, which views a placebo test as "passing" if $p > .05$ in the placebo analysis and "failing" otherwise.[6]) Then the informativeness of a placebo test can be summarized by two features: the probability of finding a significant association when the core analysis is not flawed (which we will refer to as the test's *size*, or false positive rate) and the probability of finding a significant association when the core analysis is flawed in a particular way (the test's *power*). As is well known (e.g. Royall 1997, pp. 48-49), the informativeness of a significance test (i.e. the extent to which the binary reject-or-not outcome affects our beliefs about the relative probability of some null hypothesis or an alternative hypothesis) depends only on its size and power.[7] The most informative placebo tests have small size (a low probability of rejecting the null if the core analysis is sound) and large power for some relevant flaw (a high probability of rejecting the null if the flaw is present); in that case rejecting the null provides strong evidence that the flaw is present while not rejecting provides strong evidence that it is not.

Consider first the size of the placebo test, and suppose that the confidence intervals in the placebo test are constructed so that, when there is no conditional association in the placebo analysis, we would reject the null with probability α. Then the size of the placebo test is also α if, when the core analysis is sound, (i) the effect of the treatment on the outcome

[6] Hartman and Hidalgo (2018) make a convincing case for the equivalence testing approach, in which the researcher seeks to reject the null hypothesis of a significant association in the placebo analysis.

[7] The ratio of the posterior probability that the core analysis is sound vs. flawed given a failed placebo test is the ratio of the prior probabilities (sound/flawed) times the ratio of the size to the power. Note that there are two null hypotheses in play: the null hypothesis that treatment and the outcome are conditionally independent in the placebo analysis and the null hypothesis that the core analysis is sound. We define a placebo test's size and power here as the probability of rejecting the first null hypothesis as a function of whether the second null hypothesis holds.

in the placebo analysis is zero and (ii) the placebo analysis obtains an unbiased estimate of that effect. The size could be greater than α if the treatment actually affects the outcome in the placebo and/or there are biases in the placebo analysis even when the core analysis is sound.

The power of a placebo test depends on the extent to which flaws in the core analysis are mirrored in the placebo analysis. If the goal is to detect bias, we want a placebo test that would produce a biased estimate of the treatment effect whenever that bias is present in the core analysis. (It should also estimate the treatment effect sufficiently precisely to detect that bias.) When the relevant concern is about confidence intervals in the core analysis (e.g. Fowler and Hall 2018), we want a placebo test that has incorrect confidence intervals whenever the core analysis does, and that allows us to assess the proportion of significant results across many tests.[8]

In summary, an informative placebo test is one where, by altering the core analysis, one extinguishes the effect of the treatment on the outcome while retaining potential flaws; finding a significant association in the placebo analysis then suggests that those flaws may operate in the core analysis. The key design challenge is to find an alteration that simultaneously suppresses the treatment effect (keeping the size low) while still allowing us to detect relevant flaws (making the power high). In what follows we consider how this can be achieved for each type of tests and specific research designs.

4 A running example, illustrated with DAGs

To illustrate our arguments we will refer periodically to Peisakhin and Rozenas (2018), which includes an unusually large number and variety of placebo tests. Peisakhin and Rozenas (2018) aim to measure the effects of politically slanted Russian news TV on voting and political attitudes in Ukraine around an election in 2014. Before the election, Russian TV

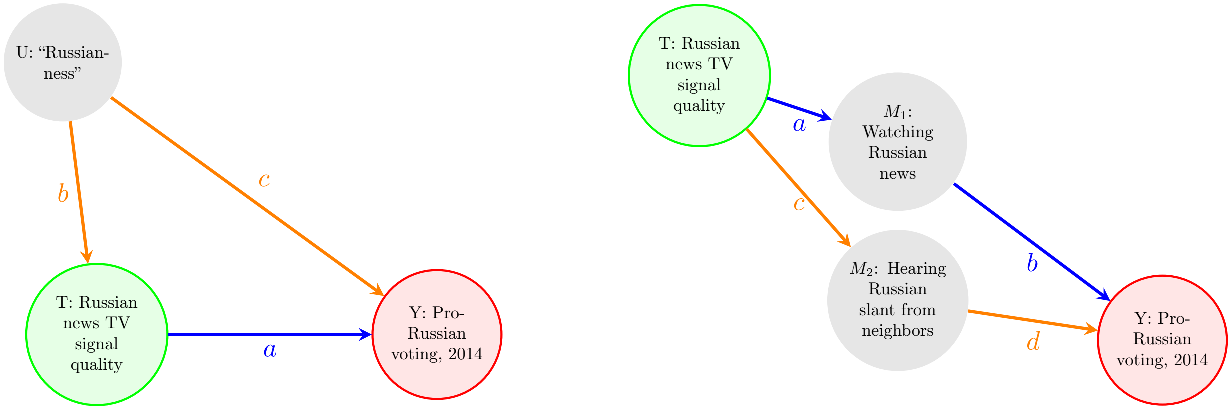
[8] In that case the distribution of estimates across tests can clarify whether bias or incorrect confidence intervals (or both) are responsible.

transmitters near the Ukrainian border broadcast Russian-language news programming that offered a pro-Russian slant on Ukrainian politics. Some voters in the northeast region of Ukraine could access these broadcasts while others could not; the quality of the signal depended on how far a particular receiver was from transmitters on the Russian side of the border, but also on the terrain lying between the receiver and the transmitters. Peisakhin and Rozenas (2018) use two types of analysis to assess the effect of slanted Russian news on perceptions in Ukraine. First they use precinct-level results from the 2014 election to measure the effect of exposure to Russian news TV on aggregate support for pro-Russian parties in the 2014 election. Next they use survey data and instrumental variables (IV) analysis to measure the effect of watching Russian news on attitudes and behavior at the individual level.

To highlight potential biases that may affect Peisakhin and Rozenas (2018)'s analysis and explain how placebo tests might address them, we use simple directed acyclic graphs (DAGs).[9] The DAG at left in Figure 3 provides a simplified representation of possible confounding bias in Peisakhin and Rozenas (2018). In their precinct-level analysis, Peisakhin and Rozenas (2018) are interested in measuring the effect of Russian news TV signal quality on support for pro-Russian parties; this effect is indicated by the arrow labeled a. They do this by regressing voting outcomes in Ukrainian precincts on a measure of estimated Russian news TV signal quality in the precinct. They include controls (county or district fixed effects and a flexible function of distance to Russia) but we omit those from the DAG for simplicity. The DAG highlights the potential concern that precincts with better reception of Russian news TV may have been more sympathetic to Russia even in the absence of Russian news broadcasts; this might occur if Russian transmitters were strategically placed to reach Ukrainian communities known to be sympathetic to the Russian perspective or if the same geographical features (e.g. altitude) affected both TV reception in 2014 and historical patterns of Russian settlement and cultural influence in Ukraine. We summarize

[9] Lipsitch, Tchetgen Tchetgen and Cohen (2010) similarly illustrate the logic of placebo tests in epidemiology with DAGs.

Figure 3: Potential bias in Peisakhin and Rozenas (2018): a simplified view via DAGs


U: "Russian-ness"
b
c
T: Russian news TV signal quality
a
Y: Pro-Russian voting, 2014
T: Russian news TV signal quality
a
M_1: Watching Russian news
c
b
M_2: Hearing Russian slant from neighbors
d
Y: Pro-Russian voting, 2014

Bias from a confounder

Bias from an alternative mechanism

these concerns with a single confounder labeled "Russian-ness" that affects both Russian TV reception (the arrow labeled b) and voting for pro-Russian parties in 2014 (the arrow labeled c). These effects could create a dependence between signal quality and pro-Russian voting that is not due to the effect of signal quality on voting, i.e. that constitutes a confounding bias.

Similarly, the DAG at right in Figure 3 provides a simplified representation of the possibility that Russian TV signal quality could affect voting behavior through alternative mechanisms, which is primarily a concern for Peisakhin and Rozenas (2018)'s IV analysis. The authors aim to measure the effect of watching Russian news on voting behavior, represented by the arrow labeled b. The choice to watch Russian news TV may depend on many possible confounders (not shown). To address this concern, Peisakhin and Rozenas (2018) use Russian news TV signal quality as an instrument for watching Russian news TV; the exclusion restriction requires that signal quality affects voting *only* by influencing consumption of Russian news TV (i.e. through the path labeled a and b). The DAG highlights an alternative mechanism by which Russian news TV signal quality could affect voting behavior: better TV signal quality could make it more likely that one hears a Russian-slanted interpretation of current affairs through one's neighbors (arrow c), which in turn may affect one's voting (arrow d). To the extent that this alternative mechanism produces a dependence between signal quality and voting behavior, it may cause bias in Peisakhin and Rozenas (2018)'s estimates.

While DAGs can be used to reflect general patterns of dependence, at some points it will be useful to use standardized versions of all variables (zero mean, unit variance) and to make the additional assumption that all causal relationships are linear, with e.g. $Y = aT + cU + \epsilon$ indicating the structural equation determining the outcome in the diagram at left in Figure 3 (where ϵ is a random error term, not shown on the DAG). In that case, and assuming all analysis is conducted with linear regression, we can express both the bias in the core analysis and the estimate in a placebo test as functions of linear path coefficients (Pearl 2013). For example, the regression of voting behavior on signal quality yields a coefficient of $a + bc$ given the DAG at left in Figure 3 (with bc indicating bias) and a coefficient of $ab + cd$ given the DAG at right.

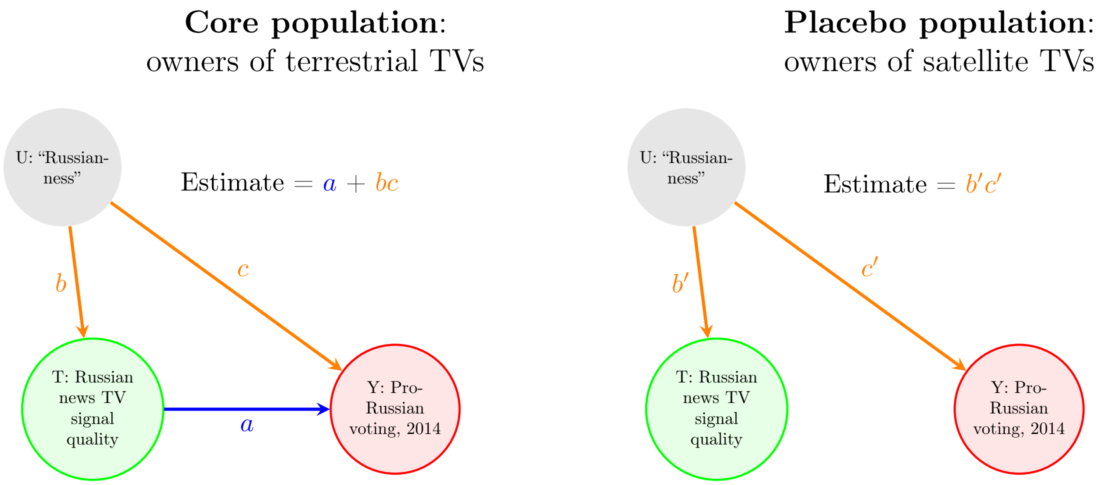
5 Placebo population tests

We now discuss the three types of placebo tests in turn, starting with the simplest.

In a placebo population test, the researcher reproduces the core analysis in an alternative population. This test is informative about bias to the extent that we believe that the postulated treatment effect does not operate in the placebo population while purported flaws would operate in a similar way. Figure 4 illustrates that logic using a placebo population test appearing in Peisakhin and Rozenas (2018).

The DAG at left in Figure 4 reproduces Peisakhin and Rozenas (2018)'s (simplified) core analysis, with possible confounding bias represented by the path from the treatment to the outcome that runs through "Russian-ness" (bc). The DAG at right shows (again in simplified form) how Peisakhin and Rozenas (2018) address this concern with a placebo population test. Peisakhin and Rozenas (2018) repeat their core analysis in the subset of survey respondents who own satellite TVs or do not watch terrestrial TV for other reasons. In this placebo

Figure 4: The logic of placebo population tests for confounding bias



population, the quality of Russian (terrestrial) TV signal should not affect consumption of Russian-slanted news; it may be, however, that "Russian-ness" would affect signal quality (as represented by path coefficient b') and pro-Russian voting (as represented by c') for satellite TV owners. Given this DAG, and assuming linearity and standardized variables, regressing Y on T in the placebo population yields in expectation $b'c'$.

It should be clear from this simple case that a placebo population test is only informative to the extent that there is a degree of similarity between the bias in the core population (here, bc) and the estimate in the placebo population (here, $b'c'$). In an ideal situation, the bias is known to be perfectly mirrored across the two populations, so that $bc = b'c'$; in that case, the placebo population test is expected to reproduce the bias in the core analysis, and we could eliminate the bias by subtracting the estimate in the placebo population test from the estimate in the core analysis. More realistically, we might suspect that $b' = kb$ for some unknown $k \neq 0$, so that e.g. "Russian-ness" is linked to signal quality among terrestrial TV owners if and only if it is linked to signal quality among satellite TV owners, though the two links may differ in degree or direction; if in addition $c = c'$ (i.e. "Russian-ness" matters in the same way for the two groups), then the expected estimate in the placebo population test is kbc, a linear function of the bias in the core analysis (though with unknown slope). In that case the test's size should be small (because the placebo test yields no association in expectation when the core analysis is unbiased); its power depends on k but also on the sample size and the amount of variation in the treatment in the placebo population.

In the case of Peisakhin and Rozenas (2018)'s placebo population test, it seems plausible that "Russian-ness" and other potential confounders would play a similar role for terrestrial TV owners and other Ukrainians. It may be, however, that satellite TV owners are richer and more mobile than terrestrial TV owners, and this could make the link between Russian TV signal and cultural "Russian-ness" (b') weaker in the placebo population. If so, the placebo population test may not be very powerful.

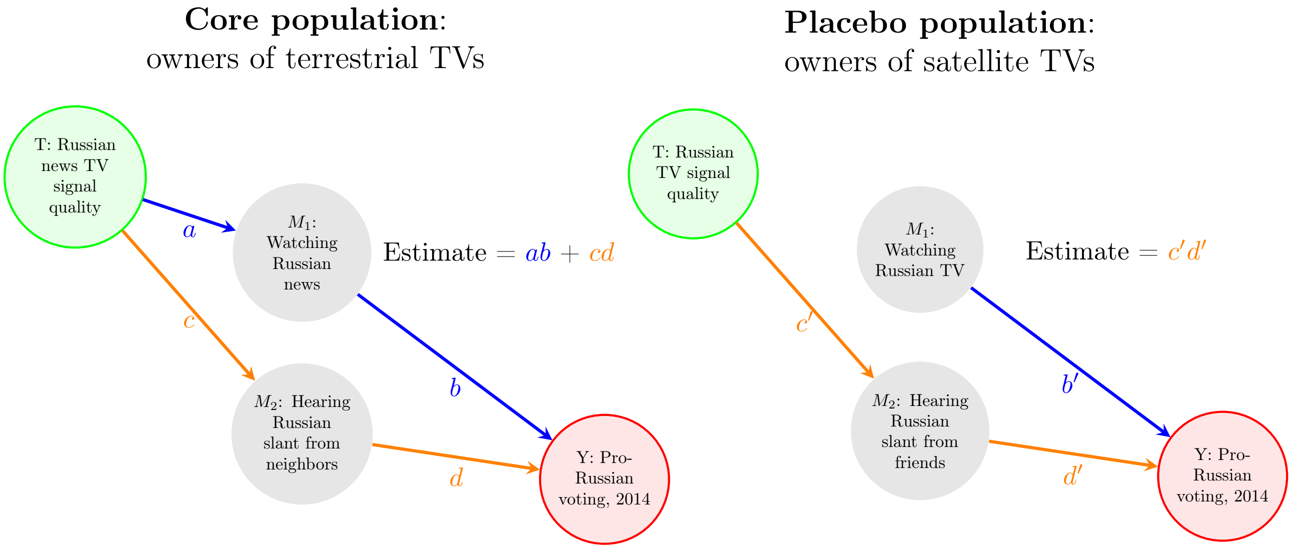
Peisakhin and Rozenas (2018)'s placebo population test can also be viewed as a test of alternative mechanisms through which the treatment might affect the outcome. Figure 5 highlights this logic.[10] The left diagram shows the concern that Russian news TV signal quality could affect voting behavior in part by exposing people to arguments their neighbors picked up from Russian TV; this would constitute a violation of the exclusion restriction in Peisakhin and Rozenas (2018)'s IV analysis. Peisakhin and Rozenas (2018)'s placebo population test helps us assess this alternative channel, because we might expect it to operate among satellite TV owners, whose own consumption of Russian news TV should not be affected by signal quality.[11]

Table 1 summarizes three more examples of placebo population tests from our survey. (Appendix A has several other examples.) The placebo population test in Acharya, Blackwell and Sen (2016*a*) is similar to Peisakhin and Rozenas (2018)'s in that it tests the exogeneity and exclusion assumptions of an instrumental variables analysis by reproducing the reduced form regression in a different population. The authors study how the proportion of enslaved people in a southern U.S. county in 1860 affects racial attitudes among white citizens in that county in recent decades. As an instrument, they use the county's suitability for growing

[10] One could also consider the treatment in Peisakhin and Rozenas (2018) to be "watching Russian news", in which case other mechanisms through which signal quality could affect the outcome are effectively confounders.

[11] Of course, if my neighbors watch Russian news because of the good signal, I might watch it on my satellite TV to keep up.

Figure 5: The logic of placebo population tests for alternative mechanisms


Core population:
owners of terrestrial TVs
T: Russian news TV signal quality
M_1: Watching Russian news
M_2: Hearing Russian slant from neighbors
Y: Pro-Russian voting, 2014
a
b
c
d
Estimate = $ab + cd$
Placebo population:
owners of satellite TVs
T: Russian TV signal quality
M_1: Watching Russian TV
M_2: Hearing Russian slant from friends
Y: Pro-Russian voting, 2014
b'
c'
d'
Estimate = $c'd'$

cotton, arguing that this would affect the concentration of enslaved people in 1860 (the treatment) and would not affect subsequent white attitudes through other channels. The authors address concerns about exogeneity and exclusion using a placebo population test that reproduces their reduced-form regression in *northern* counties, where slavery was already illegal by 1860. The logic is that confounding bias and exclusion restriction violations that might afflict the core analysis (which focuses on the South) would also likely arise in the placebo analysis (which focuses on the North), while cotton suitability could not affect the prevalence of slavery in the North; thus a significant association would raise questions about exogeneity and exclusion in the reduced form analysis. As with Peisakhin and Rozenas (2018)'s test, this placebo test could lack power if the sample from the placebo population is small (e.g. too few northern counties in the dataset) or has little variation in the treatment (e.g. cotton suitability uniformly low throughout the North), so one should at least compare the sample size and standard errors in the core analysis and placebo test before interpreting the results.[12]

[12] In Peisakhin and Rozenas (2018, online Appendix, section 11.6), the placebo population is larger than the core population and the standard errors are similar in magnitude; in Acharya, Blackwell and Sen (2016*a*, Table A.5), the placebo population is less than half as big as the core population but the standard errors are again similar in magnitude. This suggests that both tests had adequate power against bias as large as the estimated treatment effect in the core analysis.

Table 1: Examples of placebo population tests (more in Appendix A)

Paper	**Core analysis**			**Placebo population**
	Population	Treatment	Outcome	
Acharya, Blackwell and Sen (2016*b*)	White Americans living in the U.S. South	County's suitability for cotton production	Attitudes towards African-Americans today	White Americans living in the U.S. North
Chen (2013)	Households who applied for FEMA aid before Nov. 2004 election	Award of FEMA aid	Turnout in 2004 general election	Households who applied for FEMA aid after Nov. 2004 election
Fowler and Hall (2018) vis-a-vis Achen and Bartels (2017)	Counties in New Jersey in 1916	Beach counties vs. others	Support for Dem. pres. candidate in 1916	Counties in state-years without shark attacks

Fowler and Hall (2018) use placebo population tests in a way that is notable for two main reasons. First, unlike the examples above, Fowler and Hall (2018) use placebo tests to address possible flaws in someone else's research: to assess the soundness of Achen and Bartels (2017)'s finding that shark attacks in New Jersey beach counties lowered support for Wilson in 1916, Fowler and Hall (2018) replicate the basic analysis in other state-years in which no shark attacks took place.[13] (Other examples of what might be called an "adversarial placebo test" include Grimmer et al. (2018) and Kocher and Monteiro (2016) in political science, DiNardo and Pischke (1997) in economics, and Cohen-Cole and Fletcher (2008) in epidemiology.) Second, Fowler and Hall (2018)'s tests are used to assess a potential problem with Achen and Bartels (2017)'s confidence intervals, not bias; they find that they reject the null in 160 out of 593 placebo population tests (27%), which (assuming that patterns of dependence across beach and non-beach counties are similar in 1916 New Jersey and the other cases) suggests that Achen and Bartels (2017)'s confidence intervals are too small.

[13] Specifically, they compare support for the Democratic candidate in beach counties and others, controlling for past Democratic support.

6 Placebo outcome tests

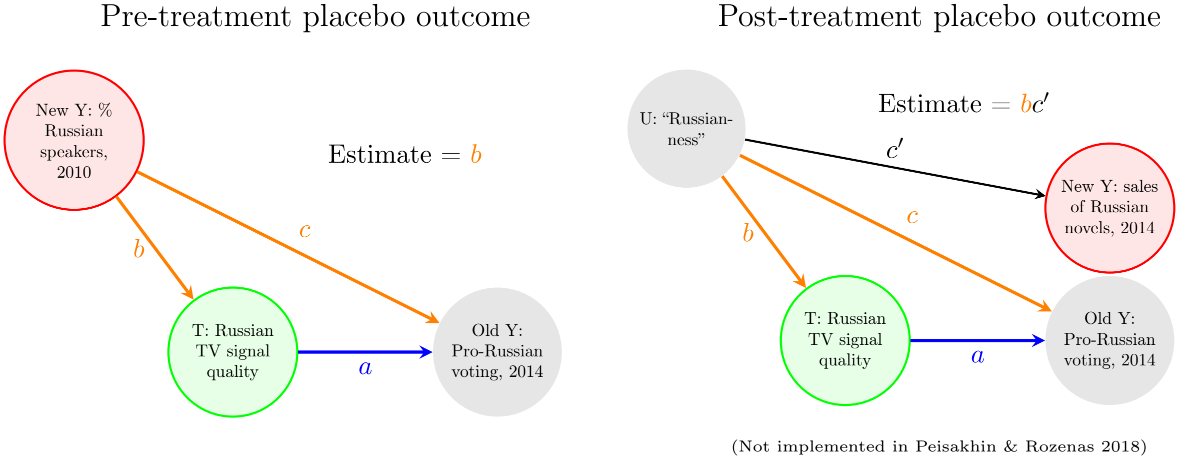
In a placebo outcome test, the researcher reproduces the core analysis with an alternative outcome variable. The logic of the test differs somewhat depending on whether the placebo outcome is a pre-treatment variable or a post-treatment variable, so we discuss the two cases separately.

6.1 Pre-treatment placebo outcomes (balance tests)

The left panel of Figure 6 illustrates a placebo outcome test using a pre-treatment variable as an outcome. (Such tests are sometimes referred to as "balance tests".) Suppose we are concerned about "Russian-ness" as a potential confounder for the relationship between Russian news TV signal quality and pro-Russian voting (as in the left panel of Figure 4). One response is to use a measure of this confounder as a placebo outcome. In their precinct-level analysis, Peisakhin and Rozenas (2018) run several such placebo outcome tests, including one where the percentage of Russian speakers in the precinct in 2010 is a placebo outcome. If places with better and worse Russian news TV signal quality differ in their inhabitants' cultural "Russian-ness", then we should find that Russian news TV signal quality appears to affect the proportion of Russian speakers in a precinct. In the DAG shown, and again assuming linearity and standardized variables, this placebo outcome test yields a coefficient of b in expectation; assuming that $c \neq 0$, the placebo outcome test thus yields 0 (in expectation) if and only if the core analysis is unbiased.

When a placebo outcome is a pre-treatment covariate, it is natural to ask why the researcher would use it as a placebo outcome rather than as a control variable. Indeed, if "% Russian speakers" were the only confounder of interest (as in the DAG at left in Figure 6), one could recover the treatment effect a (given linearity) simply by regressing the outcome on the treatment while controlling for "% Russian speakers".

Figure 6: The logic of placebo outcome tests


Pre-treatment placebo outcome
New Y: % Russian speakers, 2010
Estimate = b
b
c
T: Russian TV signal quality
a
Old Y: Pro-Russian voting, 2014
Post-treatment placebo outcome
U: "Russian-ness"
Estimate = bc'
c'
New Y: sales of Russian novels, 2014
b
c
T: Russian TV signal quality
a
Old Y: Pro-Russian voting, 2014
(Not implemented in Peisakhin & Rozenas 2018)

As the DAG shows, what we learn from a balance test using a given pre-treatment covariate X_k is closely related to what we learn from observing how the treatment effect changes when we control for X_k: in this simple case, the former yields b while the latter yields bc. The more general connection can be seen by recalling that the omitted variable bias formula (e.g. Cinelli and Hazlett 2020) expresses the change in the estimated treatment effect from controlling for X_k as the product of X_k's *impact* on the outcome (conditional on treatment and other covariates) and the *imbalance* in X_k, where the latter is precisely what the balance test measures. Thus a balance test yielding a precise null provides evidence that controlling for X_k would not affect the estimated treatment effect. On the other hand, if $c = 0$ (or the "impact" is zero in the more general OVB formulation), we could find large imbalance in a variable even though controlling for it makes no difference.

It follows that controlling for a pre-treatment covariate X_k is more informative if we want to assess bias due to X_k, but a balance test using X_k is more informative if we more specifically want to know whether X_k is balanced. (In terms of size and power, controlling for X_k lacks power as a test for imbalance in X_k, while a balance test for X_k has excessive size as a test for confounding due to X_k.) Assessing balance is the relevant goal when, due to features of the design, the researcher can plausibly assert that the treatment is *as-if random* (i.e. strongly ignorable), at least conditional on a small set of covariates that capture the treatment assignment mechanism. As-if randomness implies that any pre-treatment variable X_k that is outside the conditioning set should be balanced; thus a balance test for any such X_k can be a powerful test of as-if randomness. By contrast, controlling for X_k can only be a powerful test of as-if randomness if X_k also affects the outcome.

The choice between controlling for X_k and testing for balance in X_k mirrors the distinction between model-based inference and design-based inference as drawn by e.g. Sekhon (2009) and Dunning (2010). In model-based inference, the researcher attempts to enumerate and control for all possible confounders linking the treatment and the outcome in hopes of making the treatment conditionally independent of the potential outcomes. In design-based inference, by contrast, the researcher leverages "an experiment, a natural experiment, [or] a discontinuity" (Sekhon 2009) that makes as-if randomness plausible. In the former case the main question is whether we omitted any confounders, and the surest way to find out is to control for potential confounders. In the latter case the main question is whether as-if randomness holds, and the surest way to find out is to check for balance in pre-treatment covariates.

In cases where as-if randomness is claimed, there may be many variables that could be used in balance tests. If balance tests have not been pre-registered and one must choose which ones to conduct/report, we recommend that researchers focus on those that we might expect to be imbalanced if as-if randomness were violated in plausible ways. Thus in an RDD based on close elections, for example, as-if randomness would imply that the party of narrow election winners should be independent of all district characteristics (including e.g. the number of coffee shops), but if space is limited we should focus on characteristics like partisan control of the local electoral machinery that could be involved in departures from as-if randomness.

Table 2: Examples of post-treatment placebo outcome tests (more in Appendix A)

Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Cruz & Schneider (2017)	610 Philippines municipalities	Whether or not the municipality participated in an aid program	Number of visits to the municipality by local officials	Number of visits to the municipality by midwives
Hainmueller & Hangartner (2015)	1,400 municipalities in Switzerland, 1991-2009	Whether naturalization decisions are made by popular vote	Rate of naturalization through ordinary process	Rate of naturalization through marriage
Dube, Dube & Garcia-Ponce (2013)	Mexican municipalities located close to U.S. border, 2002-2006	Assault weapon availability from neighboring US state	Gun-related homicides	Accidents, non-gun homicides, and suicides

6.2 Post-treatment placebo outcomes

Several tests in our survey instead use post-treatment variables as placebo outcomes. The right panel of Figure 6 illustrates such a test as it might be implemented in the context of Peisakhin and Rozenas (2018):[14] to assess the possibility that more culturally Russian places got better Russian TV signal (and that this causes bias in our estimate of the effect of Russian TV signal on election outcomes), we could check for an effect of Russian news TV signal on sales of Russian novels. The test would be informative about bias due to "Russian-ness" if we believe that Russian news TV could not affect sales of Russian novels, while novel sales and voting behavior might both be affected by "Russian-ness". Table 2 provides three examples of other placebo tests using post-treatment placebo outcomes, and Appendix A contains more.

Placebo outcome tests involving post-treatment outcomes raise two issues worth discussing (both of which also apply to placebo population tests). First, unlike pre-treatment placebo outcomes, post-treatment placebo outcomes could be affected by the treatment, which raises concerns about the test's size. In each of the examples in Table 2 it is at

[14] They do not conduct this test.

least conceivable that the treatment *does* affect the placebo outcome: an aid program could attract young women to a municipality, requiring more midwife visits; changes to the naturalization procedure could affect marriage decisions or settlement decisions by international couples; assault weapon availability could affect non-gun deaths through spillovers. Thus we might expect to find some relationship between the treatment and the placebo outcome even if the core analysis is unbiased, inflating the test's size and making it less informative. Authors should help readers assess that concern by discussing possible ways the treatment could affect a post-treatment placebo outcome. A test can still be informative even when such an effect cannot be ruled out. In the cases in Table 2, it seems plausible that any effect of the treatment on the placebo outcome is smaller in magnitude than its effect on the actual outcome. (For example, assault weapon availability may affect suicides through spillovers, but that effect would likely be much smaller than the effect on gun-related homicides.) If in addition one expects that a relevant bias (if present) would be of similar magnitude in the core analysis and the placebo analysis, then finding a large effect in the core analysis and a negligible association in the placebo analysis is more consistent with the theory that the bias is small than the theory that the bias is large.

Second, unlike pre-treatment placebo outcome tests, post-treatment placebo outcome tests allow us to test alternative mechanisms by a logic similar to the one illustrated in Figure 5. If the author asserts that the treatment affects the outcome through mechanism M_1 but wonders whether mechanism M_2 might also be relevant, and if there is a placebo outcome that could not be affected by the treatment through M_1 but would be affected by the treatment through M_2 if the actual outcome is, then a placebo outcome test allows us to assess the mechanism M_2. Margalit (2013)'s placebo outcome test is one example.[15] Margalit (2013)'s core analysis finds that losing one's job is associated with increasing support for welfare spending, which he attributes to the effect of job loss on one's personal economic circumstances. He recognizes that the effect could operate through some other mechanism,

[15] In Appendix B we illustrate the logic with three additional examples from our survey.

however: perhaps people who lose their jobs spend their new free time being exposed to new forms of media, which shapes their political preferences in a more general way.[16] He therefore conducts a placebo test in which he replaces attitudes on welfare (the outcome from the core analysis) with attitudes on climate change (the placebo outcome), arguing that this placebo outcome would be affected by the alternative mechanism but not (or at least not as much) by changes in economic circumstances. The small association between job loss and climate change attitudes could then be interpreted as evidence that the effect of job loss on welfare attitudes in fact operates mainly or entirely through people's personal economic circumstances.

Note that the two points are related: it is *because* the treatment could affect the outcome in a post-treatment placebo outcome test that we can test alternative mechanisms. We must have good reason to believe that the treatment could only affect the placebo outcome in certain ways (and in that sense these tests fit Rosenbaum (1989)'s label "tests of known effects"), but without the possibility that the treatment could affect the placebo outcome, we have no way of testing *how* the treatment affects the actual outcome.

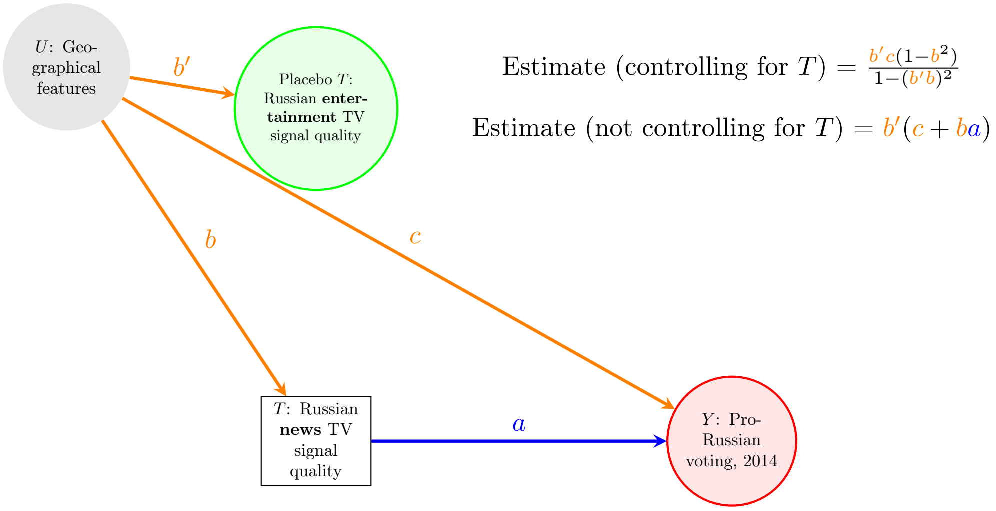
7 Placebo treatment tests

In a placebo treatment test, the researcher reproduces the core analysis with an alternative treatment variable. This test is informative about flaws in the core analysis if we believe that the placebo treatment does not affect the outcome (or does not affect it through the mechanism postulated in the core analysis) but the purported flaw would operate in a similar way.

To illustrate that logic, Figure 7 represents in simplified form a placebo treatment test appearing in Peisakhin and Rozenas (2018). According to the authors, the transmitters used

[16] Margalit (2013) does not specify an alternative mechanism, but does emphasize (p. 81) that "a change in personal material considerations, rather than a general disorientation in attitudes, accounts for the link between the experience of an economic shock and the shift in people's welfare preferences."

Figure 7: Logic of a placebo treatment test


U: Geographical features
b'
Placebo T: Russian **entertainment** TV signal quality
Estimate (controlling for T) $= \frac{b'c(1-b^2)}{1-(b'b)^2}$
Estimate (not controlling for T) $= b'(c + ba)$
b
c
T: Russian **news** TV signal quality
a
Y: Pro-Russian voting, 2014

to broadcast Russian news are different from the ones used to broadcast Russian sports and other entertainment programming; in the DAG in Figure 7, entertainment TV signal quality is affected by the same geographical features U that might confound the relationship between news TV signal quality and pro-Russian voting[17] but does not itself affect pro-Russian voting. (The key assumption is that consuming Russian-slanted news might shape political beliefs while watching e.g. Russian soccer matches would not.) In Peisakhin and Rozenas (2018)'s placebo treatment test, they regress measures of pro-Russian voting on the quality of Russian entertainment TV signal controlling for the actual treatment (Russian news TV signal quality). Given the DAG shown, the relationship between the treatment and the outcome in the placebo test reflects the dependence of Russian entertainment TV on the confounder (b') and the effect of the confounder on the outcome (c). Assuming linearity and standardized variables, the regression of Y on the placebo treatment conditional on T yields (in expectation) $b'c\frac{1-b^2}{1-(b'b)^2}$ (Pearl 2013); assuming $b' = kb$ for some $k \neq 0$, this is a rescaled version of the bias bc, so that the placebo treatment test yields an expected zero if and only if the core analysis is unbiased.

[17] For example, mountainous terrain might affect both types of TV signal and could affect voting patterns through historical migration patterns or economic links.

Table 3: Examples of placebo treatment tests (more in Appendix A)

Paper	**Core analysis**			**Placebo treatment**
	Population	Treatment	Outcome	
Jha (2013)	Towns in South Asia proximate to the coast	Whether the town was a medieval trading port	Incidence of Hindu-Muslim riots in 19th and 20th centuries	Whether the town was a colonial overseas port
Burnett and Kogan (2017)	Electoral precincts in San Diego city-wide elections in 2008 and 2010	Citizen pothole complaints before election	Incumbent electoral performance	Pothole complaints in 6 months after election
Brollo and Nannicini (2012)	Brazilian municipalities between 1997 and 2008	Partisan alignment between mayor and president (based on election RDD)	Infrastructure transfers from central government	Fake cutoffs (median margin on right and left of true threshold)

Table 3 provides three more examples of placebo treatment tests. Jha (2013)'s core analysis shows that former medieval trading ports experienced less inter-ethnic violence in 19th- and 20th-century India, which he attributes to these towns' long history of inter-ethnic cooperation in shipping. To test the idea that the difference could be explained by some other difference between port towns and other towns in South India, Jha (2013) conducts a placebo test using colonial overseas ports as the placebo treatment group; the null finding is interpreted as evidence for the author's argument that medieval ports are more peaceful because of their long history of interethnic cooperation in shipping rather than some other difference between ports and other places.

Burnett and Kogan (2017)'s placebo treatment test, like many others we found, replaces the actual treatment with a future version of the treatment. They study the effect of citizen complaints about potholes on the local incumbent's electoral performance; their placebo test uses pothole complaints *after* the election as a placebo treatment. Because future pothole complaints cannot affect current electoral outcomes, Burnett and Kogan (2017) can be assured that the result of the placebo test does not reflect the direct effect of the placebo treatment on the outcome. It seems reasonable that future pothole complaints and current pothole complaints share some causes that are potential confounders in the core analysis: perhaps pothole complaints are more common in places where citizens are generally more engaged in local politics, and maybe these citizens respond to incumbents differently for other reasons. Thus a relationship between future pothole complaints and current responses to incumbents would suggest that these persistent confounders afflict Burnett and Kogan (2017)'s core finding.

An important design question for researchers using placebo treatment tests is whether to control for the actual treatment or not. Our survey revealed considerable variation. As noted above, Peisakhin and Rozenas (2018)'s placebo treatment test controls for the actual treatment: they regress pro-Russian voting on entertainment TV signal quality, news TV signal quality, and their usual covariates. Jha (2013) does not control for medieval trading port status in his placebo treatment test. Burnett and Kogan (2017) control for current pothole complaints in their placebo treatment test, but they do so apologetically, noting that one would not do so in what they call a "traditional placebo test". Sexton (2016) and Stasavage (2014) similarly conduct placebo tests where they replace the treatment with the future treatment, but Sexton (2016) includes the actual treatment while Stasavage (2014) does not. Dasgupta, Gawande and Kapur (2017) and Fouirnaies and Mutlu-Eren (2015) both include lags and leads of the treatment in a two-way fixed effects model, but Dasgupta, Gawande and Kapur (2017) include all lags and leads in the same model (thus controlling for actual treatment) while Fouirnaies and Mutlu-Eren (2015) do not. Relatedly, Potoski and Urbatsch (2017) and Montgomery and Nyhan (2017) consider running placebo tests where they replace the treatment with the future value of the treatment, but justify not reporting the results on the grounds that the future value of the treatment is too closely related to the actual treatment; it is unclear whether they considered addressing this by including the actual treatment as a control.

To gain insight into whether a placebo treatment test should control for the actual treatment, consider again the DAG in Figure 7. As noted above, the placebo test in which we control for treatment (which we shall call the Conditional Placebo Treatment Test, CPTT) yields an expected coefficient of $b'c\frac{1-b^2}{1-(b'b)^2}$ given linearity and standardized variables; assuming $b' = kb$ for $k \neq 0$, the expected result of the CPTT is thus a linear function of the bias in the core analysis. Given linearity and standardized variables, the placebo test in which we do not control for treatment (the Unconditional Placebo Treatment Test, or UPTT) yields an expected coefficient of $b'(c + ba)$, which (again assuming $b' = kb$ for $k \neq 0$) is the sum of a linear function of the bias in the core analysis and the treatment effect in the core analysis. Which one should we prefer?

The clear disadvantage of the UPTT in this circumstance is that it combines a function of the bias in the core analysis with a function of the treatment effect in the core analysis, which suggests the test's size may be inflated: if the placebo treatment is found to be related to the outcome, it may be because the actual treatment affects the outcome and is related (through the confounder) to the placebo treatment, not because there is any bias in the core analysis. (For example, places with better Russian entertainment TV signal may see more pro-Russian voting simply because they also have better Russian news TV signal.) It also raises the risk of low power because the two components of the UPTT estimate could cancel out: supposing $a = \frac{-c}{b}$, the UPTT yields a 0 regardless of the bias in the core analysis. These considerations suggest that the UPTT may not be very informative.

The disadvantages of the UPTT may not be as serious as they first appear, however. Note that "canceling out" can only occur in the UPTT when there is a non-zero treatment effect (if canceling out is exact, when $a = \frac{-c}{b}$). If we find a non-zero treatment effect in the core analysis and a precisely estimated zero in the UPTT, we know (assuming the DAG above, plus linearity and standardized variables and $b' = kb$ with $k \neq 0$) that the true treatment effect is not zero: either the core analysis is unbiased (because $b = 0$ or $c = 0$) or there is canceling out (which requires $a = \frac{-c}{b}$). It follows that a UPTT can provide an informative rebuttal to the claim that the treatment has no effect on the outcome in the core analysis, though this requires strong assumptions.

Furthermore, canceling out can only occur when the treatment effect is much larger in magnitude than the confounding bias. To see this, note that canceling out occurs in the simple example above when the ratio of treatment effect to the bias ($\frac{a}{bc}$) is $\frac{-1}{b^2}$. If the confounder completely determines the treatment ($b = 1$, given standardized variables), then canceling out occurs in the UPTT when the bias and the treatment effect are of the same magnitude and opposite signs. In practice, b is likely to be well below 1 (Ding and Miratrix 2015), in which case the treatment effect must be much larger than the bias in magnitude. For example, if $b = .25$ (a strong effect of the confounder on the treatment), the treatment effect must be 16 times as large as the bias due to U, and opposite in sign, for canceling out to occur. This suggests that (given the assumptions above) a precise zero estimate in the UPTT is inconsistent with relatively large bias in the core analysis.

The main disadvantage of the CPTT is its power. In many cases we may believe that the placebo treatment and the actual treatment are closely related to each other. In Peisakhin and Rozenas (2018), for example, news TV signal and entertainment TV signal may be affected by similar geographical and technical features; in Burnett and Kogan (2017) pothole complaints may be highly correlated over time because road conditions and citizen characteristics are persistent. In such cases, there may not be much independent variation in the placebo treatment once we control for the actual treatment, with the result that the CPTT has low power. To see this in the simple linear case above, note that the CPTT estimate ($b'c\frac{1-b^2}{1-(b'b)^2}$) becomes more attenuated as b increases. In Appendix C we illustrate this problem with the CPTT in a simulation exercise: we show that, given the DAG above but assuming the true treatment effect is zero (which eliminates both the false positives problem and the canceling out problem in the UPTT), the CPTT is less powerful than the UPTT across a range of values for the bias in the core analysis.

Appendix C also highlights some previously unnoticed properties of the CPTT that arise when there is more than one confounder; we briefly summarize those properties here. Conditioning on the actual treatment induces collider dependence between the placebo treatment and other causes of the treatment, including confounders that are otherwise unrelated to the placebo treatment. This means that the result of the CPTT in principle reflects *all* confounding bias in the core analysis, not just bias due to a confounder that is shared between the placebo treatment and the actual treatment as in Figure 7 above. Unfortunately, these additional biases enter negatively into the CPTT while the original bias enters positively, making it difficult to interpret the result. This concern disappears in the special case where the only path from the placebo treatment to the outcome runs through the actual treatment, but in that case the placebo treatment is a valid instrument, and IV analysis would be a more straightforward means of addressing possible bias. Moreover, any path from the placebo treatment to the outcome that opens due to conditioning on the treatment runs through at least four edges of the DAG, which suggests that the CPTT will struggle to detect bias due to these additional confounders.

In sum, there is no simple answer to the question of whether to control for the actual treatment in a placebo treatment test: not controlling for the actual treatment makes the UPTT contaminated by the treatment effect, while controlling for the actual treatment effect makes the CPTT attenuated; both problems are worse when the placebo treatment and the actual treatment are more highly correlated. This suggests seeking a placebo treatment that is less closely connected to the actual treatment and, when that is not possible, focusing on other types of tests.

8 Placebo tests for specific research designs

In this section we apply insights from the previous sections to discuss appropriate placebo tests for specific research designs.

8.1 Regression discontinuity designs

The key identifying assumption in a sharp RDD is the continuity of the conditional expectation function (CEF) for the potential outcomes in the neighborhood of the threshold: if this holds, and if we can obtain a consistent estimate of the CEF at the threshold from both above and below the threshold, then the difference in these two estimates is a consistent estimator of the local average treatment effect.

Balance tests (i.e. placebo tests that use pre-treatment variables as placebo outcomes) provide an indirect way to test this identifying assumption. In a sharp RDD we observe each potential outcome on only one side of the threshold, so we cannot test the continuity assumption. We can, however, observe any pre-treatment variable on both sides of the threshold, and finding a discontinuity in one of these variables across the threshold would cast doubt that the potential outcomes are continuous across the threshold. Accordingly, RDD studies commonly perform several balance tests to assess the continuity assumption. As noted above, when space is limited these tests should be chosen to speak to a relevant threat to the continuity assumption.

In addition to balance tests, RDD studies often include placebo treatment tests in which the actual cutoff is replaced with one or more "fake cutoffs" where the treatment does not actually change,[18] as recommended by influential how-to guides (e.g. Imbens and Lemieux 2008; Cattaneo, Idrobo and Titiunik 2020). Fake-cutoff tests are described by Cattaneo, Idrobo and Titiunik as (indirect) tests of the continuity assumption: if we find jumps in the CEF at arbitrary points *away* from the threshold, we may doubt the plausibility of the continuity assumption at the threshold and therefore the reliability of our estimates (p. 89).

While there is certainly no harm in checking for discontinuities at arbitrary thresholds as part of an RDD study, we point out that such tests are not very informative about the continuity assumption. Fundamentally, the problem is that there no clear reason to think

[18] Examples include Boas, Hidalgo and Richardson (2014); Brollo and Nannicini (2012); Dinas (2014); Eggers and Hainmueller (2009); Ferwerda and Miller (2014); Hall (2015); Holbein and Hillygus (2016); Hopkins (2011).

that continuity is more likely to fail at arbitrary cutoffs away from the threshold when it fails at the threshold than otherwise. Typically the most relevant concern about the continuity assumption is that the desire to receive or avoid the treatment causes agents to attempt to sort into or out of treatment, leading either to a discontinuity in the potential outcomes or to so much local non-linearity in the CEF that (due to mis-specification) we mistakenly detect a discontinuity; in some cases we may also be worried that the main treatment is bundled with another treatment. Clearly, a placebo test at an arbitrary cutoff does not address these threats to the continuity assumption: units would have no reason to sort into or out of treatment at an arbitrary threshold, nor is there any reason to expect a bundled treatment to apply there. A fake cutoff placebo test would be more informative if we suspected the CEF to be characterized by occasional discontinuities, one of which happened to fall at the true cutoff, but this seems a remote possibility; even if it were true, the fake-cutoff placebo test would have inflated size (because there may be discontinuities at the fake cutoffs when there is none at the true threshold) and low power (because there may not be discontinuities at the fake cutoffs when there happens to be one at the true threshold).

A better justification for fake-cutoff RDDs is that they allow us to check our confidence intervals by determining how often we would find a significant "effect" when we know that none is present. For this purpose authors should check more fake cutoffs than is typical. Also, the results could be misleading if the curvature of the CEF is different near the threshold than further away or if, as is common, the tests use a more restricted sample than the core analysis.

8.2 Instrumental variables

In IV analysis, the *exogeneity* assumption implies that there are no confounders in either the first stage (i.e. the regression of the treatment on the instrument) or the reduced form (i.e. the regression of the outcome on the instrument); the *exclusion* restriction holds that the instrument affects the outcome only through the treatment. Both assumptions can be probed by adapting placebo tests discussed above.

The most straightforward placebo tests for IV designs are balance tests that check the exogeneity of the instrument. These balance tests are identical to the standard case discussed above, except that the instrument takes the role of the key causal factor (the "treatment", in the standard case). Thus the exogeneity assumption can be probed by checking for "effects" of the instrument on pre-instrument variables (which are not themselves causes of the instrument). Meredith (2013)'s study of coat-tail effects includes a balance test of this type.

The most straightforward way to test the exclusion restriction is a placebo population test that reproduces the reduced form analysis in a population in which the instrument could not affect the treatment. We considered two examples above (Peisakhin and Rozenas 2018; Acharya, Blackwell and Sen 2016*b*); Rozenas, Schutte and Zhukov (2017) is another. These tests are also informative about exogeneity to the extent that we think that possible confounding is mirrored in the placebo population and the core population. In some cases the assumption that the instrument would not affect the treatment in the placebo population can be tested (e.g. Peisakhin and Rozenas 2018). Of course, there is no guarantee that a suitable placebo population will be available in any particular case.

Some researchers similarly use post-treatment placebo outcomes to test exclusion and/or exogeneity, replacing the outcome in the reduced-form regression with an outcome that should not be affected by the treatment but might be affected by the instrument through alternative channels and by confounders that would also operate in the core analysis. Laitin and Ramachandran (2016) is an example that highlights the logic of this approach while also revealing some difficulties. The authors aim to study how a country's economic development is affected by adopting an official language that is more linguistically distinct from its native language. As an instrument they use the spatial distance from the country to one of the sites where writing was developed. Because spatial distance from writing's birthplace is likely related to many background determinants of development and could affect development through other channels, it is natural to worry about both both exogeneity and exclusion. In response, Laitin and Ramachandran (2016) repeat the reduced form analysis using measures of state capacity as the outcome. To the extent that state capacity would be affected by some of the same confounders and alternative mechanisms that might plague the core analysis, a significant association in this placebo outcome test could cast doubt on the core results. The problem is that the treatment (adopting an official language distinct from one's native language) could also affect state capacity, perhaps even more than it affects the actual outcome. In that sense Laitin and Ramachandran (2016)'s placebo test might be informative about which outcomes the official language affect (given the validity of the IV assumptions), but it provides weak evidence for the validity of the IV assumptions.

8.3 Diff-in-diff and panel studies

In studies where treatment occurs at a point in time (which may vary across units) and data is collected for periods before and after treatment assignment, researchers often design placebo tests that check whether the treatment appears to affect previous values of the outcome. Usually this placebo test is carried out as a placebo outcome test where the placebo outcome is the lagged value of the outcome variable. One common case is the parallel trends test in a classic two-period, two-group diff-in-diff (e.g. Bechtel and Hainmueller 2011): although it is not often presented in this way, the test can be viewed as a placebo test using the lagged outcome as the placebo outcome.

In some cases researchers instead check whether the outcome appears to be affected by future values of the treatment. (We cited examples above when considering whether to control for the actual treatment in a placebo treatment test.) In two-way fixed effects studies, Angrist and Pischke (2008, pg. 237) recommend re-running the core analysis while including a series of lags and leads of the treatment, which can be seen as a placebo treatment test with several placebo treatments and the actual treatment. Dasgupta, Gawande and Kapur

(2017) is an example: in a study of the effect of an anti-poverty program on Maoist violence in India, the authors regress violence in quarter t on indicators for whether the program was implemented in quarter $t+8, t+7, \ldots, t+1, t, t-1, \ldots, t-7$, showing that violence was lower (conditional on covariates, and all other lags and leads of treatment) when the program had been implemented in the past (e.g. in quarter $t-6$) but not when it was due to be implemented in the future (e.g. in quarter $t+6$).[19] Similarly, Kuziemko and Werker (2006)'s study of the effect of a U.N. Security Council (UNSC) seat on a country's foreign aid receipts includes a regression of aid receipts on indicators for whether the country was going to be elected to a seat in the next year, elected in that year, elected the previous year, elected two years previously, etc. In both cases, insignificant coefficients on the *leads* of treatment are taken as evidence that, conditional on covariates, treated and untreated units were similar before treatment was assigned, which suggests that selection bias does not account for the apparent effect of treatment in the core analysis. As discussed when we considered whether to control for the actual treatment in placebo treatment tests, these tests may be underpowered, but testing one lead at a time would likely have excessive size.

9 Conclusion

We have offered a framework for assessing the informativeness of placebo tests and used it to discuss the design and interpretation of three main test types (placebo population tests, placebo outcome tests, and placebo treatment tests). Table 4 offers a summary of the main points researchers should consider for each type of placebo test.

Although we have focused on ways in which placebo tests can improve interpretation of specific studies, the wider and better use of placebo tests could also help combat p-hacking

[19] Fouirnaies and Mutlu-Eren (2015) presents similar results in which the outcome at t is regressed on the treatment at $t+3, t+2, \ldots, t-6$, but unlike Dasgupta, Gawande and Kapur (2017) they do this in separate regressions. It is somewhat surprising that they find no effect of treatment at e.g. $t+1$ in these regressions (not controlling for treatment at t): given the persistence of treatment over time, places due to be treated in the next period were probably also more likely to be treated in this period.

Table 4: Main considerations in the design and interpretation of placebo tests

Type of test	Feature	Relevant considerations
Placebo population tests	Size	• Could the treatment affect the outcome in this population? • Does moving the analysis to this new population introduce new flaws, e.g. biases or estimation difficulties?
	Power	• Explain why we should expect flaws that might operate in the original population to operate similarly in the placebo population. • Is the placebo population large enough to produce precise estimates? • Do the treatment and outcome vary sufficiently in the placebo population to produce precise estimates?
Placebo outcome tests w. pre-treatment variables	Power	• If choosing which outcomes to test, choose based on which outcomes might be imbalanced given plausible departures from as-if randomness.
Placebo outcome tests w. post-treatment variables	Size	• Could the placebo outcome be affected by the treatment? If testing alternative mechanisms, could the placebo outcome be affected by the treatment through the main postulated mechanism?
	Power	• Explain why the possible flaw in the core analysis would apply to this outcome – e.g. a confounding variable or an alternative mechanism.
Placebo treatment tests	Size	• Could the placebo treatment affect the outcome? • (If not controlling for treatment, i.e. UPTT:) Could the placebo treatment be associated with the outcome through the treatment?
	Power	• Explain why the possible flaw in the core analysis would apply to this treatment. • (If controlling for treatment, i.e. CPTT) Are the original and placebo treatment so closely related that you could not detect even substantial bias?

and related systemic problems in social science research (e.g. Humphreys, De la Sierra and Van der Windt 2013). If researchers are consciously or unconsciously running through many specifications to find one that produces an appealing result, then the expectation to present satisfactory placebo tests constitutes an extra hurdle that is more likely to be cleared by true discoveries than by spurious findings.

Of course, placebo tests themselves are subject to some of the same pressures that lead to *p*-hacking. Researchers who carry out placebo tests for others' designs face the usual incentive to find statistically significant results, with the same possible pitfalls. Researchers who present placebo tests for their own designs (the much more common case, currently) face the opposite incentive, which may push them toward finding insignificant results or "null-hacking" (Graham et al. 2019). Moreover, in some cases a researcher could conduct a piece of analysis first and then decide later whether it belongs with the core analysis (if the results are significant) or with the placebo tests (if not), altering the causal theory accordingly; the placebo tests section of paper then becomes simply a repository of null results rather than a place to test a research design.

Addressing these problems (like addressing *p*-hacking and related problems more generally) requires effort on several fronts. Researchers can limit the possibility of "hacking" (intentional or not) by including placebo tests in their pre-analysis plans. Editors and referees can try to improve researchers' incentives by placing more emphasis on research design (including the design of placebo tests) and less emphasis on results. Finally, the research community can establish clearer expectations about which placebo tests should be included, how they should be designed, and how they should be evaluated, a project to which we hope this paper contributes.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105(490).

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2015. "Comparative politics and the synthetic control method." *American Journal of Political Science* 59(2):495–510.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016*a*. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016*b*. "The political legacy of American slavery." *The Journal of Politics* 78(3):621–641.

Achen, Christopher H and Larry M Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government.* Vol. 4 Princeton University Press.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Bechtel, Michael M and Jens Hainmueller. 2011. "How Lasting Is Voter Gratitude? An Analysis of the Short-and Long-Term Electoral Returns to Beneficial Policy." *American Journal of Political Science* 55(4):852–868.

Berry, Christopher R and Anthony Fowler. 2021. "Leadership or luck? Randomization inference for leader effects in politics, business, and sports." *Science Advances* 7(4).

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics* 119(1):249–275.

Boas, Taylor C, F Daniel Hidalgo and Neal P Richardson. 2014. "The spoils of victory: campaign donations and government contracts in Brazil." *The Journal of Politics* 76(2):415–429.

Brollo, Fernanda and Tommaso Nannicini. 2012. "Tying your enemy's hands in close races: The politics of federal transfers in Brazil." *American Political Science Review* 106(04):742–761.

Burnett, Craig M and Vladimir Kogan. 2017. "The politics of potholes: Service quality and retrospective voting in local elections." *The Journal of Politics* 79(1):302–314.

Cattaneo, Matias D, Nicolás Idrobo and Rocío Titiunik. 2020. *A practical introduction to regression discontinuity designs: Foundations.* Cambridge University Press.

Chen, Jowei. 2013. "Voter partisanship and the effect of distributive spending on political participation." *American Journal of Political Science* 57(1):200–217.

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

Cochran, William G and S Paul Chambers. 1965. “The planning of observational studies of human populations.” *Journal of the Royal Statistical Society. Series A (General)* 128(2):234–266.

Cohen-Cole, Ethan and Jason M Fletcher. 2008. “Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis.” *Bmj* 337.

Dasgupta, Aditya, Kishore Gawande and Devesh Kapur. 2017. “(When) do antipoverty programs reduce violence? India’s rural employment guarantee and Maoist conflict.” *International organization* 71(3):605–632.

De Craen, Anton JM, Ted J Kaptchuk, Jan GP Tijssen and Jos Kleijnen. 1999. “Placebos and placebo effects in medicine: historical overview.” *Journal of the Royal Society of Medicine* 92(10):511–515.

DiNardo, John E and Jörn-Steffen Pischke. 1997. “The returns to computer use revisited: Have pencils changed the wage structure too?” *The Quarterly Journal of Economics* 112(1):291–303.

Dinas, Elias. 2014. “Does choice bring loyalty? Electoral participation and the development of party identification.” *American Journal of Political Science* 58(2):449–465.

Ding, Peng and Luke W Miratrix. 2015. “To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias.” *Journal of Causal Inference* 3(1):41–57.

Dunning, Thad. 2010. “Design-based inference: Beyond the pitfalls of regression analysis?” *Rethinking social inquiry: Diverse tools, shared standards* pp. 273–311.

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

Eggers, Andrew C and Jens Hainmueller. 2009. “MPs for sale? Returns to office in postwar British politics.” *American Political Science Review* 103(4):513–533.

Ferwerda, Jeremy and Nicholas L Miller. 2014. “Political devolution and resistance to foreign rule: A natural experiment.” *American Political Science Review* 108(03):642–660.

Fouirnaies, Alexander and Hande Mutlu-Eren. 2015. “English bacon: Copartisan bias in intergovernmental grant allocation in England.” *The Journal of Politics* 77(3):805–817.

Fowler, Anthony and Andrew B Hall. 2018. “Do shark attacks influence presidential elections? Reassessing a prominent finding on voter competence.” *The Journal of Politics* 80(4):1423–1437.

Gilardi, Fabrizio. 2015. “The temporary importance of role models for women’s political representation.” *American Journal of Political Science* 59(4):957–970.

Graham, Matthew H, Gregory A Huber, Cecilia Hyunjung Mo et al. 2019. "Observational Open Science: An Application to the Literature on Irrelevant Events and Voting Behavior.".

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

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(1):18–42.

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

Healy, Andrew and Gabriel S Lenz. 2017. "Presidential voting and the local economy: Evidence from two population-based data sets." *The Journal of Politics* 79(4):1419–1432.

Holbein, John B and D Sunshine Hillygus. 2016. "Making young voters: the impact of preregistration on youth turnout." *American Journal of Political Science* 60(2):364–382.

Hopkins, Daniel J. 2011. "Translating into Votes: The Electoral Impacts of Spanish-Language Ballots." *American Journal of Political Science* 55(4):814–830.

Humphreys, Macartan, Raul Sanchez De la Sierra and Peter Van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* 21(1):1–20.

Imbens, Guido W. and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2):615–635.

Jha, Saumitra. 2013. "Trade, institutions, and ethnic tolerance: Evidence from South Asia." *American Political Science Review* 107(04):806–832.

Kocher, Matthew A and Nuno P Monteiro. 2016. "Lines of demarcation: Causation, design-based inference, and historical research." *Perspectives on Politics* 14(4):952.

Kuziemko, Ilyana and Eric Werker. 2006. "How much is a seat on the Security Council worth? Foreign aid and bribery at the United Nations." *Journal of political economy* 114(5):905–930.

Laitin, David D and Rajesh Ramachandran. 2016. "Language policy and human development." *American Political Science Review* 110(3):457–480.

Lipsitch, Marc, Eric Tchetgen Tchetgen and Ted Cohen. 2010. "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies." *Epidemiology* 21(3):383–388.

Margalit, Yotam. 2013. "Explaining social policy preferences: Evidence from the Great Recession." *American Political Science Review* 107(01):80–103.

Meredith, Marc. 2013. "Exploiting friends-and-neighbors to estimate coattail effects." *American Political Science Review* 107(04):742–765.

Montgomery, Jacob M and Brendan Nyhan. 2017. "The effects of congressional staff networks in the us house of representatives." *The Journal of Politics* 79(3):745–761.

Neumayer, Eric and Thomas Plümper. 2017. *Robustness tests for quantitative research.* Cambridge University Press.

Pearl, Judea. 2013. "Linear models: A useful "microscope" for causal analysis." *Journal of Causal Inference* 1(1):155–170.

Peisakhin, Leonid and Arturas Rozenas. 2018. "Electoral effects of biased media: Russian television in Ukraine." *American Journal of Political Science* 62(3):535–550.

Pizer, Steven D. 2016. "Falsification testing of instrumental variables methods for comparative effectiveness research." *Health services research* 51(2):790–811.

Potoski, Matthew and R Urbatsch. 2017. "Entertainment and the Opportunity Cost of Civic Participation: Monday Night Football Game Quality Suppresses Turnout in US Elections." *The Journal of Politics* 79(2):424–438.

Rosenbaum, Paul R. 1984. "From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment." *Journal of the American Statistical Association* 79(385):41–48.

Rosenbaum, Paul R. 1989. "The Role of Known Effects in Observational Studies." *Biometrics* 45:557–569.

Rosenbaum, Paul R. 2002. *Observational studies.* Springer.

Rosenbaum, Paul R. 2010. *Design of Observational Studies.* Springer.

Rosenbaum, Paul R and Donald B Rubin. 1983. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society: Series B (Methodological)* 45(2):212–218.

Royall, Richard. 1997. *Statistical evidence: a likelihood paradigm.* Routledge.

Rozenas, Arturas, Sebastian Schutte and Yuri Zhukov. 2017. "The political legacy of violence: The long-term impact of Stalin's repression in Ukraine." *The Journal of Politics* 79(4):1147–1161.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12:487–508.

Sexton, Renard. 2016. "Aid as a tool against insurgency: Evidence from contested and controlled territory in Afghanistan." *American Political Science Review* 110(4):731–749.

Stasavage, David. 2014. “Was Weber Right? The Role of Urban Autonomy in Europe’s Rise.” *American Political Science Review* 108(02):337–354.

Thachil, Tariq. 2014. “Elite parties and poor voters: Theory and evidence from India.” *American Political Science Review* 108(2):454–477.

Appendix A: Placebo tests in political science

This appendix includes the survey of placebo tests that served as the empirical basis of the paper. It is a nearly exhaustive list of every observational study mentioning a "placebo test'," "balance test," or "falsification test" in the *APSR*, *AJPS*, *JOP*, and *IO* between 2009 and 2018. The main exception is that we have only included a small sample of the simplest types of placebo tests (balance tests and fake-cutoff tests from RDD studies). We have also omitted two cases that did not neatly fit our typology.[20] Papers that have more than one type of placebo test were included under all relevant types; within types, we only include one test per paper.

We identified a total of 110 of placebo tests, including 64 placebo outcome tests, 34 placebo treatment tests, and 12 placebo population tests. To summarize each test, we report the population, treatment, and outcome used in the core analysis followed by the alteration made in the placebo (e.g. the placebo outcome, for placebo outcome tests).

[20] These are model 6 in Thachil (2014), which puts the treatment on the left hand side, and Gilardi (2015), which changes both the treatment and the outcome (in a manner similar to Cohen-Cole and Fletcher (2008)).

Placebo Outcome Tests

Paper	Core analysis			Placebo outcome
	Population	Treatment	Outcome	
Alexander, Berry and Howell (2016)	US counties, 1984-2010	Ideological distance between county's congressional representative and chamber median	Federal outlays of non-formula grants to county in given year	Formula-based grants; direct disability and retirement payments
Alt, Marshall and Lassen (2016)	6000 Danish survey respondents	Unemployment expectations (instrumented by information provided experimentally)	Intention to vote for (left-wing) government parties; trust in government	Intention to vote for left-wing parties not in gov't; preferences on redistribution
Arceneaux et al. (2016)	Roll call votes by members of US Congress, 1997-2002	The presence of Fox News in a member's district	Voting with one's party in a partisan vote in Congress	Votes in period before Fox News introduction (1991-1996)
Ariga (2015)	Candidates in Japanese elections between 1958 and 1993	Winning (close) election	Election results in subsequent election	Pre-treatment covariates
Bateson (2012)	Survey respondents from 70 countries across the world	Past crime victimization	Political participation, variously measured	Past turnout
Bayer and Urpelainen (2016)	Countries (up to 112) in the period 1990-2012	Having democratic political institutions	Adoption of feed-in tariffs (FIT) to combat climate change	Adoption of less politically attractive policies[21]
Bechtel and Hainmueller (2011)	German electoral districts	Being affected by the 2002 Elbe River floods	SPD's proportional representation vote share in a given district	Lagged dependent variable (parallel trends test)
Bhavnani and Lee (2018)	Indian political districts	Local bureaucrats being "embedded" (i.e. from the district)	The proportion of villages with high schools	The number of landline phones
Boas and Hidalgo (2011)	Brazilian city council candidates in 2000 & 2004 who applied for a radio license	Winning a city council election	Success of applications filed after the election	Success of applications decided before the election[22]
Boas, Hidalgo and Richardson (2014)	Brazilian federal deputy candidates in the 2006 election	Winning election (narrowly)	Government contracts for the candidate's donor firms	Gov't contracts for firms that donated only to other candidates

[21] The adoption rate of these policies is low, resulting in low nominal power. In one test the coefficient estimate is larger than the treatment effect in the core analysis, but the test "passes" because the standard error is large.

[22] The population is also shifted to include *all* city council candidates. The low rate of applications decided before the election may limit the nominal power.

Placebo outcome tests - continued from previous page

Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Braun (2016)	Registered Jews in 439 municipalities in the Netherlands in 1941	Proximity to minority churches (instrumented by distance to Delft, the base of influential 17th century Catholic vicar/missionary)	Evasion of deportation during WWII	Pre-treatment municipality covariates
Brollo and Nannicini (2012)	Brazilian municipalities between 1997 and 2008	Partisan alignment between the mayor and the president (via RDD)	Infrastructure transfers from central government	Formula-based transfers
Chaudoin (2014)	Tariffs imposed by the US in response to anti-dumping petitions by US firms	Domestic factors in US theorized to affect support for free trade (unemployment, election year)	Initiation of WTO dispute by country targeted by tariff	Unilateral removal of tariff[23]
Clinton and Enamorado (2014)	Members of Congress (US)	Entrance of Fox News in congressional district	Change in "presidential support score" (roll call voting agreement w. president)	Previous change in presidential support score
Cooper, Kim and Urpelainen (2018)	Roll-call votes on environmental issues by northeastern U.S. House reps in 2003/4 (pre-shale boom) & 2010/11 (post-shale boom)	The presence of shale gas resources interacted with post-shale boom dummy (diff-in-diff)	Casting a pro-environmental vote	District characteristics in two pre-shale boom periods [24]
Cox, Fiva and Smith (2016)	Norwegian electoral districts before and after the election reform of 1919	Winning margin in election before reform	Turnout change between election before reform to election after reform	Turnout change between elections that both took place before/after the reform
Cruz and Schneider (2017)	610 Philippines municipalities	Whether or not the municipality participated in the KALAHI aid program	Number of visits to the municipality by local officials	Number of visits to the municipality by midwives

[23] Unilateral removal and initiation of WTO dispute are "competing risks", which implies that the placebo analysis retains the treatment effect from the core analysis (reversed in sign).

[24] The balance tests compare changes in district characteristics for shale and non-shale districts between 2000 and 2005, which differs from the main diff-in-diff estimation strategy in form but is similar in spirit.

Placebo outcome tests - continued from previous page

Paper	Core analysis			Placebo outcome
	Population	Treatment	Outcome	
Dasgupta, Gawande and Kapur (2017)	Districts in Indian states where most Maoist conflict occurs	National Rural Employment Guarantee Scheme (NREGS) adopted in a district	Maoist conflict violence, measured in terms of violent incidents and deaths	Pre-treatment covariates
De Kadt and Larreguy (2018)	Wards (political unit) in and just outside of all Bantustans in South Africa	Alignment between local chief and ANC candidate (changes with the election of Jacob Zuma in 2006)	ANC vote share	Lagged dependent variable
de Benedictis-Kessner (2018)	Candidates in mayoral elections in U.S. cities, 1950-2014	Whether the candidate wins at time t (via RDD)	Whether the candidate runs at time $t+1$, whether the candidate wins at time $t+1$	Lagged dependent variable(s)
Dube, Dube and García-Ponce (2013)	Mexican municipalities located close to U.S. border, 2002-2006	Assault weapon availability from neighboring US state (federal ban expires in 2004 but does not affect CA)	Gun-related homicides	Accidents, non-gun homicides, and suicides
Egan and Mullin (2012)	U.S. citizens	Local temperature	Belief in climate change	Assessment of the decision to invade Iraq; assessment of George W. Bush's presidency
Eggers and Hainmueller (2009)	Candidates to the British House of Commons	Winning office	Wealth at death	Pre-treatment covariates (e.g. education)
Feigenbaum and Hall (2015)	U.S. House members, 1990-2010	District's exposure to Chinese imports	Voting on trade bills	Voting on other bills
Folke, Hirano and Snyder (2011)	U.S. states, 1885-1995	State's adoption of civil service reforms	Party control of legislature and statewide offices	Lagged dependent variable(s)
Folke and Snyder (2012)	U.S. states, 1882-2010	Election of a Democratic Governor at time t	Change in proportion of seats held by Democrats, t to $t+1$	Lagged dependent variable
Fouirnaies and Hall (2018)	Members of U.S. Congress	Being a member of specific congressional committee	Campaign contributions from donors affected by committee	Campaign contributions from other donors

Placebo outcome tests - continued from previous page

Paper	Core analysis			Placebo outcome
	Population	Treatment	Outcome	
Fukumoto and Horiuchi (2011)	Japanese municipalities in 2003	Municipal election in 2003	Population change three months before the elections	Population change more than three months before the election and after the election
Gerber and Huber (2009)	Counties in 26 U.S. states, 1992-2004	Election of president matching county's partisanship[25]	Growth rate of consumption after election (measured by tax data)	Growth rate of consumption before election (lagged dependent variable)
Gerber and Hopkins (2011)	U.S. cities (largest 120)	Election of a Democratic (vs. Republican) mayor	Spending on public safety, tax policy, social policy	Pre-treatment covariates
Grossman (2015)	Countries in Sub-Saharan Africa	Population proportion of Renewalist Christians	Political salience of LGBT issues	Political salience of agriculture, corruption
Hainmueller and Hangartner (2015)	1,400 municipalities in Switzerland, 1991-2009	Whether naturalization decisions are made by popular vote	Rate of naturalization through ordinary process	Rate of naturalization through marriage
Hajnal, Kuk and Lajevardi (2018)	US voters	Presence of voter ID laws	Voter turnout among racial and ethnic minorities	Lagged dependent variable
Hall (2015)	Primary elections for the U.S. House, 1980-2010, involving a moderate candidate and an extremist candidate	Nomination of an extremist candidate	Party vote share; party victory; voting ideology of winning general-election candidate	Pre-treatment covariates
Hayes and Lawless (2015)	US voters in 2010 (CCES survey)	News coverage of district's House race	Respondent's ratings of incumbent & candidates' ideologies; respondent's vote intention	Respondent's political knowledge; ratings of Congress as a whole
Henderson and Brooks (2016)	Member-congresses in US House of Representatives, 1956-2008	Democratic win margin (instrumented by rainfall)	Ideal points on roll call votes (estimated one per member-congress)	Lagged ideal points (for reduced form), lagged Democratic vote margin (for first stage)

[25] In their regressions, the key coefficient is an interaction between county partisanship and partisanship of president who is elected.

Placebo outcome tests - continued from previous page

Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Holbein and Hillygus (2016)	Young adults in the 2012 Florida voter file who were marginally eligible or ineligible to vote in 2008	Whether the individual was pre-registered to vote in 2012 election (instrumented by whether the individual was 18 in 2008)	Whether the individual votes in 2012	Pre-treatment covariates (i.e. balance tests)
Holland (2015)	Districts in three Latin American capital cities	Being poor	Enforcement against street vendors	Police action against violent crimes
Jha and Wilkinson (2012)	Districts in South Asia around partition of India	Average combat exposure of WWII recruits from the district	Degree of violence and ethnic cleansing during partition	Prewar covariates and outcomes
Knutsen et al. (2017)	92,762 Afrobarometer survey respondents	The presence of an active or inactive mine	Perceptions of (and experience of) local corruption	Perceptions of national-level corruption
Ladd and Lenz (2009)	British voters	Reading a newspaper that switched to endorsing Labour in 1997 election	Voting for Labour in the 1997 election	Vote intention in 1996 (before shift)
Laitin and Ramachandran (2016)	All countries worldwide	Linguistic distance between official language and local language(s)[26]	Human development	State capacity
Levendusky (2018)	US citizens (interviewed in 2008 NAES)	Heightened sense of American identity close to July 4	Attitude towards presidential candidates of the opposite party	Attitude towards presidential candidate of own party
Malesky, Nguyen and Tran (2014)	Vietnamese communes covered by government surveys, 2006-2010	Abolition of District People's Council (DPC)	Public service delivery (30 measures)	Lagged dependent variable (parallel trends test)
Malhotra, Margalit and Mo (2013)	US survey respondents in areas with strong high-tech presence	Measures of economic threat from high-skilled immigrants (working in high tech, feeling insecure about job)	Support for high-skilled immigration	Support for Indian immigration in general
Margalit (2011)	US counties in 2000 and 2004	Trade-related job dislocations from foreign competition	Change in Republican presidential vote share, 2000-2004	Lagged outcome (change in Republican presidential vote share, 1996-2000)

[26] This is instrumented by the country's distance from a site where writing was independently developed.

Placebo outcome tests - continued from previous page

Paper	Core analysis			Placebo outcome
	Population	Treatment	Outcome	
Margalit (2013)	3,000 US respondents in panel survey (2009, 2010, 2011)	Economic shock (loss of job, job insecurity, income drop)	Support for social spending	Attitudes on climate change, immigration
Mendelberg, McCabe and Thal (2017)	64,924 college students	Attending an affluent college or university	Support for higher taxes on the wealthy	Support for other conservative political positions (e.g. restricting abortion)
Meredith (2013) (first stage)	U.S. counties during gubernatorial elections	Whether a local candidate runs for governor as Democrat, Republican, or both/neither	Vote share of Dem. gubernatorial candidate	Vote share of Dem. presidential candidate in most proximate presidential election (past or future)
Mo and Conn (2018)	Teach for America (TFA) applicants	Serving in TFA (which depends partly on applicant's selection score)	Attitudes on injustice, inequality, closeness to people of different races	How close respondents feel to "the elderly" and "Christians"
Nellis and Siddiqui (2018)	Electoral constituencies in Pakistan, 1988-2011	Share of seats occupied by secular-party politicians[27]	Incidence and severity of militant and sectarian attacks	Lagged dependent variable
Peisakhin and Rozenas (2018)	Ukrainian election precincts	Russian news TV reception	Vote share for pro-Russian parties	Lagged dependent variable, other pre-treatment covariates
Pierskalla and Sacks (2018)	Indonesian electoral districts	Electoral year (local elections)	Level of local government capital expenditure	Shifts in revenue
Pietryka and DeBats (2017)	Voters in a campus election in 2010	Proximity to candidate	Turnout and vote choice	Lagged dependent variable
Potoski and Urbatsch (2017)	US survey respondents 1970-2014 (CPS and NES)	Quality and local-ness of Monday Night Football game on night before election	Self-reported turnout	Early/absentee voting, pre-election day voter registration
Querubin and Snyder (2013)	First-time candidates to the U.S. House, 1845-1875	Winning office	Wealth accumulation after candidacy	Wealth accumulation before candidacy
Rueda (2017)	Polling stations in Colombia	Size of polling station	Reported vote buying	Reported turnout suppression
Samii (2013)	Burundian military officers	Participation in an integrated Burundian military	Levels of prejudicial behavior and ethnic salience	Various pre-treatment covariates

[27] This is instrumented by the outcome of close elections between secular and religious candidates.

Placebo outcome tests - continued from previous page

Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Sekhon and Titiunik (2010)	Election precincts in Texas	Being assigned to a new congressional district	Incumbent vote share after redistricting	Incumbent vote share before redistricting
Stokes (2016)	Ontario districts (ridings) where wind projects were proposed or operational	Precincts in which a turbine project was proposed or operational in 2011[28]	Vote share for Liberal Party (incumbent in province) in 2011 provincial election	Vote share for Liberal Party in 2003 election
Szakonyi and Urpelainen (2014)	1,094 manufacturing firms in India	Bribes reported paid by the firm; experience of lobbying through a business association	Change in firm's (subjective) power quality 2002-2005	Change in perceived quality of other services (rail, phone, internet) over same 2002-2005 period
Thachil (2014)	Indian states, 1996-2004	Provision of welfare by religious organizations	Support for the BJP among non-elites	Support for the BJP among elites
Vernby (2013)	183 Swedish municipalities in 1970s	Change in proportion of non-citizen voters (triggered by law enfranchising non-citizens)	Change in spending on policies of particular interest to non-citizens (education, social/family services)	Change in spending on policy not of particular interest to non-citizens (waste handling)

[28] In IV analysis, this is instrumented by average wind power in the district.

Placebo Treatment Tests

Paper	**Core analysis**			**Placebo treatment**
	Population	Treatment	Outcome	
Archer (2018)	American partisan-affiliated newspapers between 1932 and 2004 (aggregated by pres. election year)	Vote margin of the Republican presidential candidate	Change in total circulation of Republican- vs. Democratic-aligned local newspapers	Vote margin of the Republican presidential candidate in prior or subsequent elections
Barber (2016)	Legislators in lower houses of U.S. states	Limits on campaign contributions from PACs and individuals	Ideological polarization of each state legislator	Future contribution limits
Brollo and Nannicini (2012)	Brazilian municipalities between 1997 and 2008	Partisan alignment between mayor and president (based on election RDD)	Infrastructure transfers from central government	Fake cutoffs (median margin on right and left of true threshold)
Broockman (2013)	6,928 U.S. state legislators asked by (evidently) African-American for help with unemp. benefits	Recipient's race	Response rate and response quality	Recipient's partisanship, recipient's gender
Burnett and Kogan (2017)	Electoral precincts in San Diego city-wide elections in 2008 and 2010	Citizen pothole complaints before election	Incumbent electoral performance	Pothole complaints in 6 months after election
Condra and Shapiro (2012)	Iraqi districts from 2004 to 2009	Change in civilian casualties in previous period	Change in attacks on coalition forces by insurgents	Change in civilian casualties in future period
Dasgupta, Gawande and Kapur (2017)	Districts in Indian states where most Maoist conflict occurs	Anti-poverty program adopted in a district	Violent incidents and deaths due to Maoist conflict	Leads and lags of treatment
Dinas (2014)	Americans born around 1947, and thus eligible to vote around 1968	Voting in 1968 (instrumented by being born before eligibility cutoff in 1947)	Subsequent strength of party identification	Being born before another date in 1947 (fake cutoff)
Eggers and Hainmueller (2009)	Candidates to the British House of Commons	Winning office (election RDD)	Wealth at death	Fake cutoffs

Placebo treatment tests - continued from previous page

Paper	Core analysis			Placebo treatment
	Population	Treatment	Outcome	
Enos, Kaufman and Sands (2017)	Precincts in LA	Proximity to riot activity in 1992	Difference in support for spending on public schools between 1990 and 1992[29]	Proximity to areas with large African-American population where there was no riot activity
Ferwerda and Miller (2014)	1371 French communes around the Vichy demarcation line	Being on German side of demarcation line	Resistance activity	Being on one side of false lines on either side of true line (fake cutoff)
Fouirnaies and Mutlu-Eren (2015)	Local governments in England, 1992-2012	Being governed by the same party as the central government	Grants allocated from the central government	Future value of treatment
Franck and Rainer (2012)	Survey respondents in 18 African countries	Having a co-ethnic serve as national leader during one's primary school years	Attending/completing primary school	Having a co-ethnic serve as national leader eight years after one's primary school years
Garfias (2018)	Mexican municipalities 1920s-1940s	Commodity potential in the municipality[30]	Local presence of state officials; degree of asset expropriation/land redistribution	Commodity potential one decade in the future
Gerber and Huber (2009)	Counties in 26 U.S. states, 1992-2004	Election of president matching county's partisanship[31]	Growth rate of consumption after election (measured by tax data)	Future election of president matching county's partisanship
Gordon (2011)	U.S. Congressional districts	Designation by White House Office of Political Affairs as a priority district in 2007 (prior to the 2008 election)	Federal (GSA) contracts in district (new buildings and rental contracts)	Hypothetical treatment assigned before or after the true date of the designation
Grimmer et al. (2018)	US voters	Presence of voter ID laws (interacted with respondent race)	Turnout	Future value of treatment

[29] The authors argue that spending on public schools is "associated with African Americans and racial minorities more generally and is often implicated in the social welfare demands of riot participants".

[30] This is computed based on the relative suitability and price of a set of crops.

[31] In their regressions, the key coefficient is an interaction between county partisanship and partisanship of president who is elected.

Placebo treatment tests - continued from previous page

Paper	**Core analysis** Population	 Treatment	 Outcome	**Placebo treatment**
Hall (2015)	Contested primary elections for the U.S. House, 1980-2010, involving a moderate candidate and an extremist candidate	Nomination of an extremist candidate	Party vote share; party victory; voting ideology of winning general-election candidate	Fake cutoffs
Healy and Lenz (2017)	Zip codes in California	Change in proportion of mortgages delinquent before the 2008 election	Democratic share of the two-party vote for president in 2008	Change in proportion of mortgages delinquent after the 2008 election
Holbein and Hillygus (2016)	Young adults in the 2012 Florida voter file who were marginally eligible or ineligible to vote in 2008	Being pre-registered to vote in 2012 election (instrumented by being 18 in 2008)	Voting in 2012	Fake age cutoffs
Hopkins (2011)	4,330 Latino-Americans in 2004 survey	Provision of Spanish-language election materials, which depends on language-minority population in county being above a cutoff	Turnout; support for CA Prop 227, which restricted bilingual education	Fake population cutoffs
Jha (2013)	Towns in South Asia proximate to the coast	Whether the town was a medieval trading port	Incidence of Hindu-Muslim riots in 19th and 20th centuries	Whether the town was a colonial overseas port
Kim (2017)	Swedish municipalities, 1921-44	Having a population above 1500 (which requires a representative council rather than direct democracy)	Gender gap in voter turnout in Sweden	Having a population above 1000 (fake population cutoff)
Kogan, Lavertu and Peskowitz (2016)	Local school tax referendums in Ohio from 2003 to 2012	State government determination of whether the district has made adequate yearly progress (AYP)	Passage of proposed school tax	Future AYP failure
Ladd and Lenz (2009)	British voters	Reading a newspaper that switched to endorsing Labour in 1997 election	Voting for Labour in the 1997 election	Reading the Labour-endorsing papers in the past (but stopping before the Labour endorsement)

Placebo treatment tests - continued from previous page

Paper	**Core analysis**			**Placebo treatment**
	Population	Treatment	Outcome	
Lindgren, Oskarsson and Dawes (2017)	Swedes born between 1943 and 1955	Being in a cohort facing longer compulsory schooling[32]	Running for political office 1991-2010	Being in a cohort two to six years too old to face longer compulsory schooling
Lindsey and Hobbs (2015)	US president-months from 1946-1993	Impending presidential election (shown to reduce president's attention to foreign policy)	Level of conflict within the American bloc	Impending midterm election (shown not to reduce president's attention to foreign policy)
Malik and Stone (2018)	World Bank projects between 1994 and 2013	Participation by multinational companies/Fortune 500 companies as contractors	World Bank loan disbursement rates	Foreign Direct Investment (FDI) flows and stocks
Malhotra, Margalit and Mo (2013)	US survey respondents in areas with strong high-tech presence	Working in high tech	Support for high-skilled immigration	Being a white collar worker not in high tech
Montgomery and Nyhan (2017)	Members of the House of Representatives during the 105th to 111th Congresses	Votes by members "adjacent" to a given member, where adjacency reflects how many senior staff have recently served for both members	The member's own votes	Adjacency alternatively defined by looking at shared junior staff, or at senior staff serving in future
Peisakhin and Rozenas (2018)	Ukrainian election precincts	Russian news TV reception	Vote share for pro-Russian parties	Reception of Russian entertainment channels
Potoski and Urbatsch (2017)	US survey respondents 1970-2014 (CPS and NES)	Quality and local-ness of Monday Night Football game on night before election	Self-reported turnout	Quality and local-ness of game in week after election
Sexton (2016)	All districts in Afghanistan	Commander's Emergency Response Program (CERP) spending in a specific district	Violence	Future CERP spending
Stasavage (2014)	173 Western European cities with population of at least 10,000 by 1500 (unit of analysis is city-century)	Being an autonomous city, and time since autonomy	Economic growth (proxied by population growth)	Autonomous city and time since autonomy in the *next* century (i.e. leads of treatments)

[32] This is interacted with parents' class background to test for inequality-reducing effects of the reform.

Placebo treatment tests - continued from previous page

Paper	**Core analysis**			**Placebo treatment**
	Population	Treatment	Outcome	
Weaver and Lerman (2010)	15,170 adolescents from "Add Health" survey between ages of 18 and 26 years old	Interactions with the criminal justice system	Political involvement: voter registration, turnout, civic participation, etc.	Future criminal contact

Placebo Population Tests

Paper	Core analysis			Placebo population
	Population	Treatment	Outcome	
Acharya, Blackwell and Sen (2016*b*) (reduced form)	Americans living in the U.S. South	County's suitability for cotton production	Attitudes towards African-Americans today	Americans living in the U.S. North
Braun (2016)	Registered Jews in 439 municipalities in the Netherlands in 1941	Proximity to minority churches (instrumented by distance to Delft, the base of influential 17th century Catholic vicar/missionary)	Evasion of deportation during WWII	Registered Jews in the predominantly Catholic southern part of the Netherlands
Chen (2013)	1.1 million households who applied for FEMA aid before Nov. 2004 election[33]	Award of FEMA aid	Turnout in 2004 general election	Households who applied for FEMA aid after the November election
Erikson and Stoker (2011)	260 draft-eligible, college-bound men[34]	Lottery draft number in 1969	Attitude toward Vietnam War in 1973[35]	Non-college bound men; college-bound women
Flavin and Hartney (2015)	US teachers, as surveyed by the American National Election Survey	Being in a state with a mandatory collective bargaining law for teachers	Political participation level (donating, volunteering, etc)	Non-teachers
Gailmard and Jenkins (2009)	Members of the U.S. Senate in presidential election years	Being directly elected (after passage of 17th amendment in 1913)	Members' responsiveness to mass electorate and discretion[36]	Members of the U.S. House of Representatives
Jenkins and Monroe (2012)	Members of the majority party in 107th-110th Congress (2001-2009)	Being in the center-most wing of the party caucus ideologically	Campaign contributions from majority-party leaders	Members of the minority party
Novaes (2018)	Brazilian mayors eligible for re-election	Court ruling restricting elected officials' ability to switch parties	Ability of mayors to affect higher-level election results (measured via close-election RDD)	Brazilian mayors not eligible for re-election (purportedly unaffected by court ruling)

[33] Voters also needed to be registered to vote in both 2002 and 2004 and registered as either Democrat or Republican (pp. 204-205).

[34] "College bound" respondents identified based on college prep courses taken in 1965, and not yet being in military service as of 1969.

[35] Table 3 also investigates vote choice, presidential candidate evaluations, and issue attitudes.

[36] Responsiveness measured by correlation of roll-call voting record with state-wide electoral results; discretion measured by within-delegation differences in voting records.

Placebo population tests - continued from previous page

Paper	**Core analysis**			**Placebo population**
	Population	Treatment	Outcome	
Peisakhin and Rozenas (2018)	Ukrainian survey respondents who watch analog TV	Watching Russian news TV	Vote choice for pro-Russian parties; opinion on post-Maidan Ukrainian government; trust in Putin	Survey respondents who do not have access to terrestrial TV
Rozenas (2016)	Elections in autocracies, 1947 to 2008	Economic crisis (a proxy for office insecurity), instrumented by an index of economic shocks in nearby countries	Electoral manipulation	Elections in countries with closed economies
Rozenas, Schutte and Zhukov (2017)	Oblasts in western Ukraine	Deportations during the 1940s (instrumented by distance to railways)	Pro-Russian vote share between 2004 and 2014	Oblasts in the southwestern corner of Ukraine, annexed to USSR after main wave of deportations

Appendix B: Illustrations of placebo tests of alternative mechanisms

Because we suspect that most readers are more accustomed to thinking about placebo tests for confounding than about placebo tests for alternative mechanisms, we discuss a few examples of the latter type to further illustrate the logic.

Cruz and Schneider (2017) carries out a placebo test that addresses both selection bias and alternative mechanisms in a study of local politicians in the Philippines. They measure the effect of a municipality being chosen to participate in a foreign aid program on how often local officials visit that municipality; because local politicians had nothing to do with determining whether the municipality was chosen, they interpret their treatment effect as a measure of "undeserved credit claiming". But these municipalities may differ at baseline in other ways (selection bias), and the award of foreign aid could also attract local officials for other reasons (alternative mechanisms); for example, the program may attract new residents, making the municipality a more attractive place to seek votes. In a placebo test, Cruz and Schneider (2017) use visits by *midwives* as a placebo outcome. Suppose that we are concerned that the program attracts local politicians not because it creates opportunities for credit claiming but because it attracts new residents.[37] Then checking for an effect of the program on visits by midwives would be an informative test of this alternative mechanism to the extent that we think that midwives are also attracted by new residents (e.g. if some of these new residents are pregnant women).

Malhotra, Margalit and Mo (2013) use a survey of Americans living in areas with strong high-tech sectors to study the effect of job market competition on support for immigration. The treatment is whether the respondent works in the high-tech sector and the outcome is support for high-skilled immigration (specifically, the H-1B visa program); the mechanism of interest is the threat posed by high-skilled immigration for the respondent's own job. Working in the high-tech sector could affect support for the H-1B visa program through other mechanisms – for example, by exposing the respondent to a more diverse set of co-workers, or by increasing the respondent's income. To assess the role of these alternative mechanisms (as well as baseline differences between

[37] Given counts of local residents, that could be tested directly, of course.

tech workers and others, i.e. selection bias), Malhotra, Margalit and Mo (2013) replace the outcome with support for Indian immigration in general, which might be affected by these other mechanisms but (they argue) should be less affected by job market competition. Malhotra, Margalit and Mo (2013) find that high-tech workers are less supportive of the H-1B program but not less supportive of Indian immigration in general, which they argue provides evidence that the H-1B difference is due to job market competition and not other channels through which working in tech might affect attitudes towards immigrants. Put differently, their placebo test provides evidence of hostility to immigration among tech workers not generally (as might be predicted if it operated simply through exposure to diverse people), but narrowly where you would expect to see it under their theory (i.e. toward high-skilled immigrants).

Mendelberg, McCabe and Thal (2017) present evidence that people who attend "affluent colleges" are less supportive of higher taxes on the wealthy; they argue that this is because campus affluence "set[s] in motion norms of affluence and activat[es] latent class interests". The authors conduct a placebo outcome test showing that attending an affluent college is not associated with support for abortion and other conservative social positions. Unlike most authors, Mendelberg, McCabe and Thal (2017) provide an explicit justification for their test, arguing that "if campus affluence works by setting in motion norms of affluence and activating latent class interests, it would not affect opinion on issues that do not implicate those interests". This seems to put the theoretical point too strongly, however, because one can easily imagine a mechanism by which norms of affluence and latent class interests would indirectly affect abortion attitudes. A better justification, perhaps, is that if the apparent association between campus affluence and tax preferences is due to a broader process of political socialization, we might expect campus affluence to also be related to abortion attitudes, whereas if the association between campus affluence and tax preferences is due to norms of affluence (as hypothesized) the association with tax preferences should be stronger than the association with the abortion attitudes; the weak relationship between campus affluence and abortion attitudes thus favors the latter interpretation.[38]

[38] It also can be considered evidence against selection bias, assuming that selection bias would have worked similarly with respect to support for abortion.

Appendix References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "The political legacy of American slavery." *The Journal of Politics* 78(3):621–641.

Alexander, Dan, Christopher R Berry and William G Howell. 2016. "Distributive politics and legislator ideology." *The Journal of Politics* 78(1):214–231.

Alt, James E, John Marshall and David D Lassen. 2016. "Credible sources and sophisticated voters: when does new information induce economic voting?" *The Journal of Politics* 78(2):327–342.

Arceneaux, Kevin, Martin Johnson, René Lindstädt and Ryan J Vander Wielen. 2016. "The influence of news media on political elites: Investigating strategic responsiveness in Congress." *American Journal of Political Science* 60(1):5–29.

Archer, Allison MN. 2018. "Political advantage, disadvantage, and the demand for partisan news." *The Journal of Politics* 80(3):845–859.

Ariga, Kenichi. 2015. "Incumbency disadvantage under electoral rules with intraparty competition: evidence from Japan." *The Journal of Politics* 77(3):874–887.

Barber, Michael J. 2016. "Ideological donors, contribution limits, and the polarization of American legislatures." *The Journal of Politics* 78(1):296–310.

Bateson, Regina. 2012. "Crime victimization and political participation." *American Political Science Review* 106(03):570–587.

Bayer, Patrick and Johannes Urpelainen. 2016. "It is all about political incentives: democracy and the renewable feed-in tariff." *The Journal of Politics* 78(2):603–619.

Bechtel, Michael M and Jens Hainmueller. 2011. "How Lasting Is Voter Gratitude? An Analysis of the Short-and Long-Term Electoral Returns to Beneficial Policy." *American Journal of Political Science* 55(4):852–868.

Bhavnani, Rikhil R and Alexander Lee. 2018. "Local embeddedness and bureaucratic performance: evidence from India." *The Journal of Politics* 80(1):71–87.

Boas, Taylor C and F Daniel Hidalgo. 2011. "Controlling the airwaves: Incumbency advantage and community radio in Brazil." *American Journal of Political Science* 55(4):869–885.

Boas, Taylor C, F Daniel Hidalgo and Neal P Richardson. 2014. "The spoils of victory: campaign donations and government contracts in Brazil." *The Journal of Politics* 76(2):415–429.

Braun, Robert. 2016. "Religious minorities and resistance to genocide: the collective rescue of Jews in the Netherlands during the Holocaust." *American Political Science Review* 110(1):127–147.

Brollo, Fernanda and Tommaso Nannicini. 2012. "Tying your enemy's hands in close races: The politics of federal transfers in Brazil." *American Political Science Review* 106(04):742–761.

Broockman, David E. 2013. "Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives." *American Journal of Political Science* 57(3):521–536.

Burnett, Craig M and Vladimir Kogan. 2017. "The politics of potholes: Service quality and retrospective voting in local elections." *The Journal of Politics* 79(1):302–314.

Chaudoin, Stephen. 2014. "Audience features and the strategic timing of trade disputes." *International Organization* 68(4):877–911.

Chen, Jowei. 2013. "Voter partisanship and the effect of distributive spending on political participation." *American Journal of Political Science* 57(1):200–217.

Clinton, Joshua D and Ted Enamorado. 2014. "The national news media's effect on Congress: How Fox News affected elites in Congress." *The Journal of Politics* 76(4):928–943.

Condra, Luke N and Jacob N Shapiro. 2012. "Who takes the blame? The strategic effects of collateral damage." *American Journal of Political Science* 56(1):167–187.

Cooper, Jasper, Sung Eun Kim and Johannes Urpelainen. 2018. "The broad impact of a narrow conflict: how natural resource windfalls shape policy and politics." *The Journal of Politics* 80(2):630–646.

Cox, Gary W, Jon H Fiva and Daniel M Smith. 2016. "The contraction effect: How proportional representation affects mobilization and turnout." *The Journal of Politics* 78(4):1249–1263.

Cruz, Cesi and Christina J Schneider. 2017. "Foreign aid and undeserved credit claiming." *American Journal of Political Science* 61(2):396–408.

Dasgupta, Aditya, Kishore Gawande and Devesh Kapur. 2017. "(When) do antipoverty programs reduce violence? India's rural employment guarantee and Maoist conflict." *International organization* 71(3):605–632.

de Benedictis-Kessner, Justin. 2018. "Off-cycle and out of office: Election timing and the incumbency advantage." *The Journal of Politics* 80(1):119–132.

De Kadt, Daniel and Horacio A Larreguy. 2018. "Agents of the regime? Traditional leaders and electoral politics in South Africa." *The Journal of Politics* 80(2):382–399.

Dinas, Elias. 2014. "Does choice bring loyalty? Electoral participation and the development of party identification." *American Journal of Political Science* 58(2):449–465.

Dube, Arindrajit, Oeindrila Dube and Omar García-Ponce. 2013. "Cross-border spillover: US gun laws and violence in Mexico." *American Political Science Review* 107(03):397–417.

Egan, Patrick J and Megan Mullin. 2012. "Turning personal experience into political attitudes: The effect of local weather on Americans' perceptions about global warming." *The Journal of Politics* 74(3):796–809.

Eggers, Andrew C and Jens Hainmueller. 2009. "MPs for sale? Returns to office in postwar British politics." *American Political Science Review* 103(4):513–533.

Enos, Ryan D, Aaron R Kaufman and Melissa L Sands. 2017. "Can violent protest change local policy support? evidence from the aftermath of the 1992 Los Angeles riot." *American Political Science Review* pp. 1–17.

Erikson, Robert S and Laura Stoker. 2011. "Caught in the draft: The effects of Vietnam draft lottery status on political attitudes." *American Political Science Review* 105(2):221–237.

Feigenbaum, James J and Andrew B Hall. 2015. "How legislators respond to localized economic shocks: evidence from Chinese import competition." *The Journal of Politics* 77(4):1012–1030.

Ferwerda, Jeremy and Nicholas L Miller. 2014. "Political devolution and resistance to foreign rule: A natural experiment." *American Political Science Review* 108(03):642–660.

Flavin, Patrick and Michael T Hartney. 2015. "When government subsidizes its own: Collective bargaining laws as agents of political mobilization." *American Journal of Political Science* 59(4):896–911.

Folke, Olle and James M Snyder. 2012. "Gubernatorial midterm slumps." *American Journal of Political Science* 56(4):931–948.

Folke, Olle, Shigeo Hirano and James M Snyder. 2011. "Patronage and elections in US states." *American Political Science Review* 105(03):567–585.

Fouirnaies, Alexander and Andrew B Hall. 2018. "How Do Interest Groups Seek Access to Committees?" *American Journal of Political Science* 62(1):132–147.

Fouirnaies, Alexander and Hande Mutlu-Eren. 2015. "English bacon: Copartisan bias in intergovernmental grant allocation in England." *The Journal of Politics* 77(3):805–817.

Franck, Raphael and Ilia Rainer. 2012. "Does the leader's ethnicity matter? Ethnic favoritism, education, and health in sub-Saharan Africa." *American Political Science Review* 106(2):294–325.

Fukumoto, Kentaro and Yusaku Horiuchi. 2011. "Making outsiders' votes count: Detecting electoral fraud through a natural experiment." *American Political Science Review* 105(3):586–603.

Gailmard, Sean and Jeffery A Jenkins. 2009. "Agency problems, the 17th Amendment, and representation in the Senate." *American Journal of Political Science* 53(2):324–342.

Garfias, Francisco. 2018. "Elite competition and state capacity development: Theory and evidence from post-revolutionary Mexico." *American Political Science Review* 112(2):339–357.

Gerber, Alan S and Gregory A Huber. 2009. "Partisanship and economic behavior: Do partisan differences in economic forecasts predict real economic behavior?" *American Political Science Review* 103(3):407–426.

Gerber, Elisabeth R and Daniel J Hopkins. 2011. "When mayors matter: estimating the impact of mayoral partisanship on city policy." *American Journal of Political Science* 55(2):326–339.

Gordon, Sanford C. 2011. "Politicizing agency spending authority: Lessons from a Bush-era scandal." *American Political Science Review* 105(04):717–734.

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

Grossman, Guy. 2015. "Renewalist Christianity and the political saliency of LGBTs: Theory and evidence from Sub-Saharan Africa." *The Journal of Politics* 77(2):337–351.

Hainmueller, Jens and Dominik Hangartner. 2015. "Does direct democracy hurt immigrant minorities? evidence from naturalization decisions in Switzerland." *American Journal of Political Science* .

Hajnal, Zoltan, John Kuk and Nazita Lajevardi. 2018. "We all agree: Strict voter ID laws disproportionately burden minorities." *The Journal of Politics* 80(3):1052–1059.

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(1):18–42.

Hayes, Danny and Jennifer L Lawless. 2015. "As local news goes, so goes citizen engagement: Media, knowledge, and participation in US House Elections." *The Journal of Politics* 77(2):447–462.

Healy, Andrew and Gabriel S Lenz. 2017. "Presidential voting and the local economy: Evidence from two population-based data sets." *The Journal of Politics* 79(4):1419–1432.

Henderson, John and John Brooks. 2016. "Mediating the Electoral Connection: The Information Effects of Voter Signals on Legislative Behavior." *The Journal of Politics* 78(3):653–669.

Holbein, John B and D Sunshine Hillygus. 2016. "Making young voters: the impact of preregistration on youth turnout." *American Journal of Political Science* 60(2):364–382.

Holland, Alisha C. 2015. "The distributive politics of enforcement." *American Journal of Political Science* 59(2):357–371.

Hopkins, Daniel J. 2011. "Translating into Votes: The Electoral Impacts of Spanish-Language Ballots." *American Journal of Political Science* 55(4):814–830.

Jenkins, Jeffery A and Nathan W Monroe. 2012. "Buying negative agenda control in the US House." *American Journal of Political Science* 56(4):897–912.

Jha, Saumitra. 2013. "Trade, institutions, and ethnic tolerance: Evidence from South Asia." *American Political Science Review* 107(04):806–832.

Jha, Saumitra and Steven Wilkinson. 2012. "Does Combat Experience Foster Organizational Skill? Evidence from Ethnic Cleansing during the Partition of South Asia." *American Political Science Review* 106(04):883–907.

Kim, Jeong Hyun. 2017. "Direct Democracy and Women's Political Engagement." *American Journal of Political Science* .

Knutsen, Carl Henrik, Andreas Kotsadam, Eivind Hammersmark Olsen and Tore Wig. 2017. "Mining and local corruption in Africa." *American Journal of Political Science* 61(2):320–334.

Kogan, Vladimir, Stéphane Lavertu and Zachary Peskowitz. 2016. "Performance federalism and local democracy: Theory and evidence from school tax referenda." *American Journal of Political Science* 60(2):418–435.

Ladd, Jonathan McDonald and Gabriel S Lenz. 2009. "Exploiting a rare communication shift to document the persuasive power of the news media." *American Journal of Political Science* 53(2):394–410.

Laitin, David D and Rajesh Ramachandran. 2016. "Language policy and human development." *American Political Science Review* 110(3):457–480.

Levendusky, Matthew S. 2018. "Americans, not partisans: Can priming American national identity reduce affective polarization?" *The Journal of Politics* 80(1):59–70.

Lindgren, Karl-Oskar, Sven Oskarsson and Christopher T Dawes. 2017. "Can Political Inequalities Be Educated Away? Evidence from a Large-Scale Reform." *American Journal of Political Science* 61(1):222–236.

Lindsey, David and William Hobbs. 2015. "Presidential effort and international outcomes: Evidence for an executive bottleneck." *The Journal of Politics* 77(4):1089–1102.

Malesky, Edmund J, Cuong Viet Nguyen and Anh Tran. 2014. "The impact of recentralization on public services: A difference-in-differences analysis of the abolition of elected councils in Vietnam." *American Political Science Review* 108(1):144–168.

Malhotra, Neil, Yotam Margalit and Cecilia Hyunjung Mo. 2013. "Economic explanations for opposition to immigration: Distinguishing between prevalence and conditional impact." *American Journal of Political Science* 57(2):391–410.

Malik, Rabia and Randall W Stone. 2018. "Corporate influence in World Bank lending." *The Journal of Politics* 80(1):103–118.

Margalit, Yotam. 2011. "Costly jobs: Trade-related layoffs, government compensation, and voting in US elections." *American Political Science Review* 105(1):166–188.

Margalit, Yotam. 2013. "Explaining social policy preferences: Evidence from the Great Recession." *American Political Science Review* 107(01):80–103.

Mendelberg, Tali, Katherine T McCabe and Adam Thal. 2017. "College socialization and the economic views of affluent Americans." *American Journal of Political Science* 61(3):606–623.

Meredith, Marc. 2013. "Exploiting friends-and-neighbors to estimate coattail effects." *American Political Science Review* 107(04):742–765.

Mo, Cecilia Hyunjung and Katharine M Conn. 2018. "When Do the Advantaged See the Disadvantages of Others? A Quasi-Experimental Study of National Service." *American Political Science Review* 112(4):721–741.

Montgomery, Jacob M and Brendan Nyhan. 2017. "The effects of congressional staff networks in the us house of representatives." *The Journal of Politics* 79(3):745–761.

Nellis, Gareth and Niloufer Siddiqui. 2018. "Secular party rule and religious violence in Pakistan." *American political science review* 112(1):49–67.

Novaes, Lucas M. 2018. “Disloyal brokers and weak parties.” *American Journal of Political Science* 62(1):84–98.

Peisakhin, Leonid and Arturas Rozenas. 2018. “Electoral effects of biased media: Russian television in Ukraine.” *American Journal of Political Science* 62(3):535–550.

Pierskalla, Jan H and Audrey Sacks. 2018. “Unpaved road ahead: The consequences of election cycles for capital expenditures.” *The Journal of Politics* 80(2):510–524.

Pietryka, Matthew T and Donald A DeBats. 2017. “It’s not just what you have, but who you know: Networks, social proximity to elites, and voting in state and local elections.” *American Political Science Review* 111(2):360–378.

Potoski, Matthew and R Urbatsch. 2017. “Entertainment and the Opportunity Cost of Civic Participation: Monday Night Football Game Quality Suppresses Turnout in US Elections.” *The Journal of Politics* 79(2):424–438.

Querubin, Pablo and James M Snyder. 2013. “The control of politicians in normal times and times of crisis: Wealth accumulation by US Congressmen, 1850-1880.” *Quarterly Journal of Political Science* .

Rozenas, Arturas. 2016. “Office insecurity and electoral manipulation.” *The Journal of Politics* 78(1):232–248.

Rozenas, Arturas, Sebastian Schutte and Yuri Zhukov. 2017. “The political legacy of violence: The long-term impact of Stalin’s repression in Ukraine.” *The Journal of Politics* 79(4):1147–1161.

Rueda, Miguel R. 2017. “Small aggregates, big manipulation: Vote buying enforcement and collective monitoring.” *American Journal of Political Science* 61(1):163–177.

Samii, Cyrus. 2013. “Perils or promise of ethnic integration? Evidence from a hard case in Burundi.” *American Political Science Review* 107(03):558–573.

Sekhon, Jasjeet S and Rocio Titiunik. 2010. “When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote.” *American Political Science Review Forthcoming* .

Sexton, Renard. 2016. “Aid as a tool against insurgency: Evidence from contested and controlled territory in Afghanistan.” *American Political Science Review* 110(4):731–749.

Stasavage, David. 2014. “Was Weber Right? The Role of Urban Autonomy in Europe’s Rise.” *American Political Science Review* 108(02):337–354.

Stokes, Leah C. 2016. “Electoral backlash against climate policy: A natural experiment on retrospective voting and local resistance to public policy.” *American Journal of Political Science* 60(4):958–974.

Szakonyi, David and Johannes Urpelainen. 2014. “Who benefits from economic reform? Firms and distributive politics.” *The Journal of Politics* 76(3):841–858.

Thachil, Tariq. 2014. "Elite parties and poor voters: Theory and evidence from India." *American Political Science Review* 108(2):454–477.

Vernby, Kåre. 2013. "Inclusion and public policy: evidence from Sweden's introduction of noncitizen suffrage." *American Journal of Political Science* 57(1):15–29.

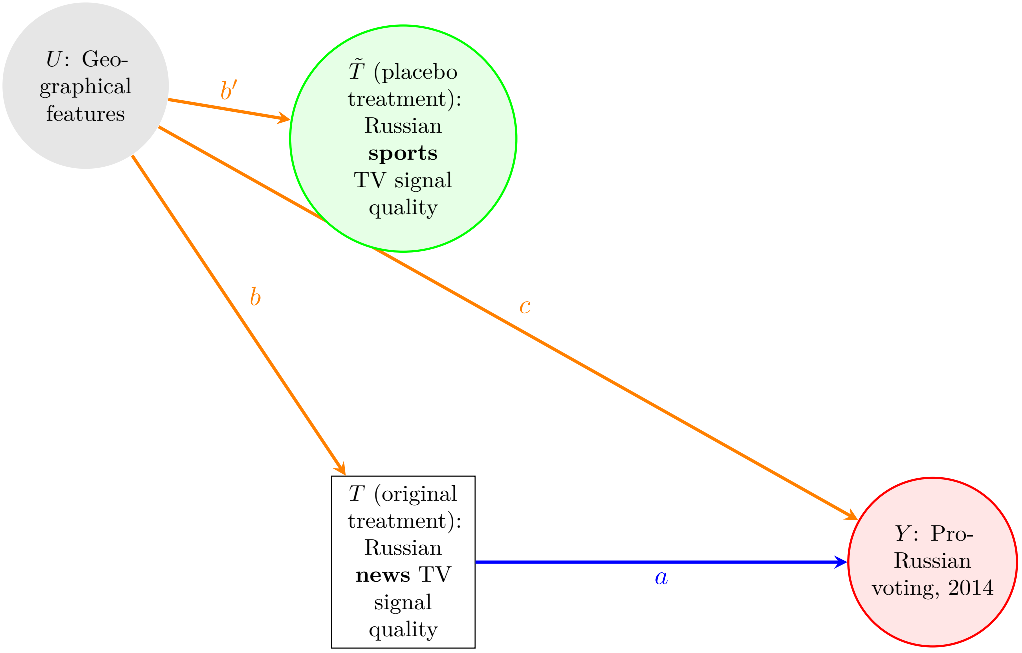
Weaver, Vesla M and Amy E Lerman. 2010. "Political consequences of the carceral state." *American Political Science Review* 104(04):817–833.

Appendix C: Placebo treatment tests

February 15, 2021

One confounder

First we consider a case where there is one confounder of interest and we have a placebo treatment that is plausibly affected by it, as in the DAG below.



U: Geographical features
b'
$\tilde{T}$ (placebo treatment): Russian **sports** TV signal quality
b
c
T (original treatment): Russian **news** TV signal quality
a
Y: Pro-Russian voting, 2014

We assume throughout that all variables are standardized (mean zero and unit variance) and all effects are linear; this allows us to express the population regression coefficient in terms of the path coefficients a, b, b', and c according to the methods outlined in Pearl (2013).

In the core analysis we regress Y on T, yielding (in expectation)

$$\beta_{YT} = a + bc,$$

where bc is bias. The purpose of the placebo treatment test is to assess this bias. Perhaps the control strategy chosen (omitted from the DAG for simplicity) is sufficient to address confounding due to U, in which case $b = 0$ or $c = 0$ or both. We want a placebo test that yields different results depending on the severity of the bias. If we think in NHST terms, and we want the null to correspond to "no bias", we want a test that rejects the null hypothesis at a rate that is low when there is no bias (low false positive rate) and increases in the magnitude of the bias.

Unconditional placebo treatment test

In an unconditional placebo treatment test (UPTT), we regress the outcome on the placebo treatment (and possibly other control variables, in the general case) but we do not control for the actual treatment. This test yields (in expectation, and assuming linearity and standardized variables as noted above)

$$\beta_{Y\tilde{T}} = b'c + b'ba. \tag{1}$$

Assuming $b' = kb$ (i.e. the effect of U on the treatment is linearly related to the effect of U on the placebo treatment), this becomes

$$\beta_{Y\tilde{T}} = k(bc + b^2a), \tag{2}$$

where the first term in the parentheses (bc) is the bias due to U. Because the actual treatment and the placebo treatment are both affected by the confounder, and because we don't control for the actual treatment in the UPTT, the UPTT estimate (assuming $b' = kb$) combines the bias from the core analysis (bc) with an attenuated version of the treatment effect (b^2a). (We know that b^2a is an *attenuated* version of the bias a because, given standardized variables, all path coefficients must be less than 1 in magnitude.)

If our goal is to test for bias due to U, the presence of the treatment effect in Equation 1 is concerning: the UPTT may produce a non-zero result (in expectation) even when the bias is zero (e.g. when $c = 0$ but $b' = b \neq 0$ and $a \neq 0$), and we can get an expected zero in the UPTT even when the bias is not zero (when $c = -ba$). Thus there is the risk of both an inflated false positive rate and an inflated false negative rate.

One redeeming aspect of the UPTT is that "canceling out" can only occur when there is a non-zero treatment effect. So if we use a UPTT to evaluate the competing theories that "the treatment affects the outcome" vs. "the treatment does not affect the outcome", then finding a non-zero treatment effect in the core analysis and a zero in the UPTT is consistent with the first theory but not the second, even if we don't know from the UPTT whether the bias is zero or canceling out has occurred.

Moreover, because of attenuation of the treatment effect, "canceling out" requires a treatment effect that in realistic situations is much larger in magnitude than the confounding bias. To see this, note that canceling out occurs when $a = \frac{-c}{b}$; substituting this into the ratio of the treatment effect to the bias ($\frac{a}{bc}$) yields $\frac{-1}{b^2}$. If the confounder completely determines the treatment ($b = 1$), then canceling out occurs in the UPTT when the bias and the treatment effect are of the same magnitude and opposite signs. In practice, b is likely to be well below 1 (Ding and Miratrix 2015), in which case the treatment effect must be much larger than the bias in magnitude. For example, if $b = .25$ (a strong effect of the confounder on the treatment), the treatment effect must be 16 times as large as the bias due to U, and opposite in sign, for canceling out to occur. Given a precise zero estimate in the UPTT (and assuming the DAG shown and assuming that $b' = kb$ for some $k \neq 0$), one is probably justified in inferring that the true treatment effect is close to the estimate from the core analysis: either there is no bias or there is canceling out but the bias is very small relative to the treatment effect.

The UPTT provides a good illustration of how the result of the core analysis and the result of an imperfect placebo test could be considered in conjunction. Let $\hat{\beta}_{YT}$ denote the result of the core analysis. Consider two interpretations of the result. The "no bias" interpretation is that $bc = 0$ so that $\hat{\beta}_{YT} \approx a$. Under this interpretation, the UPTT yields $kb^2\hat{\beta}_{YT}$ in expectation. The "all bias" interpretation is that $a = 0$ so that $\hat{\beta}_{YT} \approx bc$. Under this interpretation, the UPTT yields $k\hat{\beta}_{YT}$ in expectation. The UPTT is informative between these two interpretations to the extent that b^2 is distinct from 1, i.e. to the extent that factors other than the confounder determine the treatment.[1] Assuming we expect b to be well below 1, then if the UPTT result is close to the core analysis result we favor the "all bias" interpretation, but if it is closer to zero we favor the "no bias" interpretation. This could all be formalized.

Still, there is no doubt that the contamination of the UPTT by the treatment effect makes the UPTT less informative: it increases the risk of false positives and raises the false negative rate against some relevant biases.

[1] The same logic applies in the more common case when we suspect the treatment may affect the outcome in the placebo test, but we believe that the effect is smaller than in the core analysis. The core analysis yields $a + bc$ while the placebo outcome test yields $a' + bc'$; even if $a' \neq 0$, the test might be informative because we believe $a' \ll a$ but $c' \approx c$. Would like to write this into the draft.

Conditional placebo treatment test

In the case of one confounder, the conditional placebo treamtent test (CPTT) yields in expectation

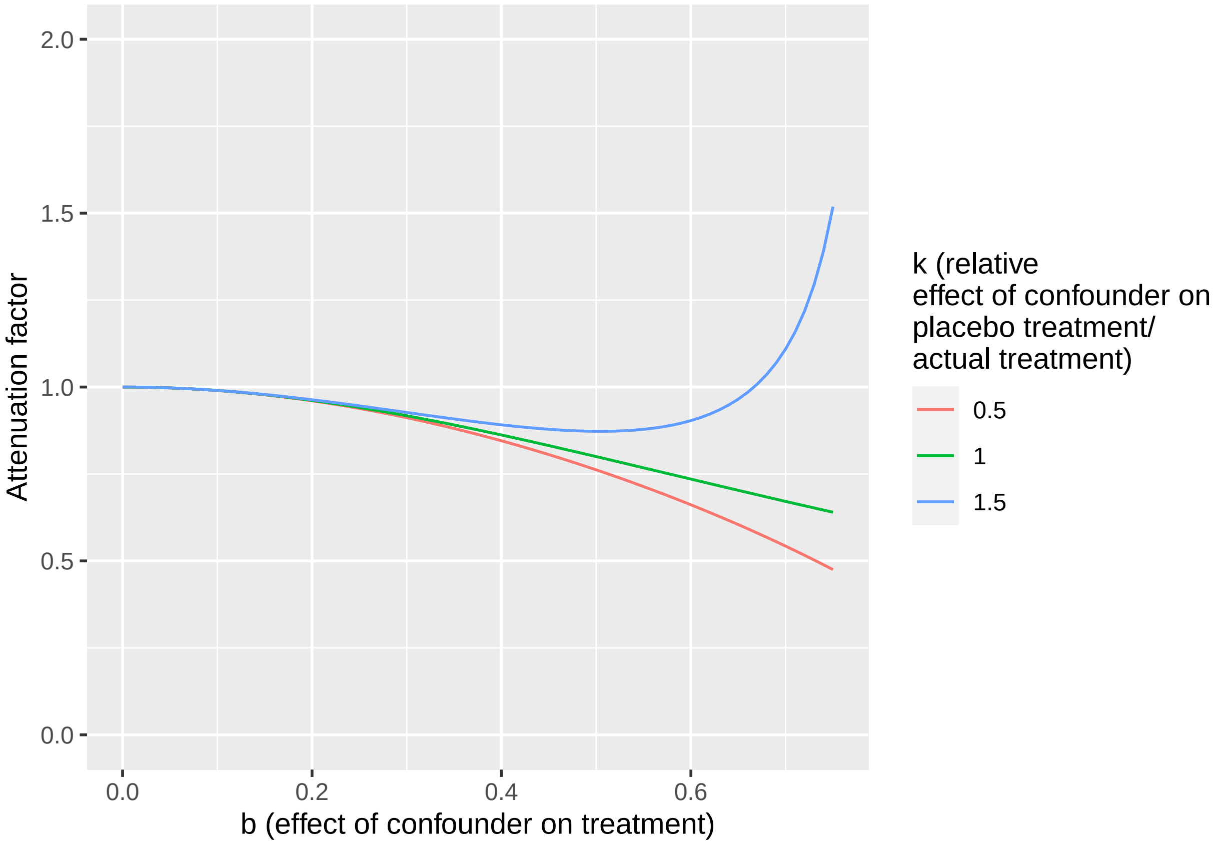
$$\beta_{Y\tilde{T}\cdot T} = \frac{b'c(1-b^2)}{1-(b'b)^2}, \tag{3}$$

which if $b' = kb$ becomes

$$\beta_{Y\tilde{T}\cdot T} = kbc\frac{(1-b^2)}{1-k^2b^4} \tag{4}$$

where bc is the bias. (See Pearl (2013) for derivation of partial regression coefficients in linear models with standardized variables.) Assuming $k \neq 0$ and $|b| < 1$, the CPTT yields zero in expectation if and only if the bias is zero, which is ideal for a placebo test.

Equation 4 expresses the expected CPTT result as the product of k (the relative effect of the confounder on the placebo treatment vs. the actual treatment), bc (the bias in the core analysis), and a term whose value depends on k and b. If the confounder very strongly affects the treatment (i.e. if $|b|$ is close to 1), this term could be far from 1, either attenuating or amplifying the kbc component. When $|b|$ is below .5 or so, however (which Ding and Miratrix (2015) argue is usually the case in observational studies), this term is close to 1 (and does not depend much on the value of k). The figure below shows the value of $\frac{(1-b^2)}{1-k^2b^4}$ as a function of b for three values of k. (Note that for $k > 0$ the maximum value of b is $\sqrt{(}1/k)$.) For $b \approx .5$ (a very strong effect of the confounder on the treatment), the attenuation is between .7 and .9 (depending on k), which may produce a noticeable decrease in the test's power; for $b \approx .2$ (a more plausible but still strong effect) the attenuation is barely noticeable.



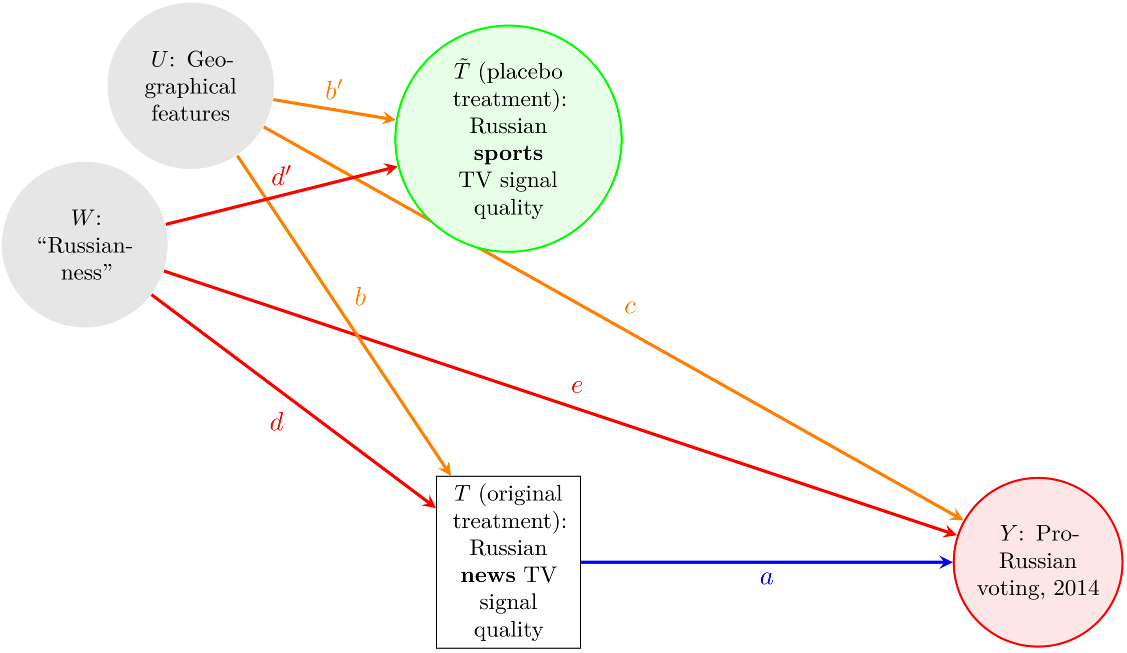
Summary

With a single confounder and a placebo treatment that is affected by the confounder but does not affect the outcome, both the unconditional placebo treatment test and the conditional placebo treatment test (i.e. the test that does not control for the actual treatment and the test that does) are informative about bias in the core analysis. The UPTT tends to be less informative because it partly tracks the treatment effect; this raises the risk of false positives and false negatives. The result of the CPTT may become attenuated as the confounder's effect on the treatment increases, but this attenuation is small in realistic cases.

Two confounders

Nothing really changes if there are two confounders and each of them affects the placebo treatment: assuming that the effect of each confounder on the placebo treatment is k times the effect of that confounder on the actual treatment, the UPTT gives us k times the bias in the core analysis plus an attenuated version of the treatment effect, and the CPTT gives us approximately k times the bias in the core analysis.

More specifically, consider the figure below, which adds a second confounder W. The path coefficients from W to T, $\tilde{T}$, and Y are labeled d, d', and e respectively.


U: Geographical features
b'
$\tilde{T}$ (placebo treatment): Russian **sports** TV signal quality
d'
W: "Russianness"
b
c
e
d
T (original treatment): Russian **news** TV signal quality
a
Y: Pro-Russian voting, 2014

The core analysis yields

$$\beta_{YT} = a + bc + de,$$

so there may be bias from both confounders.

The unconditional placebo treatment test yields

$$\beta_{Y\tilde{T}} = b'c + d'e + a\left(b'b + d'd\right),$$

which assuming $b' = kb$ and $d' = kb$ approximates the bias plus an attenuated version of the treatment effect. But the conditional placebo treatment test yields

$$\beta_{Y\tilde{T}\cdot T} = \frac{\overbrace{b'c}^{\text{bias}_U^*}\left(1-b^2\right)+\overbrace{d'e}^{\text{bias}_W^*}\left(1-d^2\right)-\overbrace{bc}^{\text{bias}_U}d'd-\overbrace{de}^{\text{bias}_W}b'b}{1-\left(b'b+d'd\right)^2}.$$

This doesn't just double up the bias terms, it introduces a new component – the collider bias that arises because, by conditioning on T, we create a new path from $\tilde{T}$ to Y. (Actually, we create two new paths: $\tilde{T} \leftarrow W \rightarrow T \leftarrow U \rightarrow Y$ and $\tilde{T} \leftarrow U \rightarrow T \leftarrow W \rightarrow Y$.) Assuming $b' = k_b b$ and $d' = k_d d$, we can rewrite the previous expression as

$$\beta_{Y\tilde{T}\cdot T} = \frac{\overbrace{bc}^{\text{Bias}_U}\left(k_b(1-b^2) - k_d d^2\right) + \overbrace{de}^{\text{Bias}_W}\left(k_d(1-d^2) - k_b b^2\right)}{1 - \left(k_b b^2 + k_d d^2\right)^2}$$

or

$$\beta_{Y\tilde{T}\cdot T} = \overbrace{bc}^{\text{Bias}_U} \frac{k_b(1-b^2) - k_d d^2}{1 - \left(k_b b^2 + k_d d^2\right)^2} + \overbrace{de}^{\text{Bias}_W} \frac{k_d(1-d^2) - k_b b^2}{1 - \left(k_b b^2 + k_d d^2\right)^2}. \tag{5}$$

Thus we get the two bias components from the core analysis, each multiplied by a factor that depends on k_b, k_d, b, and d.

To get a sense of what these factors might be, suppose that $k_b = k_d = 1$ and $b = d = .2$. Then both terms are .926.

Special case where the placebo treatment is not affected by one confounder

Now consider the case where the placebo treatment is not affected by W, i.e. where $d' = 0$ in the figure above (or $k_d = 0$, assuming $d' = k_d d$). In the case of Peisakhin and Rozenas (2018), this might occur because mountains affect both news TV signal quality and entertainment TV signal quality, but only news broadcasters strategically position their transmitters to access pro-Russian Ukrainian communities.

In that case, the *de* term drops out of the UPTT, so that it picks up the bias due to U plus the attenuated treatment effect but misses the bias due to W. This separability holds for placebo population tests and placebo outcome tests, too: if only one of the two confounders operates in the placebo population, the population placebo test will only pick up bias due to that confounder; if only one of the two confounders affects the placebo outcome, the population outcome test will only pick up bias due to that confounder.

The interesting thing (arguably) is that this separability does not hold for the conditional placebo treatment test. Substituting $k_d = 0$ into equation 5, we find that the CPTT yields

$$\beta_{Y\tilde{T}\cdot T} = \overbrace{bc}^{\text{Bias}_U} \frac{k_b(1-b^2)}{1 - \left(k_b b^2\right)^2} - \overbrace{de}^{\text{Bias}_W} \frac{k_b b^2}{1 - \left(k_b b^2\right)^2}, \tag{6}$$

where the *de* term now enters with an opposite sign.

Why does the result of the CPTT depend on a confounder W that does not affect the placebo treatment? Because of *collider dependence*: U and W are independent unconditionally in the DAG, but conditioning on T induces a dependence between them. In this specific example, when we focus on places in Ukraine that have the same quality of news TV signal we should find that the ones with less favorable geography tend to have higher levels of Russian-ness. (Put differently, a place can achieve good news TV signal either by having favorable geography or a high level of Russian-ness; therefore places with good news TV signal but unfavorable geography probably have a high level of Russian-ness.)

Equation 6 highlights a difficulty with using the CPTT to assess bias in the core analysis, whether our goal is to assess all bias in the core analysis or only the bias due to U: the bias due to U enters positively and the bias due to W enters negatively, meaning that we can neither measure the total bias nor isolate the bias due to U. It could be that the two biases point in the same direction in the core analysis (i.e. $\text{sign}(bc) = \text{sign}(de)$) but cancel each other out in the placebo treatment test, yielding a false negative.

Further examination suggests that collider bias in placebo treatment tests may not be of much practical importance because it is a "higher-order bias" (see Ding and Miratrix 2015 for a similar analysis of the practical importance of M-bias and butterfly bias). Supposing that $b' = kb$, we can rewrite the result of the conditional placebo treatment test as

$$\beta_{Y\tilde{T}\cdot T} = \frac{k}{1-k^2b^4}\left(bc - b^3c - b^2de\right). \tag{7}$$

First recall that all path coefficients must have magnitude less than one (because we have assumed that all variables are standardized). Path coefficients in social science applications will typicaly be well below 1 in magnitude (as argued by Ding and Miratrix 2015): rarely do we expect e.g. a confounder to nearly completely determine a treatment or an outcome. It follows that the denominator of the leading term will typically be approximately 1, because $b^4 \approx 0$ unless b is nearly 1 (i.e. unless U almost completely determines the treatment). If b, c, d, and e are around the same magnitude (and well below 1), then the final two terms on the RHS of Equation 7 become small in magnitude and the estimate will be close to kbc, i.e. the bias due to U scaled by b'/b. For example, if $b = c = d = e = .1$ (so that U and W each contribute a bias of .01, for a total bias of .02) and $k = 1$, then the CPTT yields 0.0098, very close to the bias due to U.

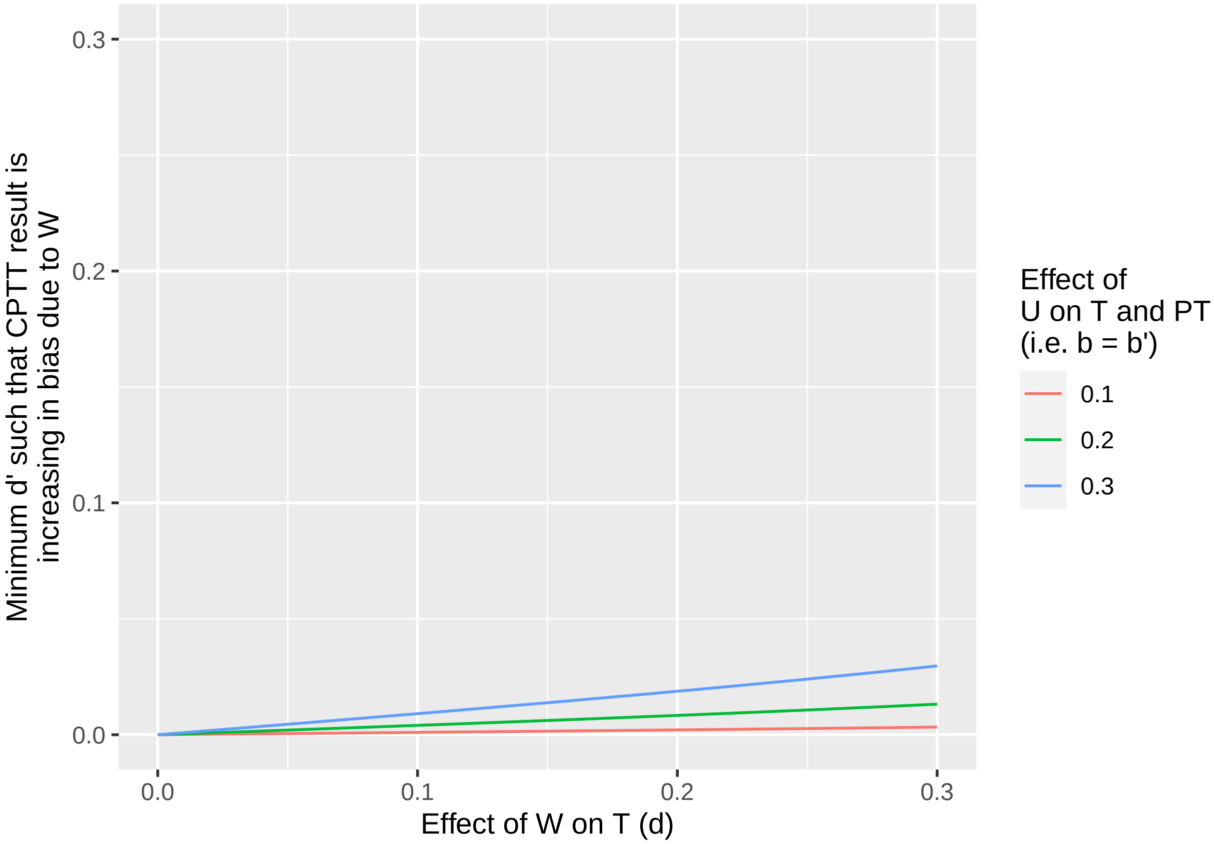
Another way to assess the importance of collider bias is to check the conditions for "canceling out", i.e. conditions under which the CPTT would yield a zero despite non-zero bias in the core analysis. Rearranging Equation 6, we see that canceling out occurs when

$$\frac{de}{c} = \frac{1-b^2}{b}.$$

When b is not close to 1, this implies that canceling out can occur only when the bias due to W is many times larger than the effect of U on Y. (For example, if $b = .2$, the multiple is 4.8; if $b = .1$, the multiple is 9.9).

Perhaps we shouldn't focus too much on the situation where $d' = 0$; it is probably more realistic to consider situations where the placebo treatment may be affected by any confounder. Then we might ask: how strong does d' need to be before the result of the placebo treatment test is positively related to both sources of bias? It should not be a surprise that, given how weak the collider dependence tends to be, the CPTT result is positively related to bias due to W even if the placebo treatment is only very slightly affected by W.

Going back to equation 5 (the general case where both confounders might affect $\tilde{T}$), we can formulate conditions under which de enters non-negatively, i.e. conditions under which the result of the placebo test is weakly increasing in both forms of bias. Generally, this condition is $k_d(1-d^2) \geq k_b b^2$, i.e. $\frac{d'}{d}(1-d^2) \geq b'b$. Now suppose that $b = b'$, i.e. $k_b = 1$. Then the condition for the CPTT result to be increasing in de is that $d' \geq \frac{db^2}{(1-d^2)}$. In the figure below we show this condition for three plausible values of b as a function of d. For quite small values of d' the CPTT works as we would hope. For example, when $b = .2$ and $d = .2$, the CPTT is increasing in bias due to W as long as $d' > 0.008$. Given how rarely we would encounter a situation where $d' = 0$, it seems unhelpful to focus on that case. Instead, we can say that a placebo treatment that tracks U much more closely than it tracks W will tend to yield an informative CPTT: the result of the CPTT won't be much affected by the bias due to W in any case, but as long as the placebo treatment tracks W even a little we can be reassured that the result of the CPTT will be increasing in the bias due to W (rather than decreasing in it as in the case where $d' = 0$).



Summary

Because the confounders we are worried about are pre-treatment, we need to consider collider bias in a conditional placebo treatment test. Collider bias may be of particular concern when the placebo treatment tracks some biases but not others: in that case the untracked biases appear in the CPTT. But the magnitude of collider bias is likely to be small (because it a higher-order bias), and the circumstance where a placebo treatment is not affected by a relevant confounder may not be very relevant; if the placebo treatment is even slightly affected by a confounder, the signal will drown out the collider bias, so that the CPTT is sensitive at least to some extent to bias due to that confounder.

The magical CPTT

Now we consider a special case of the two-confounder setup where W is the only relevant confounder but W also does not affect the placebo treatment $\tilde{T}$. In that case (i.e. the case where $c = 0$ and $d' = 0$), Equation 6 indicates that the expected result of the CPTT becomes

$$\beta_{Y\tilde{T}\cdot T} = \frac{-b'bde}{1-(b'b)^2}, \tag{8}$$

which is an attenuated (and flipped-in-sign) version of the bias due to W. In this special case we could use the CPTT to assess bias due to a confounder (here, W) that does not even affect the placebo treatment. More remarkably, if there were additional potential confounders $(W_1, W_2, W_3, \ldots)$, the CPTT would detect bias due to those confounders as well, each attenuated by $b'b$ and flipped in sign. We refer to a CPTT that produces an attenuated and flipped-in-sign measure of all bias “the magical CPTT”. Is the magical CPTT just an interesting theoretical possibility, or is it a measurement strategy of practical relevance?

First, note that the magical CPTT requires the following:

- The placebo treatment does not affect Y.
- The placebo treatment is not affected by any confounder (here, W).
- The placebo treatment and the actual treatment share a non-confounding cause (here, U).

The first requirement is an exclusion restriction common to all placebo treatment tests, but the other two are unique to the magical CPTT.

To assess the practical relevance of the magical CPTT, it may be useful to compare it to an IV. First, let's suppose that U was observed. Then U satisfies the exclusion restriction necessary for a valid instrument: it affects Y only through T. Assuming other conditions for a valid instrument are met (notably exogeneity), we would typically proceed with IV analysis using U as an instrument.

Now suppose that U is a valid instrument, but it is unobserved; it affects $\tilde{T}$, however, and $\tilde{T}$ is not affected by the confounding factors that might necessitate the instrument, nor does it affect Y directly. Then $\tilde{T}$ is also a valid instrument, though it has a weaker first stage than U would.

What the magical CPTT highlights is that, given a valid instrument ($\tilde{T}$), one can regress the outcome (Y) on the treatment (T) and the instrument ($\tilde{T}$) to assess the bias that would obtain from the naive regression of the outcome (Y) on the treatment (T). Of course, if one knew that one had a valid instrument and were concerned about bias due to W, one would typically conduct IV analysis rather than use the instrument to assess bias due to W. In most situations we are uncertain of the exclusion restriction (here, we might wonder whether U or $\tilde{T}$ affects Y), in which case a non-zero result of the regression of Y on $\tilde{T}$ conditional on T could indicate bias due to W or a violation of the exclusion restriction for $\tilde{T}$.[2] Thus the magical CPTT is really only possible in a circumstance where it would not be needed, i.e. where a valid instrument is available that would address any bias that the CPTT detected.

Summary

The magical CPTT is an interesting theoretical possibility, but it seems to have little if any practical relevance. It requires a valid IV design (in which both the placebo treatment and the relevant confounder are valid instruments), but if we had a valid IV design we could do IV analysis rather than running a CPTT to assess the need for one.

Illustrations via simulation

Basic simulation approach

To illustrate these points, we repeatedly draw datasets with a given set of causal parameters and compute for each one

- the estimated treatment effect in the core analysis (regressing Y on T)
- the unconditional placebo treatment test (regressing Y on $\tilde{T}$)
- the conditional placebo treatment test (regressing Y on $\tilde{T}$ and T),

showing that the result of the regression yields the same answer in expectation as our theory predicts.

We draw the data $\mathbf{X}$ from a multivariate normal distribution

$$\mathbf{X} = (Y, U, W, T, \tilde{T})^T \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

[2] Gerber and Green (2012) point out that one cannot test the exclusion restriction by regressing Y on the instrument controlling for the treatment, because the regression will reflect not only exclusion restriction violations but bias due to confounders (here, W) because the treatment is a collider. Our point is that one cannot test bias due to confounders (here, W) with the same regression, because the regression will reflect not only that bias but also exclusion restriction violations.

where $\mu = (0,0,0,0,0)^T$, the diagonals of $\mathbf{\Sigma}$ are all 1, and

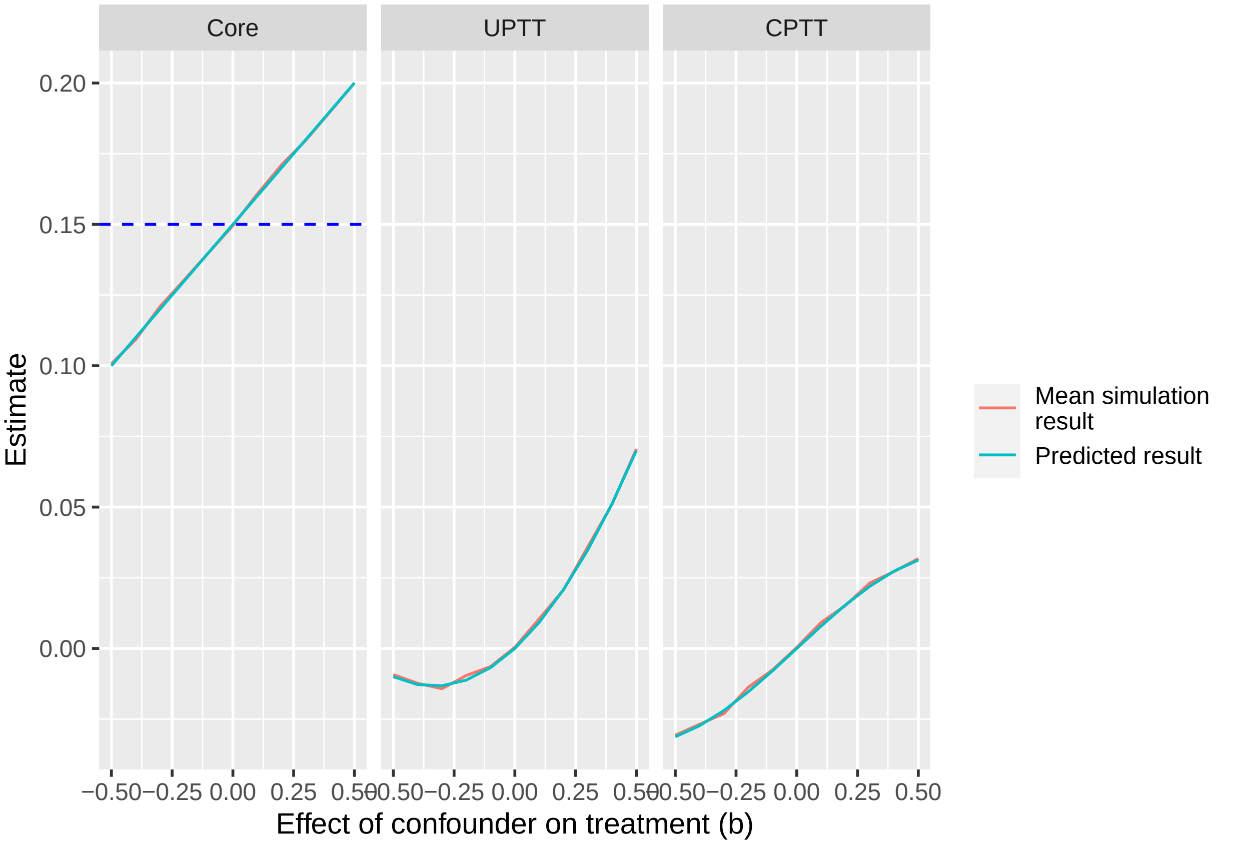
$$\begin{aligned}
\sigma^2_{TY} &= a + bc + de \\
\sigma^2_{UT} &= b \\
\sigma^2_{U\tilde{T}} &= b' \\
\sigma^2_{UY} &= c \\
\sigma^2_{WT} &= d \\
\sigma^2_{WY} &= e \\
\sigma^2_{W\tilde{T}} &= 0 \\
\sigma^2_{UW} &= 0 \\
\sigma^2_{\tilde{T}Y} &= b'(c + ab) \\
\sigma^2_{\tilde{T}T} &= b'b.
\end{aligned}$$

Our baseline set of parameters is $a = 0.15$, $c = 0.1$, $d = 0.2$, and $e = 0.3$. We will assume that $b' = kb$ with $k = .8$, and we will look at how the estimate from the placebo treatment test varies as we vary b.

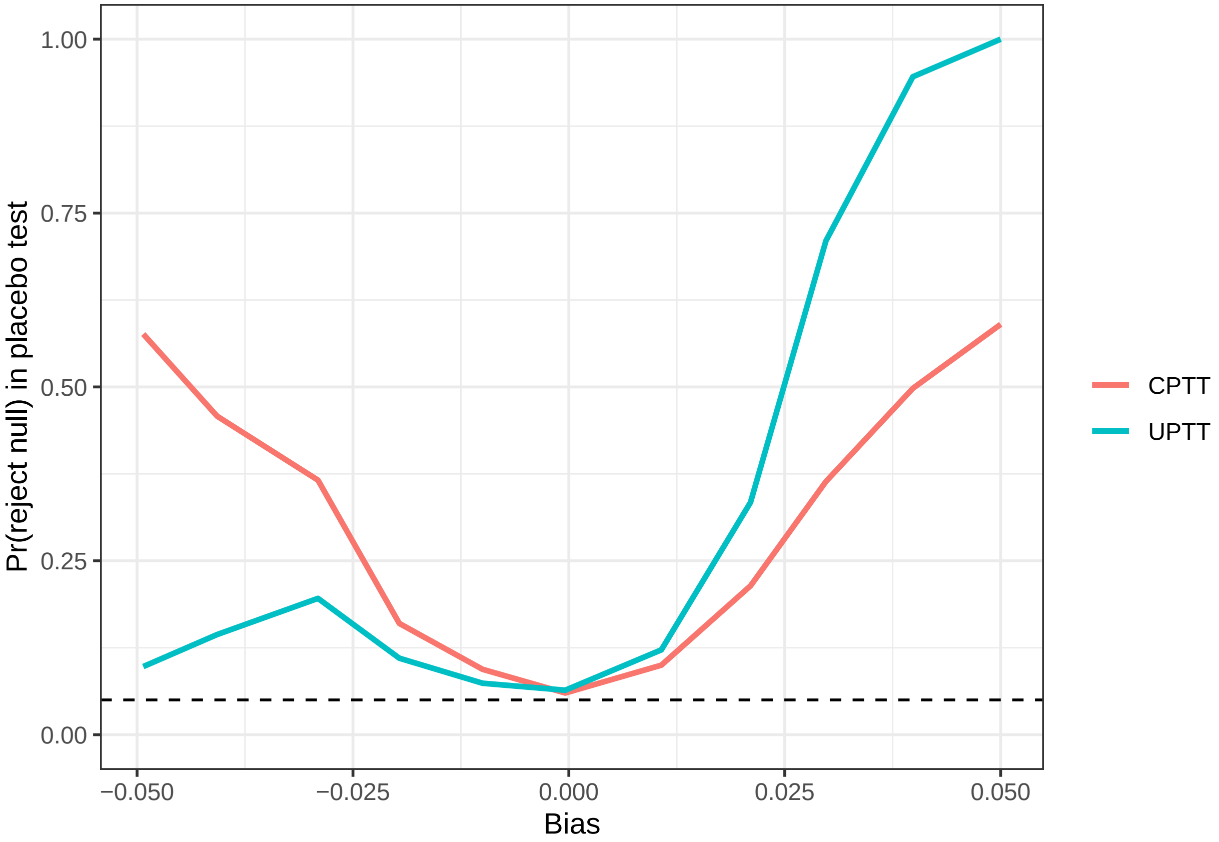
One confounder

For this simulation we set d to 0 (so that the only relevant confounder is U); we vary b from -0.5 to 0.5 by 0.1 and we assume $b' = kb$ with $k = 0.8$.

We first show that the results fit the predictions. The core analysis yields 0.15 when $b = 0$, but the estimate varies with b according to slope $c = 0.1$, reflecting confounding due to U. The UPTT basically tracks the bias, but contamination from the treatment effect (assumed to be positive) inflates the result when $b > 0$ and attenuates it when $b < 0$. The CPTT tracks the bias in a symmetric way, with some attenuation detectable as $|b|$ becomes large.

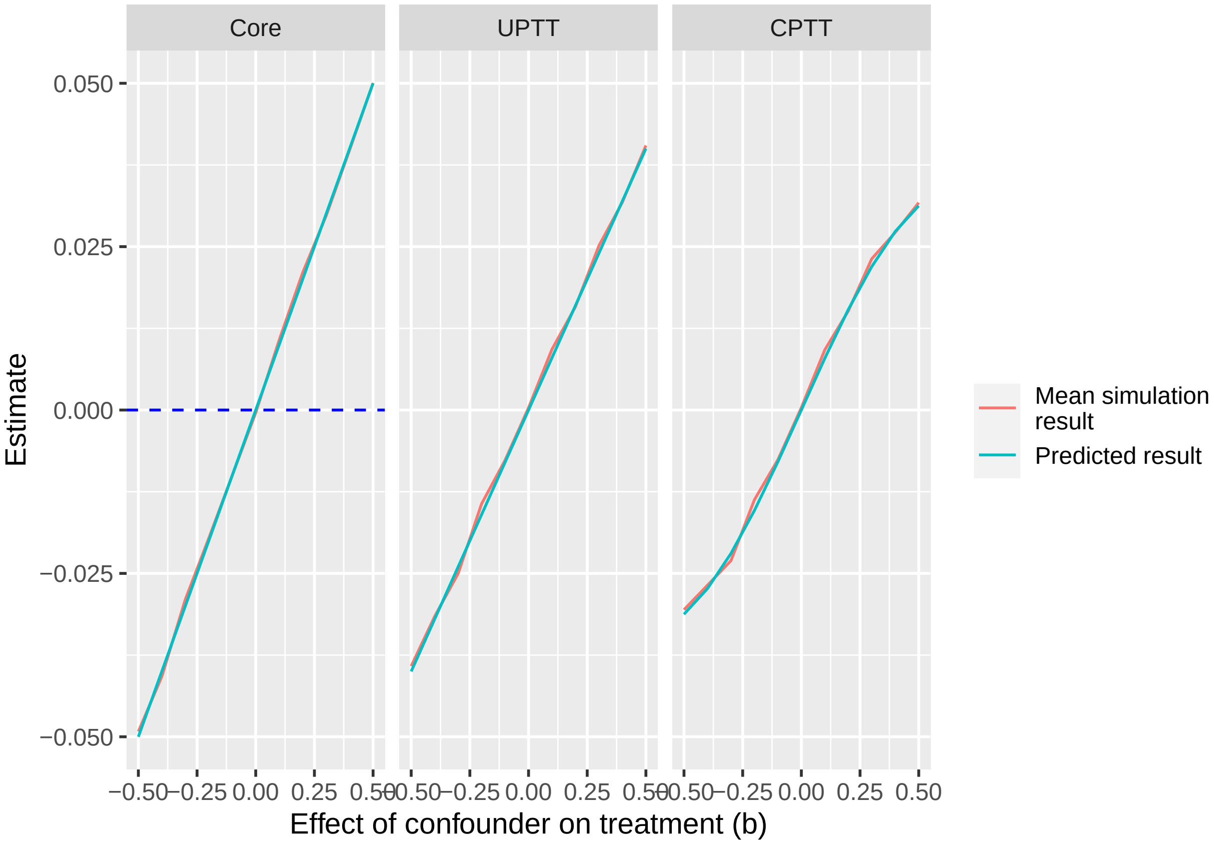


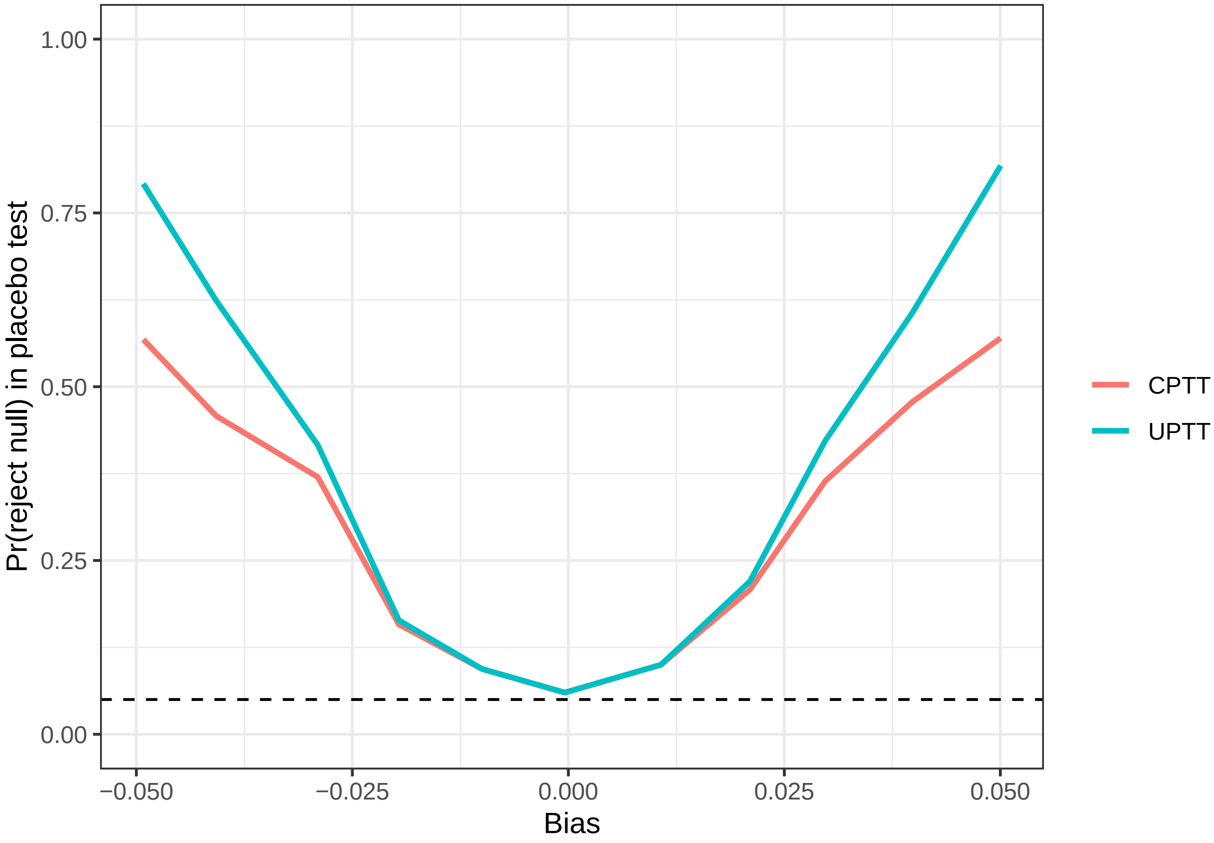
Next we observe the power curve for each test. The CPTT has the correct shape of power curve: the probability of rejecting the null is $\alpha = .05$ when the core analysis is unbiased (i.e. when $b = 0$) and increases symmetrically as the bias increases. Because we are varying b across simulations, and the UPTT yields a zero in expectation when $b = 0$, the UPTT also has the correct size when bias is zero. (This would not be true if we varied c across simulations.) When the confounder has a very strong negative effect on the treatment, however, there is canceling out due to the treatment effect; this is reflected in the downward dip of the power curve for the UPTT when bias is very strong and negative. Given that the treatment effect is $a = 0.15$ and the effect of the confounder on the outcome is $c = 0.1$, complete canceling out occurs when $b = -2/3$ (i.e. when bias is $-.066$); still, the effect on the false negative rate is clear when the bias is somewhat smaller in magnitude than that.



One confounder, no treatment effect

Next let's consider a situation that is the same except that the treatment effect is zero. This is the position a critic of the core analysis might take, i.e. that the estimate from the core analysis is all bias. This should be a favorable situation for the UPTT.

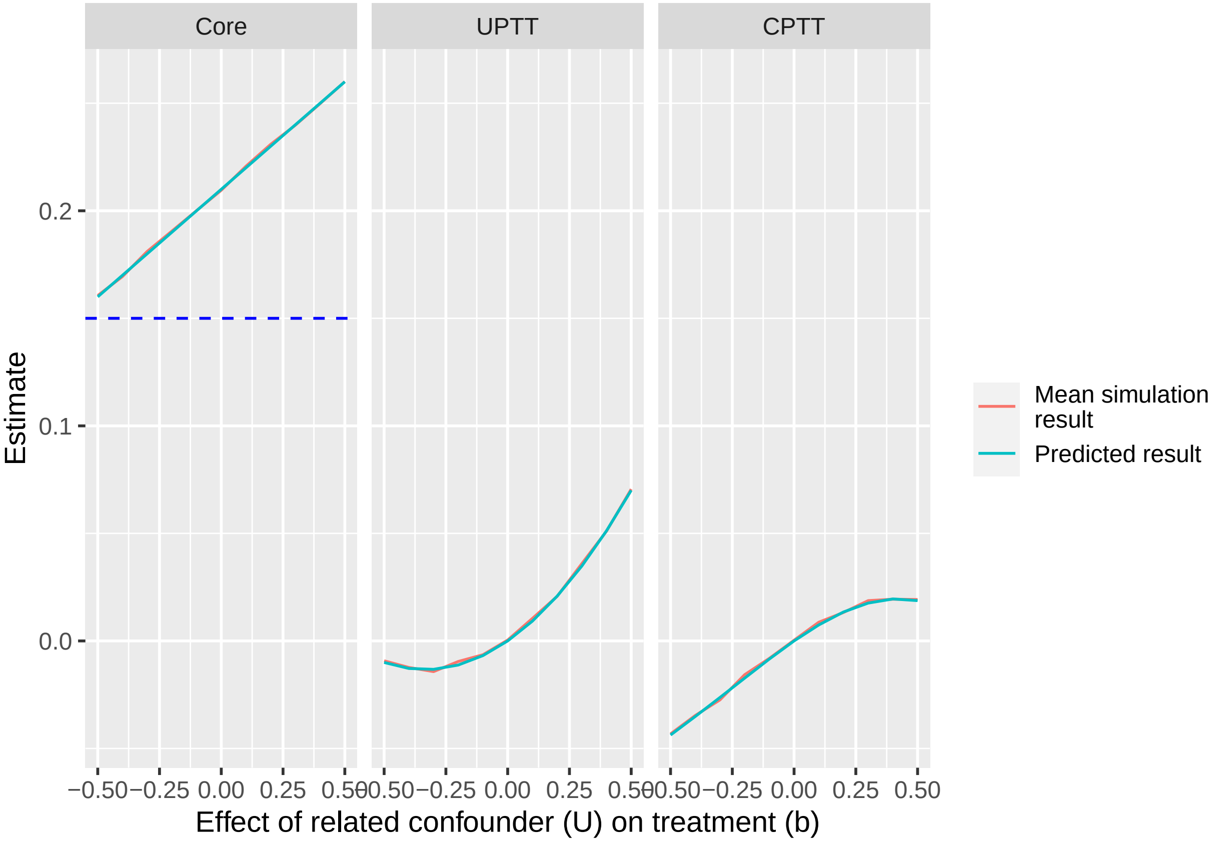

Core
UPTT
CPTT
Estimate
0.050
0.025
0.000
−0.025
−0.050
−0.50
−0.25
0.00
0.25
0.50
Effect of confounder on treatment (b)
Mean simulation result
Predicted result


Pr(reject null) in placebo test
1.00
0.75
0.50
0.25
0.00
−0.050
−0.025
0.000
0.025
0.050
Bias
CPTT
UPTT

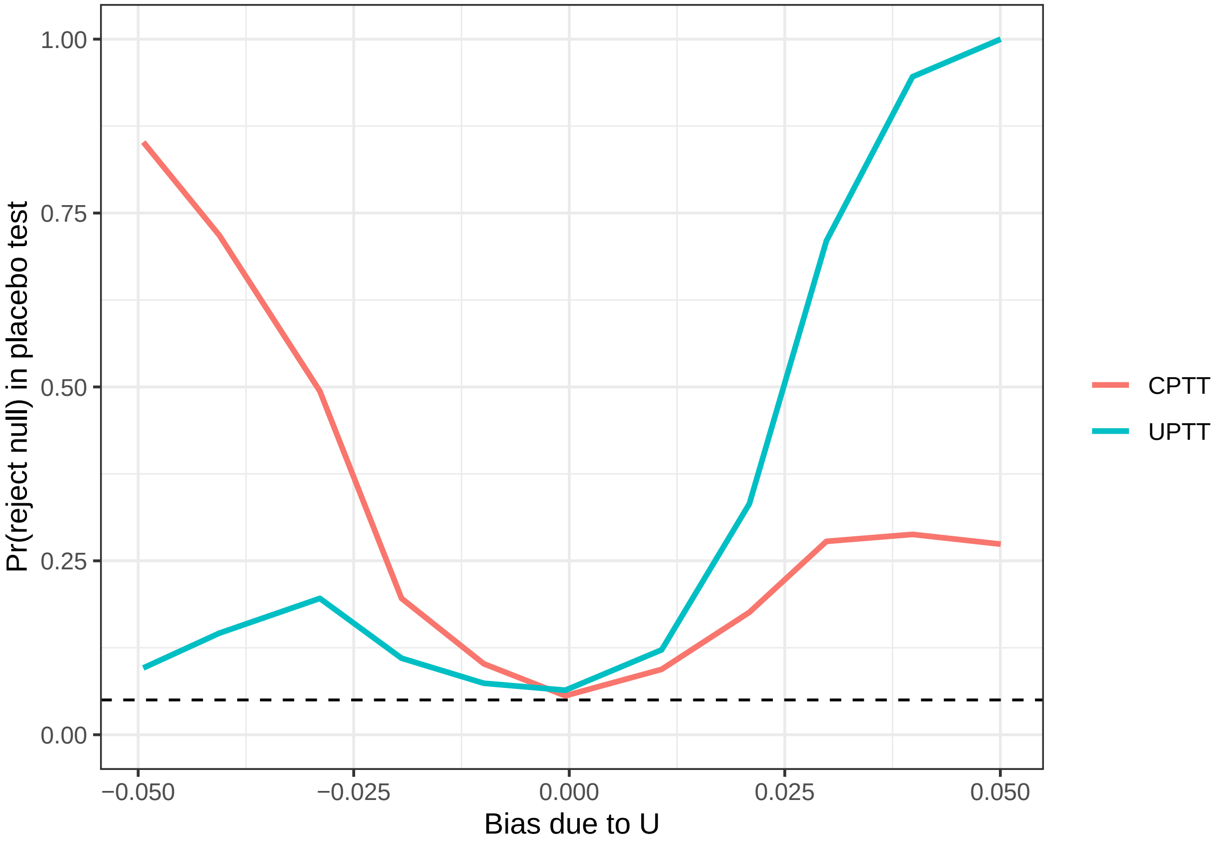
Two confounders

For this simulation we return d to 0.2; we again let b vary from -0.5 to 0.5 and b' vary in conjunction from -0.4 to 0.4. Thus there is a constant bias due to W (with size 0.06) and what varies is the bias due to U; we seek to understand how well a placebo treatment that is affected only by U tracks bias due to U.

In the figure below we see that the bias is an increasing function of b, but note that when $b = 0$ there is still bias in the core analysis (of size .06) due to W. The UPTT is the same as a function of b as when W was not relevant; the CPTT looks similar, but we see some attenuation for positive b due to the collider bias: when $b = .5$, we are subtracting approximately $.5 \times .4 \times 0.2 \times 0.3 = 0.012$.



The power curve in this case is below; we show the probability of rejecting the null as a function of the bias due to U. The UPTT curve looks the same as above: the power drops on the left side of the curve due to canceling out from the attenuated treatment effect, and is increased on the right side of the curve because the bias is amplified by the attenuated treatment effect. The CPTT is roughly the mirror image, where the diminished power on the right side of the curve (and the heightened power on the left side) is due to the collider bias (flipped in sign) rather than the attenuated treatment effect.



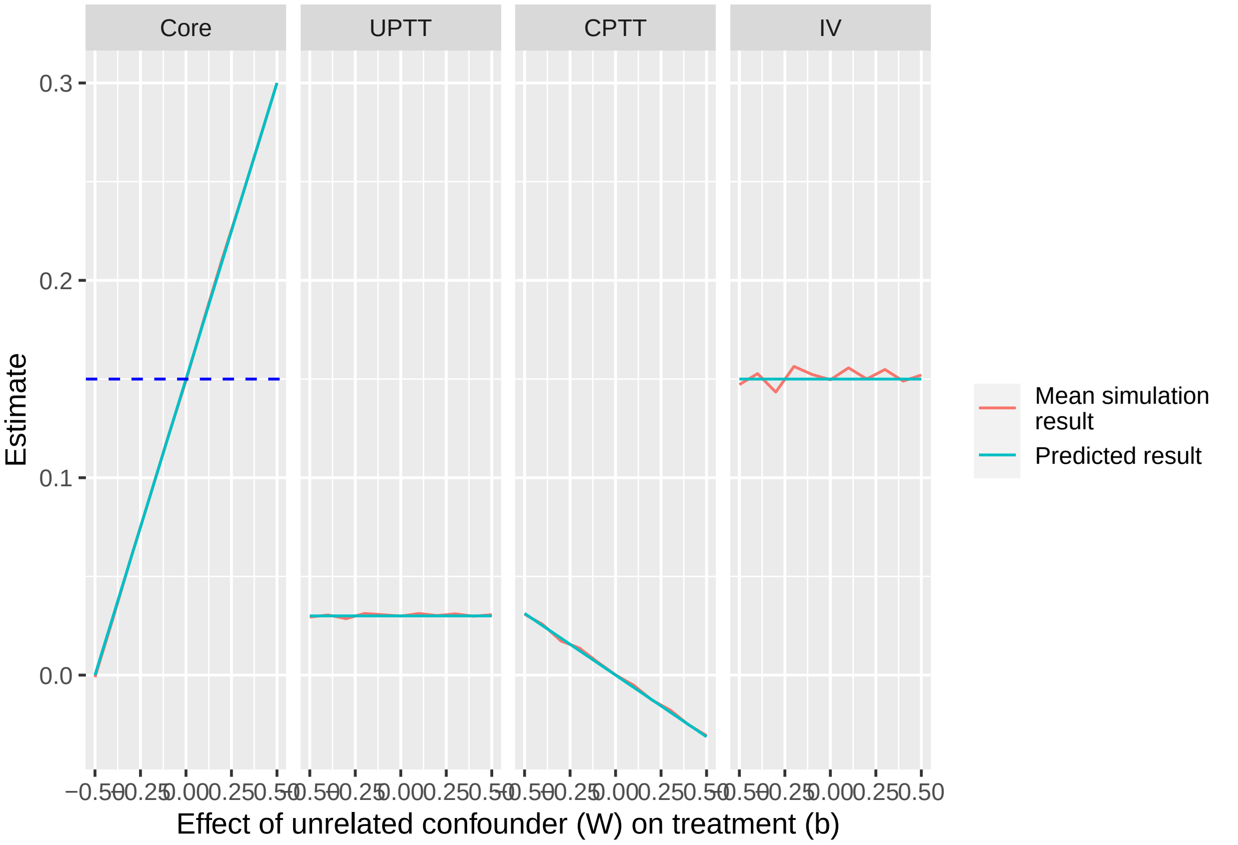
Magical CPTT illustration

For this simulation we set c to 0, so that there is no bias due to U, and we vary d from -0.5 to 0.5 by 0.1.

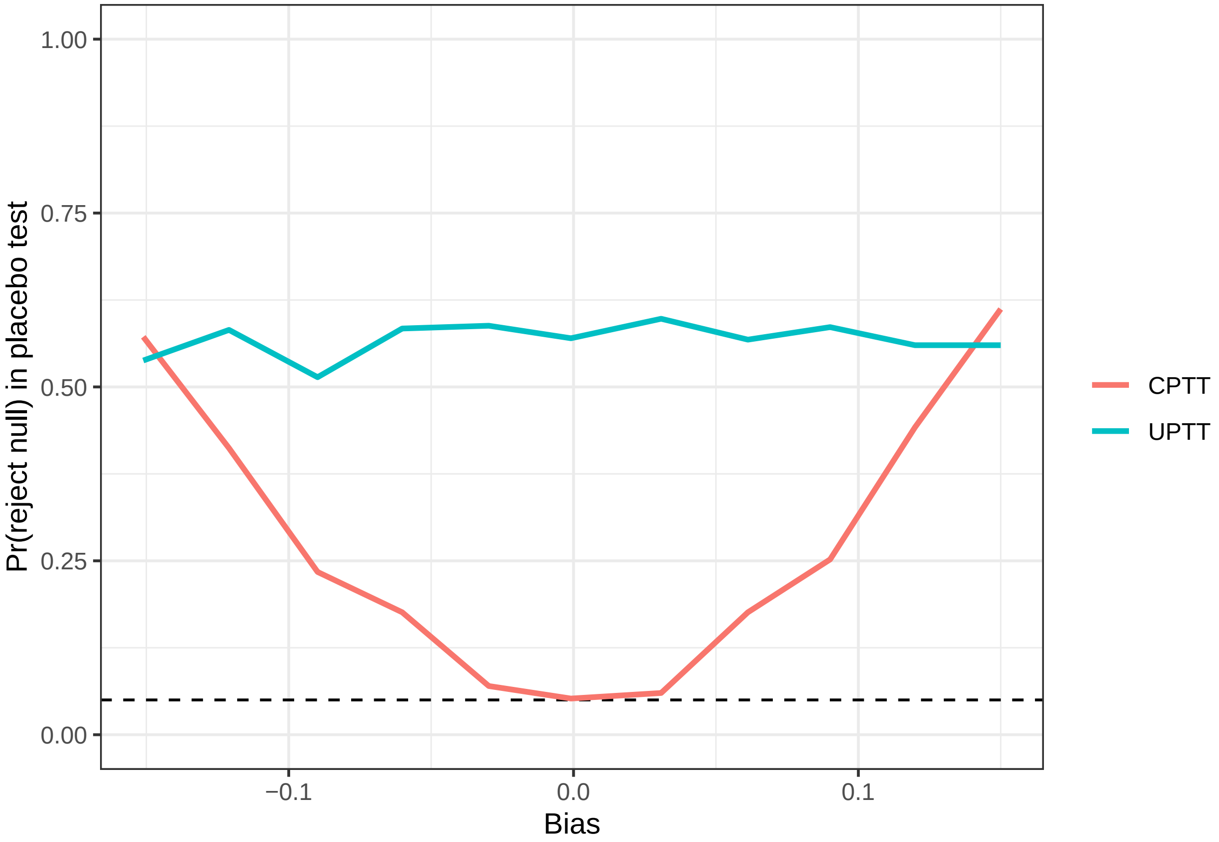
Note that

- the bias in the core analysis (which is due to W) varies with d, the effect of W on the treatment,
- the UPTT reflects the attenuated treatment effect but does not register the bias due to W,
- the (magical) CPTT does register the bias due to W (flipped in sign and attenuated), and
- the IV using the placebo treatment as the instrument eliminates the bias.

Thus when the magical CPTT is possible, one would simply use IV to recover the treatment effect.



Because the UPTT reflects the attenuated treatment effect but does not pick up the bias due to W, the power curve is flat for the UPTT (with a high false positive rate); the power curve has the correct shape for the CPTT, though it is rather flat because the bias due to W is multiplied by $b'b$, which is $0.4 \times 0.5 = 0.2$ in this case and would typically be even lower.



References

Ding, Peng, and Luke W Miratrix. 2015. "To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias." *Journal of Causal Inference* 3 (1): 41–57.

Gerber, Alan S, and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* WW Norton.

Pearl, Judea. 2013. "Linear Models: A Useful 'Microscope' for Causal Analysis." *Journal of Causal Inference* 1 (1): 155–70.

Peisakhin, Leonid, and Arturas Rozenas. 2018. "Electoral Effects of Biased Media: Russian Television in Ukraine." *American Journal of Political Science* 62 (3): 535–50.