Placebo Tests for Causal Inference

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

[1] Department of Political Science, University of Chicago

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

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

Abstract

Placebo tests are increasingly common in applied social science research, but the methodological literature has not previously offered a comprehensive account of what we learn from them. We define placebo tests as tools for assessing the plausibility of the assumptions underlying a research design relative to some departure from those assumptions. We offer a typology of tests defined by the aspect of the research design that is altered to produce it (outcome, treatment, or population) and the type of assumption that is tested (bias assumptions or distributional assumptions). Our formal framework clarifies the extra assumptions necessary for informative placebo tests; these assumptions can be strong, and in some cases similar assumptions would justify a different procedure allowing the researcher to relax the research design's assumptions rather than test them. Properly designed and interpreted, placebo tests can be an important device for assessing the credibility of empirical research designs.

Authors can be reached at aeggers@uchicago.edu, tunon@princeton.edu, and allan.dafoe@politics.ox.ac.uk. For helpful input, we thank Devin Caughey, Thad Dunning, Anthony Fowler, Don Green, Jens Hainmueller, Sophia Hatz, Seth Hill, Neil Malhotra, Luke Miratrix, Jas Sekhon, Laura Stoker, 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, UCSD, 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 the treatment effect has been estimated. Beyond assessing the probability that an association as strong or stronger could have arisen by chance (via null-hypothesis significance testing), researchers often conduct robustness checks to assess how conclusions depend on modeling choices (Neumayer and Plümper 2017), subgroup analyses 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 how remaining confounders might affect the study's conclusions (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.

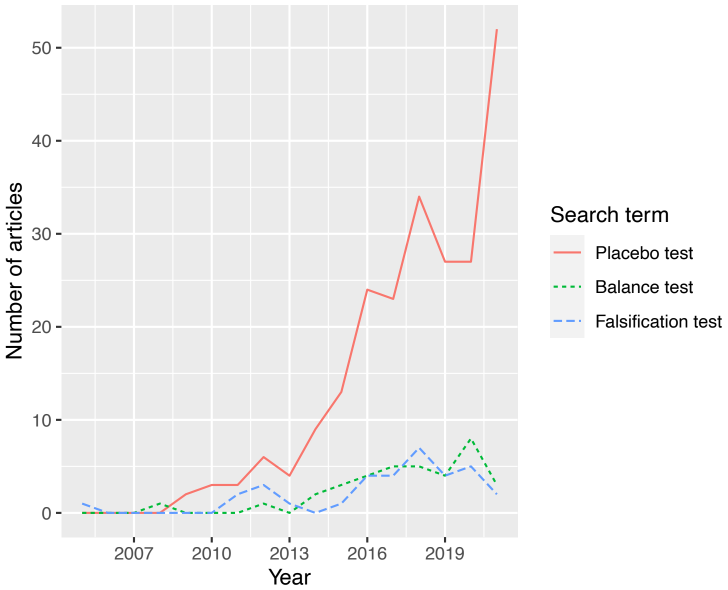
In this paper, we study placebo tests, another form of auxiliary analysis for observational studies. Like the other types just mentioned, placebo tests help assess the credibility of a research finding. The term "placebo test" has its origins in medicine, where a "placebo" originally referred to an ineffective 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 political science, economics, and other social sciences, "placebo test" now refers to a type of auxiliary analysis where the researcher checks for an association that should be absent if the assumptions underlying the design hold but might be present if those assumptions are violated in some relevant way.

As an example, consider two placebo tests presented in Peisakhin and Rozenas (2018)'s study of the effects of Russian news media in Ukraine. Before the 2014 Ukrainian election, TV transmitters in southwest Russia broadcast pro-Russian news programming into Ukraine. Peisakhin and Rozenas (2018) argue that these broadcasts substantially affected the Ukrainian election outcome, partly on the basis that Ukrainian election precincts where Russian news TV signal was stronger voted for pro-Russian parties at a higher rate (conditional on some covariates). To address concerns that precincts with better reception of Russian news broadcasts would have been more supportive of pro-Russian parties anyway, Peisakhin and Rozenas (2018) present several placebo tests. In one, they show that precincts with better Russian *sports* TV signal were not stronger supporters of pro-Russian parties; in another, they show that news TV signal quality and support for Russia were unrelated among Ukrainians who owned satellite TVs and thus did not rely on terrestrial TV signal.

Placebo tests like Peisakhin and Rozenas (2018)'s have become increasingly common in political science in recent years. Figure 1 shows the number of papers appearing on Google Scholar including "placebo test" and closely related terms that were published in seven top political science journals (*APSR*, *AJPS*, *JOP*, *IO*, *BJPS*, *QJPS*, *CPS*) each year between 2005 and 2021. We found no papers mentioning "placebo test" before 2009, but the number has increased fairly steadily thereafter, with over 50 articles in 2021 alone. (By the late 2010s, about 5% of articles mentioning "test" also mentioned "placebo test".) The growing popularity of placebo tests in political science builds on foundational work in statistics (e.g. Rosenbaum 1984, 2002) and compelling applications in adjacent disciplines (e.g. DiNardo and Pischke 1997; Cohen-Cole and Fletcher 2008); it follows Sekhon (2009) and Dunning (2012), who urged political scientists to carry out placebo tests.

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. falsification tests (Pizer 2016), tests for known effects (Rosenbaum 1989), tests of unconfoundedness using pseudo outcomes and pseudo treatments (Imbens and Rubin 2015), tests with negative controls (Lipsitch, Tchetgen Tchetgen and Cohen 2010)). Although several authors formally analyze placebo tests that assess unconfoundedness assumptions (Rosenbaum (1984), Rosenbaum (1989), Lipsitch, Tchetgen Tchetgen and Cohen (2010), Arnold et al. (2016), Imbens and Rubin (2015)), their frameworks do not encompass

Figure 1: The number of articles mentioning "placebo test" and related terms in seven top political science journals, 2005-2021



placebo tests that probe estimation assumptions; many discussions of placebo tests also address only one way of designing tests, such as using a different outcome variable. Moreover, previous discussions provide only cursory guidance about how the results of a placebo test should be interpreted. This omission is particularly important because, as noted by Hartman and Hidalgo (2018), authors tend to present null results in placebo tests as validation of a research design even when the test is severely underpowered; correspondingly, our own informal conversations suggest a widespread perception that, due to selective reporting and "null-hacking" (Protzko 2018), the evidence provided by most placebo tests is dubious at best.

This paper aims to improve the use and interpretation 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. Our main message is that placebo tests have a clear logic, and that closer attention to that logic (both in presenting and interpreting placebo tests) should lead to more informative placebo tests and a higher standard of research. To clarify the logic of placebo tests, Section 2 provides formal conditions under which the plausibility of the assumptions underlying a research design depends on the results of a set of placebo tests. (Briefly, the key requirement is that each test is more likely to "fail" if those assumptions are violated than if they hold; this will be true if the treatment does not affect the outcome in the placebo analysis, but the placebo analysis mirrors the original research design closely enough to reproduce a possible violation of the core assumptions.) We also offer a typology that classifies placebo tests according to what kind of assumption is being tested (bias assumptions, which relate to point estimates, and distributional assumptions, which roughly relate to standard errors) and what aspect of the core analysis is altered (the outcome, treatment, or population). In Section 3 and Section 4 we illustrate each type of test using directed acyclic graphs (DAGs) and examples from political science. In Section 5 we consider alternative approaches that, under similar conditions, relax the assumptions behind a research design rather than testing them. In Section 6 we discuss p-hacking, null-hacking, and other systemic problems that can make published research unreliable; placebo tests are subject to some of the same issues, but placebo tests can also help address these problems, especially if their logic is better understood. Section 7 concludes with a checklist of questions to ask about any placebo test. The concepts and recommendations contained in the paper should help researchers both interpret placebo tests and devise their own, particularly in conjunction with our library of over one hundred placebo tests gathered from recent political science research (Supporting Information).

Perhaps the most important contribution of this paper is to place both placebo tests and the research designs they probe in a statistical hypothesis testing framework. This has several benefits. First, it emphasizes that placebo tests generate false positives and false negatives by design (due to sampling variation), not just because (as Lipsitch, Tchetgen Tchetgen and Cohen (2010) point out) the placebo analysis may fail to reproduce key elements of the core analysis. Second, it allows us to handle in a unified framework placebo tests that aim to check for incorrect standard errors along with tests that aim to probe for bias; these types have previously been considered separately, with the latter attracting far more attention. Third, thinking in terms of hypothesis tests and associated rejection rates allows us to apply Bayes Rule to formalize what is learned from the results of a set of placebo tests, which facilitates discussion of multiple testing, the implications of null-hacking or p-hacking in placebo tests, and how to interpret tests whose assumptions do not exactly hold.

2 A theory of placebo tests

A placebo test is a method for probing the assumptions underlying a research design (which we call the *core assumptions*). In a placebo test, a researcher checks for an association that is more likely to be present if those assumptions are violated than if those assumptions hold. Whether (or how often) a significant association is found thus provides evidence about the validity of the research design's assumptions; doubts about these assumptions could lead to a different design or highlight the need for sensitivity analysis. In this section we formalize the Bayesian logic that we argue best explains this endeavor, specify a set of assumptions that produce an informative test, and introduce a typology of placebo tests.

2.1 What do we learn from placebo tests?

Let H_0 denote the null hypothesis that the research design's core assumptions hold, and let H_1 denote the alternative hypothesis that those assumptions are violated in some well-specified way.[1] Suppose n placebo tests are run (with $n = 1$ an important special case), and assume that each test produces a binary result: a "failing" test is one where we say that the null hypothesis is rejected; a "passing" test is one where it is not rejected. (We postpone for now the details of the rejection rule.) Let p_0 denote the probability of a failing test when H_0 is true (the false positive rate or *size* of the test) and let p_1 denote the probability of a failing test when H_1 is true (the true positive rate, sensitivity, or *power* of the test). For simplicity

[1] Hartman and Hidalgo (2018) recommend reversing the null and alternative. We discuss this proposal in Section 6.

we assume that the n tests all have the same p_0 and p_1 and are conditionally independent; this is easily generalized.

Then given x failing tests, the ratio of the posterior probability of H_0 vs. H_1 (i.e. the posterior odds ratio) is, by Bayes Rule,

$$\begin{aligned}\frac{\Pr(H_0 \mid x \text{ failures in } n \text{ tests})}{\Pr(H_1 \mid x \text{ failures in } n \text{ tests})} &= \frac{\Pr(H_0)}{\Pr(H_1)} \frac{\Pr(x \text{ failures in } n \text{ tests} \mid H_0)}{\Pr(x \text{ failures in } n \text{ tests} \mid H_1)} \\ &= \frac{\Pr(H_0)}{\Pr(H_1)} \frac{p_0^x (1-p_0)^{(n-x)}}{p_1^x (1-p_1)^{(n-x)}}. \qquad (1)\end{aligned}$$

In words, the relative plausibility of the core assumptions (H_0) vs. some departure from those assumptions (H_1), given the test results, is the prior relative plausibility times the ratio of the likelihoods (i.e. the Bayes factor) for obtaining those results under H_0 vs. H_1.[2]

We emphasize four aspects of Equation 1. First, a necessary and sufficient condition for a placebo test to be *informative*, in the sense that a failing result constitutes evidence against the core assumptions and a passing result constitutes evidence *for* those assumptions, is that $p_1 > p_0$, i.e. power > size. Broadly, the higher is p_1 and the lower is p_0 the more informative is the test, i.e. the more a single test result shifts our beliefs.[3] Thus in interpreting a placebo test we should always ask whether (and roughly to what extent) a failing result is more likely if the research design's assumptions are violated than if they hold. This requires assumptions beyond those employed in the research design itself; below we articulate one such set of assumptions in general terms before illustrating how they operate in applications.

Second, although in principle one could quantify all of the components of Equation 1, in general we view this as a heuristic for understanding the logic of placebo tests (and hypothesis tests more generally) rather than a quantitative measure to compute. Given assumptions we discuss in the next section, p_0 is close to the nominal size of the test (e.g. .05); the precise

[2] Royall (1997) p. 48-49 similarly characterizes the posterior relative odds of two simple hypotheses given a test result and the size and power of the test.

[3] More precisely, the degree to which a single test result shifts the log of Equation 1 is given by the log of $\frac{p_0}{p_1}\frac{1-p_1}{1-p_0}$, which in medical testing is called the "diagnostic odds ratio" (Glas et al. 2003). That literature uses "sensitivity" and "specificity" where we use power and (one minus) size; it uses "discriminatory ability" or "discriminatory performance" where we use informativeness.

value of p_1, by contrast, depends on the assumed data generating process under H_1. Rather than specifying that DGP and computing p_1, we typically seek to reason more heuristically about whether p_1 likely exceeds p_0 by a small or large amount.

Third, placebo tests produce false positives and false negatives, and the results should be interpreted probabilistically in light of these error rates. Assuming $p_0 > 0$, a single failing placebo test is not definitive proof that the core assumptions do not hold; assuming $p_1 < 1$, a single passing test is not definitive proof that these assumptions hold. Furthermore, there may be many ways the core assumptions could be violated, and a given test may be informative about only some of those violations. Thus placebo tests are imperfectly informative not just because (as Lipsitch, Tchetgen Tchetgen and Cohen (2010) point out) there may be flaws in the research design that the placebo test does not reproduce, or the reverse (flaws in the placebo analysis that are not present in the original research design), but also because hypothesis tests are *designed* to produce false positives (due to sampling variation) and inevitably produce false negatives due to finite statistical precision. Given enough tests, a mix of passing and failing tests may be likely under both H_0 and H_1; thus Equation 1 gives guidance about how to interpret multiple tests.

Finally, the prior plausibility of the contemplated departure H_1 matters. The results of the placebo test(s) might be more consistent with some H_1 than with H_0, but H_0 could still be more plausible than H_1 if other information strongly favors H_0 over H_1. For example, in an experiment where treatment was assigned within matched pairs by a coin flip we might detect a degree of covariate imbalance that would be more likely if the coin were weighted than if it were fair, but given the apparent impossibility of weighting a coin (Gelman and Nolan 2002) we would still tend to believe that the coin was fair. Accordingly, we should design placebo tests that are less likely to fail if the core assumptions hold than if there is some *plausible* departure from those assumptions.

2.2 Formal conditions for an informative placebo test

We now offer a set of sufficient conditions under which a placebo test can be informative about a research design's core assumptions – that is, conditions under which the test's power p_1 is greater than its size p_0. These conditions do not describe every informative test,[4] but they help to illuminate the logic behind most tests we encounter in the empirical literature.

We start with the research design itself, i.e. the core analysis, to clarify the role of the assumptions we seek to test. The core analysis produces an estimate $\hat{\delta}$ of the average effect of a treatment on an outcome using a sample from some population. This estimate trivially can be decomposed into the true average treatment effect δ, the bias $b \equiv E[\hat{\delta}] - \delta$, and sampling error $\varepsilon \equiv \hat{\delta} - E[\hat{\delta}]$, i.e.

$$\hat{\delta} = \delta + b + \varepsilon.$$

In the core analysis the researcher seeks to use the observed estimate $\hat{\delta}$ to test the null hypothesis that $\delta = 0$. Doing so requires two sets of assumptions. The *bias assumptions* $\mathcal{BA}$ jointly imply $b = 0$. In general, bias assumptions encompass assumptions about identification, estimation, measurement, and sample selection[5] that allow for unbiased estimates of δ. The *distributional assumptions* $\mathcal{DA}$ relate to the sampling distribution of ε and jointly imply $\Pr(\varepsilon \in R) \leq \alpha$ for a chosen α and corresponding two-sided rejection region R; for example $\mathcal{DA}$ could be assumptions about the (in)dependence of observations implying that ε (and therefore $\hat{\delta}$) is normally distributed with a given variance. It follows that

$$\Pr(\hat{\delta} \in R \mid \delta = 0 \wedge \mathcal{IA} \wedge \mathcal{EA}) \leq \alpha,$$

i.e. the false positive rate in testing the null hypothesis of "no effect" is at most the nominal rate α given the core assumptions $\mathcal{BA}$ and $\mathcal{DA}$.

[4] Biased but consistent estimators do not rely on the assumption that bias is exactly zero, for example.

[5] Arnold et al. (2016) discuss placebo tests in epidemiology for detecting measurement bias and selection bias.

The placebo test assesses the plausibility of these core assumptions. The placebo analysis is an altered version of the core analysis that produces an estimate $\hat{\delta}_p$ that is analogous to $\hat{\delta}$ and similarly can be decomposed into treatment effect, bias, and sampling error:

$$\hat{\delta}_p = \delta_p + b_p + \varepsilon_p.$$

The researcher seeks to use the observed estimate $\hat{\delta}_p$ to test the null hypothesis that the core assumptions hold. Doing so requires further assumptions.

Assumption 1 (i.e. No Average Treatment Effect, or NATE): $\delta_p = 0$

NATE simply states that the treatment has no average effect on the outcome in the placebo analysis. (This justifies using the term "placebo test".)

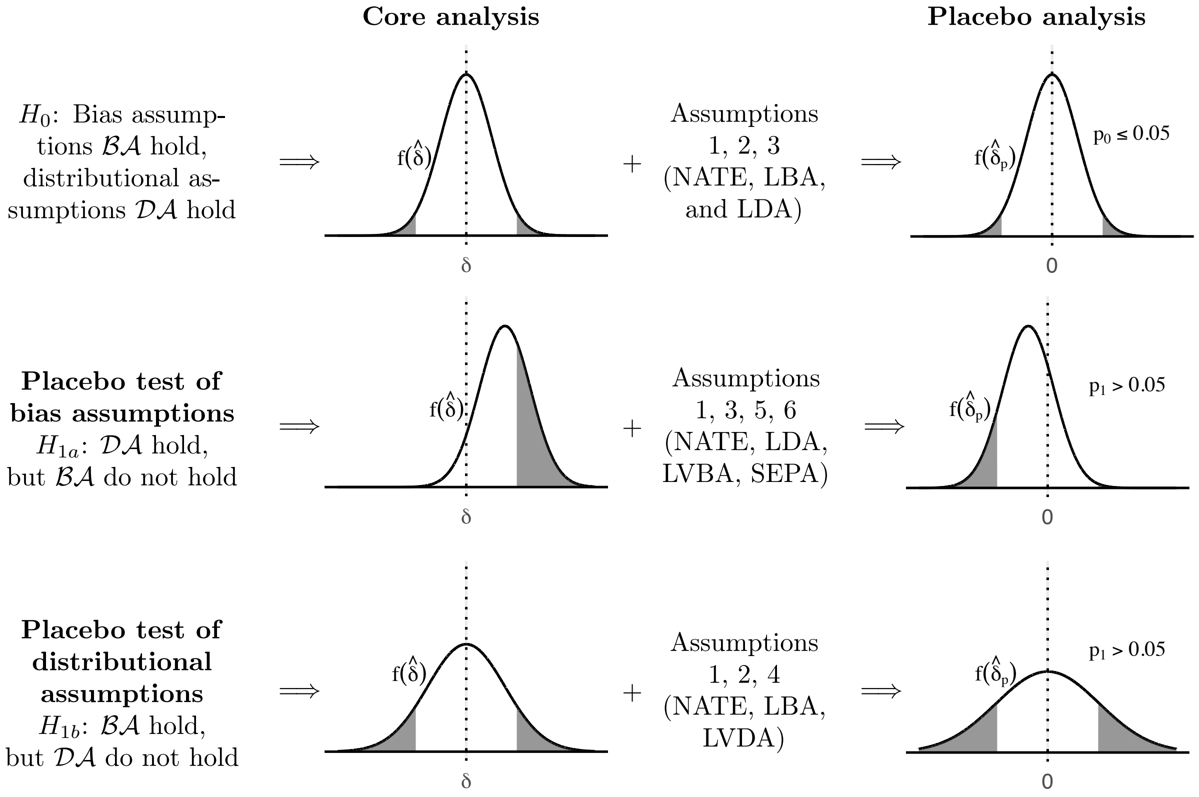
Assumption 2 (Linked Bias Assumptions, LBA): $\mathcal{BA} \implies b_p = 0$

Assumption 3 (Linked Distributional Assumptions, LDA): $\mathcal{DA} \implies \Pr(\varepsilon_p \in R_p) \leq \alpha_p$

LBA states that if the bias assumptions hold in the core analysis, then they also hold in the placebo analysis; similarly, LDA states that if the distributional assumptions hold in the core analysis, then they also hold in the placebo analysis.

NATE, LBA, and LDA jointly imply that if the core assumptions hold, then p_0 (the probability of a failing placebo test) is at most α_p, the nominal size of the test. Figure 2 illustrates the logic. (The Supporting Information contains proof of this and subsequent logical claims in this section.) Let H_0 refer to the null hypothesis that the core assumptions $\mathcal{BA}$ and $\mathcal{DA}$ hold. As shown in the diagram labeled "Core analysis" in the first row of Figure 2, the sampling distribution of the estimator $\hat{\delta}$ in the core analysis ($f(\hat{\delta})$) is centered on the average treatment effect δ (i.e. there is no bias), with mass in the tail no larger than α (here, .05). Under assumptions 2 and 3 (LBA and LDA), the sampling distribution of the estimator $\hat{\delta}_p$ ($f(\hat{\delta}_p)$) inherits the good properties of $f(\hat{\delta})$, and under NATE it is centered on zero; thus under H_0 $\hat{\delta}_p$ will be found in the rejection region with probability $p_0 \leq .05$. This is illustrated in the diagram labeled "Placebo analysis" in the first row of Figure 2.

Figure 2: Sufficient conditions for informative placebo tests


Core analysis
Placebo analysis
H_0: Bias assumptions $\mathcal{BA}$ hold, distributional assumptions $\mathcal{DA}$ hold
$\Longrightarrow$
$f(\hat{\delta})$
δ
+
Assumptions 1, 2, 3 (NATE, LBA, and LDA)
$\Longrightarrow$
$f(\hat{\delta}_p)$
$p_0 \leq 0.05$
0
Placebo test of bias assumptions
H_{1a}: $\mathcal{DA}$ hold, but $\mathcal{BA}$ do not hold
$\Longrightarrow$
$f(\hat{\delta})$
δ
+
Assumptions 1, 3, 5, 6 (NATE, LDA, LVBA, SEPA)
$\Longrightarrow$
$f(\hat{\delta}_p)$
$p_1 > 0.05$
0
Placebo test of distributional assumptions
H_{1b}: $\mathcal{BA}$ hold, but $\mathcal{DA}$ do not hold
$\Longrightarrow$
$f(\hat{\delta})$
δ
+
Assumptions 1, 2, 4 (NATE, LBA, LVDA)
$\Longrightarrow$
$f(\hat{\delta}_p)$
$p_1 > 0.05$
0

NOTE: If the bias assumptions ($\mathcal{BA}$) and distributional assumptions ($\mathcal{DA}$) behind the core analysis hold (H_0), then the sampling distribution of the estimator $\hat{\delta}$ in the core analysis ($f(\hat{\delta})$) is centered on the true value δ with the correct mass in the tails (top left diagram). Assumptions 1-3 and H_0 jointly imply that the sampling distribution of $\hat{\delta}_p$ in the placebo analysis ($f(\hat{\delta}_p)$) has the same good properties, so that the probability of a false positive in the placebo test is at most the nominal size α_p (here, .05). If $\mathcal{BA}$ fails and $f(\hat{\delta})$ is *not* centered on the true value δ (second row), then Assumptions 1, 3, 5, and 6 imply that $f(\hat{\delta}_p)$ is also not centered, producing a true positive rate $p_1 > \alpha_p$. Similarly, if $\mathcal{DA}$ fails and $f(\hat{\delta})$ has excessive mass in the tails (second row), then Assumptions 1, 2, and 4 imply that $f(\hat{\delta}_p)$ also has excessive mass in the tails, producing a true positive rate $p_1 > \alpha_p$.

We now turn to the test's true-positive rate p_1 (power). We distinguish between two types of test, depending on what alternative hypothesis is being considered. In a *placebo test of distributional assumptions*, the alternative hypothesis (call it H_{1b}) is that the bias assumptions hold but the distributional assumptions fail. The following assumption states that, if the estimation assumptions fail in the core analysis in the way contemplated by H_{1b}, they also fail in the placebo analysis:

Assumption 4 (Linked Violation of Distributional Assumptions, LVDA): $H_{1b} \implies \Pr(\varepsilon_p \in R_p) > \alpha_p$

If Assumptions 1, 2, and 4 (NATE, LBA, and LVDA) hold, then under H_{1b} the test's true-positive rate p_1 exceeds the test's nominal size. (Thus Assumptions 1, 2, 3, and 4 are jointly sufficient for an informative placebo test of distributional assumptions.) This is illustrated in the bottom row of Figure 2. H_{1b} implies an excessive false-positive rate in the core analysis due to fat tails in the sampling distribution of $\hat{\delta}$ (or, equivalently, a mis-specified rejection region): if $\delta = 0$, $\hat{\delta}$ would fall in the rejection region at a rate above $\alpha = .05$. Assumptions 1, 2 and 4 imply that $f(\hat{\delta}_p)$ will be centered on 0 but, like $f(\hat{\delta})$, will have a mis-specified rejection region, producing a true-positive rate p_1 above the test's size.

In a *placebo test of bias assumptions*, the alternative hypothesis (call it H_{1a}) is that the distributional assumptions hold but the bias assumptions fail. For these tests, we invoke the following two assumptions:

Assumption 5 (Linked Violation of Bias Assumptions, LVBA): $H_{1a} \implies b_p \neq 0$

That is, when the core analysis is biased, so is the placebo analysis.

Assumption 6 (Sampling error in placebo analysis, SEPA): The sampling distribution of $\hat{\varepsilon}_p$ (and therefore $\hat{\delta}_p$) is unimodal and symmetric, with a strictly increasing distribution function.

If Assumptions 1, 3, 5, and 6 (NATE, LDA, LVBA, SEPA) hold, then under H_{1a} the test's true-positive rate p_1 exceeds the test's nominal size. (Thus Assumptions 1, 2, 3, 5, and 6 jointly sufficient for an informative placebo test of bias assumptions.) This is illustrated in the middle row of Figure 2. H_{1a} implies an excessive false-positive rate in the core analysis because $f(\hat{\delta})$ is not centered on δ (i.e. $\hat{\delta}$ is biased). Assumptions 1, 3, and 5 imply that, under H_{1a}, $f(\hat{\delta}_p)$ will also not be centered on 0 and, given Assumption 6, this implies a true-positive rate p_1 above the test's size.[6]

To summarize and simplify, an informative placebo analysis typically exhibits two key properties. First, there is no effect of treatment (NATE). Second, the placebo analysis mirrors the core analysis in the following respect: if testing for no effect in the core analysis is reliable then it is also reliable in the placebo analysis (LBA and LDA), but if there is a problem with bias or standard errors in the core analysis then the placebo analysis would inherit that problem (LVBA or LVDA).

2.3 A typology of placebo tests

To better understand the challenges of designing and interpreting placebo tests, we examined every paper mentioning a "placebo test," "balance test," or "falsification test" in the *APSR*, *AJPS*, *JOP*, and *IO* between 2009 and 2018. In analyzing the resulting list of 110 placebo tests (which we summarize in the Supporting Information),[7] we found it useful to categorize tests according to two features.

The first feature (mentioned above) is which assumptions are being tested – bias assumptions or distributional assumptions. The second feature is how the placebo analysis differs from the core analysis. In general, the placebo analysis is a replication of the core analysis with one of three components altered. We describe a test that uses a different outcome variable as a *placebo outcome test*, we describe a test that uses a different treatment variable as a *placebo treatment test*, and we describe a test that uses a different population as a *placebo population test*. We use the terms *placebo outcome*, *placebo treatment*, and *placebo population*

[6] The sign of the biases b and b_p may be the same or different as in Figure 2.

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

Figure 3: Schematic illustrating typology and key terms

Core analysis: Estimates effect of
a **treatment**
on an **outcome**
in a **population**
based on
bias assumptions
and
distributional assumptions.

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**
to test (given additional assumptions)
bias assumptions
or
distributional assumptions.

to refer to the component that has been altered in each case.[8] The formal framework just presented helps clarify both why placebo tests alter the core analysis and why these alterations should be minimal: the alteration ideally shuts down the treatment effect, so that NATE holds; the alteration should be minimal, however, so that the placebo analysis retains key features of the core analysis that could violate the core assumptions.

While our typology (summarized in Figure 3) is helpful for analyzing and creating placebo tests, in some cases a test could arguably be classified as more than one type. Tests that examine the effect of "fake cutoffs" in regression discontinuity designs, for example, could be considered either placebo population tests or placebo treatment tests, depending in part on the estimation strategy. A simple parallel trends test can be seen as a placebo outcome test (where we replace the outcome with a lagged version of the outcome) or a placebo treatment test (where we replace the treatment by a future value of the treatment). Despite these ambiguities, we find the typology useful in making sense of the wide range of practices we observe in applied research.

[8] Rosenbaum (1984) notes that one can test the assumption of strongly ignorable treatment assignment using "unaffected responses", "essentially equivalent treatments", or "unaffected units", which are analogous to placebo outcomes, treatments, and populations in our typology. Rosenbaum presents these as "special cases of a more general formulation" (p. 44), but our survey shows that these three types account for nearly all applied tests in political science, including tests not designed to test strongly ignorable treatment assignment.

3 Designing placebo tests of bias assumptions

To illustrate how the above logic can be applied in designing informative placebo tests for bias, we use a combination of directed acyclic graphs (DAGs)[9] and examples.

3.1 The typical logic, simplified and illustrated

We begin by using simple DAGs to illustrate the typical logic of each type of test in the case where attention centers on a possible omitted variable.[10] A researcher seeks to measure the average effect of a treatment D on an outcome Y, as depicted in the top left panel of Figure 4. There is an unobserved variable U that is believed to affect Y. The researcher assumes that U does not affect D, i.e. that the dashed line connecting U and D can be erased completely. Given this assumption, the effect of D on Y is non-parametrically identified. If (contrary to the researcher's assumption) U does affect D, then dependence between D and Y may also reflect confounding due to U. The purpose of the placebo test is to assess the researcher's assumption that U does not affect D.

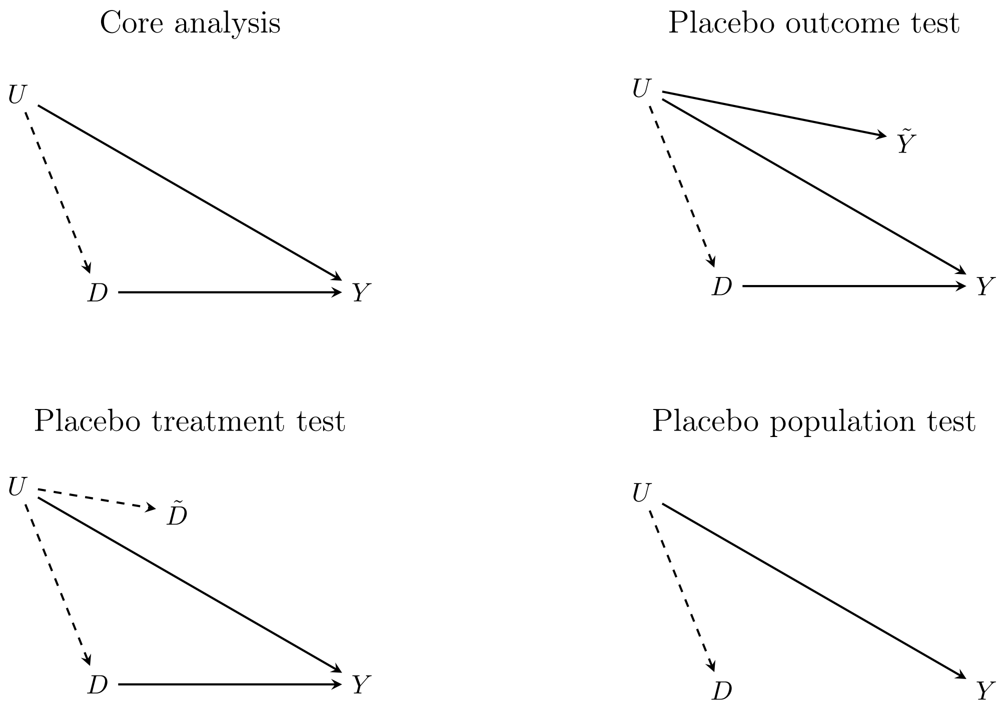
In a typical placebo outcome test (top right panel of Figure 4), the researcher locates a variable $\tilde{Y}$ that is affected by (or otherwise associated with) U but is not affected by D. The researcher then replicates the core analysis replacing Y with the placebo outcome $\tilde{Y}$. Given the assumed relationship between U and $\tilde{Y}$, finding an association between D and $\tilde{Y}$ would call into question the researcher's identification assumption.

In a typical placebo treatment test (bottom left panel of Figure 4), the researcher locates a variable $\tilde{D}$ that does not affect Y but would be affected by U in a similar way as D. The researcher then replicates the core analysis replacing D with the placebo treatment $\tilde{D}$. Given the assumed similarity between the effect of U on D and the effect of U on $\tilde{D}$, finding an

[9] Lipsitch, Tchetgen Tchetgen and Cohen (2010) similarly illustrate the logic of placebo tests in epidemiology with DAGs. For an accessible introduction see Huntington-Klein (2021).

[10] The logic is similar when concern focuses on measurement or estimation (was an observed X measured/controlled for correctly?) rather than identification (is unobserved U a confounder?).

Figure 4: The typical logic of placebo tests for bias, simplified



Core analysis

Placebo outcome test

Placebo treatment test

Placebo population test

association between $\tilde{D}$ and Y (conditional on D) would call into question the researcher's identification assumption.

Finally, in a placebo population test (bottom right panel of Figure 4), the researcher locates a placebo population where D does not affect Y but U would affect D in a similar way as in the core population. (U is also assumed to affect Y in both populations.) It follows that any systematic dependence between D and Y in this population arises from confounding due to U, which (given the assumed similarity between D's relationship to U in the placebo population and the core population) calls into question the researcher's identification assumption.

In each case, the DAG encodes the No Average Treatment Effect (NATE) assumption: there is no direct path from the treatment to the outcome in the placebo analysis. The LBA and LVBA assumptions are reflected in the assumed similarity across DAG edges: the placebo outcome $\tilde{Y}$ and Y are assumed to be similarly affected by U, as are the placebo treatment $\tilde{D}$ and D; U's effect on D and Y in the placebo population is assumed to be similar to its effect in the core population. These similarity claims are essential to an informative placebo test, and they typically require careful consideration about the substantive application and the relevant threats to inference.

3.2 Examples of placebo tests of bias assumptions

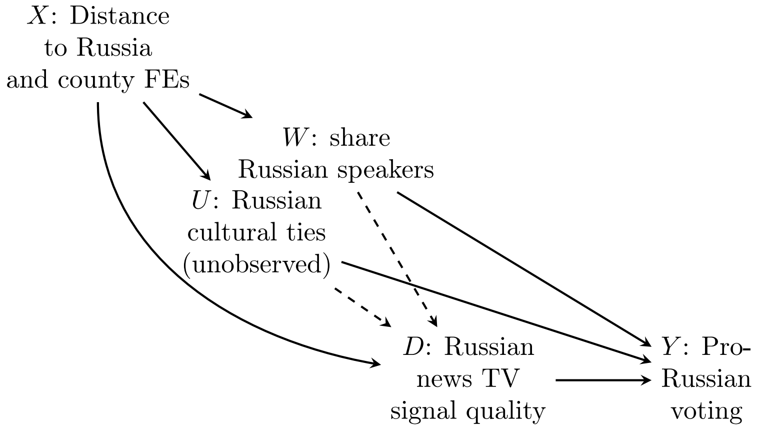
Examples help to show how this logic is used in applied research. We focus on examples from political science, making repeated reference to Peisakhin and Rozenas (2018)'s study of the effects of Russian news media in Ukraine, which includes an unusually large number and variety of placebo tests.

Placebo outcome tests

As described in the Introduction, 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. In their precinct-level analysis, Peisakhin and Rozenas (2018) seek to measure the average effect of the quality of the Russian news TV signal in Ukrainian election precincts on precinct voting outcomes. Peisakhin and Rozenas (2018)'s identifying assumption is that, conditional on a flexible function of the precinct's distance to Russia and county (or district) fixed effects, signal quality is independent of potential outcomes (i.e. of underlying political support for pro-Russian parties). The main concern is that, perhaps due to strategic transmitter location, signal quality might be better in places whose residents are more predisposed to support Russia, even conditional on Peisakhin and Rozenas (2018)'s controls.

Peisakhin and Rozenas (2018) address this concern in part with placebo outcome tests that use pre-treatment covariates (such as the percentage of Russian speakers in the precinct) as placebo outcomes. The DAG in Figure 5 illustrates the logic of the test, extending Figure 4 to include control variables. The research aim is to estimate the effect of D (Russian news TV signal quality) on Y (voting results). Concern centers on potential confounders W (the

Figure 5: Simplified DAG for Peisakhin and Rozenas (2018)'s placebo outcome test



Note: Peisakhin and Rozenas (2018) seek to estimate the effect of D on Y. Their identification assumption is that X is a sufficient conditioning set, i.e. that the dashed-line paths can be erased from the DAG. They use W as a placebo outcome.

percent of Russian speakers, observed) and U (Russian cultural ties more generally, assumed unobserved). Peisakhin and Rozenas (2018)'s identification assumption is that the dashed paths can be erased, so that it is sufficient to condition on X. In the placebo outcome test, Y is replaced by W. Given the authors' identification assumption, D and W are independent conditional on X; a significant conditional association would cast doubt on that assumption.

In general, the key assumptions necessary for a placebo outcome test using a pre-treatment placebo outcome (often called a "balance test") are uncontroversial. In terms of the formal framework above, NATE is guaranteed because W is observed before the treatment is realized, and LBA and LVBA follow trivially from the assumption that W is itself a potential confounder.

In placebo tests with post-treatment placebo outcomes, the NATE assumption is not guaranteed and should be defended. For example, Dube, Dube and García-Ponce (2013) study the impact of the federal U.S. assault weapons ban on the murder rate in adjacent Mexican states. One may be concerned that the association Dube, Dube and García-Ponce (2013) detect between assault weapon availability in the US and murders in Mexico is due to other factors that coincided with the ban and caused a drop in violence. To assess this

Table 1: Examples of post-treatment placebo outcome tests (more in Supporting Information)

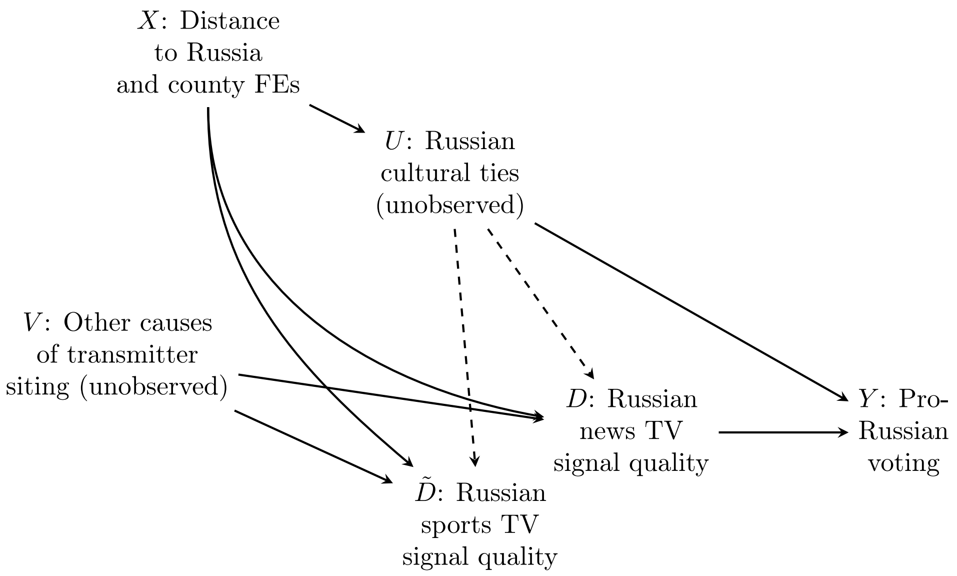
Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Dube, Dube & Garcia-Ponce (2013)	Mexican municipalities located close to U.S. border, 2002-2006	Assault weapon availability from neighboring US state	Gun-related homicides	Accidents, non-gun homicides, and suicides
Cruz & Schneider (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 in municipality are made by popular vote	Rate of naturalization through ordinary municipal process	Rate of naturalization through centralized facilitated process

concern, the authors use the rate of death by suicide and the rate of death by accidents as placebo outcomes. These placebo outcomes could be considered proxies for potential confounders that tend to produce disorder; in that sense the logic is similar to the logic motivating balance tests. But NATE is not guaranteed in this case: an assault weapons ban in the U.S. could in principle affect subsequent suicide or accident rates in Mexico. Authors using post-treatment placebo outcomes should therefore explain why NATE should hold. Table 1 lists two other examples.

Placebo treatment tests

The DAG in Figure 6 illustrates the logic of Peisakhin and Rozenas (2018)'s placebo treatment test. According to the authors, transmitters broadcasting Russian news differ from transmitters broadcasting Russian sports and other entertainment programming; in the DAG, sports TV signal quality ($\tilde{D}$) is potentially affected by the same potential confounders W and U that might confound the relationship between news TV signal quality and pro-

Figure 6: Simplified DAG for Peisakhin and Rozenas (2018)'s placebo treatment test


X: Distance to Russia and county FEs
U: Russian cultural ties (unobserved)
V: Other causes of transmitter siting (unobserved)
D: Russian news TV signal quality
Y: Pro-Russian voting
$\tilde{D}$: Russian sports TV signal quality

Note: Peisakhin and Rozenas (2018) seek to estimate the effect of D on Y. Their identification assumption is that X is a sufficient conditioning set, i.e. that the dashed-line paths can be erased from the DAG. They use $\tilde{D}$ as a placebo treatment.

Russian voting, but sports TV signal quality is assumed to not affect voting results Y. The effect of W and U on $\tilde{D}$ is assumed to be similar to the effect of W and U on D: either these variables don't affect $\tilde{D}$ and D (i.e. the dashed paths can be erased) or they affect both. This encodes LBA and LVBA. Given the DAG shown, and assuming that dashed paths can be erased, $\tilde{D}$ and Y are independent conditional on X and D; a significant conditional association would raise concerns about the authors' identification assumption.

In a placebo treatment test where the placebo treatment is realized before the outcome, the NATE assumption requires justification. In Peisakhin and Rozenas (2018)'s case the question is whether Russian sports broadcasting could affect Ukrainian political behavior; Peisakhin and Rozenas assume it does not, though sports have been found to impact politics in other settings (e.g. Depetris-Chauvin, Durante and Campante 2020). LBA would also fail to hold if e.g. sports TV transmitters were strategically sited even though news TV transmitters were not. If sports broadcasts could affect voting (violating NATE), or were subject to confounding not found in the core analysis (violating LBA), then the false positive rate p_0 could be higher than α_p, making the test less informative.

In our survey of placebo treatment tests (three of which are included in Table 2), we observed that many authors control for the actual treatment (including Peisakhin and Rozenas (2018); Burnett and Kogan (2017); Dasgupta, Gawande and Kapur (2017)) while others do not (including Jha (2013); Stasavage (2014); Fouirnaies and Mutlu-Eren (2015)). The DAG in Figure 6 highlights the main reason to condition on the actual treatment. Suppose the authors are correct that dashed paths can be erased. If we condition on X (but not on D), then $\tilde{D}$ remains connected to Y through V and D. (The path is $\tilde{D} \leftarrow V \rightarrow D \rightarrow Y$.) Although V is not a confounder for estimating the effect of D on Y, it is a confounder for estimating the effect of $\tilde{D}$ on Y (a violation of LBA). Conditioning on D closes this path. More generally, the reason to condition on the actual treatment in a placebo treatment test is that the placebo treatment and actual treatment may be correlated due to common causes that are not themselves potential confounders; if we do not condition on the actual treatment and the treatment has an effect on the outcome, we may find a significant association in the placebo test due to this correlation even when the core analysis is unbiased. If confounding is the only reason for $\tilde{D}$ and D to be related (conditional on covariates), then $\tilde{D}$ should be used as a placebo outcome instead. The reason for conducting a placebo treatment test is that $\tilde{D}$ and D share non-confounding causes; this is also the reason one should condition on the actual treatment in such a test.

Placebo population tests

Peisakhin and Rozenas (2018)'s individual-level analysis is depicted in Figure 7. Survey respondents were asked whether they watched Russian TV news (D) and how they voted (Y); the quality of Russian news TV signal is used as an instrumental variable (Z), with controls again including distance to Russia and county/district fixed effects. Concern focuses on confounders (represented here by Russian cultural ties, U) and also the exclusion restriction

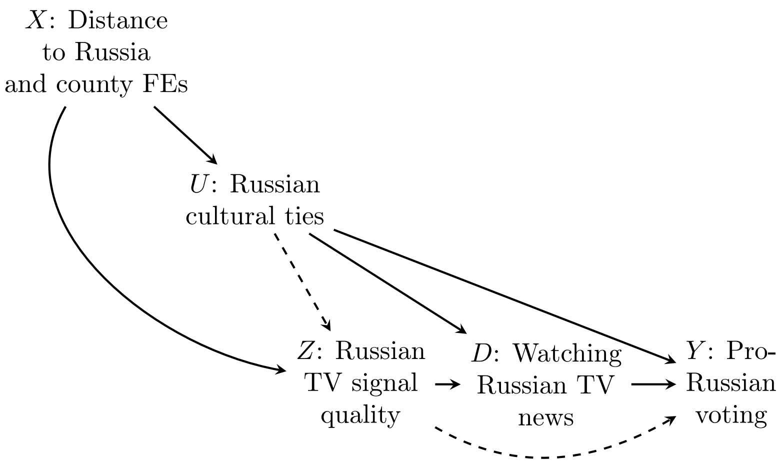
Table 2: Examples of placebo treatment tests of bias (more in Supporting Information)

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
Enos, Kaufman and Sands (2017)	Precincts in LA	Proximity to riot activity in 1992	Diff. in support for spending on public schools 1990-1992	Proximity to areas w. large African-American pop. but no riot activity

for the IV, i.e. the assumption that signal quality affects vote choice only through Russian TV consumption. The standard identification assumptions for the IV imply that the dashed-line paths can be omitted from the DAG: exogeneity requires that Russian cultural ties (U) and other potential confounders do not affect signal quality (Z), and the exclusion restriction requires that signal quality (Z) affects vote choice (Y) only through Russian TV consumption (D).

Peisakhin and Rozenas (2018)'s placebo population test addresses both concerns by shifting the analysis to a (sub-)population for whom Russian TV signal quality arguably would not affect the decision to watch Russian TV: Ukrainians who don't watch terrestrial TV because e.g. they have satellite TVs. In this population, Peisakhin and Rozenas assert, the path from Z to D in Figure 7 can be erased (NATE). It follows that, assuming the rest of the DAG is the same for the two populations and the authors' identification assumptions hold, signal quality (Z) should be independent of voting behavior (Y) conditional on covariates X in the placebo population; finding otherwise casts doubt on the exogeneity and exclusion assumptions made in the core IV analysis.

Figure 7: Peisakhin and Rozenas (2018)'s individual-level analysis (illustrating the logic of the placebo population test)


X: Distance to Russia and county FEs
U: Russian cultural ties
Z: Russian TV signal quality
D: Watching Russian TV news
Y: Pro-Russian voting

NOTE: Peisakhin and Rozenas (2018) seek to estimate the effect of D on Y in survey data. Their key identification assumptions are that X is a sufficient conditioning set for estimating the effect of Z on Y and that Z affects Y only through D, i.e. that the dashed-line paths can be erased from the DAG. They repeat the analysis in a population of Ukrainians who do not watch terrestrial TV (the "placebo population").

NATE here states that signal quality does not affect consumption of Russian news among Ukrainians without terrestrial TVs. This seems reasonable, though it could fail (increasing the rate of false positives) if Ukrainians with satellite dishes attempt to watch the same programs their neighbors are watching. LBA and LVBA are more doubtful. Perhaps satellite TV owners are richer and more mobile than terrestrial TV watchers, which could mean that there are forms of confounding in the placebo population that are not present in the core population, leading to inflated size (high p_0); it could also be that, due to higher mobility in the placebo population, confounding (if present) is weaker in the placebo population than the core population, leading to low power (low p_1). Power could also be low if there is low statistical precision due to small sample size or limited variation in the placebo analysis; similar concerns could also arise in placebo outcome and placebo treatment tests.

Table 3 summarizes three more placebo population tests from our survey. (The Supporting Information includes several others.) Chen (2013)'s core analysis compares turnout in the November 2004 U.S. election between Americans who received FEMA aid and those who

Table 3: Examples of placebo population tests of bias (more in Supporting Information)

Paper	**Core analysis**			**Placebo population**
	Population	Treatment	Outcome	
Acharya, Blackwell and Sen (2016)	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
Erikson and Stoker (2011)	Draft-eligible, college-bound men	Lottery draft number in 1969	Attitude toward Vietnam War in 1973	Non-college bound men; college-bound women

applied but did not receive FEMA aid; to assess the possibility that applicants awarded aid were inherently more likely to vote than those not awarded aid, Chen carries out the same comparison among applicants who applied *after* the election. The key assumptions are that any confounding would be similar in this population (LBA and LVBA), while the award of aid could not affect turnout decisions in an election that had already occurred (NATE).

Studies using regression discontinuity designs (RDDs) very often include a placebo test in which the basic design is replicated at arbitrarily chosen "fake cutoffs" that do not affect any treatment (e.g. Folke, Persson and Rickne 2016).[11] In many cases it is not clear what if any assumption these tests are informative about. Cattaneo, Idrobo and Titiunik (2020, p. 89) describe them as tests of the continuity of potential outcomes, which is the key identification assumption behind most RDDs. But typically the concern is not that the CEF is jumpy *everywhere*; rather, we are concerned that the CEF is discontinuous at the threshold because the treatment might induce precise sorting or might be paired with another treatment. A placebo test using cutoffs elsewhere is not informative about these threats. Fake-cutoff placebo tests are potentially more informative about the unbiasedness (or more generally coverage rates) of the estimation procedure: when we apply it to a CEF we *know*

[11] We consider such tests to be placebo population tests when they use none of the units in the core analysis; if there is overlap, they are better thought of as placebo treatment tests.

is continuous (i.e. away from the threshold), how often do we reject the null? As such, researchers should test many fake cutoffs (not just a handful as shown in Cattaneo, Idrobo and Titiunik (2020)) to report a credible estimate of the false positive rate, and they should discuss whether (due to differences in e.g. the curvature of the CEF, dependence across units, or the density of observations) the false positive rate might be different at the threshold vs. elsewhere.

4 Designing placebo tests of distributional assumptions

A less common type of placebo test checks for false positives that arise due to incorrect standard errors rather than bias. The question is typically whether the false positive rate in the core analysis is the nominal rate α or something larger.

An example in political science is Fowler and Hall (2018), who use placebo population tests to revisit Achen and Bartels (2017)'s study of the effect of New Jersey shark attacks on support for Woodrow Wilson in 1916. Achen and Bartels (2017)'s core finding is that beach counties in New Jersey experienced a sharper drop in Democratic support in 1916 compared to 1912 than other New Jersey counties did, which they attribute to voters irrationally punishing Wilson for shark attacks. Achen and Bartels assign a p-value of .01 to this occurrence: if there were no differential trend between the two sets of counties, the probability of seeing a divergence as large or larger due to a chance alignment of idiosyncratic factors is about .01. But this estimate relies on the assumption that these idiosyncratic factors are independent across counties; instead, it could be that political events often affect coastal and non-coastal counties differently, producing divergent trends more often than Achen and Bartels' independence assumption implies and possibly leading to a false positive.

To test Achen and Bartels (2017)'s inferential assumptions, Fowler and Hall (2018) reproduce Achen and Bartels's analysis for all 20 coastal states and all election years between 1872 and 2012, comparing the Democratic candidate's vote share in coastal and non-coastal counties (conditional on the previous result) and focusing on the 593 state-years in which no shark attacks took place. They reject the null hypothesis in 27% of these placebo populations (rather than the 5% they would expect if Achen and Bartels' assumptions were valid), concluding that Achen and Bartels (2017)'s result is more likely a false positive than their p-value suggests. This is a valid conclusion if we assume that excess false positives occur in these 593 state-years if the same is true in New Jersey in 1916 (and not otherwise, or not to the same extent); this need not be the case, e.g. if systematic coastal/non-coastal political discrepancies were common in the late 20th century but not in Woodrow Wilson's era.

5 Testing assumptions versus relaxing assumptions

In considering our examples, readers may wonder why the authors run a placebo test instead of some other procedure. Why do Peisakhin and Rozenas (2018) use the percent of Russian speakers as a placebo outcome rather than simply control for it, for example? This is a general feature of placebo tests: when the conditions are met to run an informative placebo test of an assumption, there is typically an alternative empirical strategy that relaxes that assumption and is valid under a similar set of conditions. Considering the choice between these procedures helps to clarify the distinctive contribution of placebo tests.

For placebo outcome tests for bias, we typically seek a placebo outcome that is (i) either a potential confounder or a descendent of one and (ii) not affected by the treatment; a variable with these characteristics could also be a good control variable, allowing us to relax the identification assumptions rather than test them.[12] The control approach is particularly appealing when (as in Peisakhin and Rozenas (2018)) there is little *a priori* reason to think that the author's conditioning set is sufficient. Still, there are at least two good reasons to withhold some covariates for placebo outcome tests rather than include all available controls. First, even if there is little theoretical support for the conjecture that X is a sufficient

[12] For the same reasons, informative placebo treatments could also be valid control variables; similar arguments apply. Our SI explores a further consideration arising in placebo treatment tests.

conditioning set (rather than X and W), a placebo outcome test using W provides evidence about that conjecture; including all available covariates makes such a test impossible (Imbens and Rubin 2015, p. 491). Second, a variable could be useless as a control variable because it does not affect the outcome but informative as a placebo outcome because of its relationship to unobserved confounders: this would be true of W in Figure 5, for example, if W has no effect on Y but affects D whenever U does.

In placebo population tests for bias, the alternative is differencing. If we are willing to assume that the bias is the same in the two populations, subtracting the estimate in the placebo population from the estimate in the core population yields an unbiased estimate of the treatment effect. (A simple diff-in-diff can be viewed in this way.) That assumption is strong, however. In the placebo testing approach, we instead start from the (potentially also strong) assumption that there is no bias in the core population, and we assume further that if there *were* bias in the core population there would also be bias in the placebo population; the key difference is that we do not assume that these two biases are equal. Which set of assumptions is more plausible will depend on the application.

The relevant alternative to a placebo test of inferential assumptions is to use the distribution of estimates across placebo outcomes, treatments, or populations to generate a p-value for the core analysis – a procedure similar to randomization inference (e.g. Rosenbaum 2002). Fowler and Hall (2018) implement both approaches. In addition to reporting that they reject the null in 27% of state-years with no shark attacks, they also report that they obtain a point estimate larger in absolute value than Achen and Bartels (2017)'s in 32% of state-years with no shark attacks, implying a p-value (.32) much higher than Achen and Bartels (2017)'s. (See also Schuemie et al. (2014).) The assumptions behind the placebo test approach imply that the false positive rate is elevated in New Jersey in 1916 iff it is also elevated in other state-years; the p-value estimate instead relies on the assumption that the distribution of point estimates across state-years approximates the null distribution of estimates for New Jersey in 1916.

6 Researcher degrees of freedom and related issues

It is well known that research findings can be unreliable when researchers, reviewers, or editors choose what analysis to run or publish in light of the results, especially when actors prefer some results over others. Researchers may intentionally or unintentionally distort the evidence in order to produce desired results, a practice variously known as p-hacking, fishing, or data dredging (e.g. Humphreys, De la Sierra and Van der Windt 2013; Gelman and Loken 2014).

Placebo tests can offer protection against these distortions. In cases where p-hacking leads to a biased estimation procedure or too-small standard errors, placebo tests may also produce a high rate of false positives and thus raise a red flag. If a false positive arose because of chance imbalance in the causes of the outcome, then placebo outcome tests using those causes could detect the problem. More broadly, the expectation to have both a significant finding in the core analysis *and* a set of insignificant findings in placebo tests helps weed out spurious results if genuine results are more likely to produce this pattern of findings than spurious ones are.

Unfortunately, placebo tests are also subject to some of the same pressures that lead to p-hacking. Researchers looking for flaws in others' designs seek statistically significant placebo tests, with the same possible pitfalls. Researchers running placebo tests on their own designs (currently the much more common case) face the opposite incentive, which may push them to massage their placebo test results to insignificance (a form of "null-hacking", as described by Protzko (2018)) and/or selectively report. A researcher could also analyze several outcomes or populations and decide later what is the core analysis and what is the placebo analysis, altering the causal theory accordingly. Moreover, when researchers face a choice between testing or relaxing assumptions (as discussed above), they might select the procedure that produces more favorable results, undermining the value of either approach.

Clearly pre-registration of placebo tests would help address all of these problems (Humphreys, De la Sierra and Van der Windt 2013). Another important safeguard against p-hacking and null-hacking in the design of placebo tests is the expectation that the core analysis and placebo analysis be as similar as possible. As discussed above, the main reason for this tight tethering is that the placebo analysis can only be informative about the validity of the core analysis's assumptions if it retains aspects of the core analysis that could violate those assumptions. But a close resemblance between the core analysis and the placebo analysis also helpfully reduces the degrees of freedom enjoyed by researchers conducting placebo tests.

Hartman and Hidalgo (2018) advocate an equivalence testing framework for placebo tests, where the null hypothesis is that the identification assumptions are violated. When null-hacking is a concern, the equivalence testing approach helpfully shifts the burden of proof, requiring researchers to actively marshal evidence in favor of their research designs; of course, this could invite p-hacking. We see equivalence testing as a reasonable way to accommodate the common tendency to misinterpret a null result in an under-powered hypothesis test as strong evidence for the null hypothesis. Our objective in this paper has instead been to use the formal logic of hypothesis testing to combat that misinterpretation. Seen properly (see Equation 1), a placebo test with low statistical precision is not very informative whether one uses a conventional null or the equivalence testing approach.

In fact, in our view the most important protection against the abuse of placebo tests is better understanding of the logic of placebo tests, to which we hope this paper contributes. Duplicitous researchers can of course produce passing placebo tests through null-hacking, but alert readers should notice if a test has low power (e.g. by examining the sample size or standard errors) and recognize such a test as uninformative. Careful readers should also be aware of the core assumptions behind a research design and the main threats to those assumptions, and they should notice if a test that would probe those assumptions is missing, or if a placebo test is included that has little to say about those core assumptions. Given the probabilistic and assumption-laden nature of the evidence provided by placebo tests, a better understanding of these tests could help reduce the incentive to selectively report or null/p-hack their results; it should also increase the incentive for authors to make their assumptions transparent and clearly explain the logic of placebo tests designed to probe those assumptions.

7 Conclusion: a placebo test checklist

To conclude, we offer a list of questions relevant to any placebo test. As explained in Section 2.1, the key overarching question to ask about a placebo test is, "Is the test more likely to fail if one of the core assumptions is violated in some relevant way than if those assumptions hold?" That question can be decomposed into the following checklist, which could be applied to any placebo test:

(i) What core assumptions — bias assumptions related to point estimation (identification, estimation, measurement, sample selection) or distributional assumptions related to standard errors — does the test probe? (Section 2.2)

(ii) What potential violations of the core assumptions are most relevant? (Section 2.1)

(iii) What component of the core analysis (outcome, treatment, population) has been altered to construct the placebo test? (Section 2.3)

(iv) Why should we think that, given this alteration, the treatment has no effect on the outcome in the placebo analysis? (NATE, Section 2.2)

(v) In what way(s) are the placebo analysis and core analysis similar, such that the placebo analysis may detect violations of the relevant core assumption(s)? (LVBA/LVDA, Section 2.2)

(vi) Might the placebo analysis suffer from violations of the core assumptions that are not present in the core analysis, raising the false positive rate? (LBA/LDA, Section 2.2)

(vii) Does the placebo test have sufficient statistical precision (judged by e.g. standard errors) to detect violations of the core assumptions? (Section 2.1)

In each case we have provided a reference to the relevant section of our formal framework, but these questions also arise in our discussion of examples in Sections 3 and 4. If readers encountering placebo tests routinely ask themselves these questions, and authors presenting placebo tests provide enough information to answer them, then placebo tests will better contribute to assessing the credibility of research designs in applied causal inference.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. “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*. Princeton University Press.

Arnold, Benjamin F, Ayse Ercumen, Jade Benjamin-Chung and John M Colford Jr. 2016. “Negative controls to detect selection bias and measurement bias in epidemiologic studies.” *Epidemiology* 27(5):637.

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.” *British Medical Journal* 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.

Depetris-Chauvin, Emilio, Ruben Durante and Filipe Campante. 2020. “Building nations through shared experiences: Evidence from African football.” *American Economic Review* 110(5):1572–1602.

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.

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.

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

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.

Folke, Olle, Torsten Persson and Johanna Rickne. 2016. “The primary effect: Preference votes and political promotions.” *The American Political Science Review* 110(3):559.

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.

Gelman, Andrew and Deborah Nolan. 2002. “You can load a die, but you can’t bias a coin.” *The American Statistician* 56(4):308–311.

Gelman, Andrew and Eric Loken. 2014. “The statistical crisis in science: data-dependent analysis - a “garden of forking paths” - explains why many statistically significant comparisons don’t hold up.” *American Scientist* 102(6):460.

Glas, Afina S, Jeroen G Lijmer, Martin H Prins, Gouke J Bonsel and Patrick MM Bossuyt. 2003. “The diagnostic odds ratio: a single indicator of test performance.” *Journal of clinical epidemiology* 56(11):1129–1135.

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

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.

Huntington-Klein, Nick. 2021. *The effect: An introduction to research design and causality.* Chapman and Hall/CRC.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

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

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.

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

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.

Protzko, John. 2018. "Null-hacking, a lurking problem in the open science movement." *PsyArXiv* .

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 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.

Schuemie, Martijn J, Patrick B Ryan, William DuMouchel, Marc A Suchard and David Madigan. 2014. "Interpreting observational studies: why empirical calibration is needed to correct p-values." *Statistics in Medicine* 33(2):209–218.

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

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

Supporting Information for "Placebo Tests for Causal Inference"

Abstract

This appendix has three components. First, we provide formal proof of the propositions stated informally in Section 2.2. Second, we discuss a special case in which one can either test identification assumptions via a placebo treatment test or relax those assumptions using instrumental variables; word limits prevented us from including this in the main text of the paper. Third, we provide 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). 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 analysis (e.g. the placebo outcome, for placebo outcome tests).

Contents

1 Formal proof of propositions in the paper

Proposition -1 (false positive rate in the core analysis): $\Pr(\hat{\delta} \in R \mid \delta = 0 \wedge \mathcal{IA} \wedge \mathcal{EA}) \leq \alpha$.

Proof: $\mathcal{BA}$ and $\delta = 0$ imply that $\hat{\delta} \leq \varepsilon$; the result then follows from $\mathcal{DA}$.

Proposition 0 (size of the placebo test): Let H_0 be that the bias assumptions $\mathcal{BA}$ and distributional assumptions $\mathcal{DA}$ hold. Then given Assumptions 1, 2, and 3, the false-positive rate p_0 of a test that rejects H_0 when $\hat{\delta}_p \in R_p$ is no more than α_p, the nominal size of the test.

Proof: Assumption 1 (NATE), Assumption 2 (LBA), and $\mathcal{BA}$ imply $\hat{\delta}_p = \varepsilon_p$: the estimate in the placebo test is purely sampling error. Then given $\mathcal{DA}$ and Assumption 3 (LDA), $\Pr(\hat{\delta}_p \in R_p) \leq \alpha_p$.

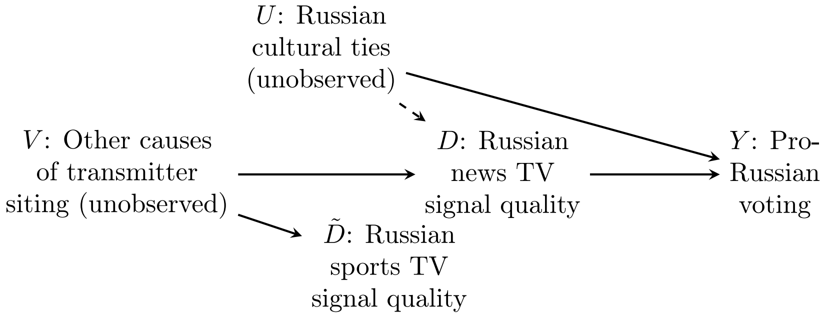
Proposition 1a (power of a placebo test of bias assumptions): Let H_{1a} be that the distributional assumptions $\mathcal{DA}$ hold but the bias assumptions $\mathcal{BA}$ fail. Given Assumptions 1, 3, 5, and 6, the true-positive rate p_1 of a test that rejects H_0 when $\hat{\delta}_p \in R_p$ is greater than α_p, the nominal size of the test.

Proof: Given Assumption 1 (NATE), $\hat{\delta} = b_p + \varepsilon_p$. Proposition 0 implies that $\Pr(b_p + \varepsilon_p \in R_p)$ is α_p where $b_p = 0$; Assumption 6 implies that $\Pr(b_p + \varepsilon_p \in R_p) > \alpha_p$ for other values of b_p. Assumption 5 (LVBA) implies that, given H_{1a}, $b_p \neq 0$.

Proposition 1b (power of a placebo test of distributional assumptions): Let H_{1b} be that the bias assumptions $\mathcal{BA}$ hold but the estimation assumptions $\mathcal{DA}$ fail. Then given Assumptions 1, 2, and 4, the true-positive rate p_1 of a test that rejects H_0 when $\hat{\delta}_p \in R_p$ is greater than α_p, the nominal size of the test.

Proof: Assumption 1 (NATE), Assumption 2 (LBA), and $\mathcal{BA}$ jointly imply $\delta_p = \varepsilon_p$. Assumption 4 (LVDA) then implies that, given H_{1b}, $\Pr(\hat{\delta}_p \in R_p) > \alpha_p$.

Figure 1: Placebo treatment or instrumental variable? (based on Peisakhin and Rozenas's placebo treatment test)


U: Russian cultural ties (unobserved)
V: Other causes of transmitter siting (unobserved)
D: Russian news TV signal quality
Y: Pro-Russian voting
$\tilde{D}$: Russian sports TV signal quality

2 Testing vs relaxing assumptions in placebo treatment tests

Digging deeper into placebo treatment tests, we discover another opportunity to choose whether to test or relax identification assumptions. Consider the DAG in Figure 1, which offers a simplified view of Peisakhin and Rozenas (2018)'s placebo treatment test under the assumption that the confounder U might affect the actual treatment D (Russian news TV signal quality) but not the placebo treatment $\tilde{D}$ (Russian sports TV signal quality); the only common cause of D and $\tilde{D}$ is V, which is not a confounder. Intuition suggests that if the placebo treatment is not affected by a potential confounder U (as in the DAG), the placebo treatment test cannot be sensitive to confounding due to U. But conditional on D, the only open path between $\tilde{D}$ and Y is $\tilde{D} \leftarrow V \rightarrow D \leftarrow U \rightarrow Y$ (because conditioning on the collider D opens the path); thus a placebo test using Russian sports TV signal as a placebo treatment can detect confounding due to Russian cultural ties (U) even if those ties affect the placement of news transmitters but not the placement of sports transmitters. More generally, a variable that shares causes with the treatment (none of which affect the outcome) but is not affected by the relevant confounders can nonetheless detect those confounders when we use it as a placebo treatment and condition on the actual treatment. This property of placebo treatment tests has apparently not previously been noticed.

Note, however, that V is a valid instrument for D in this DAG, and that the placebo treatment $\tilde{D}$ can be considered a proxy for this instrument. Thus given the DAG in Figure 1 we can use $\tilde{D}$ as an instrument to estimate the (local) effect of D on Y given confounding due to U, rather than use $\tilde{D}$ as a placebo treatment to test whether there is confounding due to U. Both procedures could be useful but the IV is more directly relevant to the research objective.

3 Library of Placebo Tests in Political Science

3.1 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 House, 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	Voting history
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[1]
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 state)	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[2]
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

[1] The adoption rate of these policies is low, resulting in low 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.

[2] The population is also shifted to include *all* city council candidates. The low rate of applications decided before the election may limit the 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[3]
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 [4]
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
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

[3] 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).

[4] 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	
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. residents	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)
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	Vote on racialized ballot initiatives from 1986 to 1990
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
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

Placebo outcome tests - continued from previous page

Paper	**Core analysis**			**Placebo outcome**
	Population	Treatment	Outcome	
Gerber and Huber (2009)	Counties in 26 U.S. states, 1992-2004	Election of president matching county's partisanship[5]	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)
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

[5] 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	
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)[6]	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)
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)

[6] 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 Population	 Treatment	 Outcome	Placebo outcome
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[7]	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	Social proximity to candidate	Turnout	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
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[8]	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

[7] This is instrumented by the outcome of close elections between secular and religious candidates.
[8] In IV analysis, this is instrumented by average wind power in the district.

Placebo outcome tests - continued from previous page

Paper	Core analysis			Placebo outcome
	Population	Treatment	Outcome	
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)

3.2 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

Placebo treatment tests - continued from previous page

Paper	Core analysis			Placebo treatment
	Population	Treatment	Outcome	
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
Eggers and Hainmueller (2009)	Candidates to the British House of Commons	Winning office (election RDD)	Wealth at death	Fake cutoffs
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[9]	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[10]	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[11]	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

[9] 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".

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

[11] 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**			**Placebo treatment**
	Population	Treatment	Outcome	
Grimmer et al. (2018)	US voters	Presence of voter ID laws (interacted with respondent race)	Turnout	Future value of 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[12]	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)
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

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

3.3 Placebo Population Tests

Paper	**Core analysis**			**Placebo population**
	Population	Treatment	Outcome	
Acharya, Blackwell and Sen (2016) (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[13]	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[14]	Lottery draft number in 1969	Attitude toward Vietnam War in 1973[15]	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[16]	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)

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

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

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

[16] 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 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.