AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Placebo Tests for Causal Inference

Andrew C. Eggers University of Chicago
Guadalupe Tuñón Princeton University
Allan Dafoe DeepMind and Centre for the Governance of AI

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.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/3RR5RJ.

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 (Cinelli and Hazlett 2020; Rosenbaum and Rubin 1983). 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 article, 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.

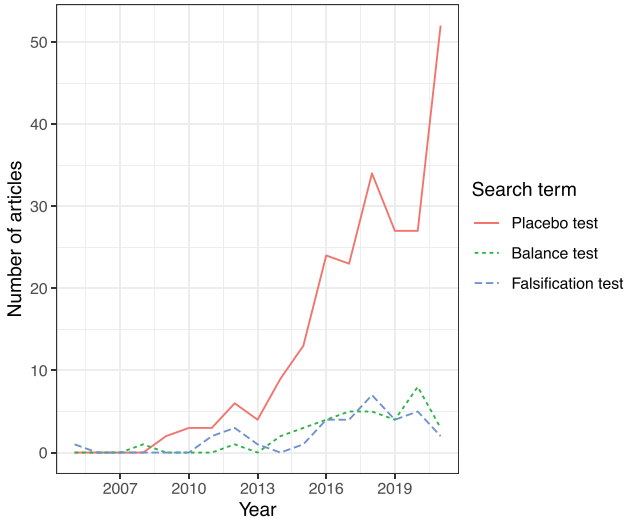
Andrew C. Eggers, Department of Political Science, University of Chicago, Pick Hall, 5828 University Ave Chicago, IL 60637 (aeggers@uchicago.edu). Guadalupe Tuñón, Department of Politics and School for Public and International Affairs, Princeton University, Robertson Hall, Princeton, NJ 08544-1013 (tunon@princeton.edu). Allan Dafoe, DeepMind and Centre for the Governance of AI, (R7) 14-18 Handyside Street, London, N1C 4DN, UK (allan.dafoe@governance.ai).

Tuñón recognizes financial support from Princeton's School for Public and International Affairs. For helpful input, the authors thank Devin Caughey, Thad Dunning, Anthony Fowler, Don Green, Jens Hainmueller, Sophia Hatz, Seth Hill, Zikai Li, John Londregan, Neil Malhotra, Luke Miratrix, Molly Offer-Westort, Jas Sekhon, Brandon Stewart, Laura Stoker, Anton Strezhnev, 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 University, New York University–Abu Dhabi, Columbia University, University of California San Diego, University of Chicago and the International Methods Colloquium provided useful feedback.

American Journal of Political Science, Vol. 00, No. 0, xxxx 2023, Pp. 1–16

DOI: 10.1111/ajps.12818

FIGURE 1 The Growing Popularity of Placebo Tests in Political Science



Note: The figure shows the number of articles mentioning "placebo test" and related terms in seven top political science journals, 2005–21.

As an example, consider two placebo tests presented in Peisakhin and Rozenas's (2018) 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 signals 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 signals.

Placebo tests like Peisakhin and Rozenas's (2018)' 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 (*American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, *British Journal of Political Science*, *Quarterly Journal of Political Science*, *Comparative Political Studies*) 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., Cohen-Cole and Fletcher 2008; DiNardo and Pischke 1997); 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 (Arnold et al. 2016; Imbens and Rubin 2015; Lipsitch, Tchetgen Tchetgen, and Cohen 2010; Rosenbaum 1984, 1989), their frameworks do not encompass 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 article 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, the section entitled "A Theory of Placebo Tests" 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 the sections entitled "Designing Placebo Tests of Bias Assumptions" and "Designing Placebo Tests of Distributional Assumptions," we illustrate each type of test using directed acyclic graphs (DAGs) and examples from political science. In "Testing Assumptions Versus Relaxing Assumptions," we consider alternative approaches that, under similar conditions, relax the assumptions behind a research design rather than testing them. In "Researcher Degrees of Freedom and Related Issues" 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. In the "Conclusion: A Placebo Test Checklist," we conclude with a checklist of questions to ask about any placebo test. The concepts and recommendations contained in the article should help researchers both interpret placebo tests and devise their own, particularly in conjunction with our library of over 100 placebo tests gathered from recent political science research (see the online supporting information, p. 3).

Perhaps the most important contribution of this article 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.

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.

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 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 versus H_1 (i.e., the posterior odds ratio) is, by Bayes Rule,

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

In words, the relative plausibility of the core assumptions (H_0) versus some departure from those assumptions (H_1), given the test results, is the prior relative plausibility times the ratio of the likelihoods (i.e.,

[1] Hartman and Hidalgo (2018) recommend reversing the null and alternative. We discuss this proposal in "Researcher Degrees of Freedom and Related Issues."

the Bayes factor) for obtaining those results under H_0 versus 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 $p_1 > p_0$, that is, power > size. Broadly, the higher is p_1 and the lower is p_0 the more informative is the test, that is, 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 each 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., 0.05); the precise 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.

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, that is, 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 $\varepsilon \equiv \hat{\delta} - E[\hat{\delta}]$, that is,

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

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*

[2] Royall (1997, 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.

[4] Biased but consistent estimators do not rely on the assumption that bias is exactly 0, for example.

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

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

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

that is, 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 $\mathcal{DA}$.

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 (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): $\mathcal{BA} \Rightarrow b_p = 0$.

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

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 online 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, 0.05). 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 0; 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.

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): $H_{1b} \Rightarrow \Pr(\varepsilon_p \in R_p) > \alpha_p$.

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 misspecified rejection region): if $\delta = 0$, $\hat{\delta}$ would fall in the rejection region at a rate above $\alpha = 0.05$. Assumptions 1, 2 and 4 imply that $f(\hat{\delta}_p)$ will be centered on 0 but, like $f(\hat{\delta})$, will have a misspecified 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:

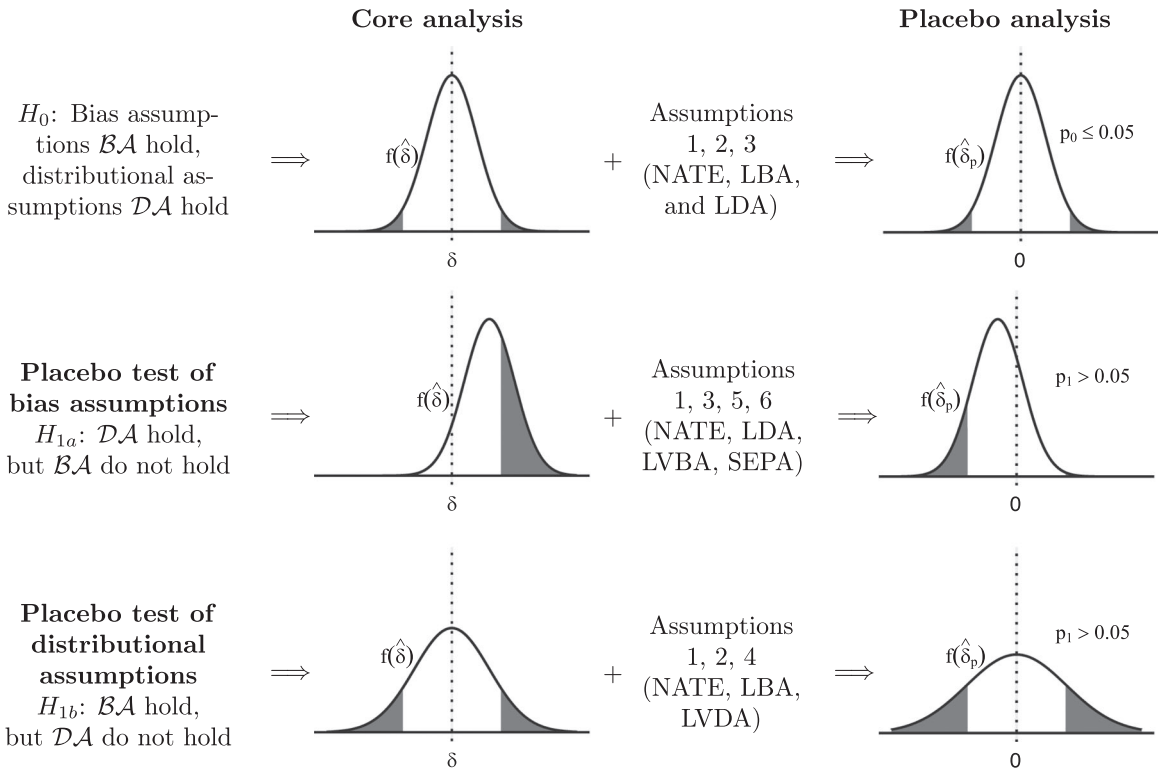
Assumption 5 (Linked Violation of Bias Assumptions, LVBA): $H_{1a} \Rightarrow b_p \neq 0$.

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

FIGURE 2 Sufficient Conditions for Informative Placebo Tests



Notes: 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, 0.05). 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 $p_1 > \alpha_p$.

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).

A Typology of Placebo Tests

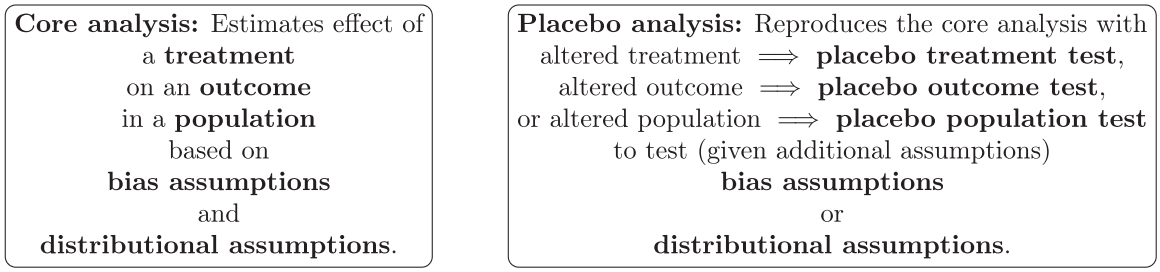
To better understand the challenges of designing and interpreting placebo tests, we examined every paper mentioning a "placebo test," "balance test," or "falsification test" in the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *International Organization* between 2009 and 2018. In analyzing the resulting list of 110 placebo tests (which we summarize in the online supporting information, p. 3),[7] we found it useful to categorize tests according to two features.

The first feature (mentioned in the previous section) 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

[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



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" 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" (1984, 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.

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.

The Typical Logic, Simplified and Illustrated

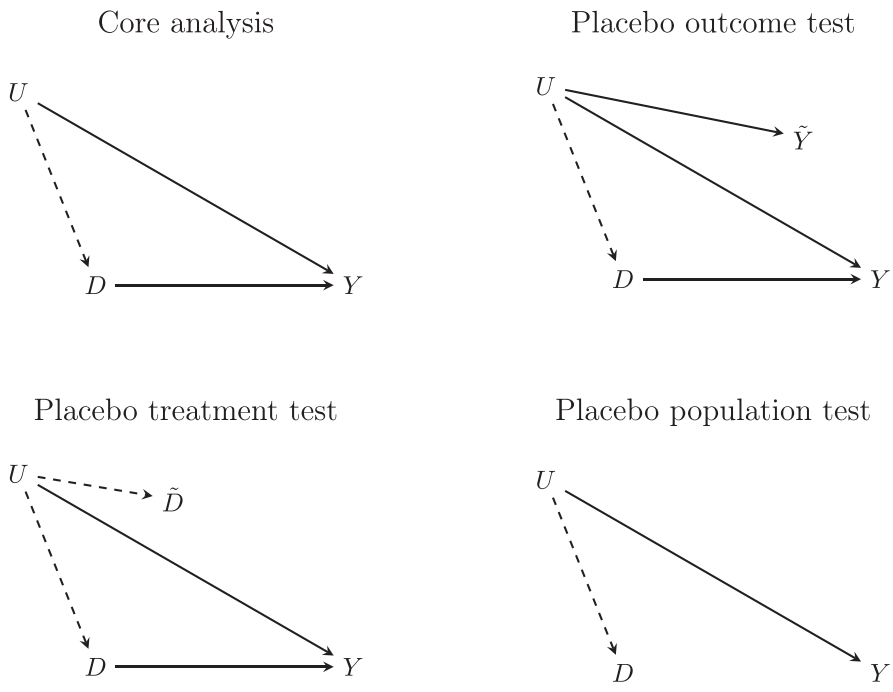
We begin by using simple DAGs to illustrate the typical logic of each type of test in the case where attention centers on a possible omitted variable.[10] A researcher seeks to measure the average effect of a treatment D on an outcome Y, as depicted in the top-left panel of Figure 4. There is an unobserved variable U that is believed to affect Y. The researcher assumes that U does not affect D, that is, that the dashed line connecting U and D can be erased completely. Given this assumption, the effect of D on Y is nonparametrically 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.

[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



Notes: In each DAG, **D** represents the treatment, **Y** represents the outcome, **U** represents a potential confounder; $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{Y}}$ represent placebo treatment and placebo outcome, respectively.

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 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.

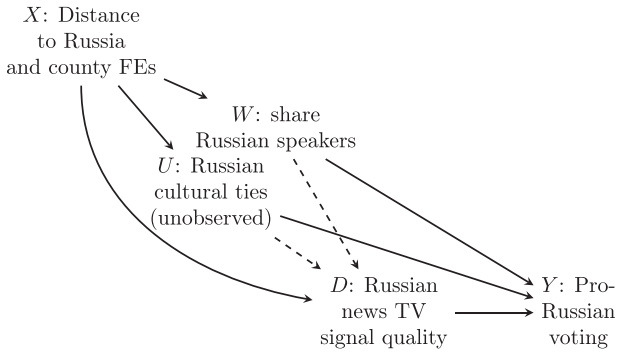
In each case, the DAG encodes the No Average Treatment Effect (NATE) assumption: there is no direct path from the treatment to the outcome in the placebo analysis. The LBA and LVBA assumptions are reflected in the assumed similarity across DAG edges: the placebo outcome $\tilde{Y}$ and Y are assumed 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.

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's (2018) 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's (2018) 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

FIGURE 5 Simplified DAG for Peisakhin and Rozenas's (2018) Placebo Outcome Test



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

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's (2018) controls.

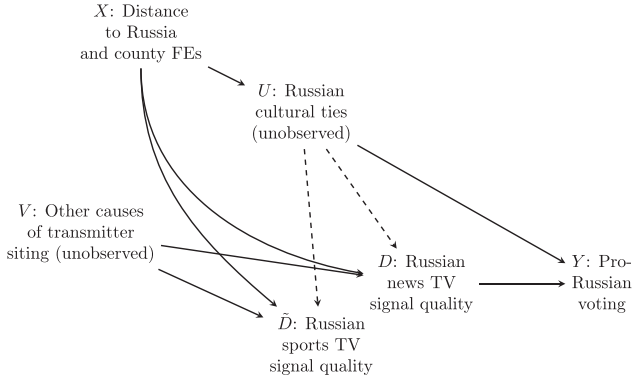
Peisakhin and Rozenas (2018) address this concern in part with placebo outcome tests that use pretreatment covariates (such as the percentage of Russian speakers in the precinct) as placebo outcomes. The DAG in Figure 5 illustrates the logic of the test, extending Figure 4 to include control variables. 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 percent of Russian speakers, observed) and U (Russian cultural ties more generally, assumed unobserved). Peisakhin and Rozenas's (2018) 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 pretreatment 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 Garcıá-Ponce (2013) study the impact of the federal US assault weapons ban on the murder rate in adjacent Mexican states. One may be concerned that the association Dube, Dube, and Garcıá-Ponce (2013) detect between assault weapon availability in the United States and murders in Mexico is due to other factors that coincided with the ban and caused a drop in violence. To assess this 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 United States 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's (2018) 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-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

FIGURE 6 Simplified DAG for Peisakhin and Rozenas's (2018) Placebo Treatment Test



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

TABLE 1 Examples of Post-Treatment Placebo Outcome Tests of Bias

Paper	Core Analysis			Placebo Outcome
	Population	Treatment	Outcome	
Dube, Dube, and Garcıá-Ponce (2013)	Mexican municipalities located close to US border, 2002–2006	Assault weapon availability from neighboring US state	Gun-related homicides	Accidents, nongun homicides, and suicides
Cruz and 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 and 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

Notes: More examples in the online supporting information.

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's (2018) case, the question is whether Russian sports broadcasting could affect Ukrainian political behavior; Peisakhin and Rozenas (2018) assume it does not, although 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, for example, 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 Burnett and Kogan 2017; Dasgupta, Gawande, and Kapur 2017; Peisakhin and Rozenas 2018) while others do not (including Fouirnaies and Mutlu-Eren 2015; Jha 2013; Stasavage 2014). 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 may share nonconfounding causes; this is also the reason one should condition on the actual treatment in such a test.

Placebo Population Tests. Peisakhin and Rozenas's (2018) 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

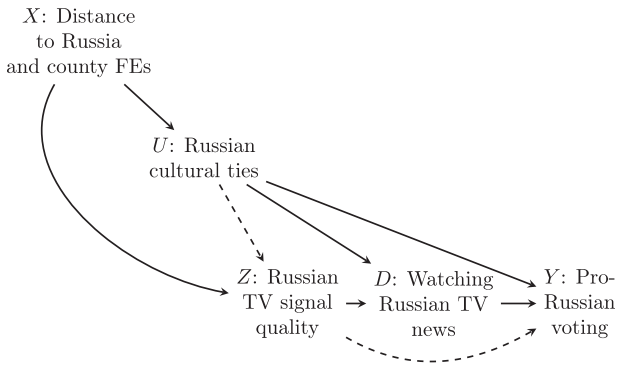
TABLE 2 Examples of Placebo Treatment Tests of Bias

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 nineteenth and twentieth 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	Difference in support for spending on public schools 1990–92	Proximity to areas with large African American population but no riot activity

Notes: More examples in Supporting Information p. 9.

ties, U) and also the exclusion restriction for the IV, that is, 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).

FIGURE 7 Peisakhin and Rozensas's (2018) Individual-Level Analysis



Notes: 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**, that is, 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").

Peisakhin and Rozenas's (2018) 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, for example, they have satellite TVs. In this population, Peisakhin and Rozenas (2018) 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.

NATE here states that signal quality does not affect consumption of Russian news among Ukrainians without terrestrial TVs. This seems reasonable, although 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

TABLE 3 Examples of Placebo Population Tests of Bias

Paper	Core Analysis			Placebo Population
	Population	Treatment	Outcome	
Acharya, Blackwell, and Sen (2016)	White Americans living in the US South	County's suitability for cotton production	Attitudes towards African Americans today	White Americans living in the US 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	Noncollege bound men; college-bound women

Notes: More examples in the online supporting information (p. 13).

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. Chen's (2013) core analysis compares turnout in the November 2004 US election between Americans who received FEMA aid and those who 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 (2013) 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 inform. Cattaneo, Idrobo, and Titiunik (2020, 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* 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, for example, the curvature of the CEF, dependence across units, or the density of observations) the false positive rate might be different at the threshold versus elsewhere.

[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.

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's (2017) study of the effect of New Jersey

shark attacks on support for Woodrow Wilson in 1916. Achen and Bartels's (2017) 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 (2017) 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 noncoastal counties differently, producing divergent trends more often than Achen and Bartels's (2017) independence assumption implies and possibly leading to a false positive.

To test Achen and Bartels's (2017) inferential assumptions, Fowler and Hall (2018) reproduce Achen and Bartels's (2017) analysis for all 20 coastal states and all election years between 1872 and 2012, comparing the Democratic candidate's vote share in coastal and noncoastal 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's (2017) assumptions were valid), concluding that Achen and Bartels's (2017) 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, for example, if systematic coastal/noncoastal political discrepancies were common in the late twentieth century but not in Woodrow Wilson's era.

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 (1) either a potential confounder or a descendent of a potential confounder and (2) 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 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, 491). Second, a variable that is useless as a control variable because it does not affect the outcome could be 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 difference-in-differences 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

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

also report that they obtain a point estimate larger in absolute value than Achen and Bartels's (2017) in 32% of state-years with no shark attacks, implying a *p*-value (.32) much higher than Achen and Bartels's (2017). (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.

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., Gelman and Loken 2014; Humphreys, De la Sierra, and Van der Windt 2013).

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 preregistration 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 underpowered hypothesis test as strong evidence for the null hypothesis. Our objective in this article 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 article 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.

Conclusion: A Placebo Test Checklist

To conclude, we offer a list of questions relevant to any placebo test. As explained in "What Do We Learn from Placebo Tests?" 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:

1. 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? (see "Formal Conditions for an Informative Placebo Test")
2. What potential violations of the core assumptions are most relevant? (see "What Do We Learn from Placebo Tests?")
3. What component of the core analysis (outcome, treatment, population) has been altered to construct the placebo test? (see "A Typology of Placebo Tests")
4. Why should we think that, given this alteration, the treatment has no effect on the outcome in the placebo analysis? (NATE, see "Formal Conditions for an Informative Placebo Test")
5. 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, see "Formal Conditions for an Informative Placebo Test")
6. 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, see "Formal Conditions for an Informative Placebo Test")
7. Does the placebo test have sufficient statistical precision (judged by, e.g., standard errors) to detect violations of the core assumptions? (see "What Do We Learn from Placebo Tests?")

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 "Designing Placebo Tests of Bias Assumptions" and "Designing Placebo Tests of Distributional Assumptions." 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):621–41.

Achen, Christopher H., and Larry M Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government.* Princeton, NJ: Princeton University Press.

Arnold, Benjamin F., Ayse Ercumen, Jade Benjamin-Chung, and John M Colford Jr. 2016. "Brief Report: Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies." *Epidemiology (Cambridge, Mass.)* 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): 302–14.

Cattaneo, Matias D., Nicolás Idrobo, and Rocıó Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations.* New York, NY: Cambridge University Press.

Chen, Jowei. 2013. "Voter Partisanship and the Effect of Distributive Spending on Political Participation." *American Journal of Political Science* 57(1): 200–17.

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): 234–66.

Cohen-Cole, Ethan, and Jason M. Fletcher. 2008. "Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analysis." *Bmj* 337: a2533.

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): 605–32.

De Craen, Anton J.M., Ted J. Kaptchuk, Jan G.P. Tijssen, and Jos Kleijnen. 1999. "Placebos and Placebo Effects in Medicine: Historical Overview." *Journal of the Royal Society of Medicine* 92(10): 511–15.

Depetris-Chauvin, Emilio, Ruben Durante, and Filipe Campante. 2020. "Building Nations through Shared Experiences: Evidence from African Football." *American Economic Review* 110(5): 1572–602.

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 Garciá-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):221–37.

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): 805–17.

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): 1423–37.

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.

Gelman, Andrew, and Deborah Nolan. 2002. "You Can Load a Die, but You Can't Bias a Coin." *The American Statistician* 56(4): 308–11.

Glas, Afina S., Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, and Patrick M.M. Bossuyt. 2003. "The Diagnostic Odds Ratio: A Single Indicator of Test Performance." *Journal of Clinical Epidemiology* 56(11): 1129–35.

Hainmueller, Jens, and Dominik Hangartner. 2019. "Does direct democracy hurt immigrant minorities? Evidence from naturalization decisions in Switzerland." *American Journal of Political Science* 63(3): 530-47.

Hartman, Erin, and F Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4): 1000–13.

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; Hall/CRC.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Jha, Saumitra. 2013. "Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia." *American Political Science Review* 107(04): 806–32.

Lipsitch, Marc, Eric Tchetgen Tchetgen, and Ted Cohen. 2010. "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies." *Epidemiology (Cambridge, Mass.)* 21(3): 383–88.

Neumayer, Eric, and Thomas Plümper. 2017. *Robustness Tests for Quantitative Research*. New York, NY: 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): 535–50.

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): 41–8.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York, NY: Springer.

Rosenbaum, Paul R. 1989. "The Role of Known Effects in Observational Studies." *Biometrics* 45: 557–69.

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): 212–18.

Royall, Richard. 1997. *Statistical Evidence: A Likelihood Paradigm*. New York, NY: Chapman & Hall.

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): 209–18.

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): 337–54.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Formal proof of propositions in the paper
Appendix B: Testing vs relaxing assumptions in placebo treatment tests
Appendix C: Placebo Outcome Tests
Appendix D: Placebo Treatment Tests
Appendix E: Placebo Population Tests