AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

An Equivalence Approach to Balance and Placebo Tests

Erin Hartman University of California Los Angeles
F. Daniel Hidalgo Massachusetts Institute of Technology

Abstract: *Recent emphasis on credible causal designs has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests. Yet current practice is to use statistical tests with an inappropriate null hypothesis of no difference, which can result in equating nonsignificant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data are inconsistent with a valid research design, and provide sufficient statistical evidence in favor of a valid design. When tests are correctly specified so that difference is the null and equivalence is the alternative, the problems afflicting traditional tests are alleviated. We argue that equivalence tests are better able to incorporate substantive considerations about what constitutes good balance on covariates and placebo outcomes than traditional tests. We demonstrate these advantages with applications to natural experiments.*

Replication Materials: The data, code, and any additional materials required to replicate all 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/RYNSDG.

Recent debates over the difficulties of causal inference, and the rise of causal empiricism, in the social sciences have spurred a growing literature on how to judge the quality of causal research designs (Austin 2008; Hansen 2008; Dunning 2010; Samii 2016) and a growing expectation that scholars defend the merits of their research designs with tests of empirically refutable implications of the assumptions justifying their inferences (Sekhon 2009, 503). For example, as evidence in favor of their designs, observational researchers are expected to provide evidence of covariate balance, and experimental researchers run randomization checks for balance on pretreatment covariates. The procedures used to check the assumptions justifying a design are just as important as those used to estimate causal effects (Rubin 2008).

In this article, we argue that "tests of design," such as balance and placebo tests, discussed in the next section, should be structured so that the responsibility lies with researchers to positively demonstrate that the data are consistent with their identification assumptions or theory.[1] This means that researchers should begin with the initial hypothesis that the data are *inconsistent* with a valid research design, and only reject this hypothesis if they provide sufficient statistical evidence in favor of data consistent with a valid design. The conceptual distinction between beginning with a null hypothesis of no difference, as is standard in current practice, versus beginning with a null hypothesis of a difference, as we advocate, may seem small, but the practical implications are substantial.

To implement our tests of design, we rely on the large body of literature in biostatistics on equivalence testing (Wellek 2010; Westlake 1976). We show how to apply these procedures to tests of design, discussed in the Mechanics of an Equivalence Test section. We pay particular attention to the selection of an equivalence

Erin Hartman is Assistant Professor, Departments of Political Science and Statistics, UCLA, 4289 Bunche Hall, 405 Hilgard Avenue/315 Portola Plaza, Los Angeles, CA 90095-1472 (ekhartman@ucla.edu). F. Daniel Hidalgo is Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E53-470, Cambridge, MA 02142 (dhidalgo@mit.edu).

Thanks to Santiago Olivella, Jasjeet Sekhon, Philip Stark, and Hans Noel for their comments and encouragement and to Kosuke Imai's Research Group, and SLAMM! 2016 Conference, and the International Methods Colloquium participants for valuable feedback. We also thank Shiyao Liu for research assistance. We are grateful for support for the R package development provided by the EGAP 2018 Standards Grant. Author order reflects contribution.

[1]Identification assumptions are assumptions about the data-generating process that allow for identification of causal effects, and which are usually inherently untestable, but often have testable, observable implications.

American Journal of Political Science, Vol. 62, No. 4, October 2018, Pp. 1000–1013

 DOI: 10.1111/ajps.12387

range, the range within which differences are deemed inconsequential, as it is a key distinction between equivalence and conventional hypothesis testing. We expand on the equivalence testing literature by considering randomization inference versions of common equivalence tests. We also introduce the "equivalence confidence interval," akin to a confidence interval, which is the minimum range that is supported by the data at the α-level. This range addresses many concerns in the literature about selecting an equivalence range by providing a transparent metric on which researchers should defend their claims. We suggest that researchers focus on defending this range rather than on the p-value associated with the test. We also discuss how equivalence tests can be used in conjunction with multiple testing corrections in the literature.

We provide applications of equivalence tests in the Examples section. First, we discuss a natural experiment conducted by Brady and McNulty (2011) on the cost of voting associated with distance to a polling place. Following that, we look at a battery of tests by applying our approach to the Dunning and Nilekani (2013) study of ethnic quotas. Further examples are included in Appendix SI-6 in the supporting information.

Throughout this work, we focus on tests of design; however, equivalence tests are related to the literature on "negligible effects" (Gross 2014; Rainey 2014). This important work, building on many others, shows why a lack of statistically significant difference is not sufficient evidence for showing substantive insignificance. We discuss the relationship to this literature, and the increased statistical power of the equivalence t-test focused on in this article, further in the next section.

Tests of Design

Balance and Placebo Tests

Before discussing how to conduct a balance test, arguably the most common test of design, we first explore why researchers are ultimately interested in balance on observable covariates. The goal of researchers is to provide evidence that their data are consistent with the identifying assumptions in their causal research design.

Many causal identification strategies require an assumption that the treatment assignment is unconfounded. In experimental settings, this assumption is met by the randomization conducted by the researcher, but in observational settings this necessary assumption is *inherently untestable in any direct manner*. Researchers relying on observational data can unemphasize prove their design is unconfounded. As discussed in Imbens and Rubin (2015, chap. 21), tests of design can be used to test the plausibility of the unconfoundedness assumption, even though we cannot directly test the assumption. If these analyses fail to provide evidence in favor of an unconfounded design,

> then the unconfoundedness assumption will be viewed as less plausible than in cases [...] supported by the data. How much the results of these analyses change our assessment of the unconfoundedness assumption depends on specific aspects of the substantive application at hand, in particular on the richness of the set of pre-treatment variables, their number and type. (Imbens and Rubin 2015)

So, while researchers must assume unconfoundedness, our aim is to formulate a statistical test that provides further evidence for the plausibility of the unconfoundedness assumption.

We thus frame this as a hypothesis testing problem of the following form:

H_0: The data are *inconsistent* with the observable implications of an unconfounded research design.

H_1: The data are *consistent* with the observable implications of an unconfounded research design (1)

To formulate a statistical test based on the observable data, we rely on the fact that the identifying assumptions of many causal research designs often have testable implications that can provide credibility to the research design. For example, unconfoundedness, when used in the natural experiment or matching framework, implies that the distributions of the potential outcomes for both treatment and control are identical. Although we cannot directly test the distribution of the potential outcomes, we can test how similar the groups look on pretreatment covariates, called a "balance test." Similarity across a large number of pretreatment covariates provides strength to the credibility of the design. The literature argues that by testing these observable implications, we are providing evidence consistent with the hypothesis defined in Equation (1).

Similarly, whereas the key identifying assumption for experiments, unconfoundedness via randomization, is true by design, randomization does not guarantee that any given treatment assignment will result in a treatment effect estimate sufficiently close to the "truth." Ensuring balance on key prognostic variables, by either blocking or

stratifying, can increase the precision of an estimator. Researchers conduct randomization checks to help defend against a "bad draw," in which there is severe imbalance on key prognostic covariates and the estimate is likely far from the truth.[2] These tests can also be used in rerandomization procedures to help improve covariate balance (Morgan and Rubin 2012).[3]

Balance tests[4] check whether the means, or distributions, of pretreatment variables are approximately the same among treatment and control units. There also exist omnibus tests for overall balance (Caughey, Dafoe, and Seawright 2017; Hansen and Bowers 2008). A related test is a placebo test, which examines the effect of the intervention on a postreatment variable known to be unaffected by the cause of interest (Rosenbaum 2002, 214).[5] If the intervention were to show a statistically significant correlation with the placebo outcome, then the validity of the research design is called into question. A common feature of these two standard tests is that it is incumbent upon the researcher to demonstrate that the difference between treated and control units on the pretreatment covariate or the placebo outcome is substantively small and thus not indicative of a severely flawed design. For the purpose of exposition, we will primarily focus on balance tests in the text of this article.

Current Practice: Lack of Difference versus Equivalence

To conduct a test of design, we argue that researchers should begin with the initial hypothesis that the data are *inconsistent* with the observable implications of an unconfounded design– for example, that there is substantial imbalance in the pretreatment covariates. Only with sufficient data should they reject the null hypothesis of imbalance in pretreatment covariates and posttreatment placebo outcomes. That is, they should provide *statistically* significant evidence to reject their data are inconsistent with a valid design, which they encode as a lack of *substantively* significant differences. However, common current practice is for researchers to use a statistical test that employs null of no difference[6] between the two groups as an indirect way of testing whether the data are consistent with an unconfounded design.[7] A design is deemed consistent with a valid research design if the statistical test fails to provide evidence in favor of a difference (i.e., a large p-value).[8] This approach could be loosely described as incorrectly equating "non-significant difference with significant homogeneity" (Wellek 2010, 3). A high p-value from such a test fails to reject the null that the two groups are different, which is only indirectly related to providing evidence that they are the same. This is not a flaw of the statistical test itself, but rather the common (mis)interpretation of the test when used as a test of design. While most researchers understand failure to reject a null hypothesis does not imply acceptance or preference for the alternative, current practice implies this nonetheless.

We propose that researchers use a statistical test consistent with the null in Equation (1), called an *equivalence test*. These tests are designed to provide statistical evidence under a null of difference, against an alternative of equivalence, which is consistent with the null and alternative hypotheses of Equation (1). The practice of equivalence testing remains largely absent from hypothesis testing in the social sciences, and for tests of design in particular.[9] There does exist, however, a large statistical literature investigating the properties of precisely these types of tests. Wellek (2010) and Berger and Hsu (1996) provide a review of the theory and main uses of equivalence testing.

[2] "Balance" is, of course, a sample property. In the case of experiments, the null hypothesis of equivalence is true by design. However, as Student (1938) put it, "it would be pedantic to continue with [a treatment assignment] known beforehand to be likely to lead to a misleading conclusion" (Morgan and Rubin 2012).

[3] In the case of rerandomization, researchers may wish to maximize balance on nonblocked variables, which could be achieved by requiring that randomization schemes do not exceed a set p-value as a metric for balance.

[4] Balance tests are also referred to as randomization checks in the experimental design literature.

[5] The definition of a placebo test is less well settled in the literature than the definition of a balance test. Some scholars appear to use balance and placebo tests interchangeably. In almost all cases, the known effect in a placebo test is 0. Another type of placebo test, which we do not consider, is the use of an alternate treatment, related to the treatment of interest, but whose effect on the outcome is known. A classic example of such a placebo test is Di Nardo and Pischke (1997).

[6] Some authors, such as Hansen (2008), do note that the actual null hypothesis researchers wish to test is not one about difference in the means of some super-population, but rather a statement about confounding.

[7] These are typically t-tests or KS-tests.

[8] There is no concrete rule for sufficient balance. While this is a clear misinterpretation of the results of a null hypothesis test of difference, this interpretation is pervasive in the literature. Authors do, implicitly, acknowledge that these tests are controlling for the incorrect error, and look for p-values to be higher than typical statistical significance, with a p-value of 0.15 or 0.2 considered evidence of good balance.

[9] There is a healthy literature on the drawbacks of the null hypothesis test across the social and natural sciences (see reviews in Gill 1999; Imai, King, Stuart 2008; and Gross 2014), but that literature did not traditionally provide many practical solutions for balance tests for applied researchers.

Fortunately for applied researchers, focusing on equivalence tests allows them to quantify and encode the strength of their design. Applied researchers will not have to significantly change their workflow while benefiting from transparent, statistical evidence supporting the strength of their design.

The ambiguity that results from using lack of statistical significance as evidence in favor of substantive equivalence is a well-documented problem (Gill 1999). The main issue is that people tend to incorrectly conflate low power with inconsequential difference or statistical significance with substantive difference. For example, consider Brady and McNulty (2011), who exploit a natural experiment in which the polling places of millions of voters in Los Angeles were moved to study the impact of the physical cost of distance to polling place on turnout. The authors employ a matching algorithm to match voters on a few important covariates to control for small imbalances noticed within the natural experiment, and the authors report balance statistics on variables not used in the matching algorithm as well as the mean differences at the precinct level.

Brady and McNulty (2011) then note that the magnitude of the differences are very small and unlikely to be indicative of hidden confounders, yet the size of their sample makes the traditional tests overly sensitive to these minute differences.[10] However, their argument would be strengthened with statistical evidence supporting the strength of their design. We will return to this example later, using an equivalence test to evaluate whether their data provide statistical evidence in favor of their design. We argue this reflects a conflict between the purpose for which the conventional null hypothesis t-test was designed and the goal of tests of design, namely, showing that differences on pretreatment covariates are substantively unimportant.

Equivalence Testing

Operationally, the most important difference between equivalence testing and tests of difference is whether or not one needs to make an ex ante decision over what range of values to define as "similar" versus "different." When using equivalence tests, the researcher must specify what is called an *equivalence range*, the set of values within which the difference between the two variables is substantively inconsequential. One example of a test for equivalence, which provides the easiest intuition, is the two one-sided test (TOST), which is set up as follows:

$$H_0\text{: } \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L$$

$$\text{versus} \qquad H_1\text{: } \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U,$$

where μ_T and μ_C refer to the mean of the treated and control groups, respectively, for a given covariate, and σ is the common standard deviation. The terms ϵ_U and ϵ_L refer to the upper and lower bounds for which two groups are considered equivalent. Choosing appropriate values for ϵ_U and ϵ_L is the most important aspect of equivalence testing, and this is discussed in detail in the Selecting an Equivalence Range section. The test is conducted using two one-sided t-tests, and the null of difference is rejected in favor of equivalence if the p-value for both one-sided tests is less than α. This test controls the Type I error of classifying the two sample means as equivalent (as defined by the equivalence range) when, in fact, they are not. This is one illustrative example of an equivalence test.

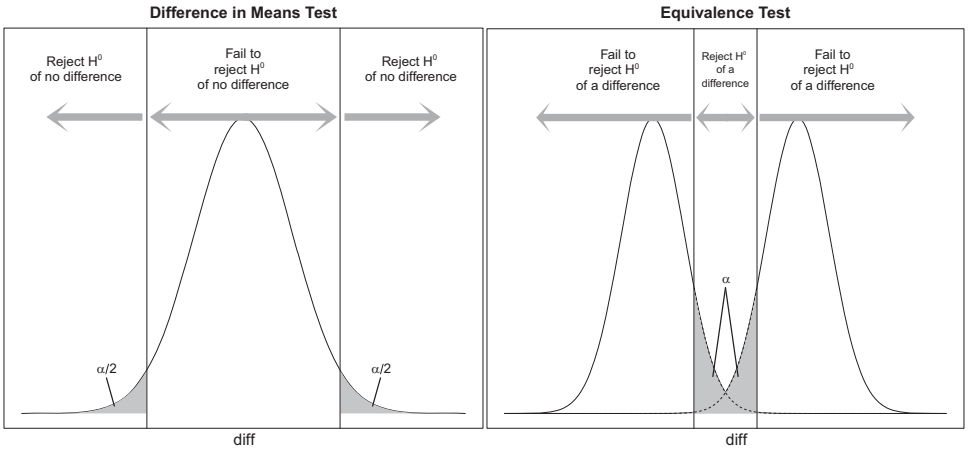
Figure 1 depicts, graphically, how the traditional balance tests and equivalence tests differ. In traditional balance tests, depicted in the left panel, we fail to reject the null hypothesis that means of two groups are different if the observed difference falls between the critical values. The shaded region corresponds to the region in which the two groups are classified as different when they are, in fact, the same, and the area corresponds to the level of the test. However, it is easy to see that this procedure is not controlling the proper Type I error implied by the null of a test of a design. In the panel on the right, the equivalence test will reject the null of a difference of at least a prespecified size in favor of the alternative of a difference less than that size when the difference lies in the shaded region for both tests. We discuss the mechanics and interpretation of equivalence testing in detail in the next section, including an equivalence version of the t-test. Alternative versions, which are designed for different types of data or sensitive to different departures of the null, are presented in Appendix SI-1.

Some recent literature in political science has suggested the practice of reversing the standard setup to make *difference* the null hypothesis and *sameness* the alternative hypothesis (Esarey and Danneman 2015; Gross 2014; Rainey 2014) for the study of negligible, or substantively insignificant, effects.[11] The negligible, or substantive

[10] "For the rest of the results, it does not make a great deal of sense to present *t*-statistics because the large sample ensures that most of these differences are statistically significant. Rather, we focus on their size" (Brady and McNulty 2011, 123).

[11] The difference between determining null, or negligible, effects, and the notion of "substantive significance" is nuanced. "Substantive significance" addresses the notion that the effect must

FIGURE 1 Tests of Equivalence versus Tests of Difference



Note: The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of one type of equivalence test—the two one-sided t-test (TOST)—under the null hypothesis of difference.

significance, approach evaluates the confidence range of the parameter and determines whether it lies entirely within ("negligible") or outside ("substantively significant") the null effect range. Both Rainey (2014) and Gross (2014) recommend the use of the 100(1–2α)% confidence interval and determining whether this interval lies entirely within a substantively defined equivalence range. This interval inclusion method is effectively the same as the TOST (Berger and Hsu 1996). Asymptotically, our suggested test and the interval inclusion method are the same and are effectively indistinguishable with reasonable sample sizes; however, the equivalence t-test described in this article is more powerful in smaller samples (Wellek 2010).[12] We show, in Appendix SI-5, why the interval inclusion approach can allow researchers to construct a statistical test with zero power in some scenarios. We build on the equivalence tests presented by Rainey (2014) and Gross (2014) by presenting additional equivalence tests appropriate for different distributions, and departures from the null, as well as randomization inference versions.

Sample Size and Traditional Balance Tests. The most common argument against traditional balance tests revolves around the common conflation of low power with an incorrect acceptance of the null hypothesis. The problem arises from the fact that the standard tests are designed to control for a Type I error of classifying the two group means as different when they are, in fact, the same.

A desirable property for a statistical test is that the power to detect the alternative increases in sample size, yet by conducting balance tests using tests of difference, the probability of rejecting the null of difference is inversely related to sample size. In equivalence tests, however, if the sample size is small, holding all else constant, the t-statistic will move toward zero. This will increase the p-value of at least one of the one-sided tests, depending on whether the observed difference is above or below zero, thus making it less likely that we will reject a null of difference. Therefore, the power of the test behaves as we would expect with respect to sample size. If a researcher wants to put a higher burden on the tests of design, and thus signal increased strength in the validity of the design, then the equivalence range should be decreased. Importantly, regardless of the researcher's chosen equivalence range, the equivalence confidence interval gives the smallest equivalence range supported by the data at the α-level, which the author should defend as substantively inconsequential to support

lie outside a range of theoretically unmeaningful values (Gross 2014), and "negligible effects" involve providing evidence that an effect lies within a range of theoretically unmeaningful values (Rainey 2014). In the parlance of equivalence tests, "negligible effects" are a straightforward application of an equivalence test, typically centered on zero, whereas "substantive significance" is often operationalized as showing that a $100(1-2\alpha)\%$ confidence interval lies entirely outside of an equivalence range. Both of these types of effects are conceptually similar to "placebo tests," a type of equivalence test conducted on a posttreatment variable that is hypothesized to lie within a specified range.

[12]The additional power in the equivalence t-test described here comes from accounting for the noncentral t distribution in the testing procedure.

her design. In Appendix SI-4, we provide simulations showing that equivalence tests are less likely to tempt researchers to conflate low power with evidence in favor of equivalence.

The main argument in defense of traditional hypothesis testing for validity tests is that although small sample sizes tend to make passing balance tests easier, small sample sizes also make finding significant treatment effects less likely. Hansen (2014) discusses how the dependence on sample size (i.e., the $n^{1/2}$ factor in the standard error calculations), appears in both the balance and outcome tests. Therefore, if one artificially inflates the p-values of the balance tests with small sample sizes, then the p-values associated with the outcomes will also be large, leading to nonsignificant findings. This logic, while correct for outcomes in which there is a theorized nonzero effect of an intervention, would not hold if a researcher hypothesized a negligible effect. While it is incorrect to accept a null of no difference in a low-power situation, an advantage of equivalence tests that are consistent with the implied hypotheses in Equation (1) is they give researchers a means by which to convey the strength of the design while avoiding the issue of the ambiguity of lack of statistical power.

Mechanics of an Equivalence Test

Implementing an equivalence test requires that a researcher define a few parameters, most importantly the equivalence range.[13] This section discusses a common test of equivalence to explicate the intuition behind this type of statistical test. We start with practical guidance for researchers about how to select an equivalence range, followed by the mechanics of the most common equivalence test, how to interpret the findings, and finally how these tests can be used with false discovery rate correction methods.

Selecting an Equivalence Range

Conducting an equivalence test requires the definition of an equivalence range—$[\epsilon_L, \epsilon_U]$ — in which we can consider the parameter of interest in the two groups to be substantively inconsequential.[14] How should one select this interval? This is arguably the most important decision a researcher must make when conducting an equivalence test, and it should be informed by the researcher's substantive knowledge.

Substantively Chosen Equivalence Range. Researchers are best suited to define equivalence ranges based on their substantive knowledge and considerations of the data at hand. This ensures that the researcher has considered what level of difference is most acceptable for the specified application given concerns about bounding bias.[15]

Researchers who have advocated for equivalence type approaches often tout the value of requiring researchers to transparently define and defend their equivalence range on theoretical grounds. As Rainey (2014, 1085) points out,

> Scholars who are cautious about the seeming arbitrariness of m [the equivalence range] should also note that as the researchers' choice for m changes, so too does the substantive claim they are making. Researchers who hypothesize that an effect lies between –1 and +1 make a weaker claim than researchers who argue that the same effect lies between –0.1 and +0.1. By explicitly defining m, researchers alert readers to the strength of their claims.

Gross (2014, 786) argues that "to convincingly argue about what results should be deemed significant in practical terms provides incentive for creative intertwining of qualitative with quantitative knowledge of subject matter." Consistent with previous authors, we consider the ability of the authors to encode the strength of their design in their equivalence range as an advantage. More powerfully, the equivalence confidence interval, described below, provides a more transparent way for authors to encode the same information that mitigates the impact of this choice.

It should be noted that the trade-off to smaller intervals, however, is power to detect equivalence. If the intervals are very narrow, then a large amount of data will

[13]We consider analyses conducted from a frequentist perspective. Researchers may, instead, wish to use Bayesian analysis, in which case they would not have to consider the appropriate null hypothesis. These researchers could consider the posterior distribution, and its relationship to an equivalence range. Wellek (2010), particularly Sections 2.4 and 3.2, discusses Bayesian methods for equivalence.

[14]Our discussion typically assumes a symmetric equivalence range for tests of difference, and the analog for ratio tests; however, tests of equivalence do not require equivalence ranges to be symmetric.

[15]Imai, King, and Stuart (2008) argue there is no theoretical level of imbalance that is acceptable if a researcher is concerned about bias—which can be of arbitrary size and direction given even small imbalances. This concern is valid, and it is a primary reason that researchers should conduct sensitivity analyses to check for the robustness of their results.

be required to obtain sufficient power to detect differences that small. As a result, researchers specifying substantively defined equivalence ranges should ensure that they have sufficient power, under the assumption that the true difference is zero and given their sample size, to detect equivalence.[16] In judging the results of a test of design, the power of the test can inform our expectations over the likelihood of rejecting the null of difference for a given equivalence range.

Sensitivity and Default Equivalence Ranges. Although we believe that equivalence ranges are best chosen out of substantive considerations, it is useful to specify default values for when researchers do not have strong substantive priors for an appropriate range. Although this is an area in need of validation studies, we provide a set of recommendations depending on the aim of the researcher and the available data.

Inherently, researchers are interested in balance as an observable implication of their design that guards against potential bias (Hansen 2008). Therefore, we propose researchers, where feasible, consider a sensitivity approach for defining the equivalence range. When a researcher is interested in a specific outcome, we recommend the equivalence range be $\pm$ one standardized effect size, using Glass's Δ, which is standardized by the standard deviation in the control group[17] on the outcome of interest. Assuming a perfect, linear correlation between the variable of interest and the outcome, imbalance outside of this equivalence range could fully explain the effect size. While this is conservative, pretreatment covariates are rarely so highly correlated with the outcome;[18] it is an assumption similar to the one made in other sensitivity analyses (Rosenbaum and Silber 2009). If researchers are concerned about nonlinearities between the variable and the outcome, they may wish to scale the standardized effect size by some nonlinear factor.

When the researcher cannot benchmark against a standardized effect size, we recommend using $\epsilon = \pm 0.36\sigma$, where σ is the pooled standard deviation of the covariate being tested.[19] The inspiration for this default value comes from Wellek (2010) and is confirmed by the simulation studies reported in Cochran and Rubin (1973), which showed that bias of this magnitude or less tended to produce only minor levels of bias when the relationship between imbalance and bias was linear, and outcome and covariates were normally distributed.[20] Further recommended default equivalence ranges for different tests, appropriate for different data types, are discussed in (Wellek 2010, 16).

We stress, however, that these default recommendations, as well as the sensitivity approach, do not guarantee any sort of bias-bounding properties. Equivalence ranges should still be given careful, substantive consideration for any particular application, and researchers should defend their choices. Regardless of the chosen range, the researcher should defend the equivalence confidence interval as inconsequentially small.

The Equivalence Confidence Interval. Since there naturally will be disagreement over an appropriate equivalence range, we recommend inverting the equivalence test to produce an *equivalence confidence interval* (ECI), which is akin to a confidence interval. The equivalence confidence interval is a symmetric interval defined by the largest difference at which the null hypothesis of difference is rejected at a prespecified α. The equivalence confidence interval specifies the smallest equivalence range supported by the observed data.[21] In other words, the difference between 0 and the maximum of the equivalence confidence interval quantifies the degree of uncertainty we have over the true degree of imbalance, and the researcher can be assured that at least $100(1-\alpha)\%$ of the time, the truth will lie within that range.

Researchers should focus on defending differences in the equivalence confidence interval as inconsequential, rather than on the p-value associated with the equivalence test. As long as the equivalence confidence interval is reported, readers can judge for themselves whether this range constitutes equivalence on the pretreatment

[16]Maximal power for equivalence tests is achieved at a true difference of zero. Although this assumption is justified for tests of design, maximal power may not be appropriate for tests of negligible effects.

[17]We choose Glass's Δ in case the treatment has an impact on the variance. If there is no impact on variance, then this will be more conservative than a pooled standard deviation (McGaw and Glass 1980).

[18]If researchers intend to use a linear regression to estimate the effect, they may wish to use equivalence ranges based on the sensitivity analyses discussed in Hosman, Hansen, and Holland (2010).

[19]Table SI-1 in the supporting information discusses how to map from substantive to standardized ranges for each test.

[20]Cochran and Rubin (1973) show that a caliper of 0.2σ when matching reduces 99% of bias, under certain conditions, and a caliper of 0.4σ reduces 96% of bias. Ho et al. (2006, 221) recommend the strictest range of 0.2 for judging "adequate" balance. Our simulation studies found 0.2σ to be a very conservative range.

[21]With a very small observed difference, it is possible that the inverted range could support an equivalence range of near zero. In this case, we define the range with the equivalence confidence interval as the observed standardized mean difference, which is a conservative range.

covariate or placebo outcome. Unlike the p-value for the equivalence test, an advantage of the equivalence confidence interval is that it is invariant to the researcher's chosen equivalence range, and therefore it provides a transparent value that researchers and the community can consider. The advantage of this is that it removes the researcher's degree of freedom in defining the equivalence range and forces the researcher to defend the range as substantively inconsequential for bias.

Conducting the t-test for Equivalence. Just as there are a variety of tests for evaluating difference, there are many equivalence tests. We discussed the TOST, which can be conducted using the interval inclusion method (i.e., determining whether a $100(1-2\alpha)\%$ confidence interval lies entirely within the equivalence range), as one conceptually straightforward method for conducting an equivalence test. Romano (2005) shows that the TOST is the asymptotically uniformly most powerful test. However, as shown in Appendix SI-5, the test can be structured to be grossly underpowered in finite samples. For this reason, we focus on an alternative test that is more powerful in finite samples.

The most appropriate test statistic depends on the type of variable and the desired sensitivity to different types of departures of H_0. Because most difference-in-means tests are conducted using t-tests, we discuss in detail the analogous t-test for equivalence in this section. However, other common tests for equivalence that are designed for different distributions, non-normal data, and parameters of interest, and which may be more appropriate for small samples, do exist. A summary of, and suggested use in cases for, these alternative tests can be found in Appendix SI-1, and formal notation can be found in Appendix SI-2.

The equivalence range for the t-test for equivalence is typically defined in standardized differences rather than the raw difference in means between the two groups, but researchers can easily map their substantive ranges to standardized differences by scaling by the standard deviation in the covariate. The standardized difference is a useful metric when testing for equivalence because, given some difference between the means of the two distributions, the two groups are increasingly indistinguishable as the variance of the distributions grows toward infinity, and increasingly disjoint as the variance of the distributions shrinks toward zero (Wellek 2010). We also recommend the t-test for equivalence because it is the uniformly most powerful invariant (UMPI) test for two normally distributed variables (Wellek 2010, 120). For simplicity, assume that $X_{Ti} \sim N(\mu_T, \sigma^2)$, with sample size m, and $X_{Ci} \sim N(\mu_C, \sigma^2)$, with sample size n; then the equivalence t-test uses the following hypothesis test:

$$H_0\text{: } \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L$$

versus

$$H_1\text{: } \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U.$$

We choose ϵ_L and ϵ_U appropriately, preferably based on substantive knowledge. Typically, the range of equivalence is symmetric around zero. After defining an equivalence range, the realized test statistic is calculated. The test statistic is

$$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{\sum_{i=1}^{m}(X_{Ti} - \bar{X}_T)^2 + \sum_{j=1}^{n}(X_{Cj} - \bar{X}_C)^2\right\}^{1/2}}.$$

This test statistic is distributed noncentral t with $N-2$ degrees of freedom (Wellek 2010, 120). If we choose a symmetric equivalence range, it can be shown that we can conduct a one-sided test using the test statistic $|T|$, which is distributed as the square root of a noncentral F, with the following rejection rule:

$$|T| < C_{\alpha;m,n}(\epsilon)$$

with

$$C_{\alpha;m,n}(\epsilon) = F(\alpha; df_1 = 1, df_2 = N-2,$$
$$\lambda_{nc}^2 = mn\epsilon^2/N)^{\frac{1}{2}},$$

where $F(\alpha, df_1, df_2, \lambda_{nc}^2)$ denotes the quantile function of the noncentral F distribution with level α, degrees of freedom 1, $N-2$, and noncentrality parameter $\lambda_{nc}^2 = mn\epsilon^2/N$. If the ϵs were not symmetric, then we would have the following rejection rule:

$$C^L_{\alpha;m,n}(\epsilon_L, \epsilon_U) < T < C^U_{\alpha;m,n}(\epsilon_L, \epsilon_U),$$

where the critical values must be determined appropriately. If $|T|$ is less than our critical value (or T lies within the critical values, in the case of asymmetric ϵs), then we reject the null hypothesis of a difference between the means of the two groups in favor of the alternative of an inconsequential difference. Otherwise, we fail to reject the null of nonequivalence. In addition to the rejection decision, researchers should also analyze the equivalence confidence interval, which gives the minimum equivalence range supported by the data. In the case in which the equivalence confidence interval is small, the researcher can be confident that the data provide strong evidence against a substantial difference. If the range is large, then the researcher may call into question the equivalence of the means of the two groups. Researchers should also be aware of the power of their test. Further discussion of the power of equivalence tests is provided in Appendix SI-5.

Interpretation

Equivalence tests are not direct tests of the underlying identifying assumptions necessary in most causal designs, so how should researchers interpret the results of these tests? Unconfoundedness is never directly testable, so researchers have taken two approaches to the interpretation of balance test results.

First, we could interpret the results from a frequentist perspective, in which the results indicate how much information the data convey against the null hypothesis, which in this case is the null of a consequential difference. A research design that truly is unconfounded does not require that the treatment and control groups look identical across all covariates in any given sample, but a lack of balance in a given sample on important variables should lead observational researchers to question their identifying assumption. By making our null hypothesis that the "data are inconsistent with the observable implications of an unconfounded design," a test of equivalence will provide evidence to reject this null in favor of an alternative that the "data are consistent with the observable implications of an unconfounded design." Of course, we should note this is distinct from the alternative that the "design *is* unconfounded," which is untestable. Though we do not accept that our design is unconfounded, our p-values will now encode a metric for how much information the data have against the implications of a flawed design.

Alternatively, we could refrain from interpreting the statistical implications of the test, and rather ask, "How similar is similar enough?" Some researchers take this more extreme view and merely consider balance tests as a nonstatistical metric for balance assessment (e.g., Imai, King, and Stuart 2008; Sekhon 2007), in which the resulting p-values are used to maximize observable balance rather than conduct tests of design, such as in matching studies.

Additionally, experimentalists may appeal to p-values as a metric for balance when conducting pretreatment balance tests, in which they wish to ensure balance on key prognostic covariates on which they cannot block. These researchers are not trying to determine whether their experiment is consistent with an unconfounded design—this is true by design. However, balance on key prognostic variables can increase the likelihood that the resulting estimate will be close to the truth. The equivalence tests discussed here are consistent with this aim and should have desirable properties that low p-values encode evidence against a null of substantial difference, and researchers will not be tempted to conflate low power with similarity. Additionally, researchers conducting rerandomization can encode their notion of "similar enough" into their balance metric via the equivalence range.

Observational researchers conducting balance checks are ultimately concerned about bias, particularly as caused by unobserved confounders. Consequently, what really matters for tests of design is the unobservable mapping between covariate imbalance and bias, and covariate balance itself is only a proxy for this potential bias.[22] Because this mapping is fundamentally unobservable, our judgments about an adequate equivalence range must ultimately depend on substantive considerations. Thus, when possible, one should specify an equivalence range small enough to satisfy readers that differences between two groups contained within the interval are substantively inconsequential, and thus unlikely to lead to substantively significant bias.

There is a healthy literature on sensitivity analyses (e.g., Rosenbaum and Silber 2009; Imbens and Rubin 2015) for assessing possible remaining unmeasureable confounding in causal effect estimates, and tests of design do not negate the need for these additional analyses. Tests of design will provide information on observable imbalance, and under certain assumptions, how that imbalance could impact our estimates. They do not, however, provide any information about unobservable imbalance, and for that reason we strongly encourage practitioners to combine tests of design with sensitivity analyses on the final estimates when providing evidence to strengthen the claims of their design.

Randomization Inference Equivalence Tests

A concern of many researchers is that balance is a characteristic of the sample, and therefore that tests of design, conducted on pretreatment covariates, which reference a hypothetical super-population, are inappropriate because they are contradictory to the nonrandom nature of the observed sample (Austin 2008; Imai, King, and Stuart 2008). One solution to this issue is to conduct tests that are conditional on the realized sample using permutation-based inference, which allows for inferences about how "differences between groups can be explained by chance, rather than what differences between sample and population can be explained by chance" (Hansen and Bowers 2008, 224). In addition to being conditional on the observed sample, the permutation tests are exact and

[22] Without additional assumptions about the mapping between the covariate and the outcome, any level of imbalance could lead to bias of arbitrary magnitude and size.

do not rely on large sample approximations. These exact tests can be conducted to asses the likelihood of observed imbalances in the sample without addressing the separate goal of assessing generalizability.

Using the intersection-union principle, each equivalence test can be tested using the union of two one-sided exact tests. Permutation tests require an arguably stronger assumption of a strict null of a constant treatment effect, and they test for distributional departures from the strict null. These types of tests are designed to test for exchangeability of the two groups, a property that should be guaranteed by the random or quasi-random design of the study. Therefore, they are well suited for tests of design, such as balance and placebo tests, where we explicitly desire a test of exchangeability. They are also robust to outliers and sensitive to departures of the null above and beyond mean differences, such as differences in variability within the two groups. To conduct the permutation version of the parametric tests, we conduct one-sided tests of the strict null hypothesis equal to the bounds of the equivalence range, and the overall null hypothesis of nonequivalence can be rejected if both corresponding permutation p-values are less than the level of the test, α.[23] Formal properties of permutation equivalence tests are explored in Arboretti et al. (2018).

Multiple Testing Corrections and Equivalence Tests

One final concern for researchers conducting tests of design is that they often conduct tests across a battery of covariates. In the balance-testing framework, the more variables, particularly highly prognostic variables, that researchers can provide balance on, the more evidence they can provide about the plausibility of the validity of their design. Sometimes researchers will conduct an omnibus test for overall balance, since the observable implication of unconfoundedness is balance across the joint distribution of the pretreatment covariates. Wellek (2010) provides the equivalence version of Hotelling's T^2, and Fisherian tests, such as those in Hansen and Bowers (2008) and Caughey, Dafoe, and Seawright (2017), can be used; however, these tests should also be structured with an alternative hypothesis of equivalence. While the omnibus test is not subject to the multiple testing problem, researchers are often interested in univariate balance statistics. However, conducting multiple tests can lead to false positives. With traditional balance tests, if a researcher conducts balance tests across 20 variables and observes a significant difference for one, should she discredit that result as chance? Typically, when conducting multiple tests, researchers can adjust for the multiple testing problem by correcting for the false discovery rate—the expected proportion of falsely rejected hypotheses— or the family-wise error rate—the probability of committing any Type I errors (Benjamini and Hochberg 1995). Perhaps more importantly, if researchers are conducting placebo tests on outcomes where they expect negligible effects, an omnibus test may not be appropriate, and researchers should adjust for the multiple outcomes, placebo and not, that they are testing.

Multiple testing procedures control the Type I error rate by appropriately inflating the resulting p-values to account for the number of tests being performed to control for either the false discovery rate or the family-wise error rate. However, these procedures would be inappropriate in conjunction with the common way in which tests of design are conducted—inflating the p-value for a test-of-difference test would be making the burden of proof lower for the researcher. The researcher wishes to control the probability of incorrectly rejecting the null of difference when a difference is, in fact, present. By using equivalence tests, the hypothesis test is consistent with the researchers' aims, and multiple testing corrections can be applied directly to the resulting p-values. The ability to correct for the multiple testing problem is a strength of the equivalence approach.

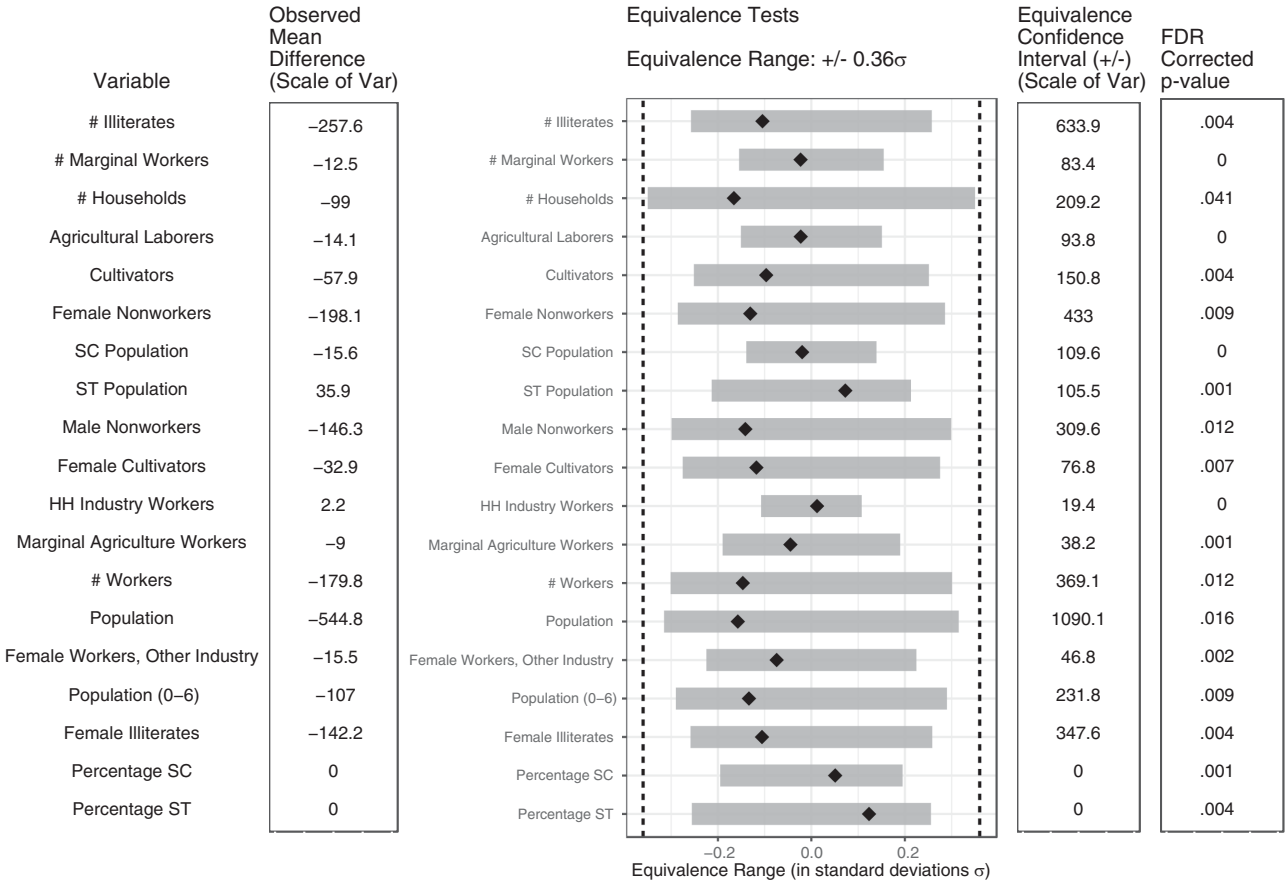
Examples

Example: Brady and McNulty (2011)

To illustrate the merits of equivalence tests, we return to the example of Brady and McNulty (2011). Recall that Brady and McNulty (2011) argue that some polling stations in Los Angeles were consolidated "as-if" random by the county registrar. Central to their argument about the quality of their design is that, prior to the consolidation, voters in treatment and control precincts had roughly equal "costs of voting," with distance between voters' residence and their polling station being their chief measure of cost. Balance on this variable is critical, yet the authors find that the pretreatment difference is "highly significant," although "substantively rather small" (2011, 123). If the conventional decision rule over adequate balance is followed, then one would question the "as-if" random identification assumption.

[23] Simulations showing properties of this test are provided in Section SI-3 in the supporting information.

FIGURE 2 Results of Equivalence Tests



Note: The observed mean difference is the mean of the treated group minus the mean of the control group. The vertical dashed lines represent the hypothesized equivalence range, defined as the standardized effect size on the outcome of interest. Gray bars represent the inverted equivalence range supported by the data, presented in standardized differences. The black diamonds represent the observed standardized difference for the variable of interest. The equivalence confidence interval is the inverted range, transformed to the scale of the variable. The p-value corresponds to the false discovery rate corrected p-value of the test of the null equivalence range of one standardized effect size.

We replicate Brady and McNulty's balance check using the two-sample t-test for equivalence. The observed average difference in distance between voters in treatment and control precincts is 0.034 miles, or 60 yards. We use an equivalence interval, based on the strict interval suggested in Ho et al. (2006) discussed earlier, of 0.2 standard deviations (amounting to about 0.055 miles, or 98 yards). Note that this is a case in which the equivalence interval used to formulate the null hypothesis could also be chosen on substantive grounds based on knowledge of factors affecting the decision to turn out that limit an acceptable distance. We also compute the equivalence confidence interval, which is the smallest equivalence interval supported by the data ($\alpha = .05$) given the observed difference between treatment and control polling stations.

Can we reject the null hypothesis that the mean difference in the distance to polling stations in 2002 is greater than $\epsilon = 0.055$ miles? This null is rejected with a p-value that is essentially zero. Given our prespecified equivalence interval, we consider the two samples to be well balanced on this variable. When we invert our test, we find that the equivalence confidence interval, supported at the $\alpha = 0.05$ level, is 0.124 standard deviations, or 0.035 miles (61 yards). Whether 0.035 miles is of concern and worthy of further adjustment, such as through regression, should be debated by subject area experts.

Example: Dunning and Nilekani (2013)

To illustrate the merits of equivalence tests over traditional tests, we reconsider the balance tests conducted in Dunning and Nilekani (2013). In this article, the authors consider a natural experiment to evaluate the effect of ethnic quotas on redistribution. Leveraging an ordered list used to determine villages in which council presidencies were reserved for scheduled castes, the authors note that villages at the bottom of the list in an earlier election period, which are assigned quotas, are indistinguishable from villages at the top of the next list that are not assigned quotas until the next election. Using purposive sampling among these villages, the authors evaluate how similar these villages are on a number of characteristics, presented as Table 2 in the original text.

The authors present balance statistics for univariate tests, and the p-values are generally high, but somewhat inconclusive for two variables in particular: "Number of households" ($p = 0.09$) and "Mean female nonworkers" ($p = 0.12$). The authors do not address these individual tests, but instead argue that an F-test of treatment assignment on all the covariates is insignificant. While the authors convincingly present a battery of evidence that the design is consistent with "as-if" randomization, the presented balance tests do not necessarily provide statistical evidence consistent with their claim. In 2, we conduct the same balance tests, this time using equivalence tests and applying an false discovery rate correction.

As can be seen in Figure 2, the equivalence tests indicate that we can reject the null of consequential difference, making the "as-if" random assumption more plausible. In this example, we conduct the test using a fairly conservative range of 0.36σ. The smallest standardized effect size in the original article is 0.43σ, which, if used as the equivalence range, yields even smaller p-values. An important contribution of the equivalence method is that rather than debating whether 0.36σ or 0.43σ is the appropriate range, we can ask whether ± 210 households, or ± 433 female nonworkers in a village–the respective ECIs–are considered substantively inconsequential differences in these data. We also see that the p-values can now be adjusted to account for the large number of tests, which we see as an alternative or supplementary approach to omnibus tests, depending on the evidence the researcher wishes to provide.

Conclusion

Researchers' need to provide evidence for equivalence between two groups, an observable implication of an unconfounded design, has always been present, but with the increased skepticism about traditional research designs in economics, political science, and sociology, we have seen more encouragement for researchers to expend great efforts in defending their effect estimates from the critique that they suffer from remaining confounding. In many areas of observational work in the social sciences, readers begin with the presumption that the observational design is flawed and must be convinced by empirical tests that this is not the case. Experimentalists are asked to defend against a "bad draw" that could lead their realized estimate to be far from the truth. Beyond the case of design, researchers are also interested in providing statistical evidence in favor of theoretical negligible effects on outcomes. This essay argues that such skepticism should be directly embedded in the hypothesis tests that are used to persuade readers over the validity of the design. By using equivalence tests, researchers begin with the assumption that the design is flawed, or that an effect is not negligible, and this hypothesis is only rejected if the data allow it. Furthermore, we believe that equivalence tests encourage researchers to directly address a substantive question about their design: What is good balance? By requiring the researcher to specify an equivalence range ex ante, equivalence tests encourage a substantive discussion about imbalances that are small enough to be tolerated versus those that are not.

Using equivalence tests for tests of designs opens up an avenue of research for methodologists. Each causal research design implies a certain test of design. Regression discontinuity designs (RDDs) imply continuity of observable variables, matching and natural experiments imply balance, and difference-in-differences or synthetic matching implies a similar time trend on pretreatment outcomes. Particularly with RDD and synthetic matching, further work must be done on the most appropriate equivalence test. Relatedly, researchers are often concerned about the "curse of dimensionality," or the fact that testing across multiple dimensions will increase the likelihood of finding an imbalanced variable (Ho et al. 2006). Further work on multivariate tests for balance that test for equivalence across a multidimensional space is necessary. The authors are also working on the development of an `R` package that will allow researchers to conduct equivalence-based tests of design.

For sample sizes typically used in natural experiments, lab experiments, and related designs in the social sciences, an equivalence approach may increase the difficulty of passing balance and placebo tests. As evidenced by our review of natural experiments in Appendix SI-6,

some studies that currently "pass" tests of design when the null is sameness will not reject a null of difference. Failing to reject a null of difference does not by itself, of course, invalidate a design or indicate hopelessly biased estimates. Many other elements of a design should go into an evaluation of its quality, such as the degree to which the assignment to treatment is exogenous or "as-if" random. For studies in which the treatment assignment mechanism is well understood and the identifying assumptions seem quite plausible, our burden of proof should be lower. In designs exploiting a discontinuity or those relying on a conditional independence assumption, more definitive evidence may be required to overcome doubt. For these cases, equivalence tests can improve on existing practice by ensuring that we encode our skepticism in the null hypothesis and require the researcher to marshall evidence against it.

References

Arboretti, Rosa, Eleonora Carrozzo, Fortunato Pesarin, and Luigi Salmaso. 2018. "Testing for Equivalence: An Intersection-Union Permutation Solution." arXiv.org.

Austin, Peter C. 2008. "A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003." *Statistics in Medicine* 27(12): 2037–49.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B: Methodological* 57(1): 289–300.

Berger, Roger, and Jason Hsu. 1996. "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science* 11(4): 283–302.

Brady, Henry E., and John McNulty. 2011. "Turning Out to Vote: The Costs of Finding and Getting to the Polling Place." *American Political Science Review* 105(1): 115–34.

Caughey, Devin, Allan Dafoe, and Jason Seawright. 2017. "Nonparametric Combination (npc): A Framework for Testing Elaborate Theories." *Journal of Politics* 79(2): 688–701.

Cochran, William, and Donald Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya: The Indian Journal of Statistics* 35(4): 417–46.

Di Nardo, John E., and Jorn-Steffen Pischke. 1997. "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?" *Quarterly Journal of Economics* 112(1): 291–303.

Dunning, Thad. 2010. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. Henry E. Brady and David Collier Lanham, MD: Rowman & Littlefield, 273–311.

Dunning, Thad, and Janhavi Nilekani. 2013. "Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils." *American Political Science Review* 107(1): 35–56.

Esarey, Justin, and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(1): 95–111.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3): 647–74.

Gross, Justin H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science*, 59(3): 775–88.

Hansen, Ben B. 2008. "The Essential Role of Balance Tests in Propensity-Matched Observational Studies: Comments on 'A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003' by Peter Austin, Statistics in Medicine." *Statistics in Medicine* 27(12): 2050–54.

Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23(2): 219–36.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2006. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Hosman, Carrie A., Ben B. Hansen, and Paul W. Holland. 2010. "The Sensitivity of Linear Regression Coefficients' Confidence Limits to the Omission of a Confounder." *Annals of Applied Statistics* 4(2): 849–70.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A: Statistics in Society* 171(2): 481–502.

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

McGaw, Barry, and Gene V. Glass. 1980. "Choice of the Metric for Effect Size in Meta-Analysis." *American Educational Research Journal* 17(3): 325–37.

Morgan, Kari Lock, and Donald B. Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *Annals of Statistics* 40(2): 1263–82.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4): 1083–91.

Romano, Joseph P. 2005. "Optimal Testing of Equivalence Hypotheses." *Annals of Statistics.*

Rosenbaum, Paul R. 2002. *Observational Studies (Springer Series in Statistics).* 2nd ed. New York: Springer.

Rosenbaum, Paul R., and Jeffrey H. Silber. 2009. "Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units." *Journal of the American Statistical Association* 104(486): 501–11.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3): 808–40.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *Journal of Politics* 78(3): 941–55.

Sekhon, Jasjeet S. 2007. "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference." *Survey Research Center, University of California*, Berkeley. http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf.

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

Student. 1938. "Comparison between Balanced and Random Arrangements of Field Plots." *Biometrika* 29(3/4): 363–78.

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

Westlake, Wilfred J. 1976. "Symmetrical Confidence Intervals for Bioequivalence Trials." *Biometrics* 32(4): 741–44.

Supporting Information

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

SI-1: Additional Statistical Tests for Equivalence
SI-2: Formalization of Additional Statistical Tests for Equivalence
SI-3: Sample Specific Versions of Parametric Tests
SI-4: Traditional vs. Equivalence Tests - A Simulation
SI-5: Negligible Effects and the 90% Confidence Interval
SI-6: Applying Equivalence Tests to Natural Experiments in the Social Sciences