An Equivalence Approach to Balance and Placebo Tests

Erin Hartman* F. Daniel Hidalgo†

March 19, 2018

Abstract

The rise of design-based inference 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 the equating of non-significant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data is *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.

***** All errors are our responsibility. *****

Total Word Count: 8566

*Department of Political Science and Statistics, UCLA, ekhartman@ucla.edu.
†Department of Political Science, Massachusetts Institute of Technology, dhidalgo@mit.edu

1 Introduction

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, p. 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 pre-treatment 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 paper, we argue that "tests of design", such as balance and placebo tests, discussed in Section 2, should be structured so that the burden of proof lies with researchers to positively demonstrate that the data is consistent with their identification assumptions or theory[1]. This means that researchers should begin with the initial hypothesis that the data is *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 Section 3. We pay particular attention to the selection of an equivalence range, the range within difference are deemed inconsequential, as it is a key distinction between equivalence and conventional hypothesis testing. We

[1]Identification assumptions are assumptions about the data generating process which are usually inherently untestable, but often have testable observable implications.

expand on the equivalence testing literature by considering randomization inference versions of common equivalence tests. We also introduce the “credible equivalence range”, 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 Section 4. 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.

Throughout this work, we focus on tests of design, however equivalence tests are related to the literature on “negligible effects” (Rainey, 2014; Gross, 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 Section 2. We are also developing an R package implementing equivalence based tests of design.

2 Tests of Design

2.1 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 is consistent with the identifying assumptions in their causal research design.

Most causal identification strategies require an assumption that the treatment assignment is unconfounded. In experimental settings, this assumption is met by the random nature of the design, but in observational settings, this necessary assumption is *inherently untestable in any direct manner*. Researchers relying on observational data can *never* prove their design is unconfounded. As discussed in Imbens and Rubin (2015, Chapter 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." 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:

$$\begin{aligned} H_0 &: \text{The data are } \textit{inconsistent} \text{ with the observable implications} \\ &\quad \text{of an unconfounded research design} \\ H_1 &: \text{The data are } \textit{consistent} \text{ with the observable implications} \\ &\quad \text{of an unconfounded research design} \end{aligned} \tag{1}$$

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. While we cannot directly test the distribution of the potential outcomes, we can test how similar the groups look on pre-treatment covariates, which we call a "balance test". Similarity across a large number of pre-treatment 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, while 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 blocking or stratifying, can increase the precision of an estimator. Researchers conduct randomization checks to help defend against the dreaded "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 re-randomization procedures to help improve covariate balance (Morgan and Rubin, 2012).[3]

Balance tests[4] check if the means, or distributions, of pre-treatment variables are approximately the same among treatment and control units. There also exist omnibus tests for overall balance (Hansen and Bowers, 2008; Caughey, Dafoe and Seawright, 2017). A related test is a placebo test, which examines the effect of the intervention on a post-treatment variable known to be unaffected by the cause of interest (Rosenbaum, 2002, p. 214).[5] If the intervention were to show a statistically significant correlation with the

[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 re-randomization, researchers may wish to maximize balance on non-stratified 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

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 pre-treatment covariate or the placebo outcome are 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.

2.2 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 is *inconsistent* with the observable implications of an unconfounded design, for example that there is substantial imbalance in the pre-treatment covariates. Only with sufficient data should one reject the null hypothesis of imbalance in pre-treatment covariates and post-treatment placebo outcomes. That is, they should provide *statistically* significant evidence to reject their data is inconsistent with a valid design, which they encode as a lack of *substantively* significant differences.[6] However, common current practice is for researchers to use a statistical test that employs null of no difference[7] between the two groups as an indirect way of testing if the data are consistent with an unconfounded design. A design is deemed consistent with a valid research design if the statistical test[8] fails to provide evidence in favor of a difference, i.e. a large p-value.[9] This

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]Here, we use the idea of substantive significance to indicate theoretically meaningful differences, whereas statistical significance indicates evidence against a null hypothesis.

[7]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.

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

[9]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 balance being declared if the p-value is higher than 0.15 or 0.2.

approach could be loosely described as incorrectly equating "non-significant difference with significant homogeneity" (Wellek, 2010, p. 3). A high p-value from such a test fails to reject 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.[10] 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. 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 of using lack of statistical significance as evidence in favor of statistical equivalence is a well-documented problem (Gill, 1999). The main issue is that people tend to incorrectly conflate low power with inconsequential 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 the physical cost of distance to polling place on turnout. The authors employ a matching algorithm to match

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

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.

The authors 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.[11] However, their argument would be strengthened with statistical evidence supporting the strength of their design. We will return to this example in Section 4.1 using an equivalence test to evaluate if their data provides 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 pre-treatment covariates are substantively unimportant.

2.3 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 are 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:

$$H_0 : \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \quad \text{or} \quad \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L \quad \text{versus} \quad H_1 : \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U$$

where μ_T and μ_C refer to the mean of the treated and control groups, respectively, for a given variable. ϵ_U and ϵ_L refer to the upper and lower bounds for which two groups

[11] "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, p. 123)

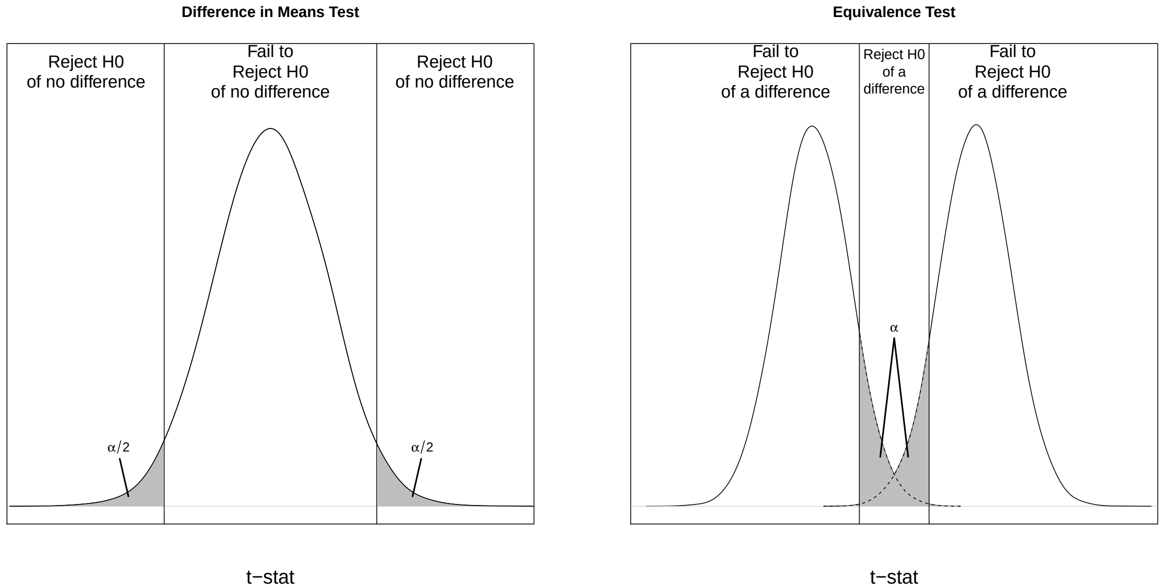

Figure 1: Tests of equivalence versus tests of difference. The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of one type of equivalence test–the Two One Sided t-test (TOST)–under the null hypothesis of difference.

are considered equivalent. Choosing appropriate values for ϵ_U and ϵ_L is the most important aspect of equivalence testing, and is discussed in detail in Section 3.1. 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 samples 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 t-statistic 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 pre-specified size in favor of the alternative of a difference less than that size when the critical value lies in the shaded region. We discuss the mechanics and interpretation of equivalence testing in detail in Section 3, 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 (Rainey, 2014; Gross, 2014; Esarey and Danneman, 2015) for the study of negligible, or substantively insignificant, effects.[12] The negligible, or substantive significance, approach evaluates the confidence range of the parameter, and determines if it lies entirely within ("negligible") or outside ("substantively significant") the null effect range. Both Rainey (2014) and Gross (2014) recommend the use of the 90% confidence interval, and determining if this interval lies entirely within a substantively defined equivalence range. This is effectively the same as the TOST. We show, in Appendix SI-5, why this approach can allow researchers to construct a statistical test with zero power.[13] For this reason, we recommend the equivalence t-test approach, which is more powerful, particularly in data sets with smaller sample sizes.[14] We build on the equivalence tests presented by Rainey (2014) and Gross (2014) by presenting additional equivalence tests appropriate

[12]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 lie outside a range of theoretically unmeaningful values (Gross, 2014), and "negligible effects" involve proving that an effect lies within a range of theoretically unmeaningful values (Rainey, 2014). In the parlance of equivalence tests, "negligible effects" are a straight forward application of an equivalence test, typically centered on zero, whereas "substantive significance" is often operationalized as showing that an α-level 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 post-treatment variable that is hypothesized to lie within a specified range.

[13]Briefly, the problem lies in the fact that 90% confidence intervals have a minimal interval length for a given sample size and standard deviation, and researchers can set an equivalence range that is smaller than this length.

[14]The additional power in the equivalence t-test describe here comes from accounting for the non-central t distribution in the testing procedure.

for different distributions, departures from the null, as well as randomization inference versions.

2.3.1 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 away from zero, which will increase the p-value of at least one of the one-sided tests, depending on if 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 credible equivalence range gives the smallest equivalence range supported by the data at the α-level, which the author should defend as substantively inconsequential to support their 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 (2008) 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 non-significant findings. This logic, while correct for outcomes in which there is a theorized non zero effect of an intervention, would not hold if a researcher theorized a negligible effect unless an equivalence test is used on the outcome. While it is incorrect to accept a null of no difference in a low power situation, and 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 skirting the issue of the ambiguity of lack of statistical power.

3 Mechanics of an Equivalence Test

Implementing an equivalence test requires that a researcher define a few parameters, most importantly the equivalence range.[15] 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 and how these tests can be used with false discovery rate correction methods.

3.1 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

[15]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.

equal.[16] 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.

3.1.1 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 given application given concerns about bounding bias.[17]

Researchers that have advocated for equivalence type approaches often tout the value of requiring researchers to transparently define and defend their equivalence range on theoretical grounds, which is not possible in traditional balance tests. As Rainey (2014, p. 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, p. 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 credible equivalence range, described below, provides a more transparent way for authors to

[16]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.

[17]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 is a primary reason that researchers should conduct sensitivity analyses to check for the robustness of their results.

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 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 0 and given their sample size, to detect equivalence.[18] 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 at a given equivalence range.

3.1.2 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.[19] While 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

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

[19]For most of the tests described in this paper it is fairly simple to choose the equivalence range based on substantive knowledge of the data. For some tests, the range can be specified in terms of standardized differences, such as for the t-test for equivalence and the Mann-Whitney test. For these tests, ranges are defined in standardized differences since they rely on a test statistic that is standardized by the variance, such as the t-statistic. This is similar to why critical constants for the t-test for difference rely on standardized deviations. Other tests, which test directly for equivalence of the raw difference in means, can be specified on the scale of the variable. In the case of the TOST ratio test, the range of equivalence can be defined in terms of the ratio of the means. Table SI-1.1 provides either a standard range of equivalence used in the literature or an equation for translating a substantively defined ϵ on the scale of the variable into a standardized difference.

range. When a researcher is interested in a specific outcome we recommend the equivalence range be $\pm$ one standardized effect size, using Glass' Δ, which is standardized by the standard deviation in the control group[20], 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, pre-treatment covariates are rarely so highly correlated with the outcome[21]; it is an assumption similar to the one made in other sensitivity analyses (Rosenbaum and Silber, 2009). If researchers are concerned about non-linearities between the variable and the outcome, they may wish to scale the standardized effect size by some non-linear 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. 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.[22] Further recommended default equivalence ranges for different tests, appropriate for different data types, are discussed further in (Wellek, 2010, pg. 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 credible equivalence range as inconsequentially small.

[20]We choose Glass' 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).

[21]If researchers intend on using 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).

[22]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, Imai, King and Stuart (2006, p. 221) recommend the strictest range of 0.2 for judging "adequate" balance. Our simulation studies found 0.2σ to be a very conservative range.

3.1.3 The Credible Equivalence Range

Since there naturally will be disagreement over an appropriate equivalence range, we recommend inverting the equivalence test to produce a "credible equivalence range", which is akin to a confidence interval. The credible equivalence range is a symmetric interval defined by the largest difference at which the null hypothesis of difference is rejected at a pre-specified α. The credible equivalence range specifies the smallest equivalence range supported by the observed data.[23] In other words, the difference between 0 and the maximum of the credible equivalence range quantifies the degree of uncertainty we have over the true degree of imbalance, and the researcher can be assured that at least (1-α)% of the time the truth will lie within that range.

Researchers should focus on defending differences in the credible equivalence range as inconsequential rather than on the p-value associated with the equivalence test. As long as the credible equivalence range is reported, readers can judge for themselves whether this range constitutes equivalence on the pre-treatment covariate or placebo outcome. Unlike the p-value for the equivalence test, an advantage of the credible equivalence range is that it is invariant to the researcher's chosen equivalence range, and therefore provides an objective value that researchers and the community can consider. The advantage of this is that it removes the researcher degree of freedom in defining the equivalence range, and forces the researcher to defend the range as substantively inconsequential for bias.

3.2 Conducting the t-test for Equivalence

Just as there are a variety of tests for evaluating difference, there are many equivalence tests. The most appropriate test statistic depends on the type of variable and the desired

[23] In the case of a very small observed difference, it can be the case that the inverted range can support an equivalence range of near zero. In this case, we define with credible equivalence range as the observed standardized mean difference, which is a conservative range.

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 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 towards infinity, and increasingly disjoint as the variance of the distributions shrinks towards 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, pg. 120). For simplicity, assume that $X_{Ti} \sim N(\mu_T, \sigma)$ and $X_{Ci} \sim N(\mu_C, \sigma)$, then the equivalence t-test uses the following hypothesis test.

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

versus

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

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

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

. This test statistic is distributed non-central t with $N-2$ degrees of freedom (Wellek, 2010, pg. 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 non-central F, with the 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 non-central F distribution with level α, degrees of freedom $1, N-2$, and non-centrality parameter $\lambda_{nc}^2 = mn\epsilon^2/N$. If the ϵs were not symmetric, then we would have the rejection rule:

$$C_{\alpha;m,n}(\epsilon_L, \epsilon_U) < T < C_{\alpha;m,n}(\epsilon_L, \epsilon_U)$$

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 two groups in favor of the alternative of an inconsequential difference. Otherwise we fail to reject the null of non-equivalence. In addition to the rejection decision, researchers should also analyze the credible equivalence range, which gives the minimum equivalence range supported by the data. In the case that the credible equivalence range is small, then the researcher can be confident that the data provides strong against a substantial difference. If the range is large, then the researcher may call in to question the equivalence of the two groups. Researchers should also be aware of the power of their test. Further discussion of the power of equivalence tests is discussed in Appendix SI-5.

3.3 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 conveys against the null hypothesis, in this case 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 is 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 is consistent with the observable implications of an unconfounded design".[24] It is important to note that this does not mean we have accepted that our design is unconfounded. Our p-values will now encode a metric for how much information the data has against 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 non-statistical metric for balance assessment (e.g. Sekhon (2007); Imai, King and Stuart (2008)), 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 pre-treatment 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 if their design is consistent with an unconfounded design–this is true by design. However, balance on key prognostic variables can increase the likelihood the resulting

[24]Note that our alternative hypothesis is not that the "data *is* unconfounded", but rather that the "data *is consistent* with an unconfounded design".

estimate will be close to the truth. The equivalence tests discussed here are consistent with this aim, and should have desireable 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 re-randomization can encode their notion of "similar enough" in to 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.[25] 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 significant bias.

There is a healthy literature on sensitivity analyses, e.g. Rosenbaum and Silber (2009) and Imbens and Rubin (2015), for assessing possible remaining unmeasureable confounding in causal effect estimates, and tests of designs 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 designs with sensitivity analyses on the final estimates when providing evidence to strengthen the claims of their designs.

[25] 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.

3.4 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 pre-treatment covariates, which reference a hypothetical super-population, are inappropriate because they are contradictory to the non-random nature of the observed sample (Imai, King and Stuart, 2008; Austin, 2008). One solution to this issue is to conduct tests that are conditional on the realized sample assignment 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, p. 224). In addition to being conditional on the observed sample, the permutation tests are exact and 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 non-equivalence can be rejected if both corresponding permutation p-values are less than the level of the test, α.[26]

[26] Simulations showing properties of this test are provided in Section SI-3.

3.5 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 a researcher 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 pre-treatment covariates. Wellek (2010) provides the equivalence version of Hotelling's T^2, and Fisherian tests, such as those in 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 twenty variables, and observes a significant difference for one, should they 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 reject hypotheses–or the family wise error rate–the probability of committing any type 1 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 conducting.

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 proportion of false discoveries–the false discovery rate–or the probability of one false discovery–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, however, 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.

4 Examples

4.1 Example: Brady and McNulty (2011)

To illustrate the merits of equivalence tests, we return to the example of Brady and McNulty (2011), discussed in Section 2. 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 pre-treatment difference is "highly significant", although "substantively rather small" (p. 123). If the conventional decision rule over adequate balance is followed, then one would question the "as-if" random identification assumption.

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 0.024 miles or 42 yards. We use an equivalence interval, based on the standardized effect size on the percentage of in-person polling place turnout, with an ϵ of ± 1.5 standard deviations (amounting to about 0.037 miles or 65 yards). Note that is a case where 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 turnout that limit an acceptable distance. We also compute the credible equivalence range which is the smallest equivalence interval supported by the data ($\alpha = 0.05$) 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 $\epsilon = 0.037$ miles? This null is rejected with a *p*-value that is essentially zero. Given our pre-specified equivalence interval, we consider the two samples to be well balanced on this variable. When we invert our test, we find that the credible equivalence range, supported at the $\alpha = 0.05$ level, is 1.16 standardized units or 0.028 miles (49 yards). Whether or not 0.028 miles is of concern, worthy of further adjustment, such as through regression, should be debated by subject area experts.

4.2 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 who are not assigned quotas until the next election. Using a 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 don’t 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

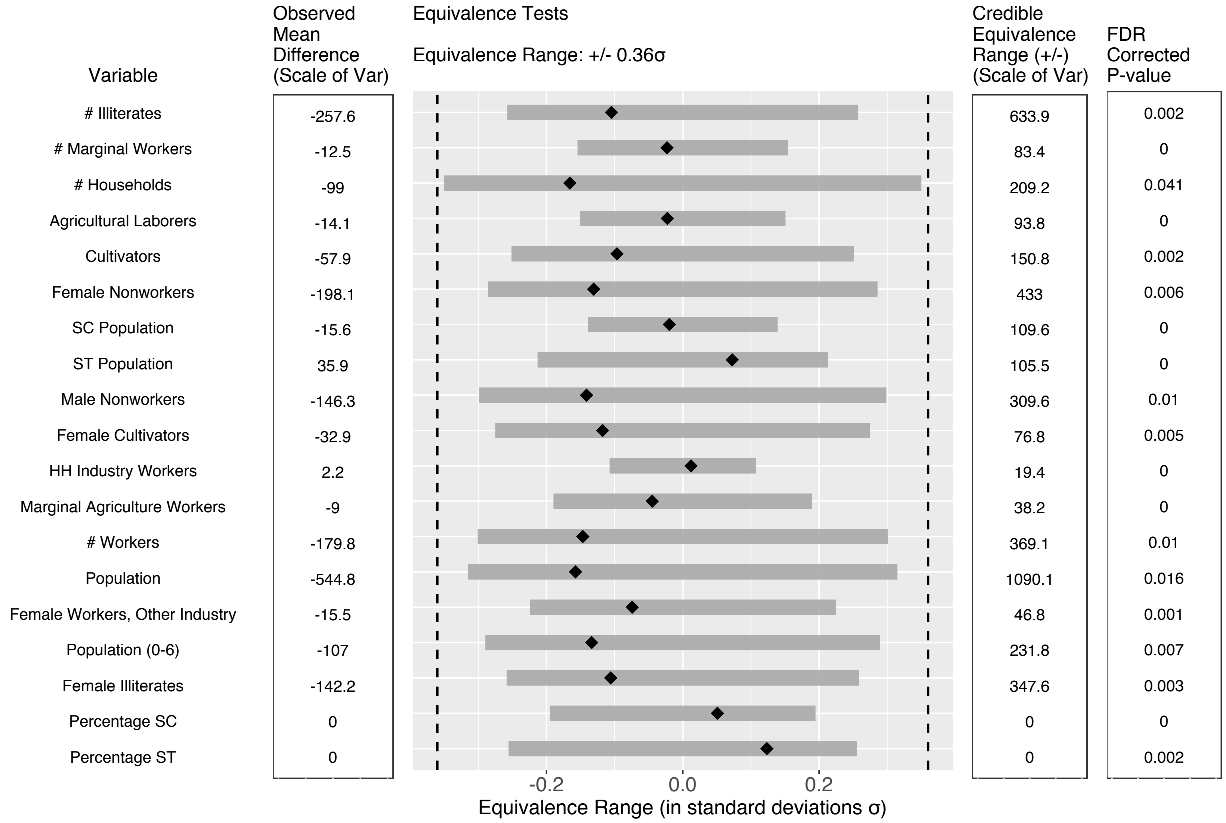

Figure 2: The figure above presents the results of equivalence tests. The "Observed Mean Difference" is the mean of the treated group minus the mean of the control group. The vertical dashed lines represents the hypothesized equivalence range, defined as the standardized effect size on the outcome of 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 "Credible Equivalence Range" 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.

that the design is consistent with as-if random, the presented balance tests do not necessarily provide statistical evidence consistent with their claim. In Figure 2 we conduct the same balance tests, this time using equivalence tests and applying an FDR correction.

As can be seen in Figure 2, the equivalence tests indicate 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 manuscript 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 is $\pm$210 households, or $\pm$433 female workers in a village a substantively inconsequential difference in this 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.

5 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. The argument of this essay is that this 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 allows 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 (RDD) imply continuity of observable variables, matching and natural experiments imply balance and synthetic matching implies a similar time trend on pre-treatment outcomes. Particularly with RDD and synthetic matching, further work must be done on the most appropriate equivalence test. Related, researchers often are 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 where the treatment assignment mechanism is well understood and the identifying assumptions seem quite plausible, our burden of proof should be lower. For example, this is true for randomization checks in experiments where the researcher controlled or knows the randomization. In other cases, such as for 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

Austin, Peter C. 2008. "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003." *Statistics in Medicine* 27(12):2037–2049.

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." *The American Political Science Review* 105(1):115–134.

Caughey, Devin, Allan Dafoe and Jason Seawright. 2017. "Nonparametric Combination (NPC): A Framework for Testing Elaborate Theories." *The 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):417–446.

Di Nardo, John E. and Jorn-Steffen Pischke. 1997. "The Returns to Computer Use Revisted: Have Pencils Changed the Wage Structure Too?" *The Quarterly Journal of Economics* .

Dunning, Thad. 2010. Design-Based Inference: Beyond the Pitfalls of Regression Analysis? In *Rethinking Social Inquiry: Diverse tools, Shared Standards*. Lanham, MD: Rowman & Littlefield, pp. 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(01):95–111.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.

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):775–788.

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):2050–2054.

Hansen, Ben B. and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23(2):219–236.

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." *The Annals of Applied Statistics* pp. 849–870.

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*. 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):325–337.

Morgan, Kari Lock and Donald B Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *The Annals of Statistics* 40(2):1263–1282.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4):1083–1091.

Rosenbaum, Paul R. 2002. *Observational Studies (Springer Series in Statistics)*. 2nd ed. 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):501–511.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2(3):808–840.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *Journal of Politics* 78(3):941–955.

Sekhon, Jasjeet S. 2007. "Alternative balance metrics for bias reduction in matching methods for causal inference." *Survey Research Center, University of California, Berkeley* .
URL: *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):363–378.

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press.

Westlake, Wilfred J. 1976. “Symmetrical Confidence Intervals for Bioequivalence Trials.” *Biometrics* 32(4):741–744.

Supplementary Information for "An Equivalence Approach to Balance and Placebo Tests" *

Erin Hartman[†] F. Daniel Hidalgo[‡]

March 19, 2018

Abstract

The rise of design-based inference 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 the equating of non-significant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data is *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.

***** All errors are our responsibility. *****

*Thanks to Jasjeet Sekhon, Philip Stark, Hans Noel, and Santiago Olivella for their comments and encouragement and to Kosuke Imai's Research Group and the SLAMM! 2016 Conference participants for valuable feedback.

[†]Department of Politics, Princeton University, ekhartman@princeton.edu.

[‡]Department of Political Science, Massachusetts Institute of Technology, dhidalgo@mit.edu

SI-1 Additional Statistical Tests for Equivalence

In many cases, researchers may be interested in testing for non-equivalence of different parameters of interest. This section outlines alternative tests for equivalence, some culled from the extant literature and others created for the problem at hand. Table SI-1 summarizes the tests. The "Type of Data" column describes the type of data each test is appropriate for and the "Randomization Inference" column describes whether the test a randomization version of a common test. The test statistic and rejection rule are also described for each test. Finally, the "Epsilon Range" column describes the recommended epsilon, or the standard in the literature where appropriate, denoted ϵ_{def}, and where available, the equation for translating substantively motivated ϵs, which are on the scale of the variable and denoted ϵ_{sub}, into the scale of the test. Δ refers to one standardized mean difference on the outcome of interest using the standard deviation in the control group, and Δ_{pooled} refers to one standardized mean difference on the outcome using the pooled standard deviation. If data is not available on the outcome of interest, researchers should use the defaults discussed in Section 3.1. The mathematical notation and steps for implementation for each test are described in detail in Appendix SI-2.

This table is intended to serve as a simple reference for practitioners, and it is not exhaustive of the types of equivalence tests available. Users should consult Wellek (2010) for a detailed discussion of the equivalence testing literature. A general method for equivalence testing is described in (Wellek, 2010, Chapter 3). For example, if researchers have paired designs, they should use the McNemar equivalence test described in (Wellek, 2010, Sec. 5.2). If they have blocking, they may wish to use the general appraoch to conduct an equivalence version of the Cochran-Mantel-Haenszel test. Blocking and clustering can be easily incorporated in to the non-parametric versions of the tests. Standard adjustments can also be made to the standard errors for the t-statistics in the equivalence t-test and the TOST.

Table SI-1: A summary of commonly used versions of equivalence tests.

Test Name	Type of Data	Test Statistic	Rejection Rule	Epsilon Range
Equivalence t	Asympt. Normal sample mean	$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{\sum_{i=1}^{m}(X_{Ti}-\bar{X}_T)^2+\sum_{j=1}^{n}(X_{Cj}-\bar{X}_C)^2\right\}^2}$	$\lvert T\rvert < C_{\alpha;m,n}(\epsilon)$	$\epsilon_{def} = \Delta$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
Two-One Sided (TOST) t	Asympt. Normal sample mean	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{SE(\bar{X}_T - \bar{X}_C)}$ and $T_L = \frac{\bar{X}_T - \bar{X}_C - \epsilon_L}{SE(\bar{X}_T - \bar{X}_C)}$	$T_U < -t_{\alpha,m+n-2}$ and $T_L > t_{\alpha,m+n-2}$	$\epsilon_{def} = \Delta_{pooled}$ $\epsilon = \epsilon_{sub}$
TOST Ratio t	Asympt. Normal sample mean	$T_L = \frac{\bar{X}_T - \epsilon_L \bar{X}_C}{S\sqrt{1/m+\epsilon_L^2/n}}$ and $T_U = \frac{\bar{X}_T - \epsilon_U \bar{X}_C}{S\sqrt{1/m+\epsilon_U^2/n}}$	$T_U < -t_{\alpha,m+n-2}$ and $T_L > t_{\alpha,m+n-2}$	$\epsilon_{def} = [0.8, 1.25]$
Exact Fisher Binomial	Binary	$\rho = p_T(1-p_T)/p_C(1-p_C)$	$p_{m,n;\epsilon}(x\vert s) < \alpha$	$\epsilon_{def} = 0.85$ $\epsilon = \frac{\log(1+2\epsilon_{sub})}{\log(1-2\epsilon_{sub})}$
Mann-Whitney	Any Continuous Distribution	$W_+ = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{I}(X_{Ti} - X_{Cj})$	$\frac{\left\lvert W_+ - 1/2 - \frac{\epsilon_1 - \epsilon_2}{2}\right\rvert}{\hat{\sigma}[W_+]} < C_{MW}(\alpha; \epsilon_1, \epsilon_2)$	$\epsilon_{def} = \Delta$ $\epsilon = \Phi(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}) - \frac{1}{2}$
Non-parametric Equivalence t	Any Continuous	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{\hat{\sigma}(\bar{X}_T - \bar{X}_C)}$ and $T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{\hat{\sigma}(\bar{X}_T - \bar{X}_C)}$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = \Delta$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
Non-parametric TOST (npTOST)	Any Distribution	$T_U = \bar{X}_T - \bar{X}_C - \epsilon_U$ and $T_L = \bar{X}_T - \bar{X}_C - \epsilon_L$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = \Delta_{pooled}$ $\epsilon = \epsilon_{sub}$
Non-parametric Mann-Whitney	Any Continuous Distribution	$T_U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{I}(X_{Ti} - X_{Cj}) - (1/2 + \epsilon_U)$ and $T_L = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{I}(X_{Ti} - X_{Cj}) - (1/2 - \epsilon_L)$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = \Delta$ $\epsilon = \Phi(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}) - \frac{1}{2}$

SI-2 Formalization of Additional Statistical Tests for Equivalence

SI-2.1 Two-One-Sided Test and Intersection Union Tests

Rather than studying the standardized difference, as is used in the equivalence t-test discussed the main body of the text, researchers may wish to conduct a test for equivalence of the raw mean difference. This can be accomplished using a Two-One-Sided-Test (TOST) (Berger and Hsu, 1996). The TOST test is conducted using two one sided t-tests centered around the bounds of the equivalence range. One advantage of the TOST is that it allows for the researcher to define the equivalence range on the scale of the variable of interest as opposed to standardizing substantive ranges. The TOST test can also be adapted to test for equivalence of the ratio of the means of the two groups, instead of the raw difference between the means. The TOST ratio test has the advantage of having an absolute scale that is independent of the scale of the variable of interest. This test is used by the FDA for declaring generic drugs as equivalent to brand-name drugs. In that case, the two drugs are declared equivalent if the ratio of the mean effect of the two drugs falls within the range [0.8, 1.25].

The TOST is a type of intersection union test, which are a way of testing multiple hypotheses at once. They are set up in the following manner:

$$H_0 : \theta \in \cup_{i=1}^{k} \Theta_i \qquad \text{versus} \qquad H_1 : \theta \in \cap_{i=1}^{k} \Theta_i^c \tag{1}$$

where θ is the parameter if interest and Θ is the parameter space. The overall null hypothesis, H_0 is rejected at the α level if all of the individual null hypotheses, H_{0i}, are rejected and the α level. Note that this can be a conservative test, depending on how the rejection region for the combined test is determined (Berger and Hsu, 1996). The typical TOST t-test is a type of intersection union test in which the hypotheses are set up as:

$$H_0 : \mu_T - \mu_C \geq \epsilon_U \cup \mu_T - \mu_C \leq \epsilon_L \qquad \text{versus} \qquad H_1 : \epsilon_L < \mu_T - \mu_C < \epsilon_U \tag{2}$$

A t-test is conducted for both of the null hypotheses, i.e. a test one sided test for $\mu_T - \mu_C \geq \epsilon_U$ and a one sided test for $\mu_T - \mu_C \leq \epsilon_L$. The overall null hypothesis is rejected at level α if the associated p-value for each of the individual hypotheses is less than α. Commonly, the null hypothesis is defined in terms of the ratio of μ_T and μ_C, thus making the hypotheses of the form:

$$H_0 : \frac{\mu_T}{\mu_C} \geq \epsilon_U \cup \frac{\mu_T}{\mu_C} \leq \epsilon_L \qquad \text{versus} \qquad H_1 : \epsilon_L < \frac{\mu_T}{\mu_C} < \epsilon_U \tag{3}$$

This test, using the ratios, is used frequently to test the bioequivalence of generic drugs versus non-generic drugs in medicine. In that case, the ϵs are chosen as $\epsilon_U = 1.25$ and $\epsilon_L = 0.8$, the current standard of the FDA. Setting up the hypotheses as a ratio has advantages such as putting the metric of difference on an absolute scale instead of on the scale of the variable. Berger and Hsu (1996) show that the ratio test is also conducted using a t-test, however the test statistic is adjusted as such:

$$T_L = \frac{\bar{X_T} - \epsilon_L \bar{X_C}}{S\sqrt{1/m + \epsilon_L^2/n}} \qquad T_U = \frac{\bar{X_T} - \epsilon_U \bar{X_C}}{S\sqrt{1/m + \epsilon_U^2/n}} \tag{4}$$

The overall null hypothesis is rejected $T_L \geq t_{\alpha,m+n-2}$ and $T_L \leq -t_{\alpha,m+n-2}$.

SI-2.2 Exact Fisher Binomial Test for Equivalence

The Fisher type exact test is well adapted to equivalence between two groups with binary outcomes. This test is based on the odds ratio as opposed to the mean difference between the two groups. Wellek (2010) discusses the advantages of choosing the odds ratio over the difference of p_T and p_C, however the basic point can be illustrated as follows. If the test statistic is defined as the difference in the probability of success between the two groups, i.e. $p_T - p_C$, then as p_T approaches 0 or 1, the range of values for which p_C could be called equivalent is diminished. If equivalence is defined as the two groups having a difference in probability of success of no more than 0.1, then if $p_T = 0$, p_C must be between 0 and 0.1. However, if $p_T = 0.5$, then p_C can be between 0.4 and 0.6. If the odds ratio is used as the test statistic, this shrinking of possibilities for p_C as p_T approaches 0 or 1, or vice versa, is not an issue. The Fisher type test for binary data tests whether the odds ratio is within a specified range, typically centered around 1. There are many other equivalence tests for binary data that focus on the raw difference in probabilities of success discussed in Barker, Rolka, Rolka and Brown (2001).

We will call the rate of units with a response value of 1 in the treatment condition p_T and the rate of units with a response value of 1 in the control condition p_C. The test statistic is the odds ratio of the two groups, $\rho = p_T(1 - p_T)/p_C(1 - p_C)$, the advantages of which are discussed in Appendix A. The hypothesis using the odds ratio as the test statistic is then set up as:

$$H_0 : 0 < \rho \leq \epsilon_L \text{ or } \epsilon_U \leq \rho < \infty \qquad \text{versus} \qquad H_1 : \epsilon_L < \rho < \epsilon_U \tag{5}$$

with $\epsilon_L < 1 < \epsilon_U$. The optimal solution to this test is based on R.A. Fisher's exact test for the homogeneity of two binomial distributions, based on the conditional distribution of the odds ratio sum of the number of successes in the treated and control groups. The distribution of this test statistic follows an extended hypergeometric distribution (Wellek, 2010). For simplicity, assume that the sample sizes are the same and that the ϵs are chosen symmetric around 1. The test rejects the null hypothesis of non-equivalence if the associated p-value of the test statistic is less than the α level of the test, where the p-value is calculated as:

$$p_{n,\epsilon}(x|s) = \sum_{j=s-\max(x,s-x)}^{\max(x,s-x)} h_s^{n,n}(j;\epsilon) \tag{6}$$

with

$$h_s^{n,n}(x;\epsilon) = \frac{\binom{n}{x}\binom{n}{s-x}\epsilon^x}{\sum_{j=\max(0,s-n)}^{\min(s,n)}\binom{n}{j}\binom{n}{s-j}\epsilon^j} \qquad , \max(0, s-n) \leq x \leq \min(s,n) \tag{7}$$

Wellek (2010) outlines the rejection rule in the case of unequal sample size and/or a non-symmetric equivalence range. We have implemented these scenarios in our accompanied R package, but the intuition behind the test is the same. In the case of binary data, multiple tests for testing the equivalence of the probabilities of success of the two groups instead of the odds ratio are also discussed in Barker et al. (2001). Most of these tests are based on the $100(1-\alpha)$ confidence interval of the t-test, which corresponds to a TOST t-test.

SI-2.3 Mann-Whitney Test for Equivalence

Researchers may prefer to use a test sensitive to differences in distribution rather than differences in means, akin to the Kolmogorov-Smirnov test (Sekhon, 2007). The Mann-Whitney test for equivalence is an asymptotically distribution free test that is sensitive to divergences between two continuous distributions (Wellek, 2010). If two distributions are equivalent then the probability that any treated observation is greater than any control observation should be approximately 1/2, thus equivalence is defined as a range around this point. Therefore, the Mann-Whitney tests uses a rank-sum statistic to test whether or this probability is within a small range around 1/2. If the two distributions are non-equivalent, then the bulk of the treated units should lie to one side of the median of the ranked treated and control observations. This test is especially advantageous because it does not depend on the underlying distributions of the treated and control groups so long as they are both continuous. This test is asymptotically distribution free and robust to outliers in the data (Wellek, 2010). Failure to reject the null of nonequivalence in this test implies not simply that the two groups differ in their means, but is designed to test for departures in other parts of the distribution as well. Lehmann (1975) originally outlined the properties of the U-statistic, and Wellek (2010) further discusses the implementation for equivalence testing. Extensions are studied in Arboretti, Carrozzo and Caughey (2015).

The basic outline of the test is as follows. Let $X_{Ti} \sim F \; \forall i = 1, \ldots, m$ and $X_{Cj} \sim G \; \forall j = 1, \ldots, n$, then they equivalence hypothesis for the non-parametric test can be set up as:

$$H_0 : \pi_+ \leq 1/2 - \epsilon_1 \text{ or } \pi_+ \geq 1/2 + \epsilon_2 \qquad \text{versus} \qquad H_1 : 1/2 - \epsilon_1 < \pi_+ < 1/2 + \epsilon_2 \tag{8}$$

where $\pi_+ = P[X_{Ti} > X_{Cj}]$. Here, π_+ is estimated using the Mann-Whitney statistic, W_+ defined as:

$$W_+ = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathcal{I}(X_{Ti} - X_{Cj}) \tag{9}$$

Intuitively, if the two samples are equivalent, then the chance that any given treated unit's value of X_{Ti} lies above any given control unit's value X_{Cj} is about one half. The epsilons, then define the tolerance around one half for which the two groups would still be equivalent. The hypothesis test is thus set up with a null hypothesis that $P[X_{Ti} > X_{Cj}]$ is either smaller or larger than the range of equivalence, and the alternative is that $P[X_{Ti} > X_{Cj}]$ lies within the range of equivalence. The statistical test is carried out with the following rejection rule:

$$\text{Reject nonequivalence iff} \qquad \frac{\left| W_+ - 1/2 - \frac{\epsilon_1 - \epsilon_2}{2} \right|}{\hat{\sigma}[W_+]} < C_{MW}(\alpha; \epsilon_1, \epsilon_2) \tag{10}$$

where

$$C_{MW}(\alpha; \epsilon_1, \epsilon_2) = \chi^{2^{-1}}(\alpha; df = 1, \lambda^2_{nc} = \tfrac{(\epsilon_1 + \epsilon_2)^2}{4\hat{\sigma}^2[W_+]})$$

The Mann-Whitney statistic is asymptotically distributed, thus allowing for the approximation of the critical value[1]. The properties of the Mann-Whitney test for equivalence are studied further in Wellek (1996).

SI-3 Sample specific versions of parametric tests

If the data is drawn from an experiment, or quasi-experiment, where the assignment mechanism is known and random, we can conduct our tests using the permutation distribution of the data, also known as randomization inference. Here we discuss generally how to conduct the permutation tests and specifically how to conduct non-parametric versions of the tests described above. Permutation tests are tests designed to test for the exchangeability of two groups and are well suited to the problem at hand of validating quasi-experimental designs. In theory, these observational designs should guarantee exchangeability between the two groups. The non-parametric versions of the above tests all use an IUT approach where the the bounds of the equivalence range are used as the strict nulls, and TOST tests are conducted based on the permutation distribution of the test statistics. If the p-value for both associated tests is less than α, then the test rejects the null of non-equivalence.

The non-parametric TOST t-test (npTOST) is set up using the same hypotheses as in (2). To test the null that $\mu_T - \mu_C \geq \epsilon_U$ the permutation distribution given the assignment mechanism and the strict null hypothesis that $\mu_T - \mu_C = \epsilon_U$ is calculated, or approximated if the number of permutations is large, using a one-sided test with the strict null of a treatment effect of ϵ_U (Rosenbaum, 2002). It is important to note that if the design includes block or cluster randomization, the permutations should be of this assignment mechanism. Then, an exact p-value corresponding to the null $\mu_T - \mu_C = \epsilon_U$ is calculated.

[1] The variance of W_+, regardless of the underlying distributions F and G is always defined as $\text{Var}[W_+] = \frac{1}{mn}\left(\pi_+ - (m+n-1)\pi_+^2 + (m-1)\Pi_{X_T X_T X_C} + (n-1)\Pi_{X_T X_C X_C}\right)$ where $\Pi_{X_T X_T X_C} = P[X_{Ti_1} > X_{Cj}, X_{Ti_2} > X_{Cj}]$ and $\Pi_{X_T X_C X_C} = P[X_{Ti} > X_{Cj_1}, X_{Ti} > X_{Cj_2}]$ (Wellek, 2010)

The p-value for the analogous test given the assignment mechanism and the strict null hypothesis that $\mu_T - \mu_C = \epsilon_L$ is also calculated. If both p-values are less than the level of the test, α, then the two groups are statistically equivalent, with the overall p-value corresponding to the maximum of the two individual one sided test p-values. The test is inverted to construct the credible equivalence range by finding the minimum (symmetric) ϵ for which $p < \alpha$. The non-parametric Mann-Whitney test is constructed analogously. However, the test statistic there is the W_+, as defined in Table SI-1, and it is tested around the strict null of $W_+ = 1/2 - \epsilon_L$ and $W_+ = 1/2 + \epsilon_U$. As before, the two one-sided permutation p-values are calculated, and the test rejects the null of non-equivalence if both p-values lie below α. For further discussion of the permutation based one sided test, see Lehmann (1975).

Figure SI-1 shows a simulation study of the npTOST. Units are drawn from a standard normal, and the constant, additive effect is set to τ. A total sample size n is selected, with complete randomization conducted of $n_t = n_c = n/2$. For power calculations, the equivalence range is set at $\epsilon_l = -0.2$ and $\epsilon_u = 0.2$. Results are shown in the solid lines. The test is underpowered when there are only 100 units in each group, but power increases as n grows.

Coverage rates are shown in dashed lines, with the the nominal coverage rate of 95% noted by a thin, solid line. As can be seen, coverage rates are close to the nominal rate, with coverage growing conservative as the truth approches zero. Because the range is defined as a symmetric range, that always includes zero, power should be conservative at, and near, zero.

SI-4 Traditional vs. Equivalence Tests – A simulation

The right panel of Figure 1 illustrates why the two-one-sided t-test (TOST) for equivalence will not conflate power with similarity as can happen in the common misinterpretation of

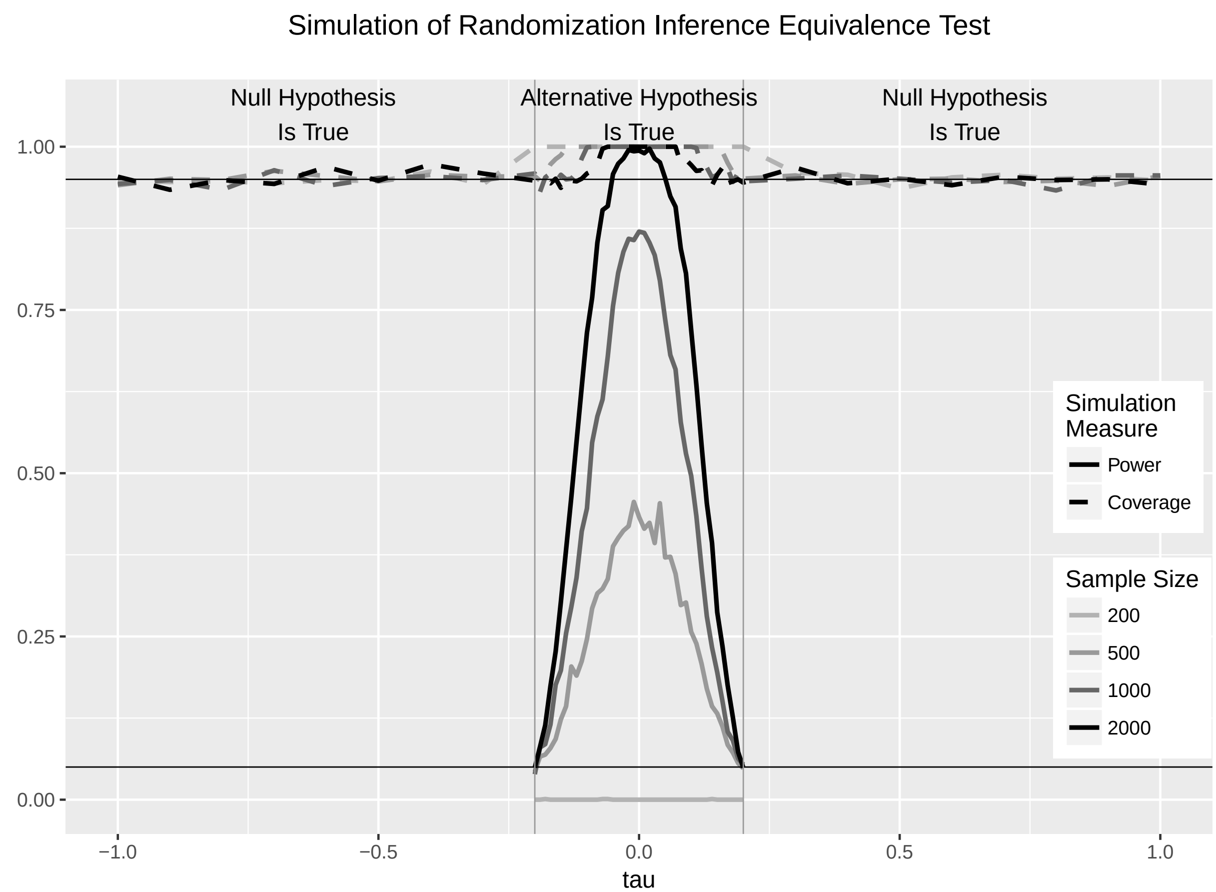

Figure SI-1: Simulations of Randomization Inference versions of the TOST. 1,000 simulations are conducted for each sample size, with 500 permutations in each simulation.

traditional balance tests using tests-of-difference.[2] In this example, the equivalence test is conducted by looking at the distribution of two null hypotheses. The lower curve is the distribution of the t-statistic around the hypothesized difference of ϵ_L and the upper curve is the distribution of the t-statistic around the hypothesized difference of ϵ_U. The two groups are considered equivalent if the observed t-statistic lies in the shaded region, i.e. the equivalence range, meaning the p-value for both tests is less than $\alpha/2$ if the ϵs are symmetric around zero. The area of the shaded region is equal to the level of the test, α. Therefore, this test controls the type I error consistent with our null hypothesis, which is rejecting a null of substantial difference when, in fact, one exists.

While it is not justified by the tests, many researchers are tempted to present large p-values on a tests-of-difference as evidence against confounding. How do equivalence tests help guard against this temptation? Recall there are three factors that can result in the t-statistic lying in either the tails or the center of the t-distribution under a null, depicted in the left panel of Figure 1. If the mean difference between the two populations is small, then the t-statistic will also be small, which is desirable for declaring the two groups equivalent. As the standard deviation grows, the t-statistic will also move towards the center, which is also desirable behavior with respect to determining equivalence. Reducing the sample size can shift the t-statistic from the center to the tail. This fact, as raised by Imai, King and Stuart, this impact of sample size can, incorrectly, tempt researchers to believe their data is consistent with an unconfounded design. The converse problem is that when one has very large sample sizes, minute differences may be statistically significant even if substantively meaningless. In equivalence tests, however, if the sample size is small, holding all else constant, the t-statistic will move away from zero, which will increase the p-value of at least one of the tests, depending on if the observed difference is above or below zero, thus making it less likely that we will reject the null of a substantial difference.

[2] In Section 3 we discuss the t-test for equivalence, which is related to the TOST, but is more powerful in small samples. The intuition that follows is the same, however.

Therefore, the power of the test behaves as we would expect with respect to sample size.

To show how equivalence tests will not tempt researchers to conflate low power with similarity, we turn to an example inspired by a simulation in Imai, King and Stuart (2008, p. 495). Imai, King and Stuart show how sample size affects the t-statistic by taking a covariate from an imbalanced observational study and conducting a t-test after randomly dropping an increasingly large percentage of the controls. They are decreasing the sample size, but in expectation they are not affecting the overall balance between the treated and control units. In this case, then, we should be unlikely to reject a null of a substantial difference. What we would like to see is that, regardless of the sample size, we fail to reject the null of a substantial difference. We'd like to see a fairly stable, flat line in which most simulations show evidence of imbalance. Using the common decision rules using traditional tests, in which individuals show evidence of balance when $p < 0.05$, however, as sample size decreases, even as imbalance remains, the p value will increase, and thus researchers may be tempted to conflate power with similarity. It should be noted that this same issue arises, although has not been addressed, when selecting the appropriate window size for regression discontinuity designs under the randomization framework (Cattaneo, Frandsen and Titiunik, 2015).

In Figure SI-2 we recreate this simulation, using data from Blattman and Annan's (2010) study on child soldiering. They examine the socioeconomic consequences of abduction by the Lord's Resistance Army, one of the main combatant groups in Uganda's civil war. In this simulation, we examine a balance test on age, which they point to as one of the most important covariates determining selection into treatment. Age is imbalanced, they argue, because the rebel army tended to target somewhat older children. The simulation study mimics Imai, King and Stuart's in that we randomly drop an increasingly large percentage of the controls (non-abductees). For each of the 5000 iterations, we conduct both a traditional and an equivalence based t-test. The figure shows the percentage of simulations that that show evidence of imbalance. For traditional tests, we use a decision

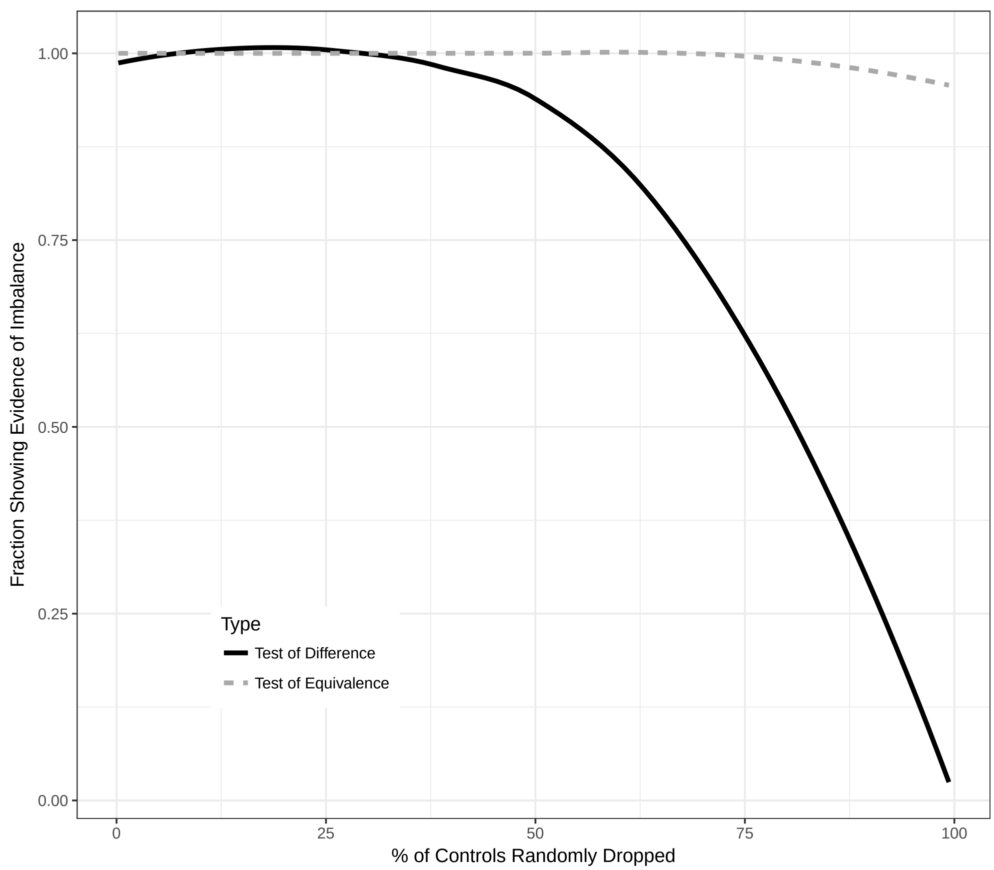

Figure SI-2: The behavior of tests of difference and equivalence when a varying percentage of the control units are dropped from the sample. The red line is the proportion of rejections of the null of no mean difference ($\alpha = .05$) using the difference in means *t*-test. The blue dashed line is the proportion of non-rejections of the null of difference using an equivalence t-test with an equivalence range of 0.2 of a standard deviation. For the difference test, as increasing numbers of control units are dropped, the share of tests falsely indicating increased balance increases. For the equivalence test, the share of tests falsely indicating increased balance are largely unaffected by sample size.

rule in which if $p < 0.05$, we say this is evidence of imbalance. For equivalence tests, if we fail to reject the null at the 5%-level, we say this is evidence of imbalance. It is important to note that the two groups are imbalanced, and randomly dropping controls does not, on average, affect the level of imbalance. Our equivalence range is 0.2 of a standard deviation in age. As can be seen, as sample size drops, the common, if incorrect, interpretation of the test-of-difference is less likely to provide evidence of imbalance, using this metric. However, equivalence tests fail to reject the null of a substantial difference, even as sample size decreases. As the percentage of the controls drops approaches 85 to 90%, the t-test for equivalence does reject the null of substantial difference in some simulations. This may be due to the fact that a few of the random draws lead to control samples that were similar to the treated group, given the very small number of controls in these draws.

SI-5 Negligible Effects and the 90% Confidence Interval

Equivalence and negligible effects are related concepts, the later of which has been addressed recently in the political science literature. Both Rainey (2014) and Gross (2014) argue that, rather than conducting the equivalence t-test, researchers should analyze the location of the the 90% confidence interval and its relation to the equivalence range. Rainey (2014) argues researchers should evaluate if the 90% confidence interval of the estimate lies entirely within the equivalence range, whereas Gross (2014) provides numerous interpretations of different relationships between the confidence interval and the equivalence range. Both argue that the best way to define the equivalence range is based on substantive knowledge.

We assert that the equivalence t-test, or a binomial analog, are superior to the 90% confidence interval range. By arguing for researchers to first define a substantive equivalence range, and conduct the 90% confidence interval test, researchers can create a test for themselves with zero power. Figure SI-3 shows simulations exemplifying this facet of

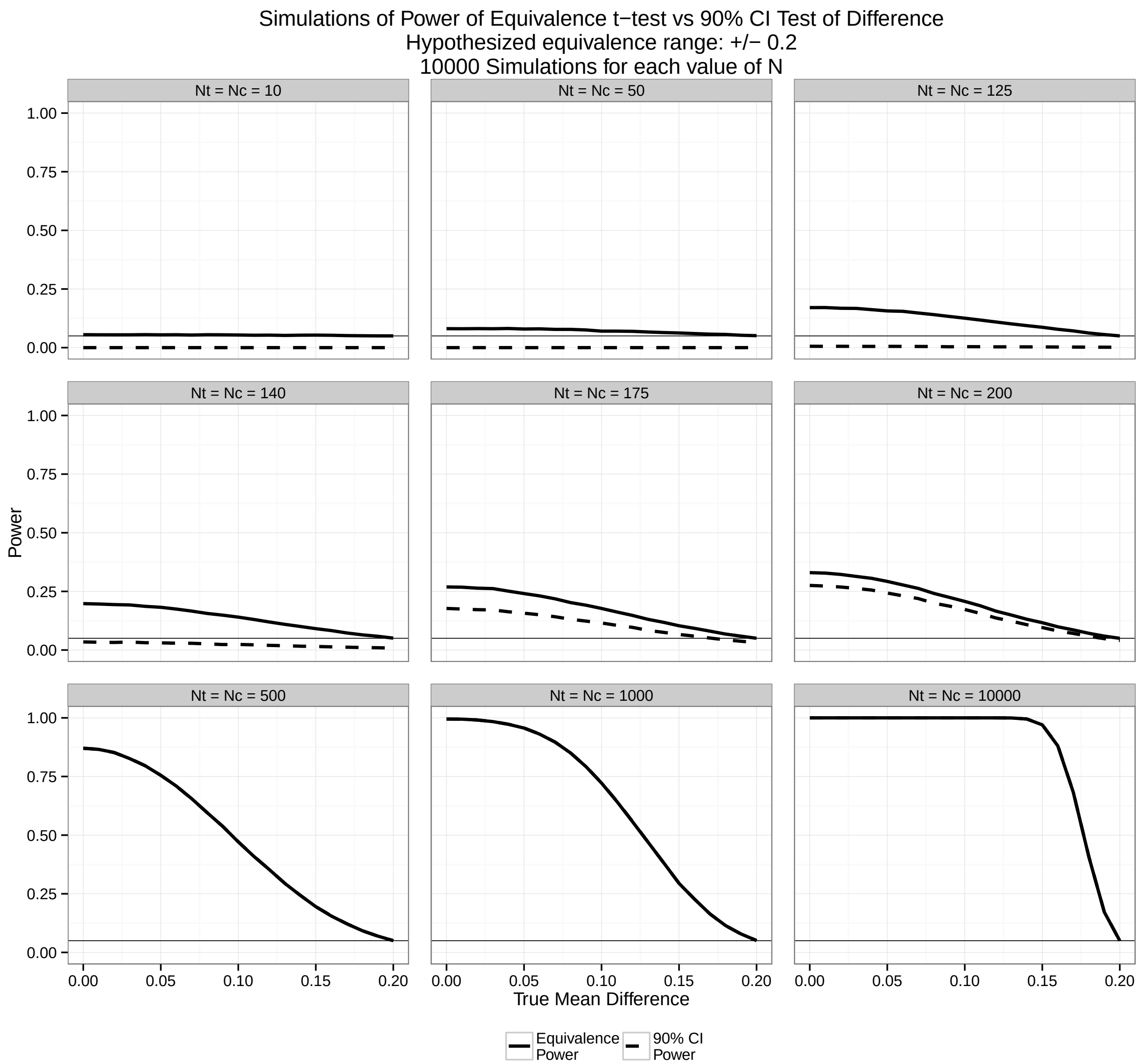

Figure SI-3: Power of the Equivalence t-test vs the 90% Confidence Interval Test. The horizontal black line is located at 0.05.

the test. The 90% confidence interval has a minimum size, conditional on α-level, the standard deviation, and the sample size. If the practitioner defines a substantive range that is smaller than this minimum possible size, then the 90% confidence interval will have zero power to declare the two groups equivalent. Note that the equivalence t-test always maintains at least α-level power. What this means, in effect, is conditional on the observed sample size, sample estimate of the standard deviation, and desired α, there is a minimum size the practitioner can define. Figure SI-4 shows the minimum sample size necessary in each group in order for a given symmetric equivalence range, assuming two $\sim$N(0,1) variables.

Even with the use of the equivalence tests for negligible effects, power remains an issue if the true effect lies close to the edge of the equivalence range. While the assumption of a true difference of zero, where the maximum power is achieved, is justified for tests of design, the point of a negligible effect test is to test if the true effect lies anywhere within the equivalence range. Figure SI-5 shows how the power of the equivalence t-test drops off as the true difference approaches the edge of the equivalence range, even for large values of n.

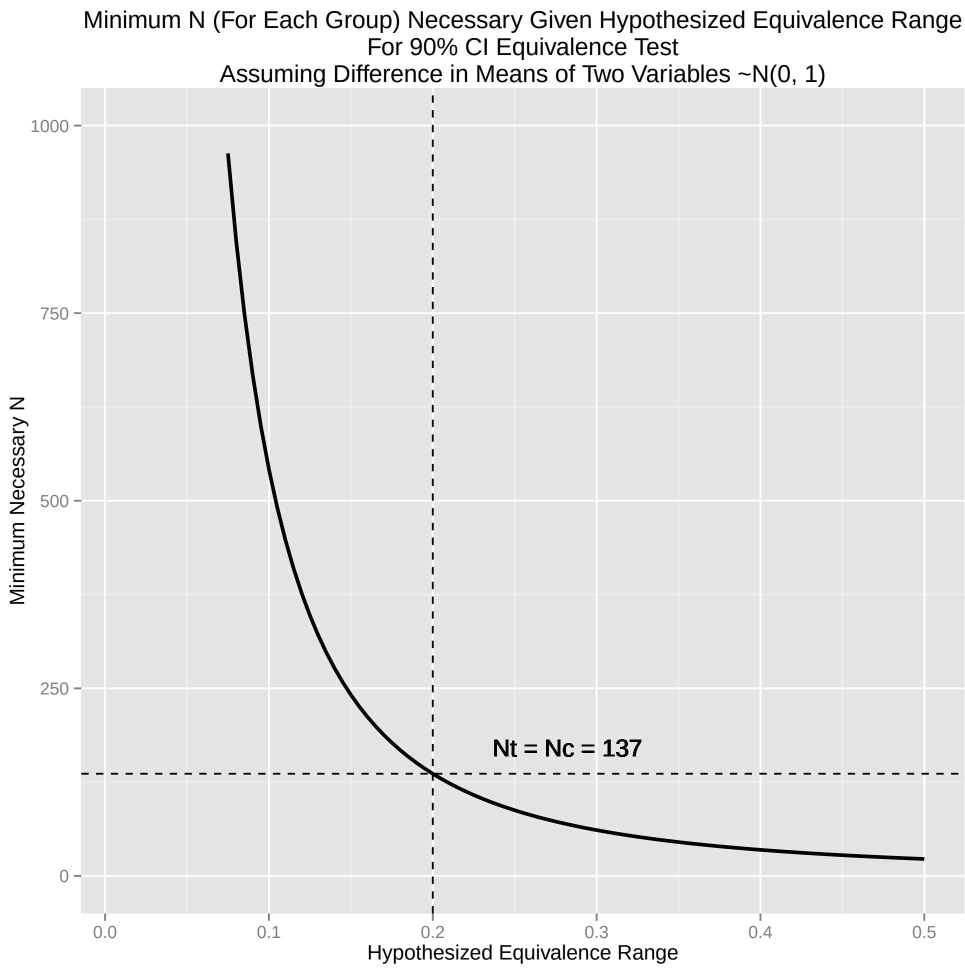

Figure SI-4: Sample size necessary in each group to maintain at least 0.05% power for the 90% confidence interval test at a given equivalence range, assuming two equal size groups both distributed $\sim$N(0,1)

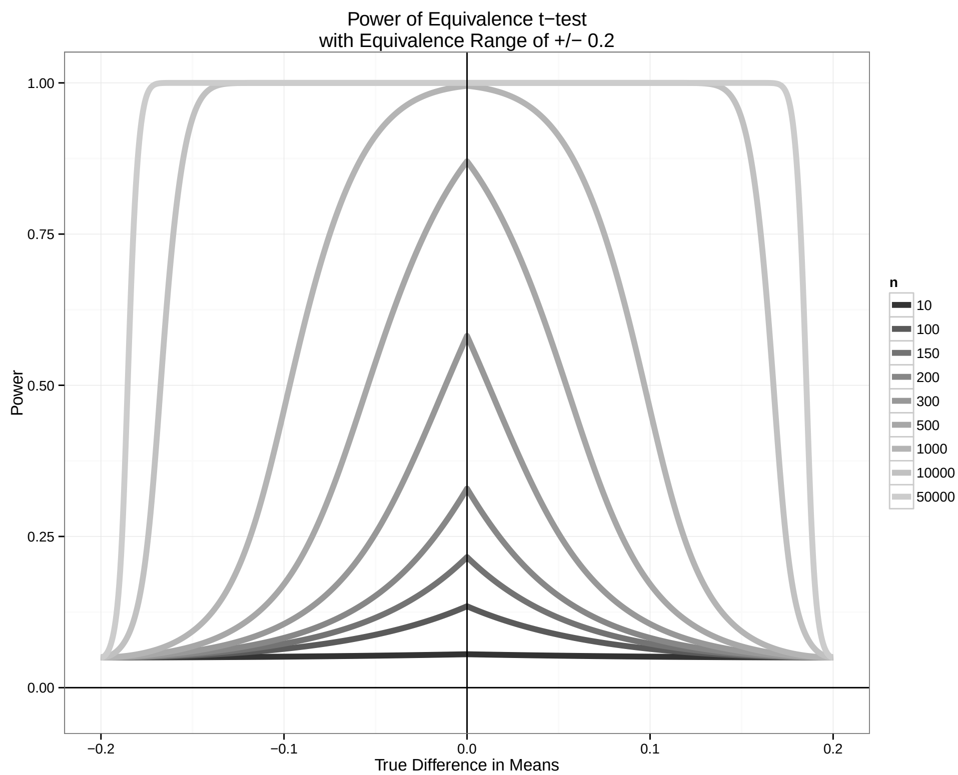

Figure SI-5: Power of the equivalence t-test with an equivalence range between two $\sim$N(0,1) variables with sample size n at different values of the true difference.

SI-6 Applying Equivalence Tests to Natural Experiments in the Social Sciences

Does the use of equivalence tests make a difference in practice? To show that it does, we apply the two sample t-test for equivalence to ten studies culled[3] from Dunning's (2010*a*) literature review of natural experiments in the social sciences. From each study, we selected one covariate that was tested for balance. Each study typically examined several covariates, so when possible we selected the pre-treatment outcome (the outcome variable as measured prior to the intervention) and, failing that, a variable that in our judgement, was closely related to the outcome of interest. The papers, which are on a diverse set of treatments in a variety of contexts, are listed in Table SI-2. For the equivalence range, we chose 0.2 of a standard deviations, following Cochran and Rubin (1973, p. 422)'s discussion.

The results of the equivalence test on a pre-treatment covariate in the ten natural experiments are shown in Table SI-2, along with the conventional difference-in-means t-test *p*-value. Nine out of the ten natural experiments reported difference-in-means t-test *p*-values greater than 0.05, thus failing to reject the null hypothesis of no mean difference and consequently "passing" their balance test. If the equivalence test is used, however, only for five[4] of the ten studies can we reject the null hypothesis of a mean difference $|\epsilon| > 0.2\sigma$ with a 0.05 level of significance, where σ is the pooled standard deviation of the covariate. Four of the studies failed to reject the null hypothesis of a difference, but also failed to reject the null hypothesis of no mean difference. Consequently, in these four

[3]In order to carry out the test, we required the mean difference, the standard error of the mean difference, and the sample size in each treatment condition. All natural experiments in Dunning's (2010*a*) list that reported this information were used.

[4]One study, Chattopadhyay and Duflo (2004) was borderline with a *p*-value of 0.1, but given the low power of the test for a study of that sample size, we would consider this covariate to be balanced.

Paper	Treat	Covariate	Mean Diff	N	P (diff)	P (equiv)	Power
Di Tella, Galiani and Schargrodsky (2007)	Property Rights	Years of Education	0.08	1080	0.75	0.00	0.88
Hyde (2008)	Observer Visit	Challengers Vote Share	0.00	1763	0.78	0.00	0.82
Annan and Blattman (2010)	Abduction	Father's Years of Schooling	-0.05	741	0.86	0.00	0.68
Ferraz and Finan (2008)	Gov. Audit	Reelection rates (2004)	0.02	373	0.69	0.05	0.29
Chattopadhyay and Duflo (2004)	Reservations	Wells	-0.02	161	0.80	0.10	0.10
Card and Krueger (1994)	Minimum Wage Increase	Employment - November 1992	-3.30	384	0.40	0.16	0.23
Dunning (2010*b*)	Reservations	Mean Scheduled Tribe Population	60.67	200	0.40	0.27	0.13
Lyall (2009)	Artillery Shelling	Rebel Presence	0.10	147	0.20	0.52	0.1
Ho and Imai (2008)	Ballot Order Position	Registered Democratic	-0.02	80	0.46	0.52	0.10
Lee (2008)	Democratic Victory	Democratic Win Prob	0.14	610	0.00	0.82	0.59

Table SI-2: Equivalence tests in ten natural experiments. Table shows the difference in means, the standard difference-in-means T-test p-value, the total number of units, the p-value from a two sample T-test of equivalence with an $\epsilon = .2$ of a standard deviation and the results of a power calculation. Studies are ordered by equivalence test p-value.

Paper	Covariate	Std. Inverted ϵ	Unstd. Inverted ϵ
Blattman (2010)	Father's Years of Schooling	0.11	0.59
Di Tella, Galiani and Schargrodsky (2007)	Years of Education	0.11	0.63
Hyde (2008)	Challengers Vote Share	0.12	0.02
Ferraz and Finan (2008)	Reelection rates (2004)	0.20	0.13
Chattopadhyay and Duflo (2004)	Wells	0.28	0.20
Card and Krueger (1994)	Employment - November 1992	0.32	17.27
Dunning (2010*b*)	Mean Scheduled Tribe Population	0.35	252.17
Lee (2008)	Democratic Win Prob	0.41	0.29
Lyall (2009)	Rebel Presence	0.48	0.33
Ho and Imai (2008)	Registered Democratic	0.77	0.13

Table SI-3: Inverted equivalence tests in ten natural experiments. Table shows upper boundary of the 95% confidence interval of the two sample T-test of equivalence in standardized and unstandardized units.

cases, the conventional decision rule would declare the natural experiments to be balanced, while our proposed test would not. Of course, failing to reject the null hypothesis of a difference by no means invalidates these studies' conclusions, but merely suggests that insufficient information exists to affirmatively declare that the treatment and control groups on these particular covariates are well balanced. At a minimum, our results suggest that these scholars could take special care to show that the design is valid using other design tests or robustness checks.

In Table SI-3, we present the maximum value of ϵ for which which we can reject the null hypothesis of non-equivalence, given the observed difference. We present both the standardized and unstandardized values of this credible equivalence range. The credible equivalence range is useful here because it can give the reader a sense of the smallest equivalence range supported by the data at a given significance level. Because researchers' opinions may differ over how small an equivalence range chosen ex-ante should be, reporting the inverted interval can allow readers to draw their own conclusion over the degree of balance evidenced in the data.

Supplementary Materials References

Annan, Jeannie and Christopher Blattman. 2010. "The Consequences of Child Soldiering." *The Review of Economics and Statistics* 92(4):882–898.

Arboretti, Rosa, Eleonora Carrozzo and Devin Caughey. 2015. "A Rank-based Permutation Test For Equivalence and Non-inferiority." *Statistica Applicata - Italian Journal of Applied Statistics* 25(1):81 – 92.

Barker, Lawrence, Henry Rolka, Deborah Rolka and Cedric Brown. 2001. "Equivalence Testing for Binomial Random Variables: Which Test to Use?" *The American Statistician* 55(4):279 – 287.

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

Card, David and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4):772–793.

Cattaneo, Matias D., Brigham R. Frandsen and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate." *Journal of Causal Inference* 3(1):1 – 24.

Chattopadhyay, Raghabendra and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72(5):1409–1443.

Cochran, William and Donald Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya: The Indian Journal of Statistics* 35(4):417–446.

Di Tella, Rafael, Sebastian Galiani and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* pp. 209–241.

Dunning, Thad. 2010*a*. Design-Based Inference: Beyond the Pitfalls of Regression Analysis? In *Rethinking Social Inquiry: Diverse tools, Shared Standards*. Lanham, MD: Rowman & Littlefield, pp. 273–311.

Dunning, Thad. 2010*b*. "Do Quotas Promote Ethnic Solidarity? Field and Natural Experimental Evidence from India.".
URL: *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.386.7070rep=rep1type=pdf*

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Releaed Audits on Electoral Outcomes." *Quarterly Journal of Economics* pp. 703–745.

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):775–788.

Ho, Daniel and Kosuke Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: The California Alphabet Lottery, 1978-2002." *Public Opinion Quarterly* 72(2):216–240.

Hyde, Susan D. 2008. "The Observer Effect in International Politics: Evidence from a Natural Experiment." *World Politics* 60(1):37–63.

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.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142(2):675–697.

Lehmann, Erich L. 1975. *Nonparametrics*. Springer.

Lyall, Jason. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks?: Evidence from Chechnya." *Journal of Conflict Resolution* 53(3):331–362.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4):1083–1091.

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

Sekhon, Jasjeet S. 2007. "Alternative balance metrics for bias reduction in matching methods for causal inference." *Survey Research Center, University of California, Berkeley* .
URL: *http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf*

Wellek, Stefan. 1996. "A New Approach to Equivalence Assessment in Standard Comparative Bioavailability Trials by Means of the Mann-Whitney Statistic." *Biometrical Journal* 38(6):695–710.

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press.