What's the Alternative?: An Equivalence Approach to Balance and Placebo Tests *

Erin Hartman† F. Daniel Hidalgo‡

April 13, 2016

***** Please do not distribute with out permission from authors. All errors are our responsibility. *****

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 some methodologists have severely criticized the current status quo of relying on hypothesis tests to assess balance on pre-treatment covariates and placebo outcomes because designs can "pass" traditional tests due solely to small sample size, thus conflating statistical insignificance with substantive insignificance. We show that these problems are due to the use of an inappropriate null hypothesis, which can result in the equating of non-significant differences with significant homogeneity. When the hypothesis test is correctly specified so that *difference* is the null and *equivalence* is the alternative, the problems afflicting traditional tests are alleviated. We leverage the statistical literature on tests of equivalence to provide an alternative framework for testing covariate and placebo balance. In addition to their superior statistical properties, 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. To demonstrate these advantages, we apply equivalence tests to a recent natural experimental design that aims to show balance to validate the design.

*Thanks to Jasjeet Sekhon and Philip Stark for their comments and encouragement and to Kosuke Imai's Research Group for valuable feedback.

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

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

1 Introduction

Recent debates over the difficulties of causal inference in the social sciences have spurred a growing literature on how to judge the quality of causal research designs (Austin, 2008; Dunning, 2010*b*; Keele, 2010). An important consequence of this debate is 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). Experimental researchers are expected to provide randomization checks, or balance on pre-treatment covariates, as evidence in favor of their design. These checks help defend against the dreaded "bad draw"[1], and these tests can be used in re-randomization procedures to help improve covariate balance (Morgan and Rubin, 2012). The emerging norm in the literature is that only if the empirical implications of the scholar's causal identification assumptions are borne out in the data, can the causal effect estimates be treated as credible. In the "design-based inference" literature, 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, 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. Current practice, however, is the opposite. The standard approach to validity checks and negligible effects is to use statistical tests which adopt the null hypothesis of no differences in pre-treatment characteristics and placebo outcomes–only rejecting this hypothesis if sufficient data exists to demonstrate otherwise. The implicit decision rule embedded in current practice is that treatment effects should only be estimated if one fails to reject the null hypothesis of no difference (at conventional levels of statistical significance) in tests of design. This approach could be loosely described as equating "a non-significant difference with significant homogeneity" (Wellek, 2010, p. 3). 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. Throughout this work, we refer to tests of design–however the notion of placebo tests can be expanded to include "negligible

[1] "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).

effects" hypothesized to be within an equivalence range by theory rather than by design (Rainey, 2014; Gross, 2014).

To implement our validity tests, we rely on the large literature in biostatistics on equivalence testing (Wellek, 2010; Westlake, 1976). The statistics of equivalence testing largely arose out of the problem in pharmokinetics of statistically demonstrating that two drugs had the same effect, an issue that arises frequently when considering regulatory approval of a generic versus brand-name drug. Drawing upon this line of research, we develop a general procedure for tests of design. First, we present the basic logic of equivalence testing and contrast it with the standard tests of difference. We show that equivalence tests are not susceptible to common critiques of balance tests. Second, we present the details of equivalence versions of frequently used parametric and nonparametric difference tests. Equivalence tests require the ex-ante selection of an "equivalence range": the interval wherein any difference between treatment and control units is declared substantively unimportant. We pay particular attention to the selection of an equivalence range as it is a key distinction between equivalence and conventional hypothesis testing[2]. We also introduce the notion of the "inverted epsilon", a minimum range, invariant to the hypothesized equivalence range, that would be supported at the α-level by the observed data. We also suggest the use of standardized coefficients for scaling the equivalence range to that of the outcome variable in an automated and intuitive way.

Finally, we discuss an advantage of equivalence tests that is often overlooked by researchers implementing tests of design–the multiple comparisons problem. Tests of design often require testing multiple pre-treatment or placebo outcomes, and researchers may also wish to test multiple outcomes for negligible effects as predicted by their theory. In all of these cases, the p-values presented from a test on a single variable will not account for the multiple comparisons being conducted, and we'd expect a false positive rate consistent with our chosen α. Since equivalence tests are designed to control the type I error rate consistent with the researchers hypothesis (i.e. rejecting two groups are equivalent when they are, in truth, equivalent), they lend themselves to the use of vast literature on multiple comparisons. We will address how to implement a multiple testing correction for tests of design.

To show that the use of equivalence tests over difference tests can alter the outcomes of

[2]This is similar to the m range defined in Rainey (2014), or the "effective null" set defined in Gross (2014)

design tests in practice, we apply our approach to Eggers, Fowler, Hainmueller, Hall and Snyder (2015)'s study of regression discontinuity designs for studying close elections. We recover some key findings, but also indicate subgroups where sample size and balance may be conflated.

2 Tests of Design

The most common tests of design are balance and placebo tests. Balance tests, the most frequently used test of design, check if the distributions of pre-treatment variables are approximately the same among treatment and control units. In the observational world, a common scenario where balance is tested in order to justify a design occurs in the analysis of natural experiments. The condition of covariate balance is implied by the "as if" randomization assumption invoked to justify analyzing the data as if it had come from a randomized experiment (Dunning, 2010*b*). 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).[3] If the intervention were to show a correlation with the placebo outcome, then the validity of the research design is called into question. A common feature of these two standard tests is that it is incumbent upon the researcher to demonstrate that the difference between treated and control units on the pre-treatment covariate or the placebo outcome are substantively small and thus not indicative of a flawed design.

Instead of beginning with the assumption that the data is consistent with a valid design, we argue that one should begin with the initial hypothesis that the data is *inconsistent* with a valid research design. Only with sufficient data should one reject the null hypothesis of imbalance in pre-treatment covariates and post-treatment placebo outcomes and declare that the design is valid. Similarly, for the notion of negligible effects, one should begin with the initial hypothesis that a subsantively significant effect exists, and only with sufficient data should one reject this null and declare an effect inconsequential. The conceptual distinction between beginning with a null hypothesis of no difference, as is standard in current practice, versus beginning with a null

[3]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. We use Rosenbaum (2002)'s definition of a placebo outcome, which is a post-treatment outcome for which the effect is known, either by design or substantive theory. In almost all cases, this known effect is 0. Another type of placebo test, which we do not consider, is the use of an alternate treatment, related to the treatment of interest, but whose effect on the outcome is known. A classic example of such a placebo test is Di Nardo and Pischke (1997).

hypothesis of a difference, as we advocate, may seem small, but the practical implications are substantial. Because one typically fixes ex-ante the probability of rejecting the null hypothesis when it is true (type I error), the consequences of increasing the sample size of a study are very different depending on the empirical content of the null hypothesis. The usual approach, as forcefully discussed by Imai, King and Stuart (2008, p. 494), has the unfortunate consequence that the fewer the units in the sample, holding all other factors constant, the more likely the validity tests will be passed. Reversing the null hypothesis, as we advocate, ensures that collecting more data does not reduce the likelihood of passing a validity test. Under our approach, when the research design is valid, additional data only improves one's ability to pass tests of design such as balance and placebo tests. This work expands upon the tests proposed by (Rainey, 2014; Gross, 2014; Esarey and Danneman, 2015) by providing additional statistical tests and use cases, recommendations for how to define the equivalence range, and methods for addressing multiple comparisons.

2.1 Sample Size and Equivalence: Two Examples

To motivate the tests of design proposed in this essay, we consider two recent studies of voter turnout in political science. The first is a small randomized experiment where a newspaper advertisement encouraging turnout was published in four treatment cities, with four cities serving as controls, as discussed in Bowers (2011). The second example is Brady and McNulty (2011), a large natural experiment in Los Angeles in which millions of voters were assigned to new polling stations due to precinct consolidation, which resulted in many voters having to travel farther to vote, thereby increasing the "costs of voting". In both studies, the authors are quite careful in assessing their identification assumptions by conducting tests of design, namely tests of pre-treatment covariate balance.

In reporting balance across treatment and control cities in the newspaper study, Bowers (p. 11) finds a particularly large imbalance on share of the population that is black: a mean difference of -14.4 percentage points. The imbalance in this covariate may indicate that the control cities are poor counterfactuals for the treatment units, but Bowers argues against that conclusion by noting that the null hypothesis of no difference is not rejected at conventional levels with a p-value

of 0.2. Despite passing the traditional balance test, a large observed difference is nevertheless troublesome in that it suggest the potential for biased effect estimates. In this case, the small size of the sample ensures that only extremely high levels of imbalance would lead to a rejection of the null of no difference. In situations like this, the researcher may incorrectly conclude the groups are equivalent due to the small sample size, the wide confidence bounds of which may contain both large differences and negligible effects.

A contrasting scenario is found in Brady and McNulty (2011), where after using a matching algorithm to match voters on a few important covariates, the authors report balance statistics on variables not used in the matching algorithm and report the mean differences at the precinct level. However, the authors do not provide t-test p-values associated with those mean differences, providing the explanation,

> "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. 9)

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. This exemplifies the notion of "substantive significance" or negligible effects–the set of values that researchers believe are unlikely to indicate bias. Their caveat, we argue, reflects a conflict between the purpose for which the typical null hypothesis t-test was designed and the goal of tests of design, namely showing that differences on pre-treatment covariates are substantively unimportant. This paper aims to outline the problems with how these tests are currently conducted and provide a set of statistical tests for equivalence, which are not subject to many of the concerns highlighted in these two examples.

3 Difference versus Equivalence

The current practice for tests of design in the social sciences is to conduct balance and placebo tests using difference-in-means tests, though test statistics which examine other aspects of covariate's distribution, such as Kolmogorov-Smirnov (K-S) tests, are becoming increasingly common.

With these tests, a high p-value corresponds to evidence that the treatment and control groups are different, either in terms of the mean for the standard *t*-test or the empirical cumulative distribution function for the K-S test. In practice, most authors declare adequate balance or placebo equivalence, when these tests show no statistically significant difference between the two groups. A high p-value from such a test fails to provide statistical evidence that the two groups are different which is only indirectly related to showing that they are the same. The problem arises from the fact that the standard tests are designed to control for a type I false positive error of classifying the two groups as different when they are, in fact, the same. If the goal is to control the false positive error rate for classifying the two groups the same when they are in fact different, then the standard test is controlling for the incorrect type of error. We argue that a statistical test showing that two groups are equivalent should control for false positives in which the two groups are classified as equivalent when they are, in fact, different.

Some recent literature has addressed 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), but the practice remains largely absent from hypothesis testing in the social sciences.[4] There does exist, however, a large statistical literature investigating the properties of precisely these types of tests, known as an "equivalence" or "bioequivalence" tests. Wellek (2010) and Berger and Hsu (1996) provide a review of the theory and main uses of 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 equivalent. One example of a t-test for equivalence, also addressed in Rainey (2014), is set up as follows:

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

where μ_T and μ_C refer to the mean of the treated and mean of the control groups, respectively,

[4]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.

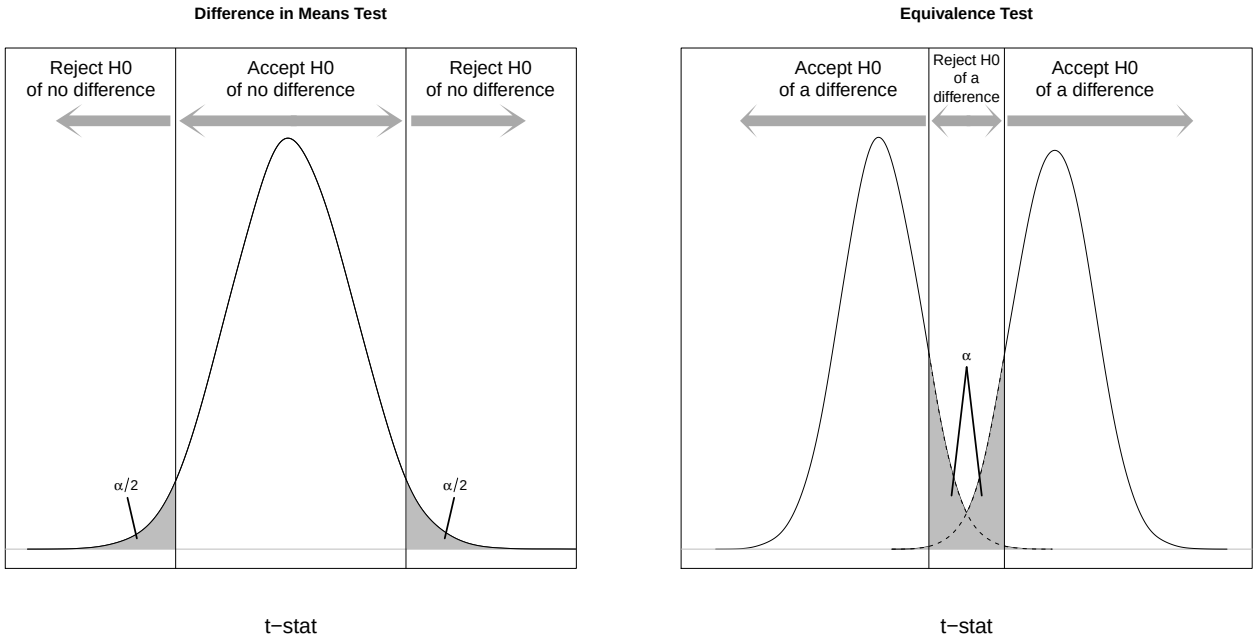

Figure 1: Tests of equivalence versus tests of difference. The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of tests of equivalence under the null hypothesis of difference.

for a given variable. ϵ_U and ϵ_L refer to the upper and lower bounds for which two groups are considered equivalent. This test controls for the type I error of classifying the two samples equivalent (as defined by the equivalence range) when, in fact, they are not. This is one illustrative example of an equivalence test. Alternative versions, which are designed for different types of data or sensitive to different departures of the null are presented in Section 4.

Figure 1 depicts graphically how the traditional balance tests and equivalence tests differ. In traditional balance tests, the means of two groups are declared to be equivalent 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 for the proper type I error as defined by the null of the equivalence test.

There are three factors that can result in the t-statistic lying in either the tails or the center of the t distribution. 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 two t-statistics will also move towards the center, which is also desirable behavior with respect to determining equivalence. More importantly, though, is the concern raised by Imai, King and Stuart (2008): holding the observed mean difference and standard deviation constant, the sample size can shift the t-statistic from the center to the tail. This is an undesirable property for balance and placebo tests because for any given mean difference, dropping observations will be beneficial in terms of "passing" the test. The converse problem is that when one has very large sample sizes, minute differences may be statistically significant even if substantively meaningless. 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. This problem is one of the reasons why Imai, King and Stuart (2008, p. 494) label the principle of balance testing as a "fallacy".

The right panel of Figure 1 illustrates why the t-test for equivalence is not subject to Imai, King and Stuart's critique of balance tests. In this example, the equivalence test is conducted by looking at the distribution of two non-central t-statistics. 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-statistics lie in the shaded region, i.e. the equivalence range. The area of the shaded region is equal to the level of the test, α. Therefore, this test controls for the correct type I error. There is an α probability that the two groups will be deemed equivalent when, in fact, they are different. Why are equivalence tests not subject to the same problems of sensitivity to sample size as the tests of difference? If the sample size is small, holding all else constant, the t-statistics will move towards the center of their respective distributions, thus making it less likely that we will call the two groups equivalent. Therefore, the power of the test behaves as we would expect with respect to sample size.

An example inspired by a simulation in Imai, King and Stuart (2008, p. 495) illustrates the effects of making the null a hypothesis about difference. 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. They then show that the t-statistic decreases, or moves towards in-

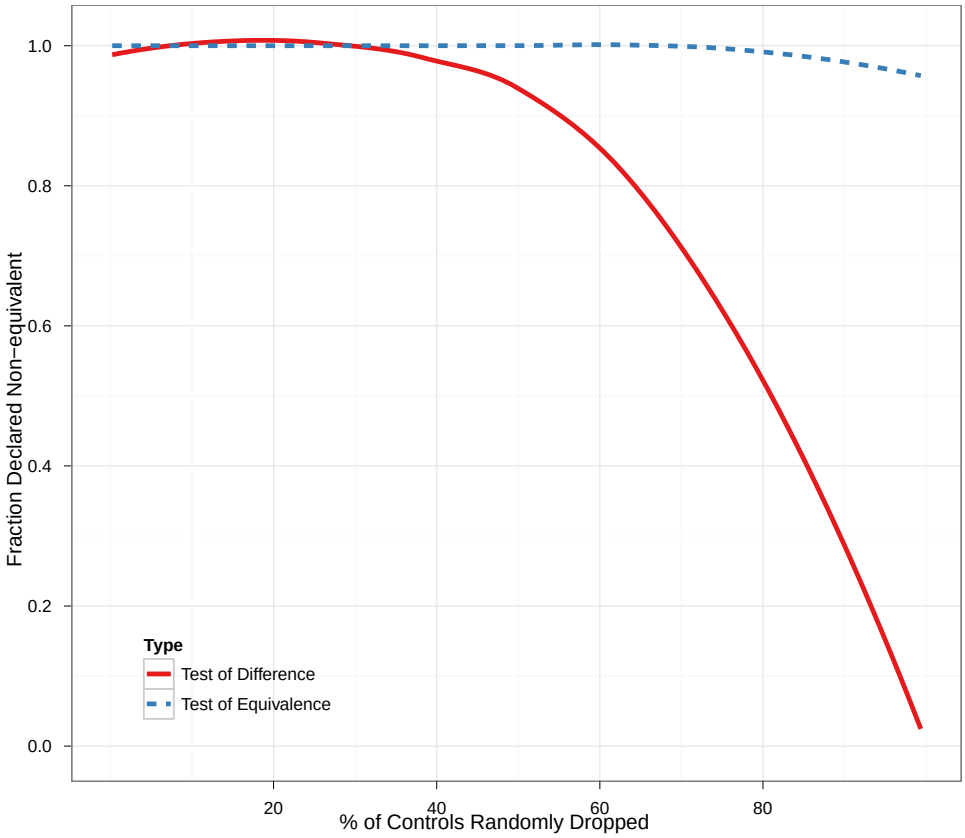

Figure 2: The behavior of tests of difference and equivalence when a varying percentage of the control units are dropped from the sample. The red line is the proportion of rejections of the null of no mean difference ($\alpha = .05$) using the difference in means t-test. The blue dashed line is the proportion 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.

significance, as more control observations are dropped, leading to the conclusion that "[t]he t-test can indicate that balance is becoming better whereas the actual balance is growing worse, staying the same, or improving". This simulation depicts the undesirable behavior of using a difference-in-means test. Austin (2008) also raises this point in justifying his claim that significance testing is inappropriate as a metric for post-matching balance because the post-matching p-values are confounded with sample size.

In Figure 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 are declared non-equivalent. It is important to note that the two groups are imbalanced, and randomly dropping controls does not, on average, affect the level of imbalance. In the case of the difference of means t-test, the groups are declared non-equivalent if they are found to be statistically different at the 5% level. For the t-test for equivalence, the two groups are declared non-equivalent if they fail to reject the null hypothesis of non-equivalence at the 5% level. Our equivalence range is 0.2 of a standard deviation in age. As was shown in the Imai, King and Stuart (2008) simulations, as the number of controls randomly dropped increases, the t-test for difference in means (red line) is increasingly likely to declare the two groups equivalent. However, the t-test for equivalence (blue line) is not subject to this problem. As the number of controls dropped increases, the t-test for equivalence still overwhelmingly declares the two groups non-equivalent. As the percentage of the controls drops approaches 85 to 90%, the t-test for equivalence does declare a few of the simulations equivalent. 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.

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, as articulated by Hansen (2008). Hansen points to the fact that 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. While it may become easier to find treatment effects that achieve significance as sample sizes increase, the associated balance tests will also be harder to pass. This argument, however, hinges on an assumption that statistically significant findings on the outcome are the goal. This argument would not hold if a negligible effect finding

was of interest. Using the equivalence tests discussed here, small sample sizes will tend to make it more difficult to call two groups equivalent, and will also lead to less significant results. Larger sample sizes will not, however, reduce the likelihood of passing the validating test. If one wants to put a higher burden on the tests of design, then the equivalence range should be decreased as sample size increases. By decreasing the range of equivalence as sample size increases, increased burden is once again placed on the equivalence tests as the power to find significant differences in the outcome increases.

3.1 Selecting an Equivalence Range

To conduct an equivalence test, one must choose, prior to conducting the test, an equivalence range or $[\epsilon_L, \epsilon_U]$, i.e the range in which we can consider the parameter of interest in the two groups to be substantively equal.[5] How should one select this interval? Researchers can rely on both design and theory.

For experimental and observational researchers conducting balance checks, one is ultimately interested in bias, yet covariate balance (or the degree of placebo equivalence) is only a proxy for bias in our estimate of the treatment effect parameter[6]. Consequently, what really matters is the unobservable mapping between covariate imbalance and bias. Because this mapping is fundamentally unobservable, our judgements 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 inconsequential.

It should be noted that the tradeoff 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, we recommend that any results from an equivalence test be accompanied by the power of the test, under the assumption that the true difference is 0.[7] In judging the results of a test of design, the power of the test will inform our

[5] 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.

[6] Without assumptions on the mapping between the covariate and the outcome, any level of imbalance could lead to bias of arbitrary magnitude and size

[7] 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.

expectations over the likelihood of rejecting the null of difference at a given equivalence range.

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. 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."

Since there naturally will be disagreement over an appropriate equivalence range, we also recommend inverting the equivalence test to produce an "inverted ϵ". The inverted ϵ is the largest difference at which the null hypothesis of difference would be rejected at a pre-specified α. Akin to a confidence interval, the inverted ϵ specifies the smallest equivalence range supported by the observed data, given the observed difference between the treatment and control group and the observed variance. In other words, the difference between 0 and the inverted ϵ quantifies the degree of uncertainty we have over the true degree of imbalance. As long as the inverted ϵ is reported, readers can judge for themselves whether this range constitutes equivalence on the pre-treatment covariate or placebo outcome. An advantage of the "inverted ϵ" 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. This inversion demonstrates the smallest difference supported by the data and can be an important metric in assessing the quality of a design.

Although we would argue that equivalence ranges are best chosen out of substantive considerations, it is useful to specify default values. While this is an area in need of validation studies, we provide a set of recommendations depending on the aim of the researcher.

When a researcher is interested in a specific outcome, or set of outcomes, we recommend the equivalence range set as one standardized mean difference on the outcome. Assuming a perfect correlation between the variable of interest and the outcome, imbalance outside of this equivalence range in the variable of interest could fully explain the effect size. Therefore, we ensure that the

inverted epsilon, or the minimal equivalence range supported by the data, is contained entirely within the standardized effect size, meaning the researcher can be least 95% confident that the true difference does not lie outside that range. While this is conservative, since the variable of interest is rarely that highly correlated with the outcome[8], it is an assumption similar to the one made in other sensitivity analyses(Rosenbaum and Silber, 2009).

When the researcher cannot benchmark against a standardized effect size, we recommend using $\epsilon = \pm 0.2\sigma$, where σ is the pooled standard deviation of the variable being tested. The inspiration for this default value for ϵ are 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. Ho, Imai, King and Stuart (2006, p. 221) also make this recommendation for judging "adequate" balance. We stress, however, that default values of equivalence should still be given careful consideration for any particular application.

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. 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 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. When possible, researchers are better off defining the equivalence range based on theory and substantive knowledge rather than using default values defined in the literature.

4 The Mechanics of Equivalence Tests

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

[8]Even if the correlation is low, we may be concerned about higher order relationships between the variable of interest and the outcome.

different types of departures of H_0. This section outlines some different types of equivalence tests that are available and the advantages of each type. Because most difference in means tests are conducted using t-tests, we discuss in detail the analogous t-test for equivalence. However, we will discuss other common tests for equivalence that are designed for different distributions and parameters of interest, and which may be more appropriate for small samples. The tests will be described briefly in this section and the formal mathematical outline of the tests can be found in the appendix.

4.1 t-test for Equivalence

As opposed to the t-test for difference, the t-test for equivalence is based on the standardized difference rather than the raw difference in means between the two groups. 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 0 (Wellek, 2010). 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.

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

versus

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

We choose ϵ_L and ϵ_U appropriately, perhaps based on substantive knowledge. Section 3.1 discussed how to choose ϵ. Typically the range of equivalence is symmetric around zero. After defining an equivalence range, the realized test statistic is calculated. The test statistic is the typical two sample t-test statistic, where we define $T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{\sum_{i=1}^{m}(X_{Ti} - \bar{X}_T)^2 + \sum_{j=1}^{n}(X_{Cj} - \bar{X}_C)^2\right\}^2}$. The only difference is now the test statistic is distributed as a non-central F distribution because it is the standardized difference instead of a raw difference. Assume that we define ϵ_U and ϵ_L to be symmetric around 0, i.e. $\epsilon = \epsilon_U = -\epsilon_L$, then the critical value can easily be obtained. The rejection rule now becomes:

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

with

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

where $C_{\alpha;m,n}(\epsilon)$ is the square root of the inverse 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 so as to control for the level of the test. 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, and declare the two groups equivalent. Otherwise we fail to reject the null of non-equivalence. In addition to the rejection decision, researchers should also analyze the inverted ϵ, which provides more detail about the minimum equivalence range that the data can support. In the case that the inverted ϵ is small, then the researcher can be confident that the data provides strong evidence that the two groups are equivalent. If the inverted ϵ is large, then the researcher may call in to question the equivalence of the two groups. In addition to the inverted range, the researcher should also look at the power of the test. If the test fails to reject the null of non-equivalence yet the sample size is small, the power of the test may provide insight in to whether it is a large difference between the two groups or a lack of statistical power which may lead to an increased p-value.

4.2 Alternative 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. The mathematical notation and steps for implementation for each test are described in detail in the appendix. Rather than studying the standardized difference, 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

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

Test Name	Type of Data	Sample Specific	Test Statistic	Rejection Rule	Epsilon Range
Equivalence t	Asympt. Normal sample mean	No	$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{\sum_{i=1}^{m}(X_{Ti}-\bar{X}_T)^2 + \sum_{j=1}^{n}(X_{Cj}-\bar{X}_C)^2\right\}^2}$	$\lvert T\rvert < C_{\alpha;m,n}(\epsilon)$	$\epsilon_{def} = 0.2$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
Two-One Sided (TOST) t	Asympt. Normal sample mean	No	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{SE(\bar{X}_T - \bar{X}_C)}$ 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 $T_L > t_{\alpha,m+n-2}$	$\epsilon_{def} = 0.2\sigma_{pooled}$ $\epsilon = \epsilon_{sub}$
TOST Ratio t	Asympt. Normal sample mean	No	$T_L = \frac{\bar{X}_T - \epsilon_L \bar{X}_C}{S\sqrt{1/m + \epsilon_L^2/n}}$ 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	Yes	$\rho = p_T(1-p_T)/p_C(1-p_C)$	$p_{m,n;\epsilon}(x\lvert s) < \alpha$	$\epsilon_{def} = 0.85$ $\epsilon = \frac{\log(1+2\epsilon_{sub})}{\log(1-2\epsilon_{sub})}$
Mann-Whitney	Any Continuous Distribution	No	$W_+ = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{I}(X_{Ti} - X_{Cj})$	$\frac{\left\lvert W_+ - 1/2 - \frac{\epsilon_1 - \epsilon_2}{2}\right\rvert}{\hat{\sigma}[W_+]} < C_{MW}(\alpha;\epsilon_1,\epsilon_2)$	$\epsilon_{def} = 0.2$ $\epsilon = \Phi(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}) - \frac{1}{2}$
Non-parametric Equivalence t	Any Continuous	Yes	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{\hat{\sigma}(\bar{X}_T - \bar{X}_C)}$ 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 $< \alpha$	$\epsilon_{def} = 0.2$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
Non-parametric TOST (npTOST)	Any Distribution	Yes	$T_U = \bar{X}_T - \bar{X}_C - \epsilon_U$ and $T_L = \bar{X}_T - \bar{X}_C - \epsilon_L$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = 0.2\sigma_{pooled}$ $\epsilon = \epsilon_{sub}$
Non-parametric Mann-Whitney	Any Continuous Distribution	Yes	$T_U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{I}(X_{Ti} - X_{Cj}) - (1/2 + \epsilon_U)$ 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 $< \alpha$	$\epsilon_{def} = 0.2$ $\epsilon = \Phi(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}) - \frac{1}{2}$

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 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).

Finally, the researcher may prefer to use a test sensitive to differences in distribution rather than differences in means, akin to the K-S 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. The rank-sum statistic is also robust to outliers in the data.

4.3 Sample Specific Equivalence Tests

Finally, 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 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 Test principle, each test described above can be tested using an exact test. 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 test a 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.

Table 1 summarizes the tests described above. The "Type of Data" column describes the type of data each test is appropriate for and the "Sample Specific" column describes whether the test is conditional on the observed sample, as opposed to referencing a super-population. The test statistic and rejection rule are also described for each test. Finally, the "Epsilon Range" column

describes the typical epsilon range defined in the literature, 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. This table is intended to serve as a simple reference for practitioners. As mentioned above, the mathematical foundations and steps for implementation are described in detail in the appendix.

5 Negligible Effects and the 90% Confidence Interval

There has long been a debate in the social sciences about how to best prove substantive significance and negligible effects, and the role that the standard null hypothesis test should play in this endeavor, which has been addressed recently in political science (Rainey, 2014; Gross, 2014; Esarey and Danneman, 2015). 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. The recent literature aims to address both of these concerns by evaluating the confidence range of the parameter, and determining if it lies entirely within ("negligible") or outside ("substantively significant") the null effect range.

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 6 in Appendix A.5 shows simulations exemplifying this facet of 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 3 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 4 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.

6 Example: Eggers et al. (2015)

To illustrate the merits of equivalence tests over traditional tests, we reconsider the balance tests conducted in Eggers et al. (2015).[9] In this paper, the authors address recent criticism over the use of regression discontinuity designs of close elections as a natural experimental design following the findings of Caughey and Sekhon (2011) that parties can sort around the cut-point in post-war House elections in the United States. Eggers et al. (2015) analyze over 40,000 additional races in different time periods, geographies within the United States, and other countries. What they find is that, while the post-war United States House does show imbalance, most close elections do not show strong evidence of sorting.

[9] Additional examples are presented in Appendix A.6.

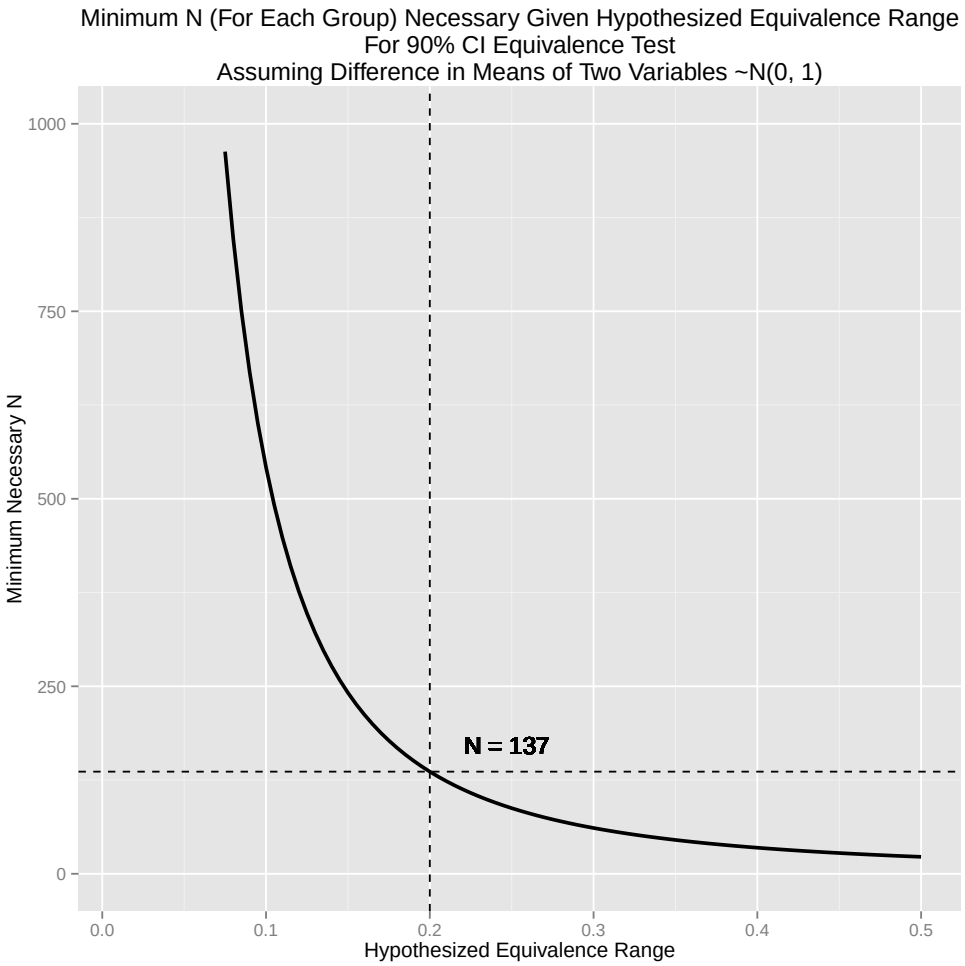

Figure 3: Sample size necessary in each group to maintain at least 0.05% power for the 90% confidence interval test at a given equivalence range, assuming two equal size groups both distributed $\sim$N(0,1)

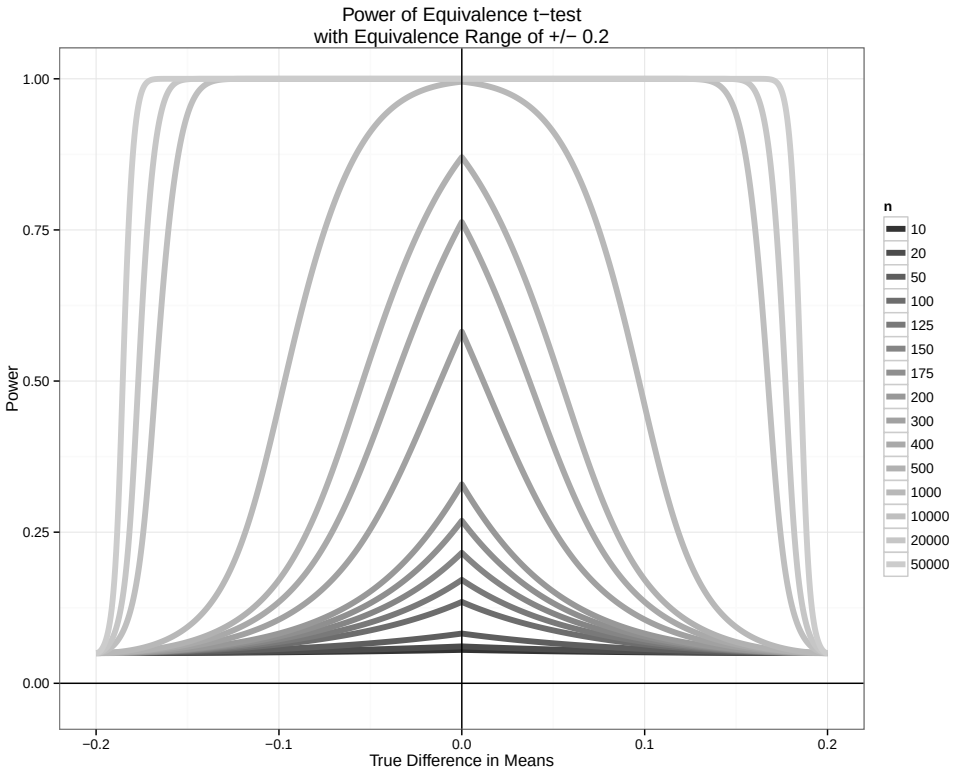

Figure 4: Power of the equivalence t-test with an equivalence range between two $\sim$N(0,1) variables with sample size n at different values of the true difference.

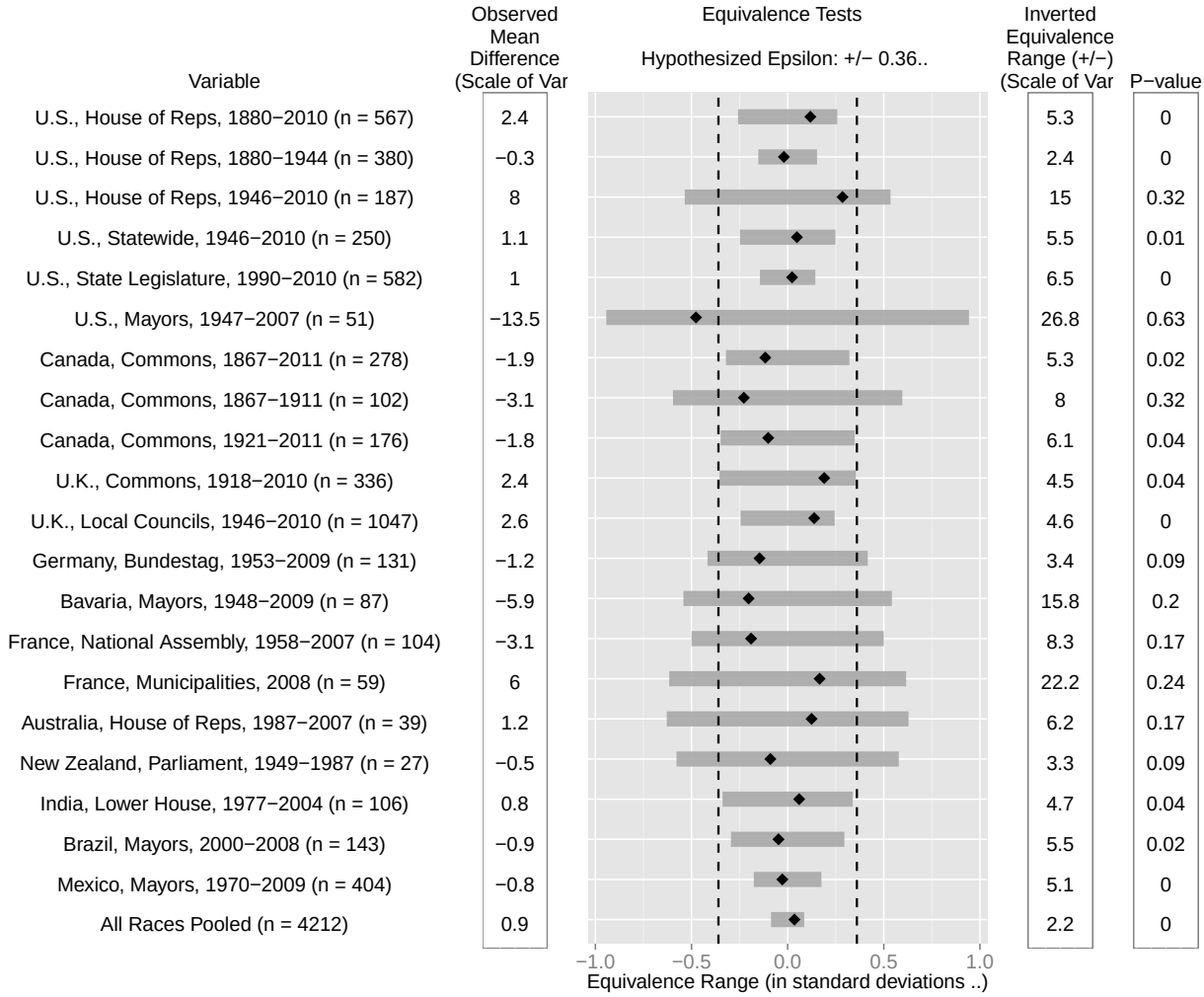

Figure 5: The figure above presents the results of equivalence tests. The "Observed Mean Difference" is the mean of the treated group minus the mean of the control group. The vertical dashed lines represents the hypothesized equivalence range, defined as the standardized effect size on the outcome of 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 "Inverted Equivalence Range" is the inverted range, transformed to the scale of the variable. The "P-value" corresponds to the p-value of the test of the null equivalence range of one standardized effect size.

Figure 5 recreates column 1 of table 2 in the original paper which presented balance statistics using the difference-in-means test and a bandwidth of +/- 0.5 percentage points around the cut-point. We use the continuous outcome, the lagged forcing variable, rather than the dichotomized lagged win, but results are substantively the same.[10] The authors do not analyze an outcome variable, so we cannot define the range as one standardized mean difference, therefore we use the strict range defined in Wellek (2010) of 0.36σ. Results are generally consistent with the original author's findings, particularly the finding about the imbalance in the post-war United States House. There is a lot of variance in the sample sizes of each of the subgroups, and the original authors do not address this in their original findings, except for excluding the four groups with sample sizes less than 60 observations. The equivalence test shows that these groups do not achieve statistical equivalence. However, there are a few groups that are not statistically equivalent that the original authors find to be equivalent, including: Canada–Commons (1867–1911), Germany–Bundestag (1953–2009), Bavaria–Mayors (1948–2009). All of these groups are relatively small, with between 87 and 131 observations, and the equivalence test allows us to avoid conflating lack of power with equivalence.

7 Conclusion

Researchers need to prove negligible effects 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 omitted variable bias. 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, direct or indirect, that this is not the case. Experimentalists are asked to defend against the notion of a "bad draw" that could lead to bias in their realized estimate. 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 convince readers over the validity of the design. By using equivalence tests, researchers begin with

[10] To be consistent with the original analysis, we use robust standard errors in the equivalence t-test.

the assumption that the design is flawed, or that an effect is not negligible, and this conclusion is only reversed 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? what is the negligible effect set? 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.

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 Section A.6, covariates or placebo outcomes in many studies that currently pass tests of design when the null is sameness will not pass when the null is difference. Not passing an equivalence test does not by itself, of course, invalidate a design or indicate hopelessly biased estimates. Many other elements of a design should go into a 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 may be 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

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

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.

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.
URL: *http://www.jstor.org/stable/2685688*

Berger, Roger and Jason Hsu. 1996. "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science* 11(4):283–302.
URL: *http://www.jstor.org/stable/2246021*

Bowers, Jake. 2011. Making Effects Manifest in Randomized Experiments. In *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.

Brady, Henry and John McNulty. 2011. "Turning Out to Vote: The Costs of Finding and Getting to the Polling Place." *American Political Science Review* 105(1):1–20.

Caughey, Devin and Jasjeet S Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008." *Political Analysis* 19(4):385–408.

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

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, T. 2010*a*. "Do Quotas Promote Ethnic Solidarity? Field and Natural Experimental Evidence from India.".

Dunning, Thad. 2010*b*. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" *Rethinking Social . . .* .

Eggers, Andrew C, Anthony Fowler, Jens Hainmueller, Andrew B Hall and James M Snyder. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59(1):259–274.

Esarey, Justin and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(01):95–111.

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, J 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* .

Hansen, B 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, D E, K Imai, G King and E A Stuart. 2006. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15(3):199–236.

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.

Keele, L. 2010. “How Much is Minnesota Like Wisconsin? States as Counterfactuals.” *Unpublished paper The Ohio State . . .* .

Krueger, AB. 2000. “JSTOR: The American Economic Review, Vol. 90, No. 5 (Dec., 2000), pp. 1397-1420.” *American Economic Review* .

Lee, DS. 2008. “Randomized experiments from non-random selection in U.S. House elections.” *Journal of Econometrics* .

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

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

Morgan, Kari Lock and Donald B Rubin. 2012. “Rerandomization to improve covariate balance in experiments.” *arXiv.org* (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. 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.

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

Sekhon, J.S. 2007. “Alternative balance metrics for bias reduction in matching methods for causal inference.” *Working Paper* .

Student. 1938. “Comparison Between Balanced and Random Arrangements of Field Plots.” *Biometrika* 29(3/4):363–378.

Wellek, S. 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.
URL: *http://dx.doi.org/10.1002/bimj.4710380608*

Wellek, S. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*.

Westlake, WJ. 1976. “Symmetrical confidence intervals for bioequivalence trials.” *Biometrics* .

A Additional Statistical Tests for Equivalence

A.1 Two-One-Sided Test and Intersection Union Tests

The Two-One-Sided Test (TOST), a type of intersection union test, is the most commonly used equivalence test for studying bioequivalence. It is the test recommended by the FDA, for instance. Intersection union tests 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 $T_L \leq -t_{\alpha,m+n-2}$.

A.2 Exact Fisher Binomial Test for Equivalence

In the case of a binary treatment with a binary response, an equivalence can take the form of an Exact Fisher type test. In this case, the data can be described by the percentage of units taking on value 1 in either the treatment or control condition. 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 Section 4. 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.

A.3 Mann-Whitney Test for Equivalence

Sometimes a test for equivalence of the means of the distribution is not sufficient to prove equivalence between the two groups. Instead, we may wish to test for equivalence of the distributions of two groups. A non-parametric equivalence test for comparing two continuous distributions, F and G, is based on the Mann-Whitney statistic. 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. 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[11]. The properties of the Mann-Whitney test for equivalence are studied further in Wellek (1996).

A.4 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). Then, an exact p-value corresponding to the null $\mu_T - \mu_C = \epsilon_U$ is calculated.

[11] The variance of W_+, regardless of the underlying distributions F and G is always defined as $\text{Var}[W_+] = \frac{1}{mn}\big(\pi_+ - (m+n-1)\pi_+^2 + (m-1)\Pi_{X_T X_T X_C} + (n-1)\Pi_{X_T X_C X_C}\big)$ 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 non-parametric equivalence t-test is set up similar to the npTOST, using the same hypotheses as for the equivalence t-test. To test that $\frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U$, the permutation distribution of the the standardized difference is computed using the strict null hypothesis that $\frac{\mu_T - \mu_C}{\sigma} = \epsilon_U$. The permutation distribution is also constructed for the strict null of $\frac{\mu_T - \mu_C}{\sigma} = \epsilon_L$, and the appropriate one-tailed permutation p-value is calculated for each distribution. The test rejects the null of non-equivalence if both p-values are less than α. The non-parametric Mann-Whitney test is constructed analogously. However, the test statistic there is the W_+, as defined in Table 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).

A.5 Power of Equivalence Tests vs 90% Confidence Interval Tests

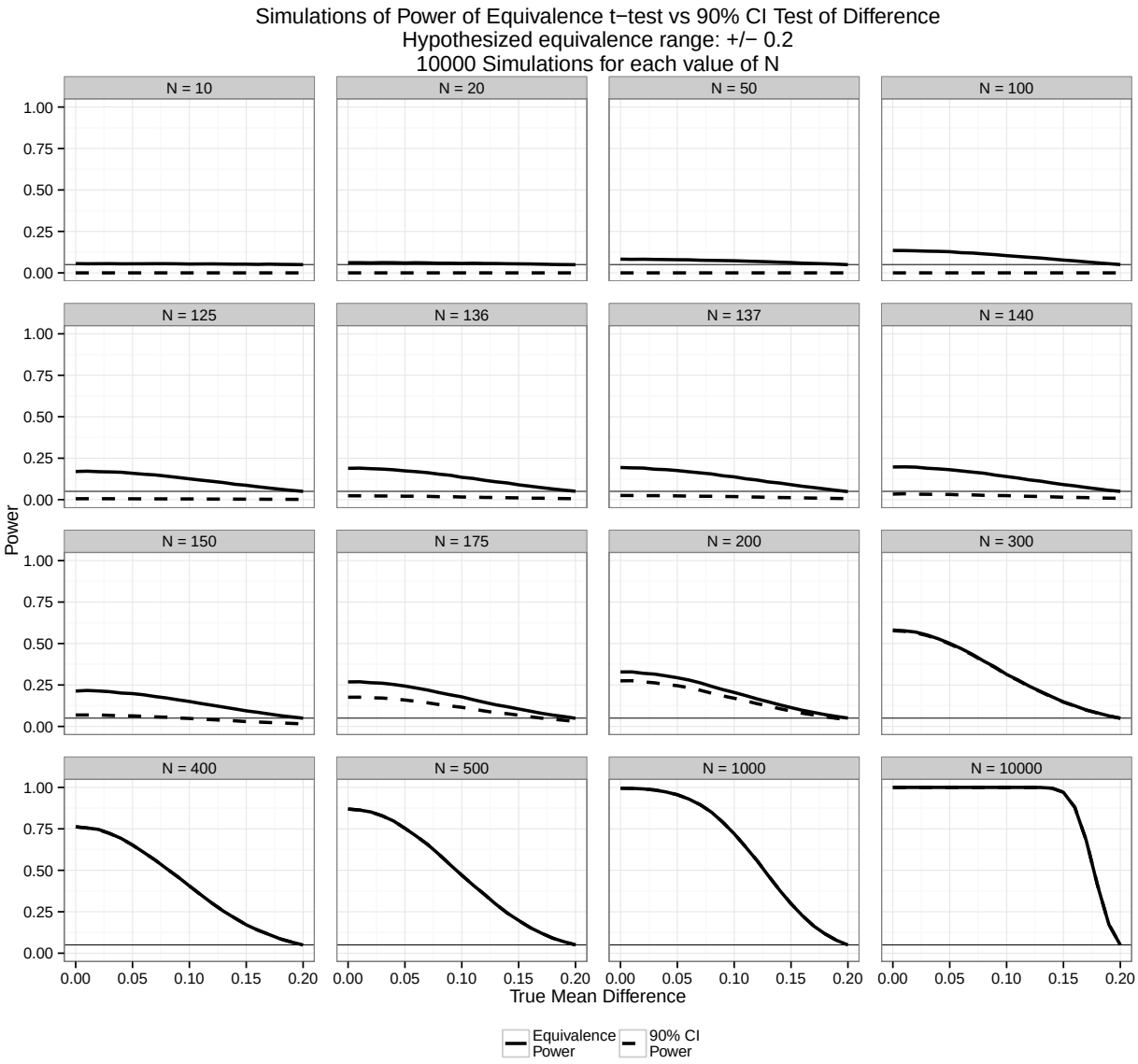

Figure 6: Power of the Equivalence t-test vs the 90% Confidence Interval Test. The horizontal black line is located at 0.05.

A.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[12] from Dunning's (2010*b*) 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 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 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[13] 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 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 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 the inverted ϵ. This inverted ϵ is useful here because it can give the

[12] 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*b*) list that reported this information were used.

[13] 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
Krueger (2000)	Minimum Wage Increase	Employment - November 1992	-3.30	384	0.40	0.16	0.23
Dunning (2010*a*)	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 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
Krueger (2000)	Employment - November 1992	0.32	17.27
Dunning (2010*a*)	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 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.

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.