Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity

Luke Keele Penn State University
Corrine McConnaughy Ohio State University
Ismail White Ohio State University

Experiments have become an increasingly common tool for political science researchers over the last decade, particularly laboratory experiments performed on small convenience samples. We argue that the standard normal theory statistical paradigm used in political science fails to meet the needs of these experimenters and outline an alternative approach to statistical inference based on randomization of the treatment. The randomization inference approach not only provides direct estimation of the experimenter's quantity of interest—the certainty of the causal inference about the observed units—but also helps to deal with other challenges of small samples. We offer an introduction to the logic of randomization inference, a brief overview of its technical details, and guidance for political science experimenters about making analytic choices within the randomization inference framework. Finally, we reanalyze data from two political science experiments using randomization tests to illustrate the inferential differences that choosing a randomization inference approach can make.

Experimentation has been a growth industry in political science over the last decade or so. This growth reflects an interest in making valid causal inferences about political phenomena. Randomly assigning subjects either to receive or not to receive a treatment that represents a causal factor of interest enables researchers to employ the assumption that they have equivalent groups, with the exception of the groups' reception of the treatment. Thus, the cause of any observed differences across the groups in the outcome of interest is validly ascribed to the treatment—but not without some uncertainty. Because random assignment delivers equivalence in expectation, there remains a chance in any one experiment that the groups were different on the outcome of interest before treatment. Thus, the experimenter is left with the question of how certain it is that any observed difference is due to the treatment, and not to the chance of random assignment, itself. We engage in this article the question of how to characterize this uncertainty through statistical tests and do so while paying attention to the frequency with which political science experiments rely upon small, nonrandom samples, typically termed convenience samples. We thus offer an overview of the logic of randomization inference and its basic implementation, as well as guidance about when particular statistical tools might be most helpful and appropriate. Randomization inference uses random assignment as the statistical basis for inference to offer estimates of uncertainty about whether observed outcomes under one random treatment assignment might have been observed under an alternative random allocation of the treatment. It is a departure from commonly used classical statistical tests in the frequentist framework, which typically rely on an assumption of random sampling and involve a

Luke Keele is Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 (ljk20@psu.edu). Corrine McConnaughy is Assistant Professor, Department of Political Science, 2018 Derby Hall, Ohio State University, Columbus, OH 43210 (mcconnaughy.3@polisci.osu.edu). Ismail White is Assistant Professor, Department of Political Science, 2008 Derby Hall, Ohio State University, Columbus, OH 43210 (white.697@osu.edu).

Authors are in alphabetical order. For helpful comments and discussion, we thank Jake Bowers, Sanford Gordon, Kosuke Imai, Cindy Kam, Walter Mebane, David Nickerson, Jasjeet Sekhon, Nicholas Valentino, Lynn Vavrek, and seminar participants at the University of Michigan, Columbia University, and Princeton University. We also thank James Fowler and Cindy Kam for generously sharing their data. For research assistance, we thank Kevin Duska. A previous version of this article was presented at the 2008 annual meeting of the Midwest Political Science Association, the 2008 annual meeting of the American Political Science Association, Boston, MA, and the 2010 Visions in Methodology Conference, Iowa City, IA. Replication materials are available at http://dvn.iq.harvard.edu/dvn/dv/ljk. An online supplement contains further details about specific randomization tests as well as information about the data used.

American Journal of Political Science, Vol. 56, No. 2, April 2012, Pp. 484–499

DOI: 10.1111/j.1540-5907.2011.00576.x

p-value as an estimate of the uncertainty about whether a sample is representative of a population. Using classical inference also relies on the assumption that the test statistic follows some parametric distribution such as the *t* or Normal. In political science applications where the central inferential task is connecting a sample to a specified population, the classical approach makes sense. In the experimental context, it very often does not.[1] Classic inferential tools do not provide direct estimates of the experimenter's quantity of interest—uncertainty about internal validity—and their assumptions of random sampling and parametric distributions can seem unwarranted or even troubling given the samples used in some experiments. The randomization inference approach enables the experimenter to proceed without such assumptions. It also offers tests that can address a number of common analytic challenges faced by political science experimenters. Yet, invoking the randomization inference framework does not necessarily entail the use of new statistical techniques. Indeed, a number of randomization tests can be approximated by standard (normal theory) tests when sample (cell) sizes are large enough. A firm understanding of the randomization inference approach, however, clearly delineates which standard tests are appropriate and when.

Randomization inference is not new. In fact, the concept dates back to the basic theory of randomization formulated by Fisher (1935). These tests have been largely absent, however, from experimental methodology in political science. We conducted a JSTOR search for all articles that contained the word "experiment" published between 1995 and 2007 in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*. This search returned 258 articles. We did not find a single example of randomization inference. We expect, however, that practitioners offered an informed choice between standard approaches and the randomization inference approach will find the latter useful.[2]

To convey the intuition of the randomization inference framework, we walk the reader through a simple example before moving on to the technical details of the approach. We cover not only the technicalities of simple hypothesis testing, but also the estimation of treatment effects and confidence intervals. We then illustrate the use and implications of randomization inference with two published experiments highlighting the differences between the inferences that would be made from these data using randomization inference rather than standard tests.

Principles of Randomization Inference: An Example

Before providing a formal description of randomization tests, we start with an example to illustrate the logic and implementation of the randomization inference approach. We introduce with this example both the general intuition of randomization inference and the use of a rank-based test, which is possible but not necessary within the randomization framework. Consider an experiment where we select seven students to play a dictator game with a computer program which asks them to allocate 100 dollars between themselves and a charitable donation.[3] Three of the students are randomly chosen to receive the treatment, a prime expected to make them more altruistic. If the treatment is effective, we would expect that the students who receive it would give away more of their money. If there is no effect of the treatment, we would expect no such difference across the two sets of students. In other words, we are interested in testing the following null hypothesis:

H_0: The prime has no effect on amount given,

against our alternative:

H_a: The prime has a positive effect on amount given.

Relying on the standard toolkit of the political scientist, we might translate these expectations into a *t*-test comparing the observed mean difference in dollars given to a null of no mean difference. We would then calculate a p-value using the critical value from a *t*-distribution. Note this implies the assumption that the test statistic follows a *t*-distribution. As an alternative, we can derive the statistical test from the randomization of treatment

[1] There are ways to recast experiments and randomization as sampling mechanisms to provide a basis for classical statistical inference. The most coherent story is to treat the subjects as a population and assume that random assignment of treatment forms a sampling mechanism for that population. Another possibility is to assume that the convenience sample is a random sample from some unknown or hypothetical population (Lehmann and Romano 2005).

[2] This is not to say that randomization tests are totally unknown to political scientists. Hansen and Bowers (2008) investigate and advocate the utility of randomization tests using a political science application. Ho and Imai (2006) provide another example with a political application. These all appear in statistics journals, which underscores the rarity of these tests in political science. We know of one article published in the main disciplinary journals that uses randomization tests to analyze an experiment: Fowler and Kam (2007), an article that fell just outside the time bounds of our initial JSTOR search.

[3] This example is adapted from one in Sprent and Smeeton (2007).

TABLE 1 Possible Combinations of Ranks in Treatment Group

1,2,3	1,2,4	1,2,5	1,2,6	1,2,7
1,3,4	1,3,5	1,3,6	1,3,7	1,4,5
1,4,6,	1,4,7	1,5,6	1,5,7	1,6,7
2,3,4	2,3,5	2,3,6	2,3,7	2,4,5
2,4,6	2,4,7	2,5,6	2,5,7	2,6,7
3,4,5	3,4,6	3,4,7	3,5,6	3,5,7
3,6,7	4,5,6	4,5,7	4,6,7	5,6,7

assignment and avoid the parametric assumption entailed in the use of the *t*-distribution.

Our expectation in this example, again, is that treated students should give away more than nontreated students. The best evidence that the prime increased altruistic behavior would be if the three students who received the treatment rank 1, 2, and 3 in terms of the amount given away. So we might ask: what is the probability that the three students in the treatment group would happen to rank 1, 2, and 3 in terms of the amount given away if the null hypothesis were true, and, in fact, there was no effect of the prime? To develop a probability statement about this hypothesis that hinges on the composition of the treatment and control groups—the feature that was assigned by a random process—we turn to basic combinatorics. Knowing that the number of ways of selecting r objects from a set of n is $n!/[r!(n-r)!]$ tells us that there are 35 ways to select three students from a set of seven. Table 1 contains all the possible combinations of ranks that we could observe for our set of treated subjects, and random allocation of treatment determines the probability for each of these cell entries. Seeing in the table that our best evidence outcome is 1 of 35 possible treatment-group compositions tells us that there is a $1/35 \approx 0.0286$ chance that we would observe the best evidence case if the null were true. That is, the chance that random assignment alone will produce this exact outcome is 1/35. This p-value indicates that the best evidence outcome enables us to be fairly confident that the prime has an effect on student behavior in our divide-the-dollar game.[4]

We can make this approach to hypothesis testing more general by introducing a summary statistic that enables us to translate ranks into a single measurement of the outcome among the treatment subjects. One possible statistic for this purpose is the sum of the ranks for the treated subjects. This statistic will be lower if the treated subjects are generally higher in their giving than the control subjects, and higher if they are not.[5] Using this statistic, we can answer the question of what the chance is of observing an outcome of a specific degree (or smaller/larger) among the treatment subjects. For example, suppose the outcome we observed among the treated subjects was the ranks 1, 2, and 7. Our summary statistic would be 10 ($1 + 2 + 7 = 10$). This seems close to the "best evidence" outcome we just considered, where the sum of the ranks would be $1 + 2 + 3 = 6$, but is it close enough to be convincing evidence of a treatment effect? To answer this question, we work out the probability of observing a rank sum statistic of the same amount or less by returning to our enumeration of all 35 possible combinations of the three ranks and calculating the rank sum for each combination. Four of the 35 possible rank combinations produce a sum of 10, three more produce a sum of 9, two result in a sum of 8, one set sums to 7, and another to 6. Thus, under random assignment, the chance of observing an outcome like the one we did or smaller would be $p = 11/35 \approx 0.314$. Put another way, if the prime has no effect, we could expect to see a value for the summed ranks as low as or lower than the one we observed 31 out of every 100 times we randomly assigned the treatment to these particular subjects. Using the traditional hypothesis testing threshold of .05, the p-value we calculated would not allow us to reject the null hypothesis; the observed outcome did not provide sufficient evidence that the prime had an effect on our subjects' behavior. Thus, the p-value in a randomization test is calculated as the probability of observing a test statistic as large as or larger than the observed test statistic.

Note that the interpretation of the p-values in this example references the information we believe the political science experimenter is after: the probability that the result observed among his or her specific set of experimental subjects can be explained away by the chance constitution of the treatment groups under one allocation of treatment. Randomization inference explicitly focuses on local inference, endeavoring to compare the responses that the subjects studied exhibited under one random allocation of treatment to the unobserved responses the same individuals would have displayed under an alternate random allocation of treatment. We do not need to assert these subjects are any sort of sample from any known population. An inference of this type is valid for any type of sample and may be particularly sensible for convenience samples. We are able to make statistical inferences because

[4] If we had a two-sided alternative hypothesis, we would double the resultant p-value. For details on other methods for calculating two-sided p-values, see Lehmann (1975).

[5] Note the inverse relationship is due to the direction of our ranking—that those who gave away more were given lower values for their ranks (highest giver being ranked 1, etc.). Having directionality—and attending to it—is what is important here.

we can use the randomization of the treatment to create a meaningful probability distribution for the null hypothesis and can calculate a p-value for any combination of values that we might observe in the experiment without ever using a parametric distribution.

Principles of Randomization Inference: Formal Matters

In this section, we offer a formal outline of randomization inference. We begin with the basic mechanics of testing that the treatment is completely without effect, including explanations of *exact* tests and *sharp null* hypotheses. We discuss the process of choosing an appropriate test statistic, including considerations about using rank-based tests. Of course, calculating the p-value for rejecting the null of no treatment effect is only one statistical quantity of interest, so we also discuss methods for generating confidence intervals and point estimates. And we discuss the possibility of approximating randomization inference tests with standard parametric tests, including caveats about turning to these statistical tools.

The basic mechanics of a statistical test built on the randomization inference framework include the same elements as any statistical test: data, a null hypothesis, a test statistic, and a distribution of the test statistic under the null hypothesis. The derivation of the last element, the null distribution, however, is unique to randomization tests. To provide a formal derivation of the null distribution, we first define **T** as a random vector that assigns subjects to either the treatment or control group. To illustrate, if the first, third, and fourth subjects out of seven subjects are selected to receive the treatment, **T** would have the following form: (1,0,1,1,0,0,0).

Next, we denote the quantity **y** as a vector of the outcomes for the subjects, and we define the teststatistic as

$$S = f(\mathbf{y}, \mathbf{T}). \tag{1}$$

That is, the test statistic, S, is the result of a function, f, that operates on both the observed outcomes and the treatment assignment. In the example in the first section, S took the form of the summed ranks for the treatment group. The function f and the test statistic S can take several possible forms; we discuss some of those possibilities in greater detail in following sections.

We further denote all possible responses under the treatment as Ω with elements s_i. This Ω contains all outcomes under all possible realizations of **T**. In our earlier example, summing each of the entries in Table 1 would form Ω for the experiment. We use Ω to calculate the probability of observing a value of S of a particular size s_i or larger (or smaller, depending on the direction of our expectations) if the null hypothesis were true. This p-value is the sum of the randomization probabilities that lead to those referenced values of S, relative to all possible values of S:

$$p = Pr(S \geq s_i | H_0) = \frac{\sum I(S \geq s_i)}{|\Omega|}, \tag{2}$$

where $I(\cdot)$ is an indicator function, and $|\cdot|$ denotes the cardinality of a set. Generating p-values in this way implies that randomization tests are *exact*. A statistical test is exact when the p-value is an always exact calculation of α, where α is the probability of making a Type I error—rejecting the null hypothesis when it is true. α is, of course, an important piece of information in the decisions we make about statistical evidence. Recall that using the conventional 0.05 significance level for hypothesis testing, for example, reflects an intention to set α to 0.05. Exactness is thus an attractive property of randomization tests. In comparison, the actual probability of a Type I error will differ from α under a traditional parametric test if the test statistic fails to meet any of the test assumptions. For example, if T, the test statistic in the t-test, diverges from a t-distribution, perhaps due to outliers in the data, the probability of a Type I error using that test will differ from α.[6]

As promising as the randomization inference approach to hypothesis testing with experimental data seems, there is one caveat, albeit a practical one. For large samples, the number of possible outcomes can be quite large and, even with modern computing, the time required to compute an exact p-value can be lengthy. For such situations, we can simulate the distribution of null outcomes and derive approximate exact p-values. With large samples, these simulated tests have been shown to very closely approximate the exact tests. In fact, in examples where the entire null distribution can be elaborated, approximate tests based on simulation produce accurate inferences even when the number of simulations is considerably smaller than the total number of permutations.[7]

[6] The same logic holds for confidence intervals. We specify the coverage for a confidence interval by selecting $1 - \alpha$. In parametric tests, a failure to meet the distributional assumption can cause this to be untrue.

[7] There are several exceptions. Fisher's exact test for a binary outcome and a binary treatment has a closed-form solution, as the permutation distribution follows a hypergeometric distribution. Other tests with closed-form solutions are the Mantel and Haenszel test and the d statistic of Hansen and Bowers (2009).

The mechanics of randomization inference mechanism bear repeating. Probability enters our calculations only through randomization of the treatment and does not rely on any parametric probability distribution or sampling mechanism. That is, the inference is entirely design-based—assuming only random allocation of units to experimental conditions—and invokes no modeling assumptions external to the study design. Here, the null distribution is completely known. Moreover, the p-values directly translate into the experimenters' original quantity of interest: the probability that what they observe as evidence of a treatment effect can be explained away by the chance process that assigned their observed subjects into treatment and control groups.

The Sharp Null

The randomization inference paradigm we have outlined was developed by Fisher (1935), who advocated making inferences about treatment effects through a test of the sharp null hypothesis. Under the sharp null hypothesis, we test whether the treatment effect is zero *for all units*. The potential outcomes framework helps clarify the distinct nature of the sharp null hypothesis. First, we note that for each unit i observed in the experiment, the indicator $T_i = t$, $t \in \{0, 1\}$, records the randomly assigned treatment status. Each unit then has a potential outcome for each treatment condition, which we denote as $Y_i(t)$. In potential outcomes notation, if the sharp null hypothesis holds, then $Y_i(1) = Y_i(0)$. That is, treatment status is irrelevant to each individual subject's outcome; each would exhibit the exact same potential outcome under the treatment as he or she would when not treated. This approach is different from the approach developed by Neyman (1923), which tests hypotheses about the average treatment effect (ATE). Under Neyman's formulation, the null hypothesis is that the ATE is zero, which may occur even if the treatment effect is not zero for all subjects—some subjects exhibiting a positive effect and others a negative effect could produce an average of "no difference" across the treated and control groups. While the Neyman approach is certainly sensible and widely embraced, there are compelling reasons to adopt Fisher's approach to hypothesis testing. Most notably, the p-value from a test of the sharp null is accorded special status due to its assumption-free nature. As Rosenbaum (2002b, 27) notes:

> In the theory of experimental design, a special place is given to the test of the hypothesis that the treatment is entirely without effect. The reason is that, in a randomized experiment, this test may be performed virtually without assumptions of any kind, that is, relying on the random assignment of treatments.

The experimenter need only assert about his or her data that he or she did, in fact, perform a randomized experiment. It is unnecessary even to make what Rubin (1986) called the "stable unit-treatment value assumption" or SUTVA, which is commonly needed to proceed with other tests. SUTVA asserts that the potential outcomes under treatment or control for each unit are fixed and notably do not depend on the treatment status of other units. The test of the sharp null is valid even if SUTVA is violated, which can easily occur through interference—when the outcome for one unit is affected by the treatment status of other units (Rosenbaum 2007).[8] And while Fisher's approach enables the analyst to proceed without assumptions, tests of average effects require additional assumptions even for simple hypothesis testing.[9]

This does not mean that tests based on p-values are of greater practical importance than point estimates and confidence intervals. Some, particularly Bayesians, have questioned the value of hypothesis tests, sharp or otherwise, and p-values (Gill 1999) and have argued that confidence intervals and point estimates along with simulations and assessments of practical significance are more valuable than a test of the null hypothesis (Gelman and Hill 2006). Certainly it is true that many researchers will wish to know more about what their results suggest about the substantive significance of their treatments. As we discuss further below, estimates of this sort require additional assumptions. When the researcher has any doubt about those assumptions, testing the sharp null seems all the more important and useful. Recent work does, however, allow one to retain the randomization inference framework but relax the assumption that the treatment effect is the same for all units in the experiment (Rosenbaum 2003, 2007). We give a very brief introduction to these techniques later. In what follows, then, when we refer to a hypothesis test, we mean Fisher's version of these tests. As we discuss later, in large samples the distinction between the Fisher and Neyman approaches vanishes.

[8] Note that Rubin (1986) argued that Fisher's framework automatically implied a special form of SUTVA held, rather than that SUTVA could be violated.

[9] In brief, the Neyman approach involves first using the mean difference as an unbiased estimator of the causal estimand. The analyst next finds an unbiased or upwardly biased estimator for the variance of the average causal effect. An appeal to the central limit theorem allows the analyst to form a confidence interval for the average causal effect based on these estimated quantities.

Test Statistics

A range of valid test statistics is available to the experimenter within the randomization framework. That is, one has choices about f, the function that operates on treatment assignment and observed outcomes to produce S, a test statistic. An ideal choice will produce a test statistic that distinguishes between the null and alternative hypothesis. In statistical terms, we want a powerful test statistic, one that will have an extreme value if the null hypothesis is false. It is difficult to offer a more specific general recommendation for choosing a test statistic because there are many alternative hypotheses, and no one test statistic is powerful in reference to all of them. We thus review here three test statistics that might be useful—each of which compares how the distribution of the potential outcomes under the alternative hypothesis differs from the distribution under the null—and provide some commentary on why each might be chosen. Importantly, we discuss the limitation of statistics as the basis for guidance in choosing a test statistic and the role that theory and substantive judgment must play.

Tests based on differences in means are a familiar tool for comparisons across treated and control groups and are available within the randomization inference framework. We might choose the average difference across the treatment and control groups:

$$\Delta_A = \frac{1}{n_T}\sum Y_i(1) - \frac{1}{n_C}\sum Y_i(0) = \bar{y}_1 - \bar{y}_0 \quad (3)$$

where n_C is the number of units in the control group and n_T is the number of units in the treatment group.[10] While this statistic is attractive for its ease of interpretation, it is sensitive to outliers and will have low power if there are larger differences in the tails of the two distributions. Other similar test statistics are possible. For example, one might use the test statistic from the t-test.

A transformation before comparing average levels across treatment and control can produce other useful statistics. For example, taking natural logarithms of the outcomes and estimating as follows,

$$\Delta_M = \frac{1}{n_T}\sum \ln(Y_i(1)) - \frac{1}{n_C}\sum \ln(Y_i(0)) \quad (4)$$

produces a statistic that we can interpret as a constant multiplicative treatment effect (Imbens and Rubin 2008). Such a transformation might be especially attractive when a constant additive effect might (incorrectly) suggest potential outcomes for some units that are outside the substantively meaningful bounds of the outcome measure, as in a count of something like income or the number of votes cast.

Test statistics based on ranks of the outcome data are also common. As we illustrated with our example in the first section, the outcomes are transformed to integers that rank where they fall in the distribution of observed values, and the test statistic is the sum of these ranks in the treatment group. Examples of these tests include the Wilcoxon rank-sum test for comparison of a control group and a single treated group and the Kruskal-Wallis test for comparison of multiple treated groups.[11] Rank tests are attractive for their ability to detect differences even in the presence of long-tails and/or outliers. Rank-based randomization tests are also convenient because they are widely available in statistical software.

Questions of how we might interpret differences between a test using the test statistic based on means and the test statistic based on ranks and how we choose the most appropriate test statistic involve subtle issues worth exploring. Suppose a researcher uses both the mean-based and rank-based test statistic. She finds that the p-value for the mean-based test statistic is well above the standard 0.05 threshold, while the rank-based statistic's p-value is below 0.05. What does this imply, other than there are some responses to treatment that are "unusual"? Answering that question involves thinking through what "unusual" means, which likely depends on the substance and theory in question. First, suppose the experiment was based on a formal theory that predicted that all agents should behave the same, perhaps that all subjects should be equally generous in a divide-the-dollar game. In the observed outcomes, however, a few subjects were unusually generous. The question for the experimenter is whether she believes a few instances of extreme generosity are due to aberrant subject-level responses or should be treated as particularly important information about the theory's validity. Perhaps she has some evidence to suggest that the extreme values were due to some subjects not fully understanding the experiment's instructions. In this case, she might choose a rank-based test, such that the extremity of those values does not exert much influence on the test of a treatment effect. If, however, she has nothing to suggest that those observations were not "true" values, she might think that the mean-based test statistic better captures what the data have to say about the theoretical prediction. Similarly, think about the challenges faced by political psychology experimenters working on questions of priming and accessibility. Many use the response time—the time in milliseconds it takes for a subject to

[10] Note this test statistic does not imply a test of an average treatment effect. It is still a test of the sharp null hypothesis, where it is assumed the effect of the treatment is constant. Why this is true will be clear when we discuss point estimation.

[11] Details on both of these tests are available in the appendix.

perform a categorization task—as a measure of accessibility. When the treatment is a prime, it should make the relevant construct more accessible and reduce response times to the corresponding items in the task. The difficulty is that any brief distraction can wildly inflate the response time since it is measured in milliseconds. One common strategy in the literature is to simply drop such observations (see Nelson, Clawson, and Oxley 1997). A rank-based statistic provides the experimenter with a principled method for dealing with the outliers that invariably occur when response times are the outcome of interest, enabling them to keep the observations but reduce their influence on the test of the treatment's effect. In general, we suggest that it can be useful to compare a statistic based on means to one based on ranks. Differences suggest a need to make judgments about the nature of unusual observations.

Estimation

Thus far, we have only discussed the calculation of an exact p-value for a test of the sharp null hypothesis. Of course, many experimenters wish to say something about the size of the treatment effect given the evidence in their data. That is, experimenters are likely to be interested in a point estimate for the treatment effect and a confidence interval for that point estimate. Both interval and point estimation are both possible within the randomization inference framework. Both, however, require assumptions that were unnecessary for calculating the exact p-value in the test of no effect.

Two additional assumptions are needed for interval and point estimation. First, we must assume SUTVA holds. Second, we must make an assumption about the nature of response to the treatment. Rosenbaum (2002b) refers to this second assumption as a model of effects. Rosenbaum (2002b, chap. 5) outlines a number of different models of effects, but the most widely used model of effects is that of a constant-additive effect, which is the effect implied by a linear model. That is, we assume that the treatment raises the response of each unit by a constant amount: $Y_i(1) = Y_i(0) + \tau$, where τ is the treatment effect.[12]

A $100(1 - \alpha)\%$ confidence interval for the additive treatment effect, τ, is formed by "inverting" a series of $100(1 - \alpha)\%$ level tests. In short, we conduct a series of hypothesis tests with nonzero values of τ_0 using the same method we outlined in the second section and keep the values not rejected at the value of α we have chosen, typically 0.05. For each test, we exploit the fact that $Y_i(1) = Y_i(0) + \tau$ under the null to adjust each observed outcome by τ_0 and then conduct a test of the sharp null hypothesis. The calculated p-values enable us to specify the values at the endpoints of the $100(1 - \alpha)\%$ confidence interval. This method of interval estimation is perhaps best understood with a simple example. Say we conduct an experiment testing whether social pressure affects intention to vote. We measure intention to vote with a 7-point scale, with higher values indicating a higher propensity to vote. We observe 12 subjects, six of whom are randomly exposed to a social pressure cue. We observe the following outcomes:

Treatment Group = (7, 7, 5, 4, 6, 5)
Control Group = (1, 4, 5, 1, 5, 5).

We use the absolute difference in means across the treated and control units as our test statistic. Thus, our test statistic has a value of approximately 2.17. First, we test the usual sharp null hypothesis that the treatment effect is zero. To calculate an exact p-value, again, we compare the number of times a test statistic value as large as or larger than the one we observe occurs relative to the universe of test statistic values computed from all possible permutations of the outcomes we observed. In this case, there are 924 possible ways to form a treatment group of six subjects from a pool of 12. Comparing the observed test statistic value of 2.17 to the values from the 924 permutations, we find that the exact p-value is 0.002, and thus we reject the sharp null hypothesis that the treatment effect is zero. Next, to construct our confidence interval, we begin by assuming a model of constant additive effects. We use that model to test a series of sharp null hypotheses, testing that $\tau_0 = 1$ and then $\tau_0 = 2$ and so on. In this example, we specify values of τ_0 in increments of 1. We then compute adjusted responses according to our model of effects, using the equation $Y_i - \tau_0 T_i$. Note that, here, adjustment is to the observed outcome, which is defined as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. In sum, we subtract the value of τ_0 for each null hypothesis from the treatment-group observations and then calculate the test statistic and corresponding exact p-value in the same way we did with the observed data. We form the confidence interval from the hypothesis tests for the values of τ_0 where we do not reject at a chosen level of α. The results from this iterative process are in Table 2.

We form a 95% confidence interval for our treatment effect, meaning we find the values of τ_0 where we reject the null at $\alpha = .05$. Given the discrete nature of exact p-values, we may be unable to form a precise $100(1 - \alpha)\%$

[12] The model of the effects may or may not correspond to the test statistic. For example, the constant-additive model of effects may be used with the average difference of means test statistic as well as the rank-sum test statistic.

TABLE 2 Inverting the Null Hypothesis to Form Confidence Intervals

95% Confidence Interval	Sharp Null Hypothesis	Exact p-value
Lower Bound	0	0.002
	1	**0.048**
	2	0.234
	3	0.701
	4	0.727
	5	0.251
Upper Bound	**6**	**0.052**
	7	0.004

confidence interval because we do not observe p-values right at the α value of 0.05. In our example, we observe that a null hypothesis of 1 has a p-value of 0.048 and the null of 6 has a p-value of 0.052. Thus, the smallest effect we would reject at the approximate 0.05 α level is a mean difference of 1, and the largest value we would fail to reject at the approximate 0.05 level is a mean difference of 6, making our confidence interval [1,6].[13] Randomization inference confidence intervals such as these maintain correct coverage regardless of the sample size often by expanding the width of the interval. For example, if the sample size is small, there may not be enough evidence to guarantee coverage at $\alpha = 0.05$. If this is the case, the confidence intervals will widen to whatever level of coverage the data will support. That is, when we form a confidence interval, it may be the case that coverage will only hold for $\alpha = .07$. The interval is exact in that coverage of the true parameter can only be guaranteed when $\alpha = .07$.

Hodges and Lehmann (1963) developed a randomization method of point estimation that invokes the same underlying structure as the calculation of confidence intervals.[14] We again focus on τ as the treatment effect that differentiates potential outcomes under treatment and control for each subject. The Hodges-Lehmann point estimator for τ is the value of $\hat{\tau}$ such that the outcomes adjusted via the chosen model of effects are exactly without treatment effect. As such, point estimation, like interval estimation, depends directly on the model of effects chosen by the analyst. For example, with the model of constant-additive effects, the Hodges-Lehmann point estimator for τ is the value of $\hat{\tau}$ such that the adjusted responses, $Y_i - \hat{\tau} T_i$, are exactly without treatment effect. Thus, if equation (3) is the test statistic, the estimating equation for the point estimate is $\bar{y}^T - \bar{y}^C - \hat{\tau} = 0$. We solve for the value of $\hat{\tau}$ which makes this equation true. For this test statistic, that will be $\hat{\tau} = \bar{y}^T - \bar{y}^C$. Note that this is an estimate of the *individual*-level treatment effect, not the average effect. Thus, while difference in means across treated and control groups is necessarily an unbiased estimator for the average treatment effect, it is only an unbiased estimator for the individual-level treatment effect when the model of constant-additive treatment effects is true.

While point estimation for other test statistics is possible, interpretation may be more complicated. For rank-based tests, in particular, the test statistic has a less direct correspondence with the point estimate. We note that the Hodges-Lehmann point estimator for ranks is closely related to test statistics that are based on a difference in medians and refer the interested reader to Hodges and Lehmann (1963) for more details. We also note that a point estimate of the treatment effect is not sensible for all experimental designs, particularly the more complex. For example, in two-way factorial designs, there is not a single parameter that summarizes all treatment effects.

Nonconstant Effects

Underlying the randomization inference approach is the notion of constant effects—that each subject has the same response to being treated. Political scientists testing formal theories of behavior or theories of psychological processes might be most comfortable with the constant effects assumption, since the theoretical mechanism is often thought to be identical across (like) subjects. As long as the design and analysis proceed by comparing subjects theorized to have the same response, the approach is valid. If theory, however, is not so precise about the expected heterogeneity or if the design failed to account for heterogeneity via a mechanism such as blocking, we may be left with the more general expectation that for some units the treatment may have produced a positive effect, while for other units the effect may be zero or negative. While we might choose to turn to the Neyman ATE estimate because it allows for such heterogeneity, we note that the ATE estimate does not reveal anything about the extent of heterogeneity. In contrast, Rosenbaum (2001,

[13] Of course, in our example, we could have used finer-grained differences than integers between the tested values of τ_0. Whether or not noninteger values are sensible will depend upon what the outcome measure is. Note that Rosenbaum (2001) calls the series of sharp nulls that are tested attributable effects. He demonstrates how this framework can be extended to any experimental design. He also proves that these attributable effects are a random variable that maps the different outcomes that might have been observed in the treatment group under the sharp null hypothesis.

[14] The Hodges-Lehmann method of point estimation was first developed in the context of rank-based test statistics but generalizes to any test statistic.

2003) demonstrates how to use randomization inference to make inferential statements about the extent of positive (or negative) effects produced by the treatment.[15] This means that not only can we allow for heterogeneity within the randomization inference paradigm, but we can also gain some insight into the extent of the heterogeneity.

In brief, the Rosenbaum approach to nonconstant effects for rank-based statistics is as follows. First, let W be the test statistic for the test based on the sum of ranks. Next, we calculate c_α, the value at which we would reject the null for a $1 - \alpha$ confidence level. We can find this value with distributional tables or with statistical software. With these two quantities, we can be $(1 - \alpha \times 100)\%$ confident that at least $W - c_\alpha + 1$ of the treated outcomes were positive (or negative) due to treatment. This value is informative but must be adjusted for positive treated-control differences that might occur due to chance. For this adjustment, we calculate the expected number of positive differences under the null hypothesis. Let $E(W_0)$ be the expected number of positive treated-control differences under the null hypothesis of no treatment effect. For the rank sum statistic, $E(W_0)$ is $m(N - m)/2$, where N is the total number of units and m is the number of treated units. As such, $(W - E(W_0))/E(W_0)$ is the expected percentage of positive differences greater than would be produced by chance. Thus, we can be $1 - \alpha \times 100\%$ confident that $W - c_\alpha + 1/E(W_0) \times 100\%$ of the positive differences were caused by treatment and not by chance fluctuations.

As an example, we return to the data on social pressure and vote intention used in the previous section. Using a rank-based test statistic under the constant-additive effects model, we find the social cue treatment increased expressed vote intention by 1.5 points on a 7-point scale. If we relax the assumption of constant effects, what inferences are possible? The rank sum test statistic, W, in this example is 35.5. In an experiment of this size, with six treated and six control subjects, if there was no treatment effect, we would expect $6(12 - 6)/2 = 18$ positive differences to occur by chance. Therefore, in all possible permuted comparisons, a treated unit had a higher expressed vote intention $(35.5 - 18)/18 = 97\%$ of the time. In some of these comparisons, however, we would expect a treated unit to have a higher outcome due to chance. With these data, we reject the sharp null hypothesis at the 0.05 level for any value of W greater than 28, so $c_{0.05}$ is 28. At least $35.5 - 28 + 1 = 8.5$ of the treated-control differences were positive because of the treatment. Therefore, we can be 95% confident that at least $8.5/18 = 47\%$ of this excess was caused by the treatment and not by accidental positive differences. In other words, we can be 95% confident that in all possible permuted comparisons, a treated unit had a higher expressed vote intention attributable to treatment 47% of the time. The estimated quantities here are rather different from the typical point estimate and confidence interval; we do not estimate the *amount* of change caused by treatment, but instead the extent of positive or negative effects. The insight into individual-level response to treatment may be particularly useful when theory has not yet specified an expectation beyond a general directional hypothesis, as it may provide empirical fodder for refinement of theoretical expectations about effect heterogeneity.

Asymptotic Approximations to Randomization Tests

While the basis for inference in the randomization inference framework is quite different from the normal theory underpinnings of the standard statistical toolkit in political science, it is nonetheless true that some standard parametric tests can be used as valid asymptotic approximations to specific randomization tests. Indeed, Fisher (1935, chap. 21) hypothesized that the t-test could be derived from the permutations of randomized treatments and that a t-test based on permutations and the t-distribution should be similar. Hoeffding (1952) later proved that asymptotically the two tests are equivalent, justifying Fisher's belief that the usual t-test can be viewed as an approximation to the distribution-free exact version of the test. The results in Hoeffding (1952) also imply that parametric tests based on the F-distribution in one- and two-way ANOVA models provide an approximation to randomization-based tests for multiple treatments. The convergence of exact and parametric distributions in large samples also implies that the distinction between tests for average treatment effects and tests of sharp null hypotheses becomes less important as the sample size grows.

Whether or not it is advisable to rely on asymptotic approximations to randomization inference tests is a question worth some consideration. The accuracy of the approximation for any given application is, of course, difficult to know unless a direct comparison is made between the exact test and its asymptotic approximation. Political science experiments often involve small numbers of subjects, however, and in these contexts asymptotic approximations provide little comfort. Note that the asymptotic approximation is based not on the overall sample size, but the group size for each treatment (i.e.,

[15] He does this specifically for the sum rank test, the sign rank test, and Fisher's exact test.

the cell size; Lehmann 1975). And while the impetus for asymptotic approximations stemmed from the practical constraints of exact calculations, modern computing enables relatively quick exact calculations for even larger experiments.

We see one other more nuanced reason to approach asymptotic approximations with caution. While there are valid asymptotic approximations for many randomization tests, simple alterations of the analysis can render these approximations incorrect. For example, a bivariate regression model with treatment status as the independent variable provides inferences about treatment effects identical to those from the *t*-distribution, which is an approximation of the randomization distribution. One might suspect that a multivariate regression model with added covariates therefore provides a similar approximation. Freedman (2008a, 2008b) proves, however, that this is not the case. He demonstrates that for the estimation of treatment effects, the multiple regression estimator is biased. The bias goes to zero as the sample size increases and is typically trivial. The bias arises from the fact that the linear model assumes treatment effects are constant across units (Freedman 2008a, 2008b). The real concern, however, is that the multiple regression model may either overstate or understate the estimates' precision by surprisingly large amounts. Why should this be the case? Recall that the usual Gauss-Markov assumptions for the linear regression model hold that the error terms are independent and identically distributed (IID). The difficulty is that this assumption is directly contradicted by the potential outcomes model of an experiment used in randomization inference: the errors will vary with treatment by definition, making the error variance nonconstant. This nonconstant error variance in a regression model is usually referred to as heteroskedasticity. Given that heteroskedasticity is the barrier to using multivariate regression to approximate the randomization distribution, one might assume that the solution is some form of robust standard errors or perhaps a multiplicative model of heteroskedasticity. Neither of these options, however, is the solution. Lin (2010) demonstrates that for the multivariate regression model to approximate the randomization distribution, one must use a fully saturated model. That is, one must include the full set of treatment-covariate interactions. So while one can use the multivariate regression model to approximate the randomization distribution, it requires alterations to the model that are not obvious.

A similar mistake can be made if logistic regression is used to approximate the randomization distribution. For experiments with a binary treatment and outcome, randomization inference is possible with either Fisher's exact test or the sign test based on binomial sampling. One might assume that a logistic regression model is an appropriate approximation to the exact test. Freedman (2008c) demonstrates that this is not the case. Again, the difficulty arises from the fact that only the treatment is stochastic, while the logistic regression model assumes the outcome is a random binomial process.

Given the possible complications, we strongly emphasize that care must be taken with asymptotic approximations. While there are valid approximations available, subtle alterations in the mode of testing can result in tests that are not approximations of the randomization distribution. We have outlined two examples where unless the analyst is careful what may seem like a harmless asymptotic approximation actually does not approximate the inference justified by randomization. If the asymptotic approximation is used, however, randomization inference brings greater clarity to the parametric test. Even though the distribution used is parametric, the justification for that distribution stems from random assignment. The inference remains local and is focused on uncertainty about the treatment, and not on sampling from a larger population.

Parametric Assumptions and Power

Next, we use a comparison of nonparametric confidence intervals with standard parametric confidence intervals to illustrate an important point about the way in which the application of randomization inference techniques can enable the data from an experiment to speak more clearly about the evidence produced by the experiment. To form parametric confidence intervals, one must assume that the data follow a particular distribution. When the sample size is small, this essentially adds information to the data, which will result in confidence intervals that may be overly narrow and fail to maintain correct coverage (Imbens and Rosenbaum 2005). The parametric assumption is analogous to using an informative Bayesian prior with the data, and informative priors are most likely to influence our answer when sample sizes are small. In comparison, the confidence intervals from Fisher-style randomization tests maintain correct coverage regardless of how many observations are used. The exact $100(1-\alpha)\%$ confidence set for the treatment effect estimate will always maintain its stated coverage of $100(1-\alpha)\%$. That includes, importantly, that when the data do not contain enough information, the interval may achieve this coverage by becoming infinite in length (Imbens and Rosenbaum 2005). That is, we may find that there are no values of the sharp null hypothesis where we are able to reject at a chosen confidence level. This, we think, is an attractive feature of these nonparametric confidence intervals: they reveal whether additional data are required to increase the power of the test in order to say something substantively meaningful.

Consider an example from an experiment we conducted. In the experiment, subjects in the treatment group viewed a story from the local newspaper about a mugging. The control group was exposed to a story from the same local newspaper about changes to the iPhone. Subjects were then asked to rate whites and African Americans on a set of stereotype items. The difference in subjects' attribution of stereotype traits to whites and African Americans measures whether the treatment caused subjects to rate African Americans lower relative to whites when primed on the topic of crime. Our experiment had 19 subjects in the control condition and 22 subjects in the treatment condition. If we proceed with a standard parametric analysis based on the t-test, the difference in mean ratings is -1.3. That is, subjects in the treatment condition rated African Americans 1.3 points lower than whites on the stereotype scale. The normal theory confidence interval for this estimate is $[-2.47, -0.19]$. The point estimate for the treatment effect from the rank sum test is seemingly similar in substantive terms, at -1.5, with an exact p-value of 0.004. The confidence interval for the nonparametric estimate, however, is $[-\infty, 0]$. Using the randomization test in this case reveals that there is not enough information in the data to say anything more about the treatment effect other than it is negative. To be more specific about the treatment effect would require us to invoke a parametric assumption or repeat the experiment with more subjects.

This somewhat minor point raises controversial issues in statistical inference. A Bayesian might argue that in small samples informative priors are necessary since the data have little to tell us. Better here to rely on substantive knowledge and impose a prior. In fact, the parametric t-test can be thought of as a Bayesian estimate with an uninformative prior. The difficulty is that, as we have demonstrated in this example, this obscures important information about statistical power. In our example, we see that by assuming the data are distributed normally adds information to the data resulting in confidence intervals that are overly narrow unless the parametric assumption is correct. With the nonparametric test, we observe that the treatment effect is clearly negative, and we can reject the null that the sharp null is zero, but we would conclude that to learn more about the treatment effect requires a larger sample size. While one might be willing to defend a flat prior, the parametric assumption, here, is troubling. We argue it is better to know that the experiment as conducted does not have enough power to rule out a variety of null hypotheses and to know for future iterations of the experiment that a larger number of subjects is needed for more precise inferences.

Statistical Tests with Randomization Inference for Political Science

Given that a randomization test needs to be built from the probability model used for treatment assignment, there are a wide variety of tests to fit various experimental designs. Clearly we cannot review them all here. The interested reader and practitioner will likely need to seek out various additional sources. Though texts on nonparametric statistics should cover many of the possible tests, they often fail to show the link between each test and the randomization mechanism in the experimental design. One notable exception is Lehmann (1975), who derives common rank-based nonparametric tests from random assignment of treatment. Higgins (2003) is a recent text that provides the randomization-based justification for tests in a wide variety of experimental designs, including both mean- and rank-based test statistics. In the statistics literature, the rank-based test statistics that we discussed earlier are quite popular, and thus many lucid discussions of randomization inference there (see Rosenbaum 2002b, chap. 2) focus almost entirely on rank-based tests. The appendix to this article contains a basic introduction to rank-based tests, including those for two-way factorial designs. The latter may be of particular interest as we have found that coverage of randomization tests for two-way factorial designs is quite rare, while such designs are relatively common in the social sciences.

Examples from Political Science Experiments

Having laid out our case for randomization tests, we now turn to applying these tests to two datasets from political science experiments. We use a dataset from Fowler and Kam (2007), whose published results represent a rare example of the use of randomization tests in political science, and part of a dataset produced by White (2003) from which results have not been previously published. Both experiments were performed on convenience samples—one relying entirely on student subjects and the other recruiting both students and nonstudent adults. We compare the results from standard statistical tests to those from randomization tests. In addition to offering a more direct estimate of the type of uncertainty that concerns the experimentalist, we find that randomization tests can produce p-values that would lead to different substantive conclusions.

Partisan Generosity

Fowler and Kam (2007) performed a series of experiments to test a set of hypotheses about individuals' propensity to give to others. The authors brought student subjects into a laboratory environment and asked them to play a dictator game, wherein each subject was given a set of 10 lottery tickets and asked to divide the tickets between themselves and an anonymous recipient. By manipulating the identity of the anonymous recipient, Fowler and Kam intended to test for differences in giving that could be attributed to the effects of social identities. Thus, three experimental conditions were employed, the treatment being the identity of the recipient: no identifying information, registered Democrat, or registered Republican.

Among Fowler and Kam's expectations was the hypothesis that subjects would display an ingroup preference and thus give more when the recipient was revealed to be in the same party as themselves. One way to use their data to assess this hypothesis is to compare the amounts given to the recipient when: (1) the subjects' partisan identities matched the recipient, (2) the subjects' identities diverged from the recipient, and (3) the subjects had no information about the identity of the recipient. The authors looked at these comparisons separately among those who strongly identified with their party label and those who weakly did so; we do the same. In the replication, we perform three different tests. First, we use the standard one-way ANOVA. We next use a randomization test where the test statistic is an F-test statistic from the same ANOVA table but the p-value is calculated from all permutations of the data. Finally, we calculated the Kruskal-Wallis test based on ranks. This allows us to compare different test statistics under the randomization testing paradigm. In Table 3, we display, side-by-side, the p-values from the three tests across the three conditions for the two partisan groups. Among the weak partisans, the differences are minimal. For the strong partisans, the permuted test provides a p-value similar to the standard test. Among strong partisans, in fact, the decrease is enough to change whether or not the null hypothesis would barely be rejected at the conventional .05 level.

TABLE 3 Exact versus Asymptotic Approximation Comparisons

ANOVA and Kruskal-Wallis p-value Comparisons

	ANOVA	Permuted F-test	Kruskal-Wallis
Strong Partisans	0.055	0.043	0.007
Weak Partisans	0.515	0.521	0.5137

***t*-test and Wilcoxon Rank Sum p-value Comparisons**

	t-test	Permuted *t*-test	Rank Sum
Ingroup vs. Control	0.050	0.053	0.015
Ingroup vs. Outgroup	0.029	0.025	0.003

Note: There are 127 subjects per cell for the tests in rows 1, 3, and 4, and 125 subjects per cell for the second-row tests.

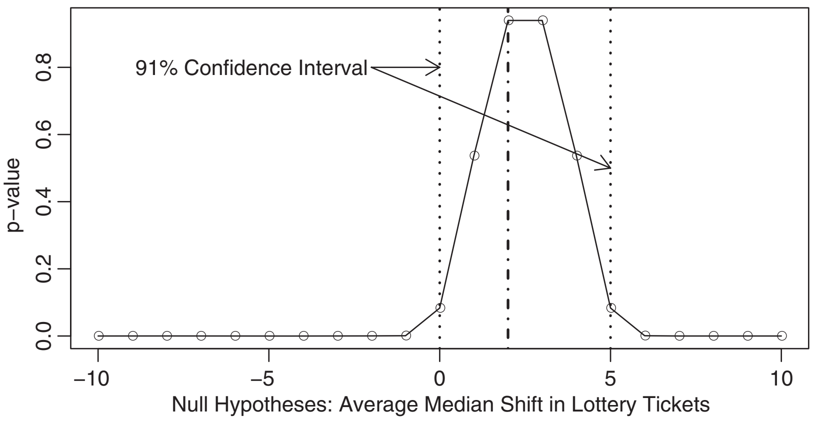
If the researchers had specific hypotheses about differences not across all three conditions, but rather across any two, we could employ either a permuted difference in means or a rank sum test as an alternative to the asymptotic approximation provided by the t-test on the difference in amounts given.[16] For example, the researchers might have specifically expected a difference in subjects' giving in the same-party recipient condition as compared to the "control" condition where no information about the recipient is given. Alternatively, they might have been interested specifically in the difference in giving when the recipient is identified with the ingroup as compared to when the recipient is identified with the outgroup. Here, we compare a standard t-test to a permuted t-test and the rank sum test for Fowler and Kam's data that would be used to test these two pair-wise comparison hypotheses among strong partisans. The results from the three tests are in Table 3. The asymptotic t-test does provide similar results to the permuted t-test. The test statistic based on ranks provides greater power once again.[17] So with a rank-based test statistic, the chance that random assignment would produce such a level of partisan generosity among strong partisans is well below 1%; therefore, this provides strong evidence against the null hypothesis of no treatment effect.

Closer examination of the data reveals that a few subjects gave away an unusually large number of lottery tickets. The rank-based statistic helps reveal this pattern in the data. How one should interpret these outliers is open to question. Is some subset of the subjects more likely to

[16] The t-test in this instance can be interpreted as either a test of whether the ATE is different from zero or an asymptotic approximation to the test of the sharp null under the randomization test.

[17] Note that if we proceeded through a series of pair-wise tests to test for differences across conditions, we would really want tests that account for the multiple comparisons we were making. Appropriate randomization tests exist (Hollander and Wolfe 1999), or one could simply use a Bonferroni correction.

FIGURE 1 **Attributable Effects Against Exact p-values for Dictator Game**



be responsive to treatment, or is some subset always going to be unusually generous in any dictator game scenario? These are not questions that can be answered here, but the rank-based test does help bring attention to the fact that some subjects gave away unusually large amounts.

For one-way factorial designs such as used here, there is not a summary statistic for the effect, just a p-value for the test, which is often reported with treatment condition means. Under the randomization inference framework, we can develop a more general point estimate with appropriate confidence intervals. We use the average median difference across the three partisan categories as a test statistic for the one-way design. That is, we calculate the median number of lottery tickets given away across the three treatments and take the average. This provides us with a measure of how behavior changed across the experimental conditions. The next step is to select a model of effects. We adopt the usual constant-additive model of effects, assuming that the treatment effect is constant and additive across each condition. We use the observed summary statistic for the data as the point estimate, which here is 2: the average median difference across treatment categories was two lottery tickets. To form a confidence interval, we specified integers from −10 to 10 as the values for *A*, the range of null hypothesis values. For each value in the range of *A*, we subtracted this value from the outcomes of the out-party condition, then calculated the approximate exact p-value based on the true null distribution. We next construct a confidence interval for this estimate. Given the discrete nature of exact p-values, we do not observe a value at the point necessary to construct a 95% confidence interval; we draw the 91% confidence interval instead. In Figure 1, we plot the exact p-value against the range of null hypotheses. Based on the randomization null distribution, this point estimate has a 91% confidence interval of [0, 5]. This example demonstrates how one can easily move beyond simply reporting a p-value for more complex designs and construct confidence intervals and point estimates for interesting features in the design.

Racial Cues

In our second example, we analyze data from White (2003). White designed an experiment to test the effects of two types of racial cues in political communication: a source cue and a racial frame. The treatment received by all subjects was a news magazine article laying out arguments for opposition to the war in Iraq; the article was manipulated across conditions to vary both the frame of the opposition argument and the source of the article. Two frames were employed in the experiment: an explicitly racial frame and an implicitly racial frame. Each of the frames was presented inside a news story appearing in either a black news magazine (*Black Enterprise*) or a mainstream news magazine (*Newsweek*). Thus, the experiment is a 2 x 2 factorial design with a total of four conditions, and subjects were randomly assigned to the conditions. Subjects were then asked to report their level of belief, on a 1–7 scale, in three arguments about the war: that the United States should wait for UN Security Council approval, that Iraq had chemical weapons, and that President Bush's handling of Iraq was approvable. The experiment was run separately on both white and black subjects, as the theory implied that blacks and whites would respond differently to the treatments.

Among the expectations was the hypothesis that blacks would be persuaded to be less supportive of the war and less receptive to arguments used to justify the war when they were exposed to any of the racial cues. Additionally, it was hypothesized that the effect of the frame might depend on the source in which the article appeared, notably that the two types of racial cues (source and frame) might have mutually reinforcing effects. Thus, the frame was postulated as the modifying factor. We concentrate on these hypotheses and only analyze the data from the black subjects. Given that this is a two-way design, we test for two "main" effects and for an interaction. We assessed these hypotheses using a standard two-way ANOVA model which provides an asymptotic approximation to the randomization distribution and the rank-based test statistic outlined in the appendix. For each of the permutation tests, we used 10,000 permutations to form the null distribution, though we found that using 1,000 permutations made little difference. That is, we took 10,000 random permutations of the data and calculated the test-statistic to form a null distribution. Table 4 contains a comparison of the p-values that resulted from both the standard two-way ANOVA and the rank-based method. We report the results for both methods across three tests for each outcome variable: a test of the effect of the racial cue for each of the two media source treatments and a test of whether those two effects differ.

In this experiment, we find that the randomization tests produce generally lower p-values for the test of the interaction, often changing whether or not the null hypotheses would be rejected. In the first example, our inference is maintained but the difference in p-values is 0.13. For the other two outcomes, we narrowly conclude that an interaction is present based on the asymptotic test. The randomization-based inference, however, provides stronger evidence that an interaction is present. While the randomization inference in this instance lowers the p-value in all three instances, such differences cannot be assumed to hold in other data sets.

TABLE 4 Comparison of Asymptotic and Permuted p-value for Two-Way ANOVA

	Parametric p-value	Approx. Exact p-value
Security Council Approval for War		
Racial Cue within Media Source Level 1	0.353	0.235
Racial Cue within Media Source Level 2	0.546	0.464
Cue × Source Interaction	0.292	0.162
Iraq Has Chemical Weapons		
Racial Cue within Media Source Level 1	0.054	0.037
Racial Cue within Media Source Level 2	0.589	0.373
Cue × Source Interaction	0.047	0.026
Approve George W. Bush		
Racial Cue within Media Source Level 1	0.051	0.013
Racial Cue within Media Source Level 2	0.585	0.361
Cue × Source Interaction	0.082	0.015

Note: Approximate exact p-values based on 10,000 permutations of the data. Cell sizes range from 23 to 28.

Conclusion

While experiments offer considerable leverage on questions of causal inference, the traditional statistical methods political science experimenters have had at their disposal have limited their potential. The standard tools, justified by the standard model of inference, are ill-suited to the experimenter's main statistical question: the estimation of the internal validity of their inferences. Adopting the randomization inference model shifts the question from characterizing uncertainty about whether a random sample is representative of a population, to that of uncertainty about how responses the subjects exhibited under one random allocation of treatment compare to the unobserved responses these same individuals would have displayed under an alternate random allocation of treatment. Thus, not only does this model of inference focus directly on the experimenter's main quantity of interest, but it also frees the experimenter from assumptions about the subject sample. Samples of convenience—even Fisher's (1935) single subject in his famous Lady Tasting Tea example—are justified as valid for providing statistical evidence about the causal question at hand. And the randomization framework even makes sense of the application of traditional statistical tools to experimental data where such tools are justified as asymptotic approximations to randomization inference tests; a standard t-test applied to experimental data gains conceptual coherence when the p-value can be interpreted as an estimate of the uncertainty introduced by random assignment. Randomization tests, we have also shown, offer the ability to avoid parametric assumptions, confidence intervals that are informative about testing power, and a capacity to change substantive inferential conclusions.

The randomization inference framework can also be extended in many ways that we did not review here. Rosenbaum (2002a) outlines a method for covariance adjustment that is fully integrated with randomization tests and is easy to implement. Such covariate adjustment can be helpful for increasing the power of a test. There are also randomization tests for block designs and within-subjects experiments. Hansen and Bowers (2009) adapt the randomization framework to an experimental design with clustering and noncompliance. Rosenbaum (2002b) uses randomization inference as a basis for sensitivity analysis in observational studies. As advances in computing continue, randomization inference tools are increasingly finding their way into common statistical software packages. Familiarity seems the only remaining impediment to the integration of the randomization inference approach into the methodological toolkit of the experimenter in political science.

References

Fisher, Ronald A. 1935. *The Design of Experiments.* London: Oliver and Boyd.

Fowler, James H., and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation." *Journal of Politics* 69(3): 813–27.

Freedman, David A. 2008a."On Regression Adjustments in Experimental Data." *Advances in Applied Mathematics* 40(2): 180–93.

Freedman, David A. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2(1): 179–96.

Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23(2): 237–49.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

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

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

Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to a Clustered Randomized Get-Out-the-Vote Campaign." *Journal of the American Statistical Association* 104(487): 873–85.

Higgins, James J. 2003. *Introduction to Modern Nonparametric Statistics.* Belmont, CA: Duxbury Press.

Ho, Daniel E., and Kosuke Imai. 2006. "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 Election." *Journal of the American Statistical Association* 101(475): 888–900.

Hodges, J. L., and E. L. Lehmann. 1963. "Estimates of Location Based on Ranks." *The Annals of Mathematical Statistics* 34(2): 598–611.

Hoeffding, W. 1952. "The Large Sample Power of Tests Based on Permutations of the Observations." *The Annals of Mathematical Statistics* 23(2): 169–92.

Hollander, Myles, and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods.* 2nd ed. New York: John Wiley and Sons.

Imbens, Guido W., and Paul Rosenbaum. 2005. "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education." *Journal of the Royal Statistical Society Series A* 168(1): 109–26.

Imbens, Guido W., and Donald B. Rubin. 2008. *Causal Inference in Statistics and the Medical and Social Sciences.* Cambridge: Cambridge University Press.

Lehmann, E. L. 1975. *Nonparametrics: Statistics Based on Ranks.* San Francisco: Holden-Day.

Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. New York: Springer.

Lin, Winston. 2010. "Agnostic Notes on Regression Adjustments to Experimental Data." Unpublished manuscript. University of California, Berkeley.

Nelson, Thomas E., Rosalee Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(3): 567–83.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(4): 465–72. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Rosenbaum, Paul R. 2001. "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot." *Biometrika* 88(1): 219–31.

Rosenbaum, Paul R. 2002a. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Statistical Science* 17(3): 286–387.

Rosenbaum, Paul R. 2002b. *Observational Studies.* 2nd ed. New York: Springer.

Rosenbaum, Paul R. 2003. "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test." *The American Statistician* 57(2): 132–38.

Rosenbaum, Paul R. 2007. "Interference between Units in Randomized Experiments." *Journal of the American Statistical Association* 102(477): 191–200.

Rubin, Donald B. 1986. "Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81(396): 961–62.

Sprent, Peter, and Nigel C. Smeeton. 2007. *Applied Nonparametric Statistical Methods.* 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

White, Ismail K. 2003. "Racial Perceptions of Support for the Iraq War." PhD dissertation. University of Michigan. Unpublished data.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Additional Information on Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity

Table A1: Asymptotic Relative Efficiency of Rank Sum Test and *t*-test

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.