Journal of Experimental Political Science, Page 1 of 24
doi:10.1017/XPS.2015.4

Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods

Diana C. Mutz*[*] and Robin Pemantle[†]

Abstract

In this essay, we more closely examine three aspects of the Reporting Guidelines for this journal, as described by Gerber and colleagues (2014, *Journal of Experimental Political Science* 1(1): 81–98) in the inaugural issue of the Journal of Experimental Political Science. These include manipulation checks and when the reporting of response rates is appropriate. The third, most critical, issue concerns the committee's recommendations for detecting errors in randomization. This is an area where there is evidence of widespread confusion about experimental methods throughout our major journals. Given that a goal of the Journal of Experimental Political Science is promoting best practices and a better understanding of experimental methods across the discipline, we recommend changes to the Standards that will allow the journal to play a leading role in correcting these misunderstandings.

Keywords: Randomization check, manipulation check, response rates, standards.

Establishing reporting guidelines for studies of any kind is an important step in the direction of improving research. The Standards Committee is to be commended for taking on this difficult and time-consuming task for experimental designs (see Gerber et al. 2014). We have no doubt that this is a positive development, something good for science as a whole, as well as for our particular discipline.

Nonetheless, in the spirit of making something that is already quite good even better, we would like to highlight some of the problems with the standards as currently written. This is not to detract from the committee's accomplishment, but is offered in the spirit of constructive suggestions for revision.

We discuss three aspects of the recommendations. The first concerns manipulation checks, a practice of great importance in experimental methodology that is not addressed by the Standards. The second is a more minor point concerning the

[*]Political Science and Communication, University of Pennsylvania, Philadelphia, PA USA; email: mutz@sas.upenn.edu
[†]Department of Mathematics, University of Pennsylvania, Philadelphia, PA USA

reporting of response rates, and how studies become classified as surveys under their suggested framework. The third issue concerns the recommendations for detecting errors in randomization. Our critique and suggestions for improvement on this front require substantial background, so we save this more involved issue for last.

MANIPULATION CHECKS

First, we recommend that manipulation checks be added to the JEPS checklist of desirable components of experiments. As is the case with many other items on the checklist, this requirement will not be relevant to every experiment, but it will be applicable to a large number of them and, most importantly, it will improve what can be learned from their results.

Manipulation checks establish that the treatment has had an effect on the theoretically relevant causal construct. In other words, manipulation checks are "a way of ensuring that an experiment actually has been conducted (i.e., that the IV has been effectively manipulated)" (Sansone et al. 2008). The majority of experiments in political science do not report manipulation checks, despite their prominence in other social science disciplines. Many social science disciplines have deemed them basic enough to be required in all but a limited number of cases. As a sociology volume on experimentation argues, "It is an essential part of an experiment to include manipulation checks. ... It is equally important to report the results of these checks" (Foschi 2007: 129). The *Handbook of Methods in Social Psychology* similarly advises, "Indeed, many editors of social psychology journals require these (manipulation checks) to be conducted as a matter of principle before accepting research for publication." While some kinds of experiments within political science do not include variable experimental treatments at all (e.g., game theoretic experiments), a majority do involve one or more randomly assigned treatments intended to induce variation in the causal variable.

In some cases, manipulation checks are unnecessary. For example, if a persuasion experiment manipulates the length of a message in order to evaluate whether long messages tend to be more persuasive than short ones, and one message has twice the number of words as another, then length has been manipulated, and it need not be the case that subjects recognize or remember the length of the argument to which they were exposed. Given that the independent variable construct and its operationalization are completely identical, a manipulation check would be unnecessary under these conditions.

The problem with assuming that the independent variable is identical to its operationalization is that this is frequently not the case. Nonetheless, in political science the experimental treatment is usually just assumed to have successfully altered the independent variable, and the results are interpreted as such. For example, when an experimental treatment suggesting "many people believe that ... trade can lead to lower prices for consumers," did not lead to more support for trade, the author concluded that it would not be worthwhile to convince people that trade lowers the costs of consumer goods in order to increase support for trade (Hiscox 2006: 756). Without knowing whether subjects actually believed this treatment, null effects cannot be distinguished from weak or ineffective manipulations. Likewise, when political scientists look for causal effects from policy threat, the salience of national identity, incivility, or innumerable other treatments, the causal construct is not identical to the operationalization of the treatment, so without a manipulation check, there is no reason to assume that the causal construct was successfully manipulated. Moreover, researchers have a tendency to underestimate the strength of treatment that is required to produce a change in the independent variable. As a result, an experiment may not actually test its hypothesis of interest. As the *Handbook* further notes, "For experiments to have the best chance of succeeding (i.e., for the IV to have an effect on the DV) the researcher needs to ensure that the manipulation of the IV is as strong as possible. Indeed, if there were a first rule of experimentation, this might be it."

Our recommendation in favor of systematically encouraging manipulation checks goes beyond consistency with the experimental traditions established in other disciplines. It also stems from our belief that (1) consistently effective treatments are a highly unrealistic assumption, and that (2) the absence of manipulation checks frequently impedes the accumulation of scientific knowledge from experimental studies in political science. We begin by delineating the conditions under which manipulation checks seem essential in political science experiments and then illustrate how their absence impedes scientific knowledge.

Independent variables in experimental studies may involve latent constructs that are manipulated only indirectly, as described above, or direct treatments in which the treatment and the operationalization are one and the same. Manipulation checks are essential to ensure construct validity when treatments are indirect manipulations of other constructs (Cozby 2009; Perdue and Summers 1986). Without verifying successful manipulation of the independent variable in the studies, even outcome effects consistent with the original hypothesis become difficult to interpret.

The reason that manipulation checks have not been emphasized in experimental political science may stem from the nature of early studies in this field, which tended to examine tangible rather than latent constructs as independent variables. Does a baby care booklet sent from one's congressional representative improve attitudes toward the elected official? (Cover and Brumberg 1982). Can information on voter registration improve turnout? (Gosnell 1942). So long as the operationalization of the treatment and the independent variable are one and the same, there was no need for a manipulation check.

But as experiments in political science have become far more ambitious, frequently using indirect strategies to manipulate latent independent variables, a smaller proportion of independent variables meet these criteria, as Newsted and colleagues (1997: 236) suggest.

Even in cases in which manipulations appear obvious, they may not be so. For example, some early research in information presentation used treatments that confounded the information form (e.g., table or graph) with other factors, such as color, making interpretations difficult. Manipulations checks can help to uncover such problems and should be as much a part of the development of a measurement strategy in an experiment as the dependent variables.

Unfortunately, even when the operationalization of a given independent variable is well-known, widely established and frequently used, there is still no guarantee that one has successfully manipulated the independent variable in any given study. For example, a widely used cognitive load manipulation appears to be responsible for highly inconsistent results in studies of how cognitive load affects charitable donations. In Kessler and Meier's (2014) careful replications of laboratory studies using the same subject pool, setting and experimental protocol, they discovered that the explanation for contradictory findings was that the manipulation varied in efficacy due to session order effects when multiple experiments were executed within a single hour-long session. The treatments produced the intended variance in the independent variable only when subjects were already somewhat fatigued. Thus, even with a well-established treatment, manipulation checks were essential to the correct interpretation of the experimental findings. In cases such as these, manipulation checks clearly contribute to researchers' ability to differentiate among competing interpretations.

Encouraging manipulation checks is particularly important in an era when survey experiments have enjoyed increased popularity. When survey experiments have participants respond from remote, unobservable locations, there is no way to know for certain if subjects were even exposed to the treatment, let alone whether they were affected in the way the investigator intended. Particularly with large heterogeneous population samples who may not pay as much attention to treatments administered online as they would in a lab, treatments can easily fail.

Simple exposure to a treatment obviously does not guarantee its effectiveness. The question is not whether "a treatment was successfully delivered," as indicated in the current guidelines, but instead whether the treatment manipulated the independent variable as intended. Subjects may doubt the veracity of information they are given, or they may not find the treatment as threatening, anxiety-inducing, or as counter-attitudinal (or whatever the treatment happens to be) as the investigator intended.

Within political science there already has been some recognition of the problem of inattentive respondents. For example, Berinsky et al. (2014) suggest that studies should include post-treatment questions about details of the stimulus in order to assess respondents' levels of awareness and attention to stimuli. However, in most studies, use of such "screeners" does not address the same question as a manipulation check. Rather than address whether a subject was *exposed* to the treatment, manipulation checks are designed to assess whether the treatment successfully induced variance in the independent variable. For this reason, even laboratory experiments in which exposure is assured require manipulation checks. While use of screeners seems reasonable, they are not a substitute for manipulation checks. Although there are some studies for which exposure to a stimulus is, in fact, the independent variable construct, most studies use treatments to induce a change in a latent construct, a change that may, or may not, have been accomplished among those who correctly answered a screener.

Surprisingly, even some studies using treatments that seem obvious rather than latent, such as the race of a person in a photograph, demonstrate that such treatments can easily fail. For example, respondents often disagree about the race of a person in a picture they are shown (Saperstein and Penner 2012). For this reason, whatever the particular manipulation was *intended* to convey should be verified before any meaningful conclusions can be drawn.

Finally, although we have made the positive case for including manipulation checks as part of the checklist in studies using latent independent variables, it is worth considering whether there is any potential harm that should be considered in encouraging them. We know of no one who has argued that they are harmful to the integrity of a study so long as they are asked after the dependent variable is assessed. Scholars across the social sciences concur that so long as manipulation checks are included after measurement of the dependent variable, there is no potential harm in including them. The only cost is in the respondent time spent on the manipulation check assessment.

But there is substantial danger if one chooses to omit a manipulation check. The risk is primarily of Type II error. The reader of a study without a manipulation check has no idea if a null finding is a result of an ineffective or insufficiently powerful treatment, or due to a theory that was simply incorrect. This is an extremely important distinction for purposes of advancing scientific knowledge. A recent article in *Science* demonstrates why this is problematic. Franco and colleagues (2014) use the TESS study database as a source of information on the file drawer problem, that is, the extent to which findings that do not achieve $p < 0.05$ are less likely to see the light of publication. In order to estimate how likely null findings were to be published, they analyzed all TESS studies tracking whether the anticipated effect on the dependent variable was found or not found, and classified as null findings those that did not produce effects on the dependent variable. However, in many, and perhaps even most of these latter cases, the independent variable was not verified as having been successfully manipulated. Thus, the lack of significant findings was not informative with respect to the theories under investigation or with respect to their anticipated effect sizes.

We would not recommend including manipulation checks as part of the JEPS checklist if they were informative on only rare occasions. But weak or ineffective manipulations are an exceedingly common problem. A study that finds no significant effects on the dependent variable and does not include a manipulation check for a latent variable is not at all informative; on the other hand, we can learn a great deal from an identical study with identical results that includes a manipulation check documenting a successful treatment. In the latter case, the theory is clearly in need of revision. Ignoring manipulation checks thus impedes the growth of scientific knowledge.

Particularly given that JEPS has publicly stated its intent to publish null results (a decision that we wholeheartedly endorse), it is essential to encourage manipulation checks whenever possible. Otherwise, a null result is not informative. More specifically, this practice inflates the possibility of Type II error and leads researchers to prematurely abandon what may be viable hypotheses. Wouldn't some other researcher recognize this potential problem and attempt a replication? There are already strong disincentives to replicate even significant experimental findings; the idea that researchers will pursue replications of null results (but this time including a manipulation check) seems improbable.

REQUIREMENT OF RESPONSE RATES FOR SURVEYS

A second issue concerns convenience samples. As currently written, the reporting standards confuse mode of data collection with the type of sample used (probability samples versus convenience samples). For purposes of applying these standards, the committee defines a survey as any study that uses survey data collection methods or that *could* conceivably have been executed as a survey, even if it was actually executed in a laboratory.[1]

For studies that qualify as surveys by virtue of their data collection method, the Reporting Standards state, "If there is a survey: Provide response rate and how it was calculated." The problem with applying this requirement to all studies that use survey data collection methods is that many survey experiments in political science use Mechanical Turk, Polimetrix or another opt-in data platform. There is nothing meaningful about a response rate when utilizing a convenience sample. Even if such a figure could be calculated, it would have no bearing on the quality of the study. When all subjects opt in to a study, this concept is meaningless. Given that the CONSORT diagram that the Standards Committee recommends (see Moher et al. 2010; Schulz et al. 2010) already codifies the practice of indicating people who

[1]An exception to this is that experiments that use video in a lab are classified as lab experiments even when they use survey methods to collect data (Gerber et al., 2014: 83). Given that videos are now also administered online as experimental treatments within surveys, this distinction is confusing.

drop out after a study has begun, attrition has already been covered in the other requirements.

If there are no claims to representativeness being made by the authors, we see no reason to require response rates. As many survey researchers have demonstrated, the representativeness of a sample is not a straightforward function of the response rate. If the authors are making claims about accurately representing some larger population, then it would make sense to ask for a demographic comparison of their sample to the population in question. But if the sample is being treated as a convenience sample for purposes of an experiment, and not as a representative one, then it is not informative to require response rates based on the means of data collection used either in a lab or in an opt-in survey.

RANDOMIZATION CHECKS/BALANCE TESTING

Finally, our chief concern with the Standards has to do with the recommendation on "Allocation Method" which addresses randomization procedure and the distribution of pre-treatment measures. As the third point under Section C states,

> If random assignment used, to help detect errors such as problems in the procedure used for random assignment or failure to properly account for blocking, provide a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pre-treatment measures (if collected) by experimental group.

This point contains a *directive*, "Provide a table . . . " as well as a *justification*, "to help detect errors . . . ." While we laud the goal of detecting errors, we find both the directive and its connection to the justification problematic.

In order to understand our criticism of this recommendation, we begin by clarifying some ambiguous uses of terms. We next discuss the role of randomization in experimental design. Finally, we discuss the proper roles, if any, for balance testing/randomization checks. Our discussion of the experimental method may seem circuitous, but it is necessary because the mistaken pairing of the directive and the justification produces potentially harmful consequences that are difficult to grasp without understanding the possible motives behind such a recommendation. Historically the adoption of randomization checks came first, while attempts at justification have been more of an afterthought. Only by understanding the common interpretation of this practice can one make sense of what is of value in this regard and what is not.

Terminology

Throughout the recommendations and the accompanying report, four terms are used to describe the "other variables" that are neither independent nor dependent measures, but are the subject of the directive described above: pre-treatment measures, covariates, demographics, and control variables. Whether or not the authors of the Standards Report had precise meanings in mind for each of these, both the report and our discussion of it will benefit from making these definitions explicit.

For purposes of our discussion, the term "pre-treatment measure" is the most self-evident, and we will assume it refers to any measure in the data set that is assessed before the treatment occurs and thus could not have been affected by the treatment. The term "covariate," on the other hand, is typically used for that subset of pre-treatment measures that are incorporated in the statistical model used to test experimental hypotheses. It is not clear from the report whether "covariate" is meant in this manner or is meant as a synonym for "pre-treatment measure." This confusion exists throughout political science (see, e.g., Arceneaux and Kolodny 2009: 760).

Importantly, this distinction becomes blurred if the model is not pre-specified; as in the Standards Committee's recommendations, we endorse the pre-specification of models and will use the term "covariate" only for measures that researchers had planned to include in the model in advance. As outlined in many sources (e.g., Franklin 1991), the purpose of a covariate is to predict variance in the dependent variable that is clearly not attributable to the treatment. For this reason a covariate must be a pretreatment variable, although not all pretreatment variables must be included as covariates. Covariates need to be selected in advance based on what one knows about the major predictors of the dependent variable in the experiment. Their purpose is to increase the efficiency of the analysis model by eliminating nuisance variance.

The term "demographic" is usually reserved for that subset of pre-treatment measures which describe characteristics of the sort found on census data: age, education, race, income, gender and the like. If they are to be used in the analysis of an experiment, they should be pre-treatment measures as well. However, there is no reason to include such measures as covariates unless one has reason to believe they are strong predictors of the dependent variable. For most outcomes in political science experiments, demographics are only weakly predictive at best. The purpose of a covariate is to increase the power of the experiment by reducing variance. As argued in Mutz and Pemantle (2011), the gain from adjusting by a weak predictor of the dependent variable does not overcome the cost in transparency and robustness. Adding an extremely weak predictor can even reduce power due to the loss of a degree of freedom.

There are other legitimate reasons to include a measure as a covariate. One is a suspected interaction. At times, demographics are included as hypothesized moderators of the treatment effect. For example, if one has reason to believe that the treatment will be greater among the poorly educated, then the moderator and its interaction with treatment are included in the analysis model.

The fourth term, "control variable," is borrowed from the observational data analysis paradigm. Control variables are variables that are included in statistical models in order to eliminate potentially spurious relationships between the independent and dependent variables that might otherwise be thought to be causal. Given that potentially spurious relationships between the independent and dependent variables are eliminated by randomization in experimental designs, we find the frequent use of this term out of place in experimental research.

Observational studies often use demographics as standard control variables. However, there is a tendency for political scientists to use the terms demographics and control variables interchangeably, regardless of whether demographics are a likely source of confounding or spurious association. The term "control variable" is rampant in published experiments in political science (for just a few examples, see Hutchings et al. 2004; Ladd 2010; Michelbach et al. 2003: 29; Valentino et al. 2002), even when there is no evidence of differential attrition or any need to "control for" other variables.

Notably, covariates serve a very different purpose from control variables and should be selected based on different criteria. Covariates are included in the statistical model for an experiment because of their anticipated relationship with the *dependent variable*, to increase model efficiency. Control variables are included in observational analyses because of their anticipated relationship with the *independent variables*, to prevent spurious relationships between the independent and dependent variables. Choosing covariates solely due to their correlation with the independent variable is problematic in experiments, as we discuss at greater length below.

Because these four terms are used more or less interchangeably in the Standards Report, the recommendation is unclear as to which and how many variables the committee would like to see broken down by condition in a table. Is it the ones included in the original (pre-specified) model? This makes little sense because those variables are already included in the model. At other times it appears they are concerned specifically with those variables not included in the model, such as demographics or other available pre-treatment measures which the experimenter had no reason to include. But if these variables are not central to the outcome of interest, it is unclear why balance on those variables is important.

As discussed further below, and as illustrated by many examples from political science journals, the recommendation in favor of displaying all pretreatment means and standard errors by experimental condition is more likely to promote confusion than clarity. Indeed the number of pretreatment variables used in balance tests has reached numbers as high as fifteen or more, and many more pre-treatment measures are often available including variables such as household internet access, party identification, age, education, race, gender, response option order, household size, household income, marital status, urbanicity, home ownership, employment status, if the respondent was head of the household, children in household and region, to cite one example (see, e.g., Malhotra and Popp 2012). As in this example, it is unclear why a particular set of pretreatment means is selected for comparison and how and why one should compare them.

The Role of Randomization

Fisher (1935) introduced randomization as a way to make treatment and control groups *stochastically* equal, meaning they are equal on average. If a researcher wants to make experimental groups as equal as possible on a *specific set* of dimensions, he or she would not use simple random assignment. Random assignment produces random deviations of relative size inversely proportional to the square root of the sample size, whereas a matched block design produces almost no deviation at all. In other words, randomization is not meant as a mindless way to implement blocking on known variables. The benefit of randomization is that it distributes all *unknown* quantities, as well as the known quantities, in a (stochastically) equal manner. This is where random assignment derives its title as the "gold standard" for causal inference: because unknown factors as well as known ones are stochastically equalized, possible confounding is ruled out by design.

The flip side of the bargain is that confounding is ruled out only stochastically. The precise inference that can be drawn is that the observed data must be caused by the treatment unless an event occurred which has probability less than p, where p is usually equal to 0.05. Historically, this is where the confusion begins to seep in: what is the nature of the "exceptional" event of probability less than 0.05, where a possible Type I error occurs?

The strong (but mistaken) intuition of many researchers is that one should be able to examine the data and see whether the randomization was unlucky. If it were possible to do this, analyses of experimental data would look very different: the rooting out of unlucky draws would be built into the analysis, in a manner specified in advance, and accompanied by precise confidence statements. There are, in fact, rejection sampling schemes that accomplish this. The downside of rejection sampling schemes is that one cannot treat the randomization that one chooses to keep as if it were the first and only one; instead, complex statistical adjustments must be made (Morgan and Rubin 2012). Notably, what is accomplished by doing so is a reduction in variance and a consequent increase in the statistical power of the experiment, not a reduction in the probability of a Type I error. The important point here is that balance is not necessary for valid inference in experiments. As Senn (2013: 1442) explains, "It is not necessary for groups to be balanced. In fact, the probability calculation applied to a clinical trial automatically *makes an allowance for the fact that groups will almost certainly be unbalanced*, and if one knew that they were balanced, then the calculation that is usually performed would not be correct" (emphasis in original).

With experimental data, judicious choice of covariates can greatly increase the power of the analysis, but this is a separate issue from confidence in the result. If one wants more confidence, he or she should use a smaller p-value. If a researcher uses a p-value of 0.05, then he or she will have to put up with a one in twenty chance that the result is mistaken. No amount of balance testing or digging into the data will eliminate or lower this uncertainty.

The Role of Covariates

Once a covariate is included in an analysis, the estimate of the treatment effect will be adjusted for this variable. Thus, there are as many potential estimates of treatment effects as there are sets of covariates that could be selected from among all available pre-treatment measures. Normatively, the model (including the precise set of covariates) is selected on the basis of theoretical considerations before the analysis is run, in which case there is one actual estimate of treatment effect. If the model is not pre-specified, the confidence statement surrounding the estimate is invalidated. For this reason, the second point under Section E in the Report—which asks researchers to be explicit about pre-specification of the model—is essential.

The most important observation to make about the many potential estimates of treatment effects is that the probability of an error is equally likely with any of the potential estimates. This is not to say that it does not matter which model is chosen. A better choice will reduce variance, increase efficiency, and lead to smaller confidence intervals. But it will not reduce the chance of Type I error. Likewise, the inclusion of other variables in the model will not increase robustness. Instead, the inclusion of covariates requires meeting additional assumptions that are not otherwise required. In particular, the relationship between the dependent variable and each of the covariates must be linear, the regression coefficient for each covariate should be the same within each treatment condition, and the treatments cannot affect the covariates, which is why they must be assessed pretreatment.

Including a large number of covariates in an analysis simply because they are demographics, or because they are available in the pretest is clearly inadvisable. With experimental data, "Rudimentary data analysis replaces scores of regressions, freeing the researcher from the scientific and moral hazards of data mining" (Green and Gerber 2002: 810–11). But the problem goes beyond the risks of data mining. Many experimental studies suggest that findings are more "robust" if they survive models that include additional covariates (e.g., Harbridge et al. 2014: 333; Sances 2012: 9). In reality, adding covariates simply because they are available reduces the robustness of the model (introducing an assumption of independent linear effects that do not interact with treatments), reduces transparency, and is unlikely to add any power.

What Purpose can Randomization Checks Serve?

It is crucial to any experiment that its random assignment be correctly accomplished. How might one detect errors in this regard? The first line of defense is a sufficiently detailed description of the randomization mechanism. Was it the RAND() function in Excel, a physical device such as a die, spinner, jar of balls, deck of cards, was it a printed random number table, or some other device? Was it pre-generated or generated as needed? If it was a blocked design, how was the blocking implemented? The randomization process is mentioned in Section C, and while we endorse this recommendation, it does not go far enough. The brief text in this recommendation and its sub-paragraph on hierarchical sampling do not cover enough bases to effectively prevent randomization errors. Because randomization is a *process* rather than an outcome, we think a more thorough description of the process is in order as recommended in Section 8a of the CONSORT (2010) checklist (Moher et al. 2010; Schulz et al. 2010).

The Report somewhat mischaracterizes our argument in saying we agree "that formal tests or their rough ocular equivalents may be useful to detect errors in the implementation of randomization." The important points are (1) that such tests are not *necessary* in order to detect randomization problems; and (2) that they are not, in and of themselves, sufficient evidence of a randomization problem. Due to the rarity of randomization failure, we believe that the impulse to check for balance is probably spurred by something other than skepticism over the functionality of the random assignment process.

The terms "balance test" and "randomization check" are typically used interchangeably to indicate a table of the distribution of pre-treatment measures across treatment groups, often along with a statistical statement concerning the likelihood of the extremity of the distribution having been produced by chance alone. Such a statistic can be reported for each variable or a joint test can be reported as a single omnibus statistic for the joint distribution of all test variables. If one tests for differences in each variable individually, a large number of such tests obviously increases the chance of finding significance. If one uses a joint test, it will take into account the number of variables, but it will still matter a great deal which particular variables are chosen for inclusion in the omnibus test. A standard example of including such a check reads, "Randomization check shows that demographics and political predispositions do not jointly predict treatment assignment ($X^2{}_{[24]} = 18.48$, $p = 0.779$)" (Arceneaux 2012: 275).

The report does not provide guidance as to what these balance variables should be, except to refer to them as "pretreatment" or "demographic" variables. In examples such as the one above, the exact variables are not mentioned. Given the many different outcome variables that are examined in political science experiments, it is unclear why demographics, in particular, are deemed particularly important when other variables may be more pertinent to the outcome under study.

To reiterate, the Standards Committee calls for tables of unspecified pre-treatment measures across treatment groups "to help detect errors" in randomization. Most importantly, it is not clear how such tables accomplish this task. The distribution of pre-treatment measures across conditions provides evidence of errors only if a faulty randomization device was used; in other words, we are testing the null hypothesis, which is the assumption that the randomization mechanism worked. If we reject the null hypothesis and conclude that the randomization device was faulty, then the study can no longer be considered an experiment nor be published as one. In practice, however, when imbalance is identified, this is seldom the course of action that is taken as we describe further below.

Other Uses of Balance Testing

The Standards Report (Gerber et al. 2014: 92) suggests that there are additional reasons to require balance tests.

> Detectable imbalances can be produced in several ways (other than chance). They include, but are not limited to, mistakes in the randomization coding, failure to account for blocking or other nuances in the experimental design, mismatch between the level of assignment and the level of statistical analysis (e.g., subjects randomized as clusters but analyzed as individual units), or sample attrition.

It is worth considering these additional rationales individually. Mistakes in coding variables do indeed occur with regularity, but why should they be more likely to occur with randomization variables than with the coding of other variables? Failure to account for blocking is already addressed elsewhere in the requirements where authors are required to describe whether and how their sample was blocked, as well as how they accomplished the random assignment process. Likewise, the description already must include mention of the unit of analysis that was randomized, so if the authors then analyze the data at a different unit of analysis, this will be evident.

The one scenario in which balance testing does make sense is when experimental studies take place over time, thus raising the possibility of differential sample attrition due to treatment. Sample attrition does not indicate a broken randomization mechanism, and it is already covered in the CONSORT diagram. Assuming a control condition is present, it sets an expectation for acceptable attrition levels. And if there is differential attrition across experimental conditions, then it makes perfect sense to conduct balance tests on pretreatment variables among post-test participants. If the post-attrition distribution of pre-treatment measures across treatment groups is distinguishable from the random pre-treatment distribution, then the experiment is clearly confounded.

For various reasons, we believe that error detection and differential attrition are not the primary reasons that balance testing is popular. Instead, as described above, we believe part of its appeal stems from researchers' strong intuition that they can unearth the unlucky draw. Further, the Report of the Standards Committee explicitly says that error detection is not the only reason for doing randomization checks. As stated on page 5 of the standards document,

> There may be other uses of summary statistics for covariates for each of the experimental groups. For instance, if there is imbalance, whether statistically significant or not, in a pretreatment variable that is thought by a reader to be highly predictive of the outcome, and this variable is not satisfactorily controlled for, the reader may want to use the baseline

sample statistics to informally adjust the reported treatment effect estimates to account for this difference.

There are several problems with this statement, which we address in order of appearance. First, the phrase "statistically significant or not" is meaningless in the context of adjustment. The only thing one can test statistically is the null hypothesis, which is the assumption that the randomization mechanism worked. If one is estimating treatment effects, then one is already assuming that the mechanism worked, so there is no question of significance. This point has been made many times in the literature in political science (Imai et al. 2008) as well as in other disciplines (Boers 2011; Senn 1994).

Further, it is important to think through the consequences of this requirement for reviewers as well as authors. This statement implies that it is acceptable and even appropriate for a reviewer to (either subjectively or based on a prescribed test) perceive an imbalance in the table, assert that it is a variable that might be related to the dependent variable, and therefore insist that something be done to address the situation.

Regardless of whether there is a statistical test, what happens next? Is it incumbent upon the author to somehow "prove" that randomization was done appropriately? How can this possibly be accomplished? And if we conclude from an author's inability to produce such evidence that randomization was not done correctly, then what? If balance tests/tables are truly being used to ascertain whether random assignment was done correctly, then the only logical response to concluding that it was not done correctly is to throw out the study altogether, or possibly analyze it as purely observational data.

Random assignment was either done correctly or it was not; there is no middle ground. This does not appear to be widely understood. As an experimental study in the *American Journal of Political Science* explained, "To test the robustness of our randomization scheme, we tested for any differences among the other observables on which we did not block.. .." (Butler and Broockman, 2011: 467). *Results* can certainly be more or less robust, but random assignment is either done correctly or it is not; there are no varying degrees of randomization.

Fixing a Broken Mechanism?

The assumption that one can "fix" a broken random assignment by the virtue of adding a covariate is commonplace throughout our top journals. For example, in a *Public Opinion Quarterly* article we are told that, "Partisanship is included in the analysis because of imbalances in the distribution of this variable across the conditions." (Hutchings et al. 2004: 521). Likewise, an article in the *American Journal of Political Science* assures us that "Every relevant variable is randomly distributed across conditions with the exception of education in Study 1. When we included education in our basic models, the results were substantially the same as those we report in the text" (Berinsky and Mendelberg 2005: 862).

There is no logic to including a "control" variable to correct for lack of true random assignment on just one or a few characteristics, a point that does not seem to be widely understood by political scientists. For example, Barabas and colleagues (2011: 21) assert that "we observed non-random treatment assignment (i.e., $p < 0.10$ differences between the treatment and control groups on partisanship, age, education, and race) which necessitates the use of statistical controls later in the paper." Of course, by "non-random," the authors probably did not mean that their randomization mechanism was faulty; therefore, they continue to treat the study as an experiment, not as an observational study resulting from a failed randomization mechanism.

Adding a variable to the statistical model for an experimental analysis because it failed a randomization check is an inferior model choice (see Imai et al. 2008; Mutz and Pemantle 2011). It is a misnomer to say that it "controls" for the lack of balance and there is no defensible reason to accept this as a "fix" for a broken random assignment mechanism, if that is indeed what we are looking for by providing such tables.

We suspect that instead of a failure to randomize, what many authors and reviewers actually have in mind is the unlucky chance that experimental conditions are unbalanced on some variable of potential interest. Of course, if it is a strong predictor of the dependent variable, a pre-treatment measure of that variable should have been used for blocking purposes or as a planned covariate in the model to increase model efficiency regardless of balance; this is the only appropriate purpose of covariates.

But more importantly, using a third variable to try to "correct" a model for imbalance ignores the fact that the alpha value used to test experimental hypotheses *already takes into account* that cells will be uneven on some characteristics due to chance. The committee report states that "we . . . do not counsel any particular modeling response to the table of covariate means that we ask the researcher to provide." However, given that the only example provided of what one might do with this information is to adjust the treatment effects by including covariates, this seems somewhat misleading. As they elaborate, "Our guiding principle is to provide the reader and the reviewer the information they need to evaluate what the researcher has done and to update their beliefs about the treatment effects accordingly." But exactly how should the reviewer or reader "update" his or her beliefs about the effects of treatment based on such a table?

If such a table truly serves as evidence (or lack thereof) that proper random assignment was accomplished, then such tables will greatly affect a study's chances of publication. By requiring such information, an editor automatically suggests to readers and authors that it is both informative and relevant because it is worth valuable journal space. If it is to be required, it seems incumbent upon the editors to inform authors as to how they will interpret such information. Will they conclude that random assignment was done incorrectly on this basis and thus automatically reject it from an experimental journal? Will they compel authors to provide evidence that random assignment was done correctly, and if so, what would be considered compelling evidence?

And are they required to present evidence of balance even on demographic variables that bear no relation to the outcome variables simply because they are widely used as control variables in observational analyses or on all pre-treatment measures because they happen to be in the study? We maintain that such practices have no scientific or statistical basis and serve only to promote further methodological confusion.

The report does not distinguish between pre-treatment measures available to the researcher but not chosen for inclusion in the model, and those chosen in advance for inclusion in the model. If, as is common with survey data, there are dozens of available pre-treatment measures, then is balance supposed to be reported for all of them? If so, why? As Thye (2007: 70) has noted, "Not all the factors that make experimental groups different from control groups are relevant to the dependent variable; therefore, not all factors must necessarily be equated. Many differences simply do not matter." To advocate such a practice is to encourage mindless statistical models, which should not be promoted by any journal. It encourages a misunderstanding of what randomization does and does not accomplish. It also promotes further confusion in the field as to the distinction between experimental and observational analysis.

To reiterate, pre-treatment variables known from previous research to be highly predictive of the outcome should always be included in the model as covariates. To fail to do so is to reduce power so that only the strongest effects will be seen. It should not take a failed balance test to reveal such a variable, and the fact that a balance test fails for a particular variable makes it no more likely that this variable is in fact related to the dependent variable.

Finally, the question of whether a variable is adequately "controlled for" is a non sequitur in experimental research. Control variables exist for good reasons in observational studies (potential spuriousness), but treating a covariate as a control variable in the experimental setting makes no sense. Nonetheless, this practice is currently widespread. Including a variable in the statistical model because it has been found to be out of balance is also precisely the wrong reason to include a variable and should not increase our confidence in findings.

Taken at face value, the Standards Report promotes randomization checks strictly as a way of "evaluating the integrity of the randomization process" (Gerber et al., 2014: 92). They suggest that imbalances due to chance are distinguishable from imbalances due to faulty random assignment mechanisms. But if a fear of faulty mechanisms is the real reason for doing them, then the typical response (adding new variables to the model) is completely inadequate; if a randomization mechanism fails, the researcher needs to start over from scratch.

To summarize, failed balance tests cast doubt on experimental results; as a result, one seldom if ever finds a published experimental study with a "failed" randomization; instead they are routinely dubbed "successful" (Malhotra and Popp 2012: 39) and even "highly successful" (Butler and Broockman 2011: 467). Moreover, if an author admits a "failed" balance test, it is strictly on one or two "unbalanced" variables that are, as a result of the balance test, included as covariates. This practice does not fix the problem if the randomization mechanism was, indeed, broken.

The real harm in this practice is the possibility of a Type II error when a skeptical referee or editor causes a correct finding to be suppressed or uses it as a reason to alter the statistical model to include more covariates in order to suggest that they have "adjusted" for a bad or unlucky randomization. This practice implies that the reader's ad hoc estimates of treatment effects and confidence might be superior to the researcher's stated estimates and confidence. As mentioned above, changing the model voids confidence statements.

At times, this kind of misunderstanding of randomization is made explicit, even within our top journals. For example, as an article in the *Journal of Politics* explains,

> In order to ensure that the experimental conditions were randomly distributed—thus establishing the internal validity of our experiment—we performed difference of means tests on the demographic composition of the subjects assigned to each of the three experimental conditions . . . . As Tables 1a and 1b confirm, there were no statistically significant differences between conditions on any of the demographic variables.. .. Having established the random assignment of experimental conditions, regression analysis of our data is not required; we need only perform an analysis of variance (ANOVA) to test our hypotheses as the control variables that would be employed in a regression were randomly distributed between the three experimental conditions (Scherer and Curry 2010: 95).

A test of mean differences across five demographic variables is not what gave this study internal validity; proper use of random assignment did. Moreover, controlling for these variables in a regression equation or using them as covariates would not have fixed a failed randomization, nor would it have increased the power of the study, unless those variables were chosen in advance for the known strength of their relationships with the dependent variable rather than for their relationships with the independent variable, as is suggested above.

Many researchers do not appear to understand that the alpha value used in statistical tests *already incorporates the probability of the unlucky draw*. As Hyde (2010: 517) suggests in another experimental study,

> In theory, the randomization should produce two groups that are equivalent except that one group was assigned to be "treated" with international election observation. Although it is unlikely, it is possible that randomization produces groups of villages/neighborhoods that

> are different in important ways, and could potentially generate misleading results. Therefore, I also check the degree to which the two groups are similar . . .

Here again, a randomization check is being used to try to uncover the unlucky draw in order to increase confidence in the findings as opposed to presenting “misleading results.” This is a well-intentioned impulse, but one should not update his or her confidence in the findings on this basis.

Using balance tests on a subset of observed variables as a way of establishing group equivalence promotes further confusion because of the current popularity of matching techniques in analyzing observational data. If, for example, a researcher matches treated and untreated subjects on five demographic characteristics, there is a tendency to see this as equivalent to an experiment in which a balance test has been performed on these same five variables. What is lost here is an understanding of the fundamental importance of random assignment. Matching techniques, no matter how complex, cannot accomplish the same strength of causal inference as a true experiment. Only random assignment collectively equates subjects on observed and unobserved characteristics.

The lure of the unlucky draw, however, goes far beyond this. There is a strong urge to believe that one can test for occurrences of the exceptional event: that not only does Type I error have a visible signature but also that we can sense it, and therefore should look at balance tests even though we are not able to prescribe an acceptable response to what we see. This may be what is responsible for the heightened concern about errors in randomization.

Sub-Optimal Statistical Models

Randomization checks notwithstanding, a more serious and widespread problem in political science experiments is confusion surrounding analyzing experimental versus observational data. By the Standards Committee’s own count, 75% of the experimental studies published in political science do not show the unadulterated effects of treatments on outcomes (Gerber et al. 2014: 88). In other words, 75% of experimental results never show the reader the dependent variable means by experimental condition or a regression including only treatment effects; instead, they present multiple regression equations in which effects of treatment are already adjusted by many other “control” variables, or they present *predicted means* as a function of a multivariate regression equations including other variables (e.g., Hutchings et al. 2004: 521–2).

For example, in one analysis of experimental results, in addition to dummy variables representing six different experimental treatments, the author includes in his experimental regression analysis nine different “control variables” including if the respondent follows politics “most of the time,” if he/she is a college graduate, age, female, minority, employed, internet connection speed, conservative ideology and liberal ideology. The rationale for this particular set of variables when predicting the dependent variable—nightly news exposure—is unclear (see Prior 2009). Likewise, in another experiment, in addition to dummy variables for experimental treatment effects, the author includes 13 additional predictors of the outcome, none of which significantly predicts the dependent variable, and the reader is instructed that these variables are included as "control variables" (Ladd 2010: 39).

What is particularly unfortunate about this practice is that reviewers and authors often seem to be under the impression that an experimental finding is more robust if it survives the inclusion of a large number of "control variables" when nothing could be further from the truth. Instead of encouraging this practice, reviewers and editors should look at such large models with suspicion and demand justifications for the particular statistical model that is used. Findings can be coaxed over the line of statistical significance by virtue of what is included or excluded.

We are not suggesting that social scientists are dishonest when such variables are included in a model. In fact, many authors find themselves compelled to include them in an analysis specifically because of a reviewer or editor's request. Even when they are aware that their finding is more valid without excessive and unjustified variables in the model, they comply in order to achieve publication. Adding variables after the fact invalidates the reporting of confidence levels. Moreover, the proper reporting of confidence is not an idle exercise; in fact, some suggest that it has large scale consequences (Ioannidis 2005).

In some cases, these additional variables in models testing experimental effects even include items assessed after the treatment. For example, in an article in the *American Journal of Political Science,* a study of income distribution norms and distributive justice promotes the inclusion of a variable assessed post-treatment as a means of strengthening confidence in the experimental findings.

> By asking participants in the post-experimental questionnaire about their own perception of the relationship between merit and income, and then entering that information as an independent variable in our regression analyses, we are able to determine that our experimental manipulations rather than participants' pre-existing perceptions explain our results. This test shows how using multiple regression analysis to enter additional controls can strengthen experimental research" (Michelbach et al. 2003: 535).

In short, what is most common within political science is for researchers to analyze experimental data as if it were observational data, often including "control variables" inappropriately. If there is no reason to think variables will increase efficiency in the estimation of treatment effects, and no reason to think that they are even correlated with the outcome, they should not be in the model, regardless of what they are called. Which other variables are or are not included is unsystematic and typically unjustified with some models including one set, and another model within the same paper including a different set, thus opening the floodgates for all kinds of foraging for results through their inclusion and exclusion.

We are not the first to note the logical problems inherent in randomization checks. Psychologist Robert Abelson (1995: 76) dubbed the practice of testing for differences between experimental groups a "silly significance test": "Because the null hypothesis here is that the samples were randomly drawn from the same population, it is true by definition, and needs no data." Senn (1994: 1716) calls the practice of performing randomization tests "philosophically unsound, of no practical value, and potentially misleading." In the context of political science, Imai and colleagues (2008: 482) echo this sentiment, suggesting that any other purpose [than to test the randomization mechanism] for conducting such a test is "fallacious."

The field of political science is populated with researchers primarily trained in observational modes of research. For those trained exclusively in observational methods, the source of confusion is obvious. If one treats experimental data as if it were observational, then of course one would be worried about "controlling for" variables, and about imbalance in any variable not used as a covariate. We believe the Standards Committee should take a stand on whether they believe "control" variables are sensible in experimental studies and/or whether they are an acceptable fix for the broken random assignment mechanisms that balance tests are supposedly designed to root out.

So, how can researchers be certain that randomization was done properly? The CONSORT guidelines already provide guidance as to the kinds of details that can help reviewers and readers judge the randomization process (see Moher et al. 2010; Schulz et al. 2010). Notably, because randomization is a *process* rather than an outcome, what is more useful than tables of means is a description of that process. Political scientists should test randomization mechanisms in advance of studies if there are concerns, and promote transparency by describing the process of randomization for each study.

When debating the utility of randomization checks, one argument we have heard a number of times is "Why not just do both and let the reader decide?" In other words, why not present both the original analysis and one adjusted by including the covariates that fail a specified balance test? Assuming the randomization mechanism is not faulty, there remain several good reasons not to do this. We elaborate on four such reasons.

1. *Incorrectness*. Significance statements and size comparisons for the estimated treatment effect will be wrong. To see why, consider the process by which the adjusted estimate is computed. After the random assignment to experimental conditions, a set of covariates exhibiting imbalance is added to the model. An estimated treatment effect is computed by regressing onto the treatment variable and this larger set of covariates. Confidence intervals and p-values for such an estimator do not coincide with confidence intervals and p-values for a model in which the same covariates are chosen before the units are assigned to conditions (see Permutt 1990).
2. *Intractability*. Computing correct confidence statements for a model in which covariate selection is not fixed in advance has, to our knowledge, never been undertaken. An idealized example is worked out in Permutt (1990). Whether or

not such a computation is feasible, it is certainly not included in any standard statistics package. We can therefore be fairly certain that the correct computation was not carried out.

3. *Inefficiency*. Even if confidence was computed correctly for the adjusted estimate, the new estimator would not be an improvement over the old one. Any selection of covariates, whether chosen in advance or based on imbalance due to random assignment, leads to an estimator. The better estimator is the one with the least variance. For any pre-treatment measure, Z, one might choose to include Z in the model, exclude it, or include it only if it is unbalanced. The last of these is never the best choice. One always does better by deciding up front whether to include Z as a covariate. The mathematical proof supporting this is discussed in greater detail in Mutz and Pemantle (2011).
4. *Irrelevance*. One might argue that presenting both estimators and allowing the reader to choose is best because it reports everything that would originally have been reported, plus one more piece of data which the reader is free to ignore. Reporting a second conclusion, however, casts doubt on the first conclusion; it does not merely add information. It leads to "the wrong impression that we need balance, which is one of the many myths of randomization" (Statisticalmisses.nl 2013). Recommendation E2 of the Standards for Experimental Research calls for an analysis to be specified prior to the experiment, and deviations from this come at a cost in credibility. Furthermore, if given a choice between two models, many would automatically choose the model with more covariates based on a (faulty) belief that such models are more robust. The researcher's job is to present the best data analysis, not to present them all and allow the reader to choose.

CONCLUSION

The goal of the *Journal of Experimental Political Science* should be not only promoting the more widespread use of experimental methods within the discipline, but also promoting best practices and a better understanding of experimental methods across the discipline. Toward that end, we hope the Standards Committee will consider changing the standards with respect to manipulation checks, reporting of response rates, and randomization checks as part of the ongoing process of making what was historically an observational discipline more appropriately diverse in its approaches to knowledge. More specifically, we suggest the following adjustments:

1. Recommend manipulation checks for latent independent variables; that is, independent variables in which the operationalization and the causal construct are not identical;
2. Require response rates only for studies that claim to be random probability samples representing some larger population;

3. If tables of pretreatments means and standard errors are to be required, provide a justification for them. (Note that the following are not suitable justifications: (a) Confirmation of "successful" randomization, (b) Supporting the validity of causal inference, and (c) Evidence of the robustness of inference.)
4. If the inclusion of balance tests/randomization checks is described as desirable as in the current document, prescribe the appropriate response and interpretation of "failed" tests.

Evidence of widespread misunderstandings of experimental methods is plentiful throughout our major journals, even among top scholars in the discipline. As a result, future generations of political scientists are often not exposed to best practices. *The Journal of Experimental Political Science* should play a lead role in correcting these misunderstandings. Otherwise, the discipline as a whole will be seen as less methodologically sophisticated than is desirable. *The Journal of Experimental Political Science* could play an important role in raising the bar within the discipline by including requirements that are both internally coherent and statistical defensible.

REFERENCES

Abelson, R. 1995. *Statistics as Principled Argument*. Hillsdale, NJ: L. Erlbaum Associates.

Arceneaux, K. 2012. "Cognitive Biases and the Strength of Political Arguments." *American Journal of Political Science* 56(2): 271–85.

Arceneaux, K. and R. Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53(4): 755–70.

Barabas, J., W. Pollock and J. Wachtel. 2011. "Informed Consent: Roll-Call Knowledge, the Mass Media, and Political Representation." Paper Presented at the Annual Meeting of the American Political Science Association, Seattle, WA, Sept. 1–4. (http://www.jasonbarabas.com/images/BarabasPollockWachtel_RewardingRepresentation.pdf), accessed September 1, 2014.

Berinsky, A. J., M. F. Margolis and M. W. Sances. 2014. "Separating the Shirkers from the Wokers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3): 739–53.

Berinsky, A. J. and T. Mendelberg. 2005. "The Indirect Effects of Discredited Stereotypes in Judgments of Jewish Leaders." *American Journal of Political Science* 49(4): 845–64.

Boers, M. 2011. "In randomized Trials, Statistical Tests are not Helpful to Study Prognostic (im)balance at Baseline." *Lett Ed Rheumatol* 1(1): e110002. doi:10.2399/ler.11.0002.

Butler, D. M. and D. E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3): 463–77.

Cover, A. D. and B. S. Brumberg. 1982. "Baby Books and Ballots: The Impact of Congressional Mail on Constituent Opinion." *The American Political Science Review* 76(2): 347–59.

Cozby, P. C. 2009. *Methods of Behavioral Research*. (10th ed.). New York, NY: McGraw-Hill.

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

Foschi, M. 2007. "Hypotheses, Operationalizations, and Manipulation Checks." Chapter 5, In *Laboratory Experiment in the Social Sciences*, eds. M. Webster and J. Sell, (pp.113–140). New York: Elsevier.

Franco, A., N. Malhotra and G. Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203): 1502–5.

Franklin, C. 1991. "Efficient Estimation in Experiments." *Political Methodologist* 4(1): 13–15.

Gerber, A., K. Arceneaux, C. Boudreau, C. Dowling, S. Hillygus, T. Palfrey, D. R. Biggers and D. J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1(1): 81–98.

Gosnell, H. F. 1942. *Grass Roots Politics*. Washington, DC: American Council on Public Affairs.

Green, D. P. and A. S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: The State of the Discipline*, 3rd Edition. eds. H. V. Milner and I. Katznelson, (pp.805–32). New York: W.W. Norton & Co.

Harbridge, L., N. Malhotra and B. F. Harrison. 2014. "Public Preferences for Bipartisanship in the Policymaking Process." *Legislative Studies Quarterly* 39(3): 327–55.

Hiscox, M. J. 2006. "Through a Glass and Darkly: Attitudes Toward International Trade and the Curious Effects of Issue Framing." *International Organization* 60(3): 755–80.

Hutchings, V. L., N. A. Valentino, T. S. Philpot and I. K. White. 2004. "The Compassion Strategy: Race and the Gender Gap in Campaign 2000." *Public Opinion Quarterly* 68(4): 512–41.

Hyde, S. D. 2010. "Experimenting in Democracy Promotion: International Observers and the 2004 Presidential Elections in Indonesia." *Perspectives on Politics* 8(2): 511–27.

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

Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8): e124. doi:10.1371/journal.pmed.0020124.

Kessler, J. B. and S. Meier. 2014. "Learning from (Failed) Replications: Cognitive Load Manipulations and Charitable Giving." *Journal of Economic Behavior and Organization* 102(June): 10–13.

Ladd, J. M. 2010. "The Neglected Power of Elite Opinion Leadership to Produce Antipathy Toward the News Media: Evidence from a Survey Experiment." *Political Behavior* 32(1): 29–50.

Malhotra, N. and E. Popp. 2012. "Bridging Partisan Divisions over Antiterrorism Policies: The Role of Threat Perceptions." *Political Research Quarterly* 65(1): 34–47.

Michelbach, P. A., J. T. Scott, R. E. Matland and B. H. Bornstein. 2003. "Doing Rawls Justice: An Experimental Study of Income Distribution Norms." *American Journal of Political Science* 47(3): 523–39.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger and D. G. Altman. CONSORT. 2010. "Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials." *Journal of Clinical Epidemiology* 63(8): e1–e37.

Morgan, K. and D. Rubin. 2012. "Rerandomization to improve covariate balance in experiments." *Annals of Statistics* 40(2): 1263–82.

Mutz, D. C. and R. Pemantle. 2011. "The Perils of Randomization Checks in the Analysis of Experiments." Paper presented at the Annual Meetings of the Society for Political Methodology, (July 28–30). (http://www.math.upenn.edu/~pemantle/papers/Preprints/perils.pdf), accessed September 1, 2014.

Newsted, P. R., P. Todd and R. W. Zmud. 1997. "Measurement Issues in the Study of Human Factors in Management Information Systems." Chapter 16, In *Human Factors in Management Information System*, ed. J. Carey, (pp.211–242). New York, USA: Ablex.

Perdue, B. C. and J. O. Summers. 1986. "Checking the Success of Manipulations in Marketing Experiments." *Journal of Marketing Research* 23(4): 317–26.

Permutt, T. 1990. "Testing for Imbalance of Covariates in Controlled Experiments." *Statistics in Medicine* 9(12): 1455–62.

Prior, M. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *Journal of Politics* 71(3): 893–908.

Sances, M. W. 2012. "Is Money in Politics Harming Trust in Government? Evidence from Two Survey Experiments." (http://www.tessexperiments.org/data/SancesSSRN.pdf), accessed January 20, 2015.

Sansone, C., C. C. Morf and A. T. Panter. 2008. *The Sage Handbook of Methods in Social Psychology*. Thousand Oaks, CA: Sage Publications.

Saperstein, A. and A. M. Penner. 2012. "Racial Fluidity and Inequality in the United States." *American Journal of Sociology* 118(3): 676–727.

Scherer, N. and B. Curry. 2010. "Does Descriptive Race Representation Enhance Institutional legitimacy? The Case of the U.S. Courts." *Journal of Politics* 72(1): 90–104.

Schulz, K. F., D. G. Altman, D. Moher, for the CONSORT Group. CONSORT 2010. "Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *British Medical Journal* 340: c332.

Senn, S. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13: 1715–26.

Senn, S. 2013. "Seven Myths of Randomisation in Clinical Trials." *Statistics in Medicine* 32(9): 1439–50. doi: 10.1002/sim.5713. Epub 2012 Dec 17.

Statisticalmisses.nl, 2013. (http://www.statisticalmisses.nl/index.php/frequently-asked-questions/84-why-are-significance-tests-of-baseline-differences-a-very-bad-idea), accessed January 21, 2015. As attributed to Senn (2013).

Thye, S. 2007. "Logic and Philosophical Foundations of Experimental Research in the Social Sciences." Chapter 3, In *Laboratory Experiments in the Social Sciences*, (pp.57–86). Burlington, MA: Academic Press.

Valentino, N. A., V. L. Hutchings and I. K. White. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes during Campaigns." *The American Political Science Review* 96(1): 75–90.