The perils of balance testing in experimental design: Messy analyses of clean data

Diana C. Mutz*
Departments of Political Science and Communication
University of Pennsylvania
and
Robin Pemantle†
Department of Mathematics, University of Pennsylvania
and
Philip Pham
Department of Mathematics, University of Pennsylvania‡

January 31, 2017

Abstract

Widespread concern over the credibility of published results has led to scrutiny of statistical practices. We address one aspect of this problem that stems from the use of balance tests in conjunction with experimental data. When random assignment is botched, due either to mistakes in implementation or differential attrition, balance tests can be an important tool in determining whether to treat the data as observational versus experimental. Unfortunately the use of balance tests has become commonplace in analyses of "clean" data, that is, data for which random assignment can be stipulated. Here, we show that balance tests can destroy the basis on which scientific conclusions are formed, and can lead to erroneous and even fraudulent conclusions. We conclude by advocating that scientists and journal editors resist the use of balance tests in all analyses of clean data.

Keywords: balance test, control, covariate, estimator, regression, variance

*Research supported in part by the Institute for the Study of Citizens and Politics
†Research supported in part by National Science Foundation grant # DMS 1209117
‡Now at Google

1 Introduction

As a widely read study begins, "There is increasing concern that most current published research findings are false" (Ioannidis, 2005). In the last ten years the same theme has been repeated (Harvey et al., 2014), questioned (Jager and Leek, 2014) and elaborated (Moonesinghe et al., 2007), leading to 3,000 citations for Ioannidis' paper alone. The crisis of confidence in experimental studies is of particular concern because randomized trials are the most credible source for causal inference. Many responses to the perceived crisis have been proposed, including replication and advanced registry of statistical models (Humphreys, 2013). Some responses come in the form of guidelines such as those set forth by Gerber et al. (2014), the EGAP members committee (2011b) and the CONSORT guidelines (Moher et al., 2010).

Following such guidelines should add to, rather than subtract from, the credibility and accuracy of findings. Indeed, this is the case with the vast majority of such guidelines. Here we focus on an exception, namely **balance testing** of experimental data. In this case, published guidelines (e.g., EGAP members committee, 2011b; Gerber et al., 2014) as well as common practices have resulted in less accurate, less credible reporting of results.

1.1 Balance testing

Consider an experiment in which N subjects are each randomly assigned to one of k experimental conditions. For each subject, a number of measures are available before the assignment of treatment. A balance test is an analysis of the degree to which the distribution of these other measures across treatment groups is near its expectation.

Balance tests can be used in one of three ways. Minimally, a *table of baseline differences* is displayed, e.g.,

	Treatment	Control
MALE	32	11
HISPANIC	4	5
50+ YEARS OF AGE	19	16
etc.	...	...
etc.	...	...
Number of subjects	50	25

Table 1: Baseline demographics by condition

Second, and most commonly, the term refers to statistical tests for the null hypothesis

that the assignment to treatment is independent of pretreatment data. Typically, this is accompanied by an interpretation, for example, "Balance tests are not statistically significant, therefore the randomization was successful." Third, it is sometimes suggested that the analysis be run after controlling for any significantly imbalanced covariates (EGAP members committee, 2011b). The main point of this article is that balance testing is inappropriate and potentially harmful whether (*i*) displaying baseline data, (*ii*) performing a statistical test on this data, or (*iii*) altering the model based on balance tests.

To motivate the thorough examination of these practices, it is important to understand that in many social sciences balance testing is the rule not the exception, that it occurs in flagship journals for many fields, and that it is encouraged and sometimes required by reviewers and editors. To illustrate these practices, we draw from recent experimental studies in diverse fields such as political science, sociology, psychology, economics, business, medicine and education.

Frequently, authors state that balance is necessary for inference in experimental designs. An article in the *Journal of Politics* article explicitly describes this line of reasoning:

> 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 ... 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, p. 95)

Likewise, articles in the *American Sociological Review* and *Social Psychology Quarterly* claim:

> Because the vignettes were randomly assigned, confounding would only be an issue if randomization [referring to a balance test] failed" (Phelan et al., 2013, 2014, p. 178, p. 306).

Sometimes it is implied that randomizations failing balance tests are not truly random. As

stated in an article in *Management International Review*,

> Analyses of variance were performed to determine if there were any significant differences across the experimental cells in the respondents' age, education, work experience, and joint venture experience. There were none, and thus the subjects were deemed to have been randomly assigned to the various treatments (Sullivan and Peterson, 1982, page 35).

Further, failed balance tests frequently lead authors to alter their statistical model to adjust for the imbalance. For example, as suggested by Hutchings et al. (2004, p. 521),

> Partisanship is included in the analysis because of imbalances in the distribution of this variable across the conditions.

Likewise, an article in the *American Journal of Political Science* suggests that such corrections should increase confindence in the result:

> 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, p. 862); see also Jerit and Barabas (2012); Prior (2009); Panagopoulos (2011).

Experimental economists also have claimed that balance testing is informative for model selection:

> Even if the randomization was carried out appropriately, it may be informative to see whether any of the key covariates were by chance relatively imbalanced between treatment and control group, so that prior to seeing the outcome data an analysis can be designed that addresses these presumably modest imbalances (Imbens and Athey, 2016, p. 23).

In a similar vein, some authors say that an adjustment *would have been* performed had a balance test failed.

Of particular interest is the attitude taken by the influential CONSORT (Consolidated Standards of Reporting Trials) guidelines (Moher et al., 2010), discouraging statistical tests for balance but requiring tables of baseline demographics by experimental condition. Most

medical journals, including the prestigious *New England Journal of Medicine*, follow CONSORT guidelines. These guidelines go a long way toward ensuring the quality of statistical analyses; however, their treatment of balance is confusing at best. Their recommendation for detatiled descriptions of the randomization mechanism is a good start. But then they require a table of "baseline demographic and clinical characteristics for each group" (p. 6). As a result, a balance table appears prominently as Table 1 in every publication in every journal adhering to CONSORT[1]. Likewise, the standards posted for the *Journal of Experimental Political Science* (APSA Standards Committee, 2014, Section C) require authors to supply "a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures by experimental group" (pps. 84, 93). In both cases, no advice is given as to which demographics or other characteristics are relevant for these purposes. Further, it is unclear what one is to make of the table. CONSORT advises against testing for statistically significant differences in the table. But the table is still required because "The study groups should be compared at baseline for important demographic and clinical characteristics so that readers can assess how similar they were" (Moher et al., 2010, page 14). How a reader decides if the groups were similar is left open, as is what to do if one suspects an imbalance of some kind. Further, all such guidelines leave open the possibility for adjusting based on an observed covariate imbalance. For example, CONSORT advises that authors "should clarify the choice of variables that were adjusted for, ... and specify whether the analysis was planned or suggested by the data" (p. 14).

1.2 Clean experiments

In an observational study, any imbalanced covariate is a possible explanation for a spurious relationship. Therefore, an analysis omitting the variable is suspect. Sometimes an experiment can be compromised in a way that renders the data essentially observational, that is, not modeled by a probability distribution on assignments to treatment, independent of all else that has occurred prior to the treatment. There is, however, a limited number of ways that random assignment can be compromised. First, the randomization mechanism could be faulty. Nowadays the prevalence of perfectly good randomizing apps on computers, tablets and even phones has nearly eliminated malfunction in the random assignment mechanism.

[1]Nonetheless, according to CONSORT, citing Assmann et al. (2000), over half of a sample of papers from prestigious journals report the statistical significance of balance tests.

Although one can certainly point to examples where randomization went awry (Heeringa, 2001; Conroy-Krutz and Moehler, 2016; Gerber and Green, 2000), in only one example of which we are aware was a balance test useful in detecting the problem (see Imai, 2005, for details). In most cases, a description of the randomization procedure would have sufficed to detect the fault (Heeringa, 2001; Conroy-Krutz and Moehler, 2016; Gerber and Green, 2000). When detected most such problems cause the data to be treated as observational. No amount of control variables will render it analyzable as experimental data.

Aside from a faulty randomization mechanism, there is one additional circumstance under which an experiment cannot be analyzed cleanly. This is when differential attrition may have occurred. This means that the likelihood of a subject dropping out of the study is affected by the treatment, with different treatment conditions leading to different rates of attrition. In this case, the assumption has been violated that the (remaining) subjects in different experimental conditions are statistically similar[2]. Importantly, only attrition that is differential over treatment groups poses a threat to internal validity. Considerable attention has been devoted to handling this case (see, e.g., Gerber and Green 2012, Chapter 7). We exclude cases subject to differential attrition from the definition of "clean data".

In all other cases, namely when randomization occurred, treatments were delivered, and differential attrition did not occur, we call it a **clean experiment**[3]. A clean experiment allows precisely quantifiable statistical inferences to be made about causal relations.

1.3 Messy analyses

Procedures are available to reduce random error and thereby increase the efficiency of experimental analyses. Each possible procedure yields a potentially different estimate, with different confidence bounds and a different interpretation. Many of these procedures can help to reduce the noise to signal ratio, but they are not necessary for the internal validity of experimental conclusions.

Consider these four reasons for including covariates in a model for an experimental analysis:

(i) Covariates widely believed to predict the dependent variable may be included to reduce

[2]In cases where the treatment was delivered differentially, e.g., medicine was not taken, literature was not read, but subjects remained in the experiment, an intent to treat analysis is still valid.

[3]There is an implicit assumption of correct execution: no violation of blinding, no interaction among subjects, all machinery in working order, and an endless number of other assumptions, without which, of course, no inference can be valid.

the noise to signal ratio, increasing efficiency.

(ii) Covariates may be included because the advanced plan for the analysis was to include everything in the data set on the off chance that these variables might provide a more efficient analysis.

(iii) Covariates may be included because their distribution is unbalanced across treatment groups.

(iv) Covariates may be included because, among several analyses, the one with that specific set of covariates was the one that produced the cleanest, most significant findings.

Most would agree that (*i*) is a good reason[4] and that (*iv*) is dishonest. The contended middle ground consists of (*ii*) and (*iii*). Our point in this paper is to address (*iii*), and to a lesser extent (*ii*). As we discuss in the next section, conducting balance tests leads to their use in model selection. This occurs because many researchers find (*iii*) to be a compelling justification, and even those who don't find it compelling receive pressure from editors and reviewers to use balance tests in model selection. Including covariates for this reason is not only unnecessary with clean data, but as we will show, it is not even helpful to do so, and may in fact be damaging. Thus, the thesis of this paper is that **one should not perform balance tests on clean data**.

1.4 The case against balance testing with experimental data

A number of criticisms of balance testing in the experimental context are old but bear repeating. First, as pointed out on multiple occasions (e.g., Senn, 1994), a statistical test for the hypothesis that the conditions were randomly assigned is meaningless if one already knows that the conditions were randomly assigned. Secondly, if there is a great number of pre-treatment variables, then one *should* find significant imbalance on at least one variable, so what is one to make of a flunked balance test? To address this concern, (e.g., Hansen and Bowers, 2008) have developed omnibus balance tests that take into account the number of variables examined for balance. But the question of what to make of a failed balance test

[4]Even this is not an open and shut case. As pointed out in Berk et al. (2016), "Still, one has to wonder if any of these covariance-based options are really worth the trouble. Simple differences in means or proportions are unbiased ATE estimators under the Neyman model or under random sampling. [...] Possible gains in precision from covariance adjustments are in principle most needed with small samples, a setting in which they currently have no formal justification."

(omnibus or otherwise), or an apparently lopsided table of pre-treatment variables, remains largely unanswered.

More fundamentally, a balance test appears to be an attempt to reduce the probability of Type-I error by checking whether the randomization was an "unlucky draw". When we make an assertion at the confidence level of, say, 0.01, we are asserting that only 1% of randomizations would have produced such results under the null hypothesis. The underlying statistical theory contains no information as to when one has encountered one of these "unlucky draws". Despite this, a great number of scientists believe that they can tell when an unlucky draw is likely to have occurred and that balance tests are instrumental in doing so. When a false positive result occurs, it is due to an anomalous distribution of the (adjusted) *dependent* variable across treatment conditions. The intuition that this coincides with unequal distribution of pre-treatment variables operates outside any model and therefore provides no scientific conclusion as to the likelihood of an unlucky draw, nor what to do in the event of one. Section 4 addresses unlucky draws in a quantitative way, exploring what sort of model one could develop and what it would say about the likelihood of detecting unlucky draws using balance tests.

Consider an experiment with two conditions (treatment and control), in which for each subject there is one outcome measure (the dependent variable) as well as p available pre-treatment measures (covariates). We may choose any subset of covariates to include in the model, so there are 2^p possible models for estimating treatment effects by linear regression. Any one of these, when specified in advance, leads to a quantifiable statistical conclusion. What does not lead to any scientific conclusion is the following procedure, which we call "balance test and adjust" (BT&A): select zero or more covariates for the model; then if the researcher observes any ways in which other pre-treatment measures are not balanced across treatment groups, include those variables as well. The main point of this article is to show why this is a bad idea, how the conclusions are weakened when researchers do so, and why this practice threatens the integrity of the scientific process. In particular we will argue that

- Confidence statements are wrong (see the Appendix of Permutt 1990).
- The wrong statistical model is chosen (Theorems 1 and 2 below).
- Absolutely any result can be obtained when there are sufficiently many covariates (Pham,

2016, Theorem 2.1).

- In typical experimental settings, BT&A can easily lead to a false boost that pushes the p-value across a threshold such as 0.05 or 0.01 (Table 2 below).

In the next section we begin by recalling the basic statistical paradigm as set forth by Fisher, Rubin and others. In Section 3 we discuss the probability models underlying various analyses and the lack of such a model for BT&A. When a probability model is introduced, it belies the confidence statements resulting from such a procedure. We then show that BT&A always produces an inferior choice of covariates relative to what could be chosen without balance testing. In addition to *post hoc* adjusting via covariates, adjustment via post-stratification is also shown to be sub-optimal. In Section 4, we analyze a formal model in which the reduction of Type-I error when weeding out unlucky draws is shown to be small compared to the introduction of Type-II error. The tradeoff is worse than would be obtained by other means. In Section 5 we discuss more quantitatively the potential harm done by the incorrect inclusion of covariates.

2 Statistical framework

2.1 Randomization

Random assignment, while not guaranteeing to distribute any one characteristic perfectly among treatment groups, stochastically distributes all characteristics, known and unknown[5]. No non-random assignment, matching included, can do this. As a result, random assignment allows scientists to make mathematically precise inferences. For this reason, randomized trials are known as the gold standard for inference about causation.

The price one pays for stochastic equalization of all factors, known and unknown, is the addition of noise. This results in a quantifiable chance of any given relation appearing to be true due to random fluctuations. The fact that the probability of such an illusion is quantifiable, irrespective of the precise pathway by which a random assignment might produce this illusion, is the crowning jewel of the method of random assignment. Any inference comes with a confidence statement asserting that some null or alternative hypothesis

[5]It is important to remember that if the only concern is balance on a known set of covariates, then the optimal procedure is some sort of matching; this avoids the errors of order $N^{-1/2}$ introduced by randomization.

would have probability less than p of producing such an illusion. What makes such analyses scientific is that there *is* an underlying model, that a mathematical statement can be made about the model, and, to the extent that the model is true, we can be certain about the (probabilistic) conclusion.

The statistical models that underlie confidence statements do not differentiate between a "good" or "bad" randomization. Anyone who talks about balance testing in terms of establishing "successful randomization" has lost track of the statistical model. The main problem with balance testing is that it leads to conclusions *outside of any statistical model*. In Section 4 we construct a probability model that could plausibly underlie the use of balance tests in weeding out potential spurious findings. In order for this to be helpful, one must assume that the researcher made a sub-optimal choice in omitting from the analysis one or more variables known or suspected of predicting the dependent variable. Even in these cases, we show that the degree to which conclusions might be altered is relatively small.

2.2 The role of conditional bias

The next most common confusion is the status of an estimator that ignores available variables: is it unbiased, conditionally unbiased or neither, and how important is this? To answer, it is important to keep in mind precisely what is being estimated. Consider, by way of example[6], the simplest experiment in which one must choose between two linear estimators, one a simple difference of means and the other adjusted by a covariate. Specifically, we suppose that N subjects (N is even) have been divided uniformly and randomly into two equally sized groups. Let $T_i, 1 \leq i \leq N$ denote which treatment is given to subject i (1 for treatment, 0 for control), let $Y_i, 1 \leq 1 \leq N$ denote the values of the dependent variable, and let $Z_i, 1 \leq i \leq N$ be the values of a single pre-treatment measure[7]. Letting $\{\xi_i\}$ denote independent mean zero noise, the linear model

$$Y = \alpha + \beta T + \xi \tag{1}$$

[6] We avoid repetition by discussing one representative scenario instead of all variations: continuous versus categorical dependent variable, number of treatment groups, group sizes distributed as binomials or precisely equal, and so forth; qualitatively, our arguments and conclusions remain unchanged across these possible scenarios.

[7] These values are taken as fixed, not random. The model does not address whether there are underlying true values, of which Z_i are noisy measures.

produces an estimator of the treatment effect given by

$$\beta_1 := \frac{C(T,Y)}{V(T)}\,. \tag{2}$$

Here and in what follows, $C(T,Y)$ denotes the empirical covariance $N^{-1}\sum_{i=1}^{N}(T_i-\overline{T})(Y_i-\overline{Y})$ and $V(T)$ denotes the empirical variance $C(T,T)$. Because we have two groups of equal size, $V(T)$ is a constant and β_1 is just the difference in sample means in the two conditions.

If the model (1) is correct, then β_1 is an unbiased estimator of β. As pointed out by Freedman (2008b), the Neyman-Rubin model provides a far more general setting in which β_1 estimates a quantity of interest and is unbiased. Following Freedman's exposition, we suppose a value of the outcome measure exists for all subjects in *both* conditions, treatment and control. This holds in the linear model: one can simply change the value of T_i to see what the outcome would have been in the other condition[8]. Clearly the Neyman-Rubin model makes fewer assumptions than does the linear model (1) [9].

In the linear model (1), β_1 estimates the model parameter β which has a clear interpretation as the amount treatment adds to the dependent variable. If instead one only assumes the Neyman-Rubin model, then β_1, the empirical difference in means, is an unbiased estimator of the mean within-subject difference between the treated and not-treated values of the dependent variable over the subject population. No linearity assumption is required. Within-subject linearity is automatic because there are only two possible values of the dependent variable[10].

Suppose now one augments the linear model to include the covariate:

$$Y = \alpha + \beta T + \gamma Z + \xi\,. \tag{3}$$

One may then estimate β by regressing on both T and Z, obtaining

$$\beta_2 := \frac{C(T,Y)V(Z) - C(T,Z)C(Y,Z)}{V(T)V(Z) - C(T,Z)^2}\,. \tag{4}$$

[8] If ξ_i is not independent of T_i, just conditionally mean zero, then this involves a possible mean zero change in ξ_i as well.

[9] Even the Neyman-Rubin model involves counterfactual suppositions whose philosophical grounds could be shaky, as is vividly explained by Hofstadter (1980, pages 633–640).

[10] In either model, relating β_1 to the mean treatment effect in some larger population would be a further step, having to do with the way the subject population was obtained, and is not one we will discuss.

As an estimator of β in (3), β_2 is unbiased. Under the Neyman-Rubin causal model, when the linear model (3) fails, β_2 is not in general an unbiased estimator of treatment effect. This is noted by Freedman (2008a), who gives conditions under which this estimator, despite having bias, outperforms the simple estimator β_1.

The notion of conditional bias or lack thereof is somewhat of a red herring. Because we know Z, it would seem that conditional unbiasedness given Z is a decided advantage, but this is not a general truth. Conditional unbiasedness means that the estimator has the correct mean given the values of Z for all values of the model parameters. If we have reason to believe the coefficient of Z is zero or very small, it does little good to demand lack of bias in the event that Z is a strong predictor. This can be seen most clearly in the case where Z and Y are independent. It is intuitively clear that the estimator β_1 that ignores Z should perform better than the estimator β_2 that uses irrelevant information. In fact, the latter is equal to the former plus a quantity that is a noisy estimate of zero. The difficulty is we do not know whether Z is independent of Y. We can choose to estimate its effect and include that in the computation or we can choose not to.[11] By making the correct choice, we hope to obtain a less noisy estimate of Y.

It cannot be stressed enough that the researcher *always* makes this choice. This choice involves judgment based on previous experience. It cannot be demonstrated to be correct within the model. The researcher can choose to include Z, exclude it, or even flip a coin. The researcher could allow pre-treatment measures to influence the model, as in BT&A [12] Section 3 of this paper is devoted to showing that no matter what the available information, the outright choice leads to a better estimate than does BT&A.

2.3 Efficiency versus credibility

We use the term "credibility" to refer to our trust that the reported findings represent the true state of affairs, or in other words, that a false positive has not occurred. Depending on one's philosophy, this could mean that replication would produce similar results, or that changing the treatment variable would cause a change in the outcome with average size in

[11] It is on this point that the generally helpful EGAP recommendations may be misleading. In (EGAP members committee, 2011a, Section 6), it is explained that including a covariate on which imbalance is observed can help to "retrieve the conditionally unbiased estimate." The implication is that adding covariates due to their imbalance is advocated. Section 7 goes on to warn against adding non-prognostic covariates, with which we heartily agree, however it is doubtful that this entirely dispels the implication in Section 6.

[12] This can only be done in a pre-specificed manner if inferences are to remain valid.

the given range, or would in the future cause such a change. A smaller p-value increases the credibility of a causal finding. A suspected flaw in the experiment decreases credibility. Credibility can also be decreased for subjective reasons, such as disbelief in the theoretical explanation or strong prior beliefs as to the likelihood of the purported findings.

Researchers care not only about false positives but also about reducing the rate of false negatives. The precision to which an effect has been estimated affects the significance of the results. Thus, greater precision goes hand in hand with greater efficiency. However, once significance is summarized in a p-value, information about precision is no longer related to credibility. The same goes for sample size, degrees of freedom and so forth; these help to determine significance, but do not further affect the credibility of a finding beyond the information captured by the p-value.

Covariates can be very helpful for purposes of increasing precision. A measure known or suspected to predict the dependent variable, when included in the design (blocking, matching) or the analysis (as a covariate), can greatly reduce the variance within experimental conditions, leading to more efficient tests and more precise estimates. Because of this, it is generally a mistake not to include in the model a variable known or strongly suspected to predict the dependent variable. This mistake affects the significance level that can be obtained for a given effect size, or the effect size that can be inferred at a given significance level. However, a conclusion with a given p-value is not less credible than any other result with the same p-value just because one did not obtain the precision that one could have by including covariates.

Unfortunately we have seen many formal and informal responses to balance tests along the following lines: "Covariate Z was not balanced across treatment groups. I will not believe your estimate of the effect of treatment T on the dependent variable Y with stated confidence level $p < \alpha$ unless you change your model to include Z as a covariate." Such a statement confuses efficiency with credibility. A better reasoned statement would be, "I bet you would have more precisely estimated the effect of T on Y had you included Z as a covariate in the model." The latter statement suggests that more convincing evidence (for example a lower p-value) could be marshalled by adding the covariate Z to the model, but correctly says nothing about the credibility of the existing model.

Interestingly, such a comment is rare when the covariate Z is well balanced across treatment groups. The inclusion of Z in the model would not in that case change the

estimate of the treatment effect, however it could potentially change the confidence interval in either direction. The difference in prevalence of such responses when Z is imbalanced versus balanced is a good proxy for the level of misunderstanding about whether covariates are included for the purposes of credibility or efficiency.

The question of whether the experimenter should have included the covariate in the model is a valid one, though it does not address the credibility of the result that was actually obtained. One should choose covariates for their anticipated relation to the *dependent* variable. To alter this choice because of a balance test is to choose based on a relation with the *independent* variable. In Section 3 we will show that *ex ante* considerations (the extent to which the covariate is expected to predict the dependent variable) are always better than using the results of a balance test when choosing covariates for a model.

3 Possible models for BT&A and their consequences

In this section we consider possible models for balance testing and adjustment and the conclusions to which they lead. In all cases, legitimacy rests on pre-determining one's actions. Changing decisions about whether to re-randomize, alter the analysis, re-interpret the conclusion and so forth will undermine any scientific basis for the conclusions.

The first possibility is balance testing without adjustment. Indeed, the guidelines from CONSORT and the Journal of Experimental Political Science (APSA Standards Committee, 2014) mandate presenting tables of pre-treatment measures; no adjustment is recommended and one is left to guess how such a table might be used. Each reader is left to change his or her opinion based on subjective assessments of the balance test. Given what we know about the propensity of readers to confuse efficiency with credibility, to confuse conditional unbiasedness with accuracy, stochastic equidistribution with balance, p-values with posteriors, and so on, providing a table of balance with no further instruction seems like a decidedly bad idea.

A second possibility is that the result of the balance test may be used to select covariates for the model. This is, for example, what is recommended by the EGAP members committee (2011a)[13]. This, at least, can be formally modeled and scientifically valid consequences computed. Subjects are recruited, pre-treatment measures gathered and random assignments made, after which covariates are selected according to a pre-specified algorithm, and

[13]This advice was modified in 2016 and now provides no indication of how to use balance test results.

the analysis conducted.

But this sequential procedure is a different probability model than any model choosing that chooses covariates in advance, and as such, has different confidence levels. Such a sequential analysis has been analyzed precisely once, to our knowledge, by Permutt (1990). This analysis concerns a model in which the treatment and a single covariate have jointly Gaussian distribution with unknown correlation, and the model includes the covariate if and only if the empirical correlation between the treatment and the covariate exceeds a certain threshold. The correct confidence statements are shown to differ from those obtained from the standard analysis of the model chosen in each case. In cases with a greater number of potential covariates, we know of neither theory nor empirical work that can guide the assessment of confidence levels.

3.1 Sub-optimality of BT&A

Another problem with choosing covariates based on the results of a balance test is that such procedures are never as good as choosing covariates *ex ante*[14]. To make this presicse, we concentrate on a model with N subjects assigned at random to one of two treatment conditions. The treatment variable X_j is 1 if subject j is in the treatment group and 0 if subject j is in the control group. There is one pre-treatment measure Z_j available for inclusion in the model if desired. The dependent variable is Y_j. Let $\overline{X}, \overline{Y}$ and $\overline{Z}$ denote the empirical means and let C denote the empirical covariance, thus,

$$C(X,Y) = N^{-1}\sum_{j=1}^{N}(X-\overline{X})(Y-\overline{Y})\,.$$

Let $V(X) = C(X,X)$ denote empirical variance. The estimators

$$\hat{\beta}_0 = \frac{C(X,Y)}{V(X)} \tag{5}$$

$$\hat{\beta} = \frac{C(X,Y)V(Z) - C(X,Z)C(Y,Z)}{V(X)V(Z) - C(X,Z)^2} \tag{6}$$

are the result of regressing Y on X alone or on both X and Z, respectively.

[14]More precisely, any sequential procedure can be dominated by an *ex ante* procedure based on the same available informtation, excluding the balance test.

Suppose that the truth is given by the linear model

$$Y = \mu + \beta X + \gamma Z + \theta\xi \tag{7}$$

where ξ_j are independent mean zero unit variance noise terms. Despite the omission of Z from the first model, both $\hat{\beta}_0$ and $\hat{\beta}$ are unconditionally unbiased estimators of β. The variance of Y explained by Z is $\gamma^2\mathbb{E}Z^2$, while the variance unexplained by either X or Z is the variance θ^2 of the noise term. The ratio

$$T := \frac{\gamma^2}{\theta^2}V(Z)$$

is a parameter of the model measuring the portion of variance of Y explained by Z. The following result is proved in the Online Appendix.

Theorem 1. *Given the X and Z variables, let $\rho^2 = C(X,Z)^2/(V(X)V(Z))$. Then $\mathbb{E}_*(\hat{\beta} - \beta)^2 < \mathbb{E}_*(\hat{\beta}_0 - \beta)^2$ if and only if if $T > N^{-1}/(1-\rho^2)$, where $\mathbb{E}_*$ denotes conditional expectation with respect to the X and Z variables.*

To understand what Theorem 1 is saying, keep in mind that ρ^2 is known before treatment, but that T is a hidden parameter, never known to the experimenter. If T were known, one could be certain whether including Z in the model increased or decreased the variance of the resulting estimator. One doesn't know T, but one can calculate the threshold $t = N^{-1}/(1-\rho^2)$. This threshold increases with ρ. The experimenter always has to guess how strongly Z predicts Y. This result says that one should be *less* inclined to include Z, not more, if you see that Z is imbalanced (ρ^2 is large). For example, in the hypothetical study in Table 1, suppose it is unclear whether the "Hispanic" variable is sufficiently predictive of the dependent variable to include in the model. The underrepresentation of Hispanics in the treatment group means you need a stronger belief in the likely effect of race on the outcome to make the case for inclusion than you would if the distribution of Hispanics had been closer to its mean.

An intuitive explanation is that collinearity between the treatment and the covariate introduces uncertainty as to which is responsible for any variation predicted in the dependent variable. The threshold, it should be noted, is small, and minimizing the variance of the estimator is far from the only goal. Theorem 1 does not dictate what variables to

include, but it does imply that a failed balance test is not a good reason for including a variable in the analysis.

What if the experimenter omits a variable from the analysis that really should have been included? Might not its appearance on a failed balance test be useful as a reminder or alert? Failure to include this variable will reduce the efficiency of the experiment. On the other hand, surely the experimenter can take responsibility for running over this mentally before covariates are chosen. Any covariate which one would be tempted to include should it flunk a balance test, should be included from the start. The contrapositive of this is, if it is not worth including from the start, then a failed balance test should not change one's mind.

3.2 Sub-optimality of BT & post-stratify

For completeness, we discuss one more possible response to a failed balance test, namely post-stratification. Our discussion will be brief because this situation is largely analogous to the selection of covariates. Post-stratification is a means of increasing *efficiency*. As such, it is not relevant to credibility. As was the case with selection of covariates, post-stratification is meant for cases where a substantial relationship between the dependent variable and the covariate is anticipated. For example, Miratrix et al. (2013) warn that "post-stratifying on variables not heavily related to outcome is unlikely to be worthwhile and can be harmful."

One can make this statement precise by considering a sequential design, in which one first tests for balance on a covariate Z, then post-stratifies on Z only if Z fails the balance test. Again, a threshold may be computed for the portion of variance of the dependent variable that a dichotomous covariate must predict in order for stratification by this variable to produce a net reduction in MSE of the treatment estimate. In the Online Appendix we prove the following result, giving the threshold as a function of the distribution of the covariate across treatment groups:

Theorem 2. *Consider two estimators of a treatment effect, one the simple difference of means estimator and one post-stratified by a covariate taking the values 0 and 1. The post-stratified estimator will have a lower MSE than the simple estimator if and only if the ratio of variance in the dependent variable predicted by the covariate to the unpredicted variance is greater than*

$$\frac{1}{n}\frac{[ab(a+b)+cd(c+d)](a+b)(c+d)}{abcd(a+b+c+d)}\,.$$

Here, a, b, c, and d are the respective sizes of the groups of controls with covariate value 0, controls with covariate value 1, treated subjects with covariate value 0 and treated subjects with covariate value 1.

The threshold is somewhat complicated, but it is minimized when the group sizes are equal, and it always grows when the distribution of the covariate within a treatment group becomes more imbalanced. This means that a failed balance test is not a good reason to perform a stratified analysis. If a balance test fails, the threshold for stratification to improve efficiency increases rather than decreases.

4 Detecting the unlucky draw

We have discussed three common reactions to a failed balance test: doing nothing, including more covariates in the model, or changing to a post-stratified analysis. A fourth possible reaction is to modify the conclusion, either by throwing it out entirely (declaring it not credible) or by altering the confidence statement accompanying the conclusion, e.g., "We found an effect significant at the level of $p < 0.01$ but because the randomization was bad, we should be less confident." This section examines the use of balance tests to weed out unlucky instances of randomization or to modify conclusions in such a case.

There is an immediate foundational problem: probability theory, as it is currently formalized, does not provide a definition of randomness for an instance of randomization. If a coin is flipped one hundred times independently, then with probability 2^{-100} it should come up heads every time, and this outcome is no less "random" than any other given sequence $HTTHTHH \cdots HT$, which should also arise with probability 2^{-100}. The probability model of independent coin flips applies only to the collective description of all possible sequences that could have occurred.

There is considerable divergence between the mathematical theory of probability, which underlies statistical inference, and the common practice and understanding of probabilities. These have been well documented and occur even among well trained users of statistical methods; see, e.g., Hastie and Dawes 2010; Kahneman et al. 1982. Most relevant here is the strong and seemingly universal urge to classify individual sequences as "successfully randomized" or "unsuccessfully randomized". Definitions for randomness of an individual sequence have indeed been put forth, mostly in the context of pseudo-random number

generation (Kolmogorov, 1998; Martin-Löf, 1966). These definitions typically apply only to infinite sequences and none is widely used in probability theory or statistics.

To analyze what happens when balance tests are used for weeding out unlucky draws, one must create a probability model for the situation where the distribution of covariates across conditions is analyzed, and a certain randomization is flagged as "unsuccessful". There are two possible scenarios, depending on whether the option exists to re-randomize. If each subject is randomly assigned and treated before the next subject arrives, then re-randomization is not possible. Likewise, re-randomization is not possible when a balance test is requested by a reviewer.

Re-randomization

Assuming the re-randomization is done by creating an entirely new random assignment schedule, the correct probability model is rejection sampling. Assignments are divided into two well defined sets, with the rejected set typically of much smaller probability than the acceptable set. If the random assignment is from the rejected set, it is discarded and a new assignment is chosen, repeating until the chosen assignment is from the acceptable set. Mathematically, all one has done is to replace the uniform measure on random assignments by the uniform measure on the acceptable set.

A treatment of re-randomization via rejection sampling can be found in Morgan and Rubin (2012). They provide philosophical arguments for and against re-randomization, along with a mathematical analysis. For example, when the treatment and control are required to be of equal size, then any rejection criteria symmetric with respect to switching the two groups leads to an unbiased estimator under rejection sampling (Morgan and Rubin, 2012, Theorem 2.1 or the more general Theorem 4.1). They give an example of a criterion involving the Mahalanobis distance between the distributions of the treatment and control groups and compute the reduction in variance obtained by rejecting when the distance is above a threshold.

Morgan and Rubin state directly, "We only advocate rerandomization if the decision to re-randomize or not is based on a pre-specified criterion." Thus, they stay within the paradigm of rejection sampling as described above (see Morgan and Rubin, 2012, page 1267). Potential drawbacks of re-randomization are that it can lead to biased estimates, incorrect confidence statements (usually ones that are too conservative) and can

invalidate Gaussian approximations. Morgan and Rubin state explicitly (cf. Section 2.3 above) that the advantage of their method is the reduction of variance, leading to "more powerful tests and narrower confidence intervals;" in other words, greater efficiency. Their paper makes no mention of credibility of findings.

Post-hoc adjustment, that is, BT&A

We turn next to the more common case of post-hoc adjustment based on the result of a balance test. Suppose that an experiment finds a treatment effect to be significant at the level of $p < 0.05$. A balance test is conducted. Should we revise our inference in light of the result of the test? The intuition is that we should believe the result more if there is balance and less if there is not. The p-value is the probability of a false positive of at least this magnitude under the null hypothesis. A false positive is, by definition, undetectable because it looks just like a true positive. Covariate imbalance, however, is detectable. If researchers are interested in using the results of the balance test to help us avoid false positives, the question becomes: will the occurrence of a false positive (undetectable) sufficiently often be indicated by covariate imbalance (detectable)?

In one sense the answer is disappointing: this depends highly on the parameters of the model. In particular, it depends on the extent to which the covariate predicts the dependent variable. One cannot, therefore, expect to compute this in advance except by imposing assumptions that are necessarily subjective. In another sense, though, the answer is surprisingly concrete. Given particular assumptions, the number of false positives ensnared by a particular balance test is precisely computable because it is a feature of the null hypothesis, a well defined probability model.

We illustrate with some computations in a theoretical model. In order to incorporate the unknown predictive value of the covariate, we allow the covariate to be a random variable. We suppose there are $2N$ subjects divided randomly into a treatment and a control group, each of size N. There is one dependent variable, Y, and one potential covariate Z. The treatment estimator $\hat{\beta}$ is the difference in means of Y between the treatment and control group. The balance statistic δ is the difference in means of Z between the treatment and control group. The treatment effect is judged significant if $\hat{\beta} > b$ (one-sided test) or $|\hat{\beta}| > b$ (two-sided test) for some constant b. The covariate is judged to be out of balance if $|\delta| > d$ for some constant d. It remains to specify the distribution of (X_j, Z_j, Y_j).

In order to examine false positives we assume the null hypothesis (that is, independence of X_j and Y_j, which, because Z is a pre-treatment measure and X is a random assignment, implies independence of X_j from the pair (Z_j, Y_j). For the distribution of (Z_j, Y_j), because this is an illustrative model, we use a distribution suited for computation, namely a bivariate normal. The correlation α is a parameter of the model; the remaining parameters are scale parameters and do not affect the computations.

Proposition 3 (bivariate normal null model, one-sided test)**.** *Suppose the common distribution of (Y_j, Z_j) is bivariate normal with correlation $\alpha \in (0, 1)$. Then the portion that a false positive occurs for a one-sided test rejecting the null hypothesis with probabity p, but it is weeded out by a two-sided balance test with rejection probability q, is given by*

$$\frac{1}{2\pi} \int_q^\infty \int_{\frac{-r-\alpha y}{\sqrt{1-\alpha^2}}}^{\frac{r-\alpha y}{\sqrt{1-\alpha^2}}} e^{-\frac{1}{2}(y^2+w^2)} \, dy \, dw \, . \tag{8}$$

□

What does this say for specific values of α, b and d? Suppose $b = d = 0.05$ as is most commonly the case in published articles. When $\alpha = 0$, this tells us something we already know: 2.5 thousandths of the time a false positive is rejected due to a (coincidental) failure of a balance test. Of course this means nothing because an equal proportion of true positives are rejected. More interesting is the fact that this does not increase all that much with α.

For example, when $\alpha = 0.3$, a false positive is weeded out 3.9 times in a thousand trials. There are 50 false positives in every thousand trials. The number of false positives weeded out when a balance test is performed with a covariate having correlation 0.3 with the dependent variable is just 1.4 per thousand higher than is obtained by witch-hunting (a balance test for a completely irrelevant variable). This represents an additional 3% of all false positives. The efficacy of this must be measured against the true positives that will be discarded. Furthermore, it should be stressed that a variable has been excluded from the original analysis whose correlation with the dependent variable is 0.3. It would be highly unusual to ignore a variable with this kind of predictive power. A much better procedure would be to include this variable in the original analysis from the beginning[15].

[15]The same reduction in false positive rate could be obtained simply by testing significance at a 4.5% level instead of a 5% level. This makes the Type-I error the same as in the balance test scenario. Meanwhile,

5 How easy is it to cheat?

Outright fishing for an analysis that will yield significance is of course quite different from a post-hoc adjustment of the model due to the specific concern of covariate imbalance. Nevertheless, any argument that parsimonious model selection helps to ensure credibility of results must examine the degree to which credibility is threatened when a model incorporates more covariates than are justified in the planning (and perhaps registration) phase. Furthermore, as discussed at the end of this section, variables that fail a balance test are particularly likely to aid in data dredging [16].

If the supply of potential covariates is unbounded, then by selecting an appropriate subset for the model one can produce an arbitrarily large apparent treatment effect, with an arbitrarily small apparent p-value (Pham, 2016, Theorem 2.1). This threat to credibility is often dismissed due to the perception that it would take a great deal of snooping through a great number of potential analyses before identifying one that altered the findings in any meaningful way. Only by sifting through more data than is feasible or by reducing the degrees of freedom to a level that rings alarm bells, it is argued, can results be rigged.

Quantitative studies of this issue are difficult to come by. Multiplying the p-value by the number of analyses that have been run (or that could have been run) gives valid confidence statements but is overly conservative because the analyses are not independent. We know of only one paper quantifying this, namely the work of of Berk et al. (2013). Their results, which apply to observational as well as experimental data, give estimates for the maximum significance over a family of estimators. Their examples are more of theoretical interest than of practical use.

In order to give some insight into the feasibility of manipulating results by selecting among potential covariates, Pham (2016) considers a model in which one dichotomous treatment variable X and a set of dichotomous potential covariates $\{Z^{(k)}\}$ are all, in fact, independent of the dependent variable Y. Synthetic data is generated for N subjects, after which a subset S of covariates is sought for which the analysis of the linear model with covariates $\{Z^{(k)} : k \in S\}$, if analyzed in the usual way, produces an apparently significant rejection of the null hypothesis (regression coefficient of Y on X is zero) at a 5% confidence

because the test is much more efficient, the Type-II error is far less than for a simple difference of means test, whereas balance testing and rejecting always produces a greater Type-II error.

[16] Gelman and Loker (2014) point out that data dredging can often be unintentional, thus the discussion concerns not only rare cases of fraud but in fact a circumstance arising in many, perhaps most studies.

level. The subset S is found via dynamic programming. Each trial records the size of S as well as the number m of variables that were examined by the program. The value of m is important because it cannot exceed the supply of potential covariates.

The simulations are summarized in two tables, the first covering all the data sets that were generated and the second restricted to a subset of borderline cases. Table 2 tabulates the means of $|S|$ and m over all of the synthetic data sets, as well as the frequency with which S has size 1. Many datasets such as those gathered under the auspices of the NSF-supported Time-sharing Experiments for the Social Sciences routinely include dozens of potential covariates for each subject. In other words, values of m ranging from 20 to perhaps 60 are not uncommon in contemporary survey-experimental data. This translates, according to Table 2, to a great likelihood of finding a set whose inclusion as covariates would create a false significant finding of a treatment effect.

N	Mean size of S	Mean of m	frequency S has size 1
50	10.8	24.6	3.5%
100	15.0	37.4	2.5%
200	20.9	53.1	1.4%
400	30.2	75.8	1.0%

Table 2: Mean size of a set S of covariates needed for a false positive, as well as mean and variance of the number, m, of potential covariates considered by the dynamic programming algorithm in order to arrive at the successful set, S

A related finding concerns the use of covariates to boost significance over a given threshold, such as the all-important $p < 0.05$ mark, termed "p-hacking" by Ellenberg (2014, page 153) [17]. Table 3 tabulates synthetic data over those instances in which the original p-value lies between 0.05 and 0.10. When we look at the use of irrelevant covariates to boost a p-value from the range $[0.05, 0.10]$ to a significant result ($p < 0.05$), the results are even more striking because it is quite common that only one covariate is needed in order to boost significance across the commonly accepted threshold.

The size of S is important because the justification of S becomes more difficult the larger and more arbitrary it becomes. However, a single covariate is unlikely to raise eyebrows. A reasonable question here would be how likely it is to turn up such a covariate as a result of a failed balance test. One might naively expect that the single covariate a cheater would like to use is significantly out of balance only 5% of the time, but this is not true. Pham

[17] Ellenberg attributes the term *p-hacking* to Uri Simonsohn.

N	Mean size of S	Mean of m	frequency S has size 1
50	2.0	7.8	49.7%
100	2.3	8.6	44.4%
200	3.2	11.4	25.2%
400	3.9	14.2	21.0%

Table 3: Mean size of a set S of covariates needed for a false positive, as well as mean and variance of the number, m, of potential covariates considered by the dynamic programming algorithm in order to boost significance of a p-value in the range $[0.05, 1]$ to $p < 0.05$.

(2016) examined this for the cases in Table 2 and found that the portion of the time that the covariate in question failed a $p < 0.05$ balance test was roughly 15%. A closer look at equation (6) explains why. The difference between $\hat{\beta}$ and $\hat{\beta}_0$ is roughly proportional to the empirical covariance between the treatment and the covariate. Therefore, the Z-score of the covariate in the cases where significance was boosted across a threshold should be size-biased samples of the unit normal. The two-sided 1.96-tail for a size-biased standard normal is $2\int_{1.96}^{\infty} \frac{|x|}{2} e^{-x^2/2}\, dx \approx 0.146$.

To summarize the results of this section, it is almost always possible to obtain a false positive by adding covariates indiscriminately, requiring a pool of only a few dozen. If one limits the cheating to adding a single covariate, the Type-I error, nominally set at 5%, jumps to between 6% ($N = 400$) and 8.5% ($N = 50$). Among those cases where the original data has a significance level between 0.05% and 0.10%, a single covariate boosts the p-value across the $p < 0.05$ threshold between 20% and 50% of the time, over the same range of N. Failing a balance test triples the chance that a covariate will accomplish this. Finally, one should keep in mind that this analysis underestimates the ease of cheating, as it fails to take into account any tweaking of the analysis other than inclusion or exclusion of available covariates. Many other practices are possible for post-hoc model selection, including creation of new variables, e.g., for curvilinear relationships, changing how a variable is binned, post-stratification or other re-weighting, and so forth.

6 Conclusion

Our central conclusion is that there is no statistical basis for advocating the Balance Test & Adjust procedure for analyzing randomized experiments. Although balance testing is widely advocated and is believed to produce more credible estimates of experimental effects, post-hoc adjustments using covariates selected on the basis of failed balance tests have no basis

in statistical theory. Covariates that are chosen after an experiment is conducted should produce greater rather than lesser skepticism about the results. In an era of grave concerns surrounding the validity and impartiality of experimental findings, it is of particular concern when researchers include covariates for reasons other than their pre-established capacity to predict the dependent variable. Furthermore, tables of baseline demographics by condition, whether or not accompanied by statistical testing, lead to subjective interpretations outside of any statistical model.

As demonstrated in this article, including covariates in clean experiments for reasons other than efficiency opens the floodgates to coaxing weak coefficients across the critical p-value threshold. The availability of large numbers of covariates makes it relatively easy to adjust findings through the use of one or more covariates. We do not mean to suggest that most researchers who engage in the practices criticized here are actively trying to cheat. However, given the rising concern about whether individual findings are replicable, the risks of including variables without justification, and the lack of statistical basis for advocating BT&A, we advocate eliminating this common practice.

Acknowledgements

Thanks to Larry Brown and Richard Berk for helpful discussions of several drafts of this paper. Thanks to Bruce McCullough for detailed suggestions and a number of citations.

References

APSA Standards Committee (2014). Recommended reporting standards for experiments. *J. Exper. Poli. Sci.*.

Assmann, S., S. Pocock, L. Enos, and L. Kasten (2000). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Lancet 355*, 1064–1069.

Berinsky, A. and T. Mendelberg (2005). The indirect effect of discredited stereotypes in judgments of Jewish leaders. *Amer. J. Pol. Science 49*, 845–864.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Ann. Statist. 41*, 802–837.

Berk, R., E. Pitkin, L. Brown, A. Buja, E. George, , and L. Zhao (2016). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*.

Conroy-Krutz, J. and D. Moehler (2016). Partisan media and engagement: a field experiment in a newly liberalized system. *Pol. Comm. 33*, to appear.

EGAP members committee (2011a). 10 things to know about covariate adjustment.

EGAP members committee (2011b). Evidence in governance and politics: research principles.

Ellenberg, J. (2014). *How not to be wrong*. New York: Penguin Press.

Freedman, D. (2008a). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat. 2*, 176–196.

Freedman, D. (2008b). On regression adjustments to experimental data. *Adv. Appl. Math. 40*, 180–193.

Gelman, A. and E. Loker (2014). The garden of forking paths. *Amer. Scientist 102*(6), 9.

Gerber, A., K. Arceneaux, C. Boudreau, S. Dowling, S. Hillygus, T. Palfrey, D. Biggers, and D. Hendry (2014). Reporting guidelines for experimental research. *J. Exper. Poli. Sci. 1*, 81–98.

Gerber, A. and D. Green (2000). The effects of canvassing, telephone calls and direct mail on voter turnout: a field experiment. *Amer. Pol. Sci. Review 94*, 653–663.

Gerber, A. and D. Green (2012). *Field Experiments*. New York: W. S. Norton and Company.

Hansen, B. and J. Bowers (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science 23*, 219–236.

Harvey, C., Y. Liu, and H. Zhu (2014). ... and the cross-section of expected returns. *NBER Working Paper Paper #20592*, DOI: 10.3386/w20592.

Hastie, R. and R. Dawes (2010). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage Publications, Inc.

Heeringa, S. (2001). The surveycraft cati systems random number generation features and their effects on analysis of the 1997 nes pilot 'group threat' experiment.

Hofstadter, D. (1980). *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Vintage.

Humphreys, M. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis 21*, 1–20.

Hutchings, V., N. Valentino, T. Philpot, and I. White (2004). The compassion strategy: race and the gender gap in Campaign 2000. *POQ 68*, 512–541.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *Amer. Pol. Sci. Review 99*, 283–300.

Imbens, X. and X. Athey (2016). The econometrics of randomized experiments. *arXiv 1607.00698*, 85.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine 8*, DOI: 10.1371/journal.pmed.0020124.

Jager, L. and J. Leek (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics 15*, 1–12.

Jerit, J. and J. Barabas (2012). Partisan perceptual bias and the information environment. *J. Politics 74*, 672–684.

Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Kolmogorov, A. (1998). On tables of random numbers. *Theor. Comp. Sci. 207*, 387–395.

Martin-Löf (1966). The definition of a random sequence. *Information and Control 9*, 602–619.

Miratrix, L., J. Sekhon, and B. Yu (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. Royal. Stat. Soc. Edin., ser. B 16:26*, 28.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). Consolidated standards of reporting trials.

Moonesinghe, R., M. Khoury, and A. Janssens (2007). Most published research findings are false – but a little replication goes a long way. *PLOS Medicine 28*, DOI: 10.1371/journal.pmed.0040028.

Morgan, K. and D. Rubin (2012). Rerandomization to improve covariate balance in experiments. *Ann. Stat. 40*, 1263–1282.

Panagopoulos, C. (2011). Thank you for voting: Gratitude expression and voter mobilization. *Journal of Politics 73*, 707–717.

Permutt, T. (1990). Testing for imbalance of covariates in controlled experiments. *Stat. Med. 9*, 1455–1462.

Pham, P. (2016). Just How Easy is it to Cheat a Linear Regression? Master's thesis, University of Pennsylvania, Philadelphia, PA.

Phelan, J., B. Link, and N. Feldman (2013). The genomic revolution and beliefs about essential racial differences. *Amer. Socio. Review 28*, 167–191.

Phelan, J., B. Link, S. Zellner, and L. Yang (2014). Direct-to-consumer racial admixture tests and beliefs about essential racial differences. *Social Psych. Quarterly 77*, 296–318.

Prior, M. (2009). Improving media effects research through better measurement of news exposure. *J. Politics 71*, 893–908.

Scherer, N. and B. Curry (2010). Does descriptive race representation enhance institutional legitimacy? *J. Politics 72*, 90–104.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine 13*, 1715–1726.

Sullivan, J. and R. Peterson (1982). Factors associated with trust in Japanese-American joint ventures. *Mgmt. Inter. Review 22*, 30–40.

7 Online Appendix: proofs of mathematical results

PROOF OF THEOREM 1: Replacing X_i by $X_i - (1/n)\sum_{j=1}^n X_j$, we may assume without loss of generality that $\sum_{i=1}^n X_i = 0$. Similarly, we may assume $\sum_{i=1}^n Z_i = 0$ and that $\beta = 1$. We establish

$$\mathbb{E}_*(\hat{\beta}_0 - 1)^2 = \frac{\theta^2}{n\,V(X)} + \frac{C(X,\gamma Z)^2}{V(X)^2}\,; \tag{9}$$

$$\mathbb{E}_*(\hat{\beta} - 1)^2 = \frac{\theta^2}{n} \cdot \frac{V(Z)}{V(X)V(Z) - C(X,Z)^2}\,. \tag{10}$$

Plugging in the structural equation (7) for Y into (5) yields

$$\begin{aligned}\hat{\beta}_0 &= \frac{n^{-1}\sum_{i=1}^n X_i(X_i + \gamma\, Z_i + \theta\, U_i)}{n^{-1}\sum_{i=1}^n X_i^2} \\ &= \frac{\sum_{i=1}^n X_i^2 + \gamma\sum_{i=1}^n X_i Z_i + \theta\sum_{i=1}^n X_i U_i}{\sum_{i=1}^n X_i^2}\,.\end{aligned}$$

Taking the conditional expectation with respect to $\mathcal{F}$, subtracting 1 and squaring gives

$$(\beta_* - 1)^2 = \frac{1}{V(X)^2}\left[C(X\gamma Z)^2 + 2C(X,\gamma Z)C(X,\theta U) + C(X,\theta U)^2\right]\,. \tag{11}$$

The identities $\mathbb{E}C(t,U) = 0$ and $\mathbb{E}C(t,U)^2 = n^{-1}V(t)$ hold for any real numbers $t_1,\ldots,t_n$. Taking the expectation of (11) and applying these identities gives

$$\mathbb{E}\left((\hat{\beta}_0 - 1)^2 \,|\, \mathcal{F}\right) = \frac{1}{V(X)^2}\left[C(X,\gamma Z)^2 + \frac{\theta^2 V(X)}{n}\right]$$

which is the same as (9).

Similarly, plugging in the structural equation (7) for Y into (6) yields

$$\begin{aligned}
\hat{\beta} &= \frac{C(X,Y)\,V(Z) - C(X,Z)\,C(Y,Z)}{V(X)\,V(Z) - C(X,Z)^2} \\
&= \frac{V(Z)(V(X) + \gamma\,C(X,Z) + \theta\,C(X,\theta U))}{V(X)\,V(Z) - C(X,Z)^2} \\
&\quad - \frac{C(X,Z)(C(X,Z) + \gamma V(Z) + \theta\,C(Z,U)}{V(X)\,V(Z) - C(X,Z)^2} \\
&= \frac{V(Z)V(X) + \theta\,V(Z)C(X,U)) - C(X,Z)^2 - \theta\,C(X,Z)C(Z,U)}{V(X)\,V(Z) - C(X,Z)^2} \\
&= 1 + \theta\,\frac{V(Z)C(X,U) - C(X,Z)C(Z,U)}{V(X)\,V(Z) - C(X,Z)^2}\,.
\end{aligned}$$

Subtracting 1, squaring, taking expectation with respect to $\mathcal{F}$, again using the identity $\mathbb{E}C(t,U)^2 = n^{-1}V(t)$, gives (10):

$$\begin{aligned}
&\mathbb{E}((\hat{\beta} - 1)^2 \,|\, \mathcal{F}) \\
&= \theta^2\,\frac{V(Z)^2\mathbb{E}(C(X,U)^2 \,|\, \mathcal{F}) + C(X,Z)^2\mathbb{E}(C(Z,U)^2 \,|\, \mathcal{F})}{(V(X)\,V(Z) - C(X,Z)^2)^2} \\
&\quad - \frac{2V(Z)C(X,Z)\mathbb{E}(C(X,U)\,C(Z,U) \,|\, \mathcal{F})}{(V(X)\,V(Z) - C(X,Z)^2)^2} \\
&= \frac{\theta^2}{n}\,\frac{V(Z)^2V(X) + V(Z)C(X,Z)^2 - 2V(Z)C(X,Z)^2}{(V(X)\,V(Z) - C(X,Z)^2)^2} \\
&= \frac{\theta^2}{n}\,\frac{V(Z)}{V(X)\,V(Z) - C(X,Z)^2}\,.
\end{aligned}$$

Comparing, $\mathbb{E}_*(\hat{\beta} - 1)^2 < \mathbb{E}_*(\hat{\beta}_0 - 1)^2$ if and only if

$$\frac{\theta^2}{n\,V(X)} + \frac{C(X,\gamma Z)^2}{V(X)^2} > \frac{\theta^2}{n}\,\frac{V(Z)}{V(X)V(Z) - C(X,\gamma Z)^2}\,.$$

Algebraic simplification, multiplying both sides by $V(X)$ and using $C(X,Z)^2/(V(X)V(Z)) = r^2$, reduces this to

$$\frac{\theta^2}{n} + \gamma^2 r^2\,V(Z) > \frac{\theta^2}{n(1-r^2)}\,.$$

Dividing by θ^2 and moving the first term on the left to the right then shows this is equivalent

to

$$\frac{\gamma^2}{\theta^2}\, r^2\, V(Z) > \frac{1}{n}\left(\frac{1}{1-r^2} - 1\right) = \frac{r^2}{n(1-r^2)}\,,$$

and dividing both sides by r^2 proves Theorem 1. □

PROOF OF THEOREM 2: We compare two estimators of the treatment effect β. The first is the usual ITT estimator of the population mean treatment effect

$$\hat{\beta} = \frac{1}{\sum_{i=1}^n X_i} \sum_{i:X_i=1} Y_i \;-\; \frac{1}{\sum_{i=1}^n (1-X_i)} \sum_{i:X_i=0} Y_i \tag{12}$$

which is the difference between the mean of the dependent variable in the treatment group and the mean of the dependent variable in the control group. The second is the stratified estimator

$$\hat{\beta}' = \frac{\sum_{i=1}^n Z_i}{n}\beta_{Z=1} + \frac{\sum_{i=1}^n (1-Z_i)}{n}\beta_{Z=0} \tag{13}$$

where

$$\beta_{Z=1} = \frac{1}{\sum_{i=1}^n X_i Z_i} \sum_{i:X_i=Z_i=1} Y_i \;-\; \frac{1}{\sum_{i=1}^n (1-X_i) Z_i} \sum_{i:X_i=0,Z_i=1} Y_i$$

is the ITT estimator for the subpopulation for whom $Z_i = 1$ and

$$\beta_{Z=0} = \frac{1}{\sum_{i=1}^n X_i (1-Z_i)} \sum_{i:X_i=1,Z_i=0} Y_i \;-\; \frac{1}{\sum_{i=1}^n (1-X_i)(1-Z_i)} \sum_{i:X_i=Z_i=0} Y_i$$

is the ITT estimator for the subpopulation for whom $Z_i = 0$.

Break the population into four groups depending on the value of the covariate and treatment variables. Denote these groups by $S_{u,v}$ where u is the value of the treatment variable and v is the value of the covariate, and denote the sizes of the four groups by

$$\begin{aligned} a &= \#S_{00} = \#\{i : X_i = Z_i = 0\} \\ b &= \#S_{01} = \#\{i : X_i = 0, Z_i = 1\} \\ c &= \#S_{10} = \#\{i : X_i = 1, Z_i = 0\} \\ d &= \#S_{11} = \#\{i : X_i = Z_i = 1\} \end{aligned}$$

. Both estimators are linear in the observed values of the dependent variables, with weights depending on the group to which the subject belongs. For $\hat{\beta}$ the weights of the variables

by group are $w_i = -1/(a+b)$ for $i \in S_{00} \cup S_{01}$ and $w_i = 1/(c+d)$ for $i \in S_{10} \cup S_{11}$. For $\hat{\beta}'$ the weights w_i' are given by

$$\begin{aligned} w_i' &= -\frac{1}{a}\left(\frac{a+c}{a+b+c+d}\right) \quad \text{for} \quad i \in S_{00} \\ w_i' &= -\frac{1}{b}\left(\frac{b+d}{a+b+c+d}\right) \quad \text{for} \quad i \in S_{01} \\ w_i' &= \frac{1}{c}\left(\frac{a+c}{a+b+c+d}\right) \quad \text{for} \quad i \in S_{10} \\ w_i' &= \frac{1}{d}\left(\frac{b+d}{a+b+c+d}\right) \quad \text{for} \quad i \in S_{11}\,. \end{aligned}$$

Thus

$$\begin{aligned} \hat{\beta} &= \sum_i w_i Y_i\,; \\ \hat{\beta}' &= \sum_i w_i' Y_i\,. \end{aligned}$$

and $MSE(\hat{\beta}) > MSE(\hat{\beta}')$ if and only if

$$\gamma^2 > \frac{\theta^2}{n}\frac{[ab(a+b)+cd(c+d)](a+b)(c+d)}{abcd(a+b+c+d)}\,. \tag{14}$$

Now it is just a matter of plugging the definition of Y into the weighted sums and simplifying. Thus,

$$\begin{aligned} \hat{\beta} &= \sum_i w_i Y_i \\ &= \frac{-1}{a+b}\sum_{i\in S_{00}} \mu + \theta U_i \;+\; \frac{-1}{a+b}\sum_{i\in S_{01}} \mu + \theta U_i + \gamma \\ &\quad + \frac{1}{c+d}\sum_{i\in S_{10}} \mu + \beta + \theta U_i \;+\; \frac{1}{c+d}\sum_{i\in S_{11}} \mu + \beta + \theta U_i + \gamma \\ &= \beta + \left(\frac{d}{c+d} - \frac{b}{a+b}\right)\gamma + \frac{1}{a+b}\sum_{i\in S_{00}\cup S_{01}} \theta U_i + \frac{1}{c+d}\sum_{i\in S_{10}\cup S_{11}} \theta U_i \end{aligned}$$

leading to

$$\text{MSE}(\hat{\beta}) = \left(\frac{d}{c+d} - \frac{b}{a+b}\right)^2 \gamma^2 + \theta^2\left(\frac{1}{a+b} + \frac{1}{c+d}\right)\,. \tag{15}$$

Similarly,

$$\begin{aligned}
\hat{\beta}' &= \sum_i w_i' Y_i \\
&= -\frac{a+c}{an} \sum_{i \in S_{00}} \mu + \theta U_i \; - \; \frac{b+d}{bn} \sum_{i \in S_{01}} \mu + \theta U_i + \gamma \\
&\quad + \frac{a+c}{cn} \sum_{i \in S_{10}} \mu + \beta + \theta U_i \; + \; \frac{b+d}{dn} \sum_{i \in S_{11}} \mu + \beta + \theta U_i + \gamma \\
&= \beta + \sum_i w_i' \theta U_i
\end{aligned}$$

leading to

$$\mathrm{MSE}(\hat{\beta}') = \theta^2 \left(\frac{(a+c)^2}{an^2} + \frac{(b+d)^2}{bn^2} + \frac{(a+c)^2}{cn^2} + \frac{(b+d)^2}{dn^2} \right) . \tag{16}$$

Subtracting, we see that $\mathrm{MSE}(\hat{\beta}) > \mathrm{MSE}(\hat{\beta}')$ if and only if

$$\gamma^2 \left(\frac{d}{c+d} - \frac{b}{a+b} \right)^2 \; > \; \theta^2 \left[\frac{(a+c)^2}{n^2} \left(\frac{1}{a} + \frac{1}{c} \right) + \frac{(b+d)^2}{n^2} \left(\frac{1}{b} + \frac{1}{d} \right) - \left(\frac{1}{a+b} + \frac{1}{c+d} \right) \right] .$$

Dividing through by the coefficient of γ^2 gives

$$\gamma^2 \; > \; \frac{\theta^2}{n} \left[\frac{a+c}{n} \, \frac{(a+c)^2}{ac} \; + \; \frac{b+d}{n} \frac{(b+d)^2}{bd} \; - \; \frac{n}{a+b} - \frac{n}{c+d} \right] \; \frac{(c+d)^2 (a+b)^2}{(ad-bc)^2}$$

which simplifies, after cancellation of $(ad-bc)^2$, to (14). □

Proof of Proposition 3: By rescaling, we can assume without loss of generality that (Y_j, Z_j) are normal with covariance matrix $\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$. We may then write $Z_j = \alpha Y_j + \sqrt{1-\alpha^2} W_j$ where W_j are standard normals, independent of Y_j's and of each other. Equation (8) then follows from applying the stated limits of integration to the bivariate standard normal density. □