The Perils of Randomization Checks in the Analysis of Experiments

Diana Mutz [1]

Robin Pemantle [2,3]

[1]Samuel Stouffer Professor of Political Science and Communication, University of Pennsylvania, 208 South 37th Street, Philadelphia, PA 19104, mutz@sas.upenn.edu

[2]Research supported in part by National Science Foundation grant # DMS 0905937

[3]University of Pennsylvania, Department of Mathematics, 209 S. 33rd Street, Philadelphia, PA 19104, pemantle@math.upenn.edu

The Perils of Randomization Checks in the Analysis of Experiments

ABSTRACT:

In the analysis of experimental data, randomization checks, also known as balance tests, are used to indicate whether a randomization has produced balance on various characteristics across experimental conditions. Randomization checks are popular in many fields although their merits have yet to be established. The grounds on which balance tests are generally justified include either 1) the credibility of experimental findings, and/or 2) the efficiency of the statistical model. We show that balance tests cannot improve either credibility or efficiency. The most common "remedy" resulting from a failed balance test is the inclusion as a covariate of a variable failing the test; this practice cannot improve the choice of statistical model. Other commonly suggested responses to failed balance tests such as post-stratification or re-randomization also fail to improve on methods that do not require balance tests. We advocate resisting reviewer requests for randomization checks in all but some narrowly defined circumstances.

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

When executing experimental designs, it has become common practice to run randomization checks or balance tests as part of the analysis of results[1]. The exact origins of this practice are unclear, but such tests are nonetheless widespread. As we will argue, they are often motivated by faulty assumptions. It would be one thing if they were harmless, but unnecessary excesses. Unfortunately, these practices often lead to inferior model choices in the analysis of findings and unjustified interpretations of findings. For these reasons, we discourage their use except under highly specific circumstances.

To establish the parameters of our argument, we begin by specifying some circumstances under which balance tests make sense. The remainder of the paper details the problems caused by the use of balance tests in all other cases. We address the argument that the credibility of experimental findings is enhanced by the use of balance tests. Next we show why failed balance tests do not tell us anything that can improve the efficiency of model choice when analyzing experimental results. In addition to the lack of benefits from balance testing, there is a downside. In the final section we discuss potential negative effects of balance testing on both the credibility and efficiency of experimental findings. We conclude with speculation on the origins of this problematic practice.

Scope

Our argument against the use of balance tests includes the majority of ways in which they are used in political science experiments. It does not, however, condemn all

[1]We use the terms "randomization check" and "balance test" interchangeably throughout the paper.

possible uses of these tests. When referring to "experiments" we assume 1) that the researcher has control over assignment to experimental conditions and can ensure exposure to treatment, and 2) that respondents do not undergo attrition differentially *as a result of* assignment to a specific experimental condition. If either of these conditions is not met, then a balance test may be warranted. For example, in a laboratory study, if some kinds of respondents were more likely to drop out of an experiment in progress after exposure to an unpleasant experimental treatment, then there would be good reason for concern about the comparability of treatment and control groups. It is important to note that attrition itself is not a concern and occurs in all kind of studies, experimental and otherwise. In order to be problematic, the non-equivalence across groups must result from an interaction between the treatment to which the respondent is assigned in combination with characteristics of the individual subject. For many types of experimental designs, this is not a serious concern.

If experiments are well designed, that is, if protocols are designed so that no experimental condition is shorter or longer than any other, or more or less likely to retain respondents than any other, then randomization accomplishes its mission. Regardless of whether it is executed using a random number table, using computer-generated random numbers such as those generated by Excel, or via programming code as is common in online experiments, the investigator should have the information necessary to assess threats to the randomization procedure without ever looking at the results of the study.

Similar considerations arise in field experiments. If researchers assign some homes to receive treatments (such as campaign brochures) but the researchers cannot ensure

that respondents in the home are actually exposed to the experimental treatment, then it is possible that more politically interested respondents would be more exposed than the politically uninterested. If the researcher compares the groups based on *intent to treat* rather than on actual treatment, statistical comparisons will be valid (although effect sizes will be underestimated). But if the researcher compares those who actually looked at the brochure to a control group, then there are obvious reasons for concern about balance. In summary, balance tests serve a useful purpose when assessing threats to the execution of the randomization protocol.

Aside from differential exposure or noncompliance, the only additional scenario in which a randomization check makes sense is when the randomization mechanism itself may not be working properly. As noted in Bowers (2009, Section 4.1.1), prior works that are critical of the practice of balance testing (Senn, 1994; Imai et al., 2008) specifically exclude cases where the mechanism may be faulty. While cases of faulty programming such as in the 1997 National Election Study may be rare, there are other examples in which faulty assumptions lead to randomization failures. For example, experiments in which randomization at the individual level is only approximate because the treatment is implemented at a different level of analysis can lead to problems (see Imai 2005; Gerber and Green 2000; Hansen and Bowers 2008).

With the exceptions noted above, randomization checks lead researchers down a misguided and fruitless path. They are irrelevant to the credibility of experimental results and have no capacity to improve efficiency. Testing for demographic or other differences between experimental groups is misguided because a lack of differences on whatever variables happen to have been tested does not mean it was "successful" any

more than finding a difference means that it was a "failure".

Credibility of Experimental Findings

Reviewers of experimental studies routinely request that authors provide randomization checks, that is, statistical tests designed to substantiate the equivalence of experimental conditions on one or more factors measured before treatment. The logic behind such requests is to examine whether random assignment to experimental conditions has "succeeded" in producing comparable experimental conditions on characteristics measured before the experimental treatment was administered. But what exactly does it mean for a randomization to "succeed"? The purpose of such tests is to reassure readers that they can trust the internal validity of a given experiment's results. Researchers compare experimental groups on variables that are not part of the central theoretical framework of the study. Often these are demographics, but frequently they include other variables as well.

The point of this exercise is not to test whether there is an error in the mechanism for random assignment. Instead, it is to convince one's audience that this particular randomization did not happen to be one of those "unlucky" draws wherein some measured variable is unequal across conditions. By comparing means across conditions for one or more variables, the investigator supports his or her assertion about the pre-treatment equivalence of experimental groups.

Unfortunately, this practice and the conclusions drawn from it are problematic. There is a temptation to run comparisons on a large number or at least a handful of

auxiliary variables, just to see if anything comes up significant. This is a statistically misguided idea in several respects. The problem is not so much the number of comparisons as the logical basis for doing them to begin with. The use of randomization checks for this purpose demonstrates a fundamental misunderstanding of what random assignment does and does not accomplish. A well-executed random assignment to experimental conditions does not promise to make experimental groups equal on all possible dimensions or on any one characteristic, or even a specified subset of them. Across many independent randomizations this is very likely to be the case, but not for any given randomization.

Doesn't this lack of across-the-board equivalence pose problems for drawing strong causal inferences? Contrary to popular belief, it does not. This idea was the single fundamental scientific contribution of R. A. Fisher. It is not necessary for experimental conditions to be identical in all possible respects (Thye, 2007). Psychologist Robert Abelson (Abelson, 1995) dubs the practice of testing for differences between experimental groups a "silly significance test." As he explains, "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, p. 1716) likewise calls the practice of performing randomization tests "philosophically unsound, of no practical value, and potentially misleading." In the context of political science, Imai et al. (2008) echo this sentiment, "Any other purpose [than to test the randomization mechanism] for conducting such a test is fallacious," and they go on to cite works in respected journals in economics, political science, sociology, psychology, education, management science, medicine, public health and statistics which succumb to this fallacy. A number of others have put forth some of the same arguments (Altman,

1985; Permutt, 1990). Even those in favor of balance testing in some circumstances (Begg, 1990) deplore the arbitrary and misleading practice of testing for significance as a means of implementing balance tests.

But then what should one conclude if one finds a pre-treatment difference in some characteristic that might be relevant to the outcome of the experiment? Does that mean that the randomization essentially "failed"? When findings indicate no significant differences on the variables chosen for balance tests, the randomizations are claimed to be "successful," so this inference seems logical. But both assertions would be wrong or, at the very least, misguided. Assuming there is no technical problem in the software that does random assignment, no errors in the random number table that is used by the research assistant, or some other concrete procedural glitch, random assignment is successful by definition, so long as it is executed correctly. Our point is not that random assignment always functions to make groups perfectly comparable (because this is demonstrably not the case). But perfect comparability is not necessary, and a lack of it is not a legitimate justification for conducting randomization checks, particularly given that the statistical significance attached to such a test is not meaningful.

Of course, if one does enough comparisons, one is bound to find a significant difference on one of them by chance alone. Some scholars have suggested procedures for combining multiple tests into one grand test (Bowers, 2009; Hansen and Bowers, 2008), thereby avoiding this problem. These procedures solve the problem of conducting too many tests, but they miss the larger point: statistical tests used to analyze experiments already take the possibility of nonequivalent conditions into account. Al-

though it is true that one can never totally eliminate the possibility of differences in experimental group composition, that probability has already been incorporated into the statistical tests used to test the null hypothesis of no difference between experimental groups. To reiterate, the infamous "$p < .05$" that tests the null hypothesis already includes the probability that randomization might have produced an unlikely result, even before the treatment was administered. So, while there is no guarantee of avoiding an unlikely fluke result (short of replication, which should happen in any case), by conducting the standard test and reporting it correctly, researchers have already done due diligence.

Thus far the points we have made are not original; they have been made by others from time to time in the statistical literature, although they have been largely ignored in practice. Instead, the contrary view, that balance tests are essential to the credibility of findings, persists. For example, as Hansen and Bowers (2008, page 13) suggest, "Comparative studies typically present a small number of covariates that must be balanced in order for the study to be convincing, along with a longer list of variables on which balance would be advantageous." In purely experimental studies, a simple hypothesis test with, say, $p < 0.001$ should be considered very convincing without a balance test[2].

[2]Those who consider balance essential might consider sequential randomization procedures such as those in (Pocock et al., 2002) which produce a blocking-type balance but over many variables simultaenously.

Efficiency of Statistical Models

If credibility is not a compelling reason to execute balance checks, a second possibility is that balance checks are useful for gaining efficiency in experimental analyses. Knowledge of an imbalance on some variables may promote efficiency if the analysis takes this knowledge into account. This might be accomplished by the addition of covariates, by post-stratification or by rejecting the randomization and re-randomizing. We consider each of these possibilities in turn.

Covariates

The inclusion of covariates is well known to improve efficiency so long as the covariates predict variance in the dependent variable. In fact, one may compute a threshold: an amount of variance in the dependent variable that a covariate must predict in order for its inclusion in the model to yield an increase, rather than a decrease, in efficiency. This threshold, in less precise language, has appeared elsewhere (e.g., Franklin 1991). The reasons behind this are quite intuitive. The data is the sum of a signal and some noise. An experimental effect can be detected only if the signal is stronger than the noise. Subtracting some of the noise that is present makes the signal more visible.

On the other hand, adding or subtracting new noise only obscures the signal further. Adjusting for a covariate subtracts a linear estimate of the portion of the noise due to the covariate. What is subtracted is therefore some portion of the noise due to this covariate. But there is new noise introduced as well, namely the noise in the linear estimate. A covariate improves efficiency only if it subtracts more old noise

than new noise. If the new noise is held fixed, one sees a threshold in the amount of old noise that must be subtracted in order to produce a gain in efficiency. In the limiting case, when the covariate does not explain variance in the dependent variable, there is no old noise and the effect of linear adjustment is entirely the introduction of new noise.

Given these considerations, are covariates an appropriate response to failed balance tests? Covariates should be selected because they are expected to explain variance in the *dependent* variable, whereas a failed balance tests indicate a relationship between the covariate and the *independent* variable. To date, a glance through the literature reveals considerable divergence of opinion on the issue of "correcting" for imbalance in experimental designs. There is a mathematical result, a version of which is proved in Feldstein (1973) but which appears in other places (e.g., Franklin 1991), which should put to rest the issue of whether a balance test can help to identify a covariate whose inclusion in the model would be advantageous.

> Let t denote the threshold for the fraction of the variance of the dependent variable that the covariate must predict in order that including the covariate increases efficiency rather than decreasing it. Then, conditional on the values of the covariate, the proportion t is an increasing function of r^2, the squared empirical correlation between the covariate and the independent variable.

One may interpret this as follows. We have a number of potential covariates measured prior to the experimental treatment. In deciding which to include in the analysis we have the option to see whether the randomization distributes the values of these

covariates evenly among the treatment and control groups. We may decide, for example, to include only those covariates for which there is a significant imbalance. Alternatively, we may include only those covariates known from previous research or suspected for theoretical reasons to be strong predictors of the dependent variable. A third strategy would combine these: including known predictors of the outcome together with any other measured covariates that happen to be assigned to treatment groups in an imbalanced way.

The above result tells researchers to follow the second strategy. When a covariate fails a randomization test, r^2 increases, hence the threshold for inclusion goes up rather than down. In other words, if inclusion of this variable as a covariate in the model will increase the efficiency of an analysis, then it would have done so, and to a slightly greater extent, had it *not* failed the balance test. This renders the balance test uninformative when it comes to the selection of covariates.

To understand at a purely conceptual level why balance tests do not help in selection of covariates, it is useful to think about why this particular variable has been chosen for inclusion in the analysis to begin with. In the case of balance testing, a variable is included strictly because of its significant relationship with the independent variable. Thus when both variables are used simultaneously as predictors of the dependent variable, some of the variance in the dependent variable that would have been attributed to the experimental treatment will be attributed to the covariate instead, just as collinear variables in observational data may fight over the same variance in the dependent variable.

Whether this adjustment is helpful or harmful depends on the proportion of shifted

attribution that is meaningless noise and is therefore a harmful correction. The more collinearity between the covariate and the treatment variable, i.e., the more imbalance, the more meaningless noise will show up in the adjustment to the estimate of treatment effect. This means the portion of the variance in the dependent variable predicted by the covariate must be higher in order for the helpful part of the adjustment to outweigh the harmful part. If a variable was deemed insufficiently predictive to include in the analysis before a balance test was done, then including it in the analysis after a "failed" balance test cannot help and can, as we describe shortly, undermine the integrity of the analysis.

The more general question of whether to include a given covariate is complicated and we are not taking a position on this broader issue. Factors that go into this include: prior knowledge of the relation of the covariates to the dependent variable, whether maximal efficiency will be needed to obtain significance, desire to keep the findings transparent and interpretable, or to give an assurance that the reported significance is for a single analysis and not a large number of unreported analyses. Our main result is comparative: "failed" randomization with respect to a covariate should not lead a researcher to include that covariate in the model. If the researcher plans to include a covariate for the sake of efficiency, it should be included in the model regardless of the outcome of a balance test.

Post-stratification

If not as an aid in the selection of covariates, could balance test be useful in another fashion? The term *post-stratification* has been used in the literature to refer to a

collection of practices meant to enforce balance after the data has been collected. The "strata" are typically subpopulations corresponding to different values of one or more covariates. The classic work by Holt and Smith (1979) uses the term post-stratification to mean re-weighting (see also Little 1993). The main use of weighting is in extrapolating experimental results to the general population. However, weighting can also be used to force the composition of treatment groups and control groups to coincide. For example, if the treatment group has 27 males and 23 females while the control group has 20 males and 30 females, weighting females more heavily in the treatment group and weighting males more heavily in the control group can equalize the compositions of the groups. If the design did not block on gender to begin with, this is the next best way to force exact comparability of treatment and control groups along the lines of gender. As a means of increasing efficiency, however, there is no demonstrable gain. If one wishes one had blocked on gender from the start, then inclusion of gender as a covariate is equally effective as post-stratifying on gender and the results will be more transparent.

Post-stratification can refer more generally to any alteration of the analysis that gets around imbalance of strata in treatment versus control groups. The logical extreme is a subpopulation analysis, breaking the data set into groups depending on the values of imbalanced covariates. Hansen and Bowers (2008, page 14), for instance, suggest that "such imbalances can be remedied by post-stratification: if treatments are on the whole older than controls, then compare older treatments only to older controls, and also compare younger subjects only amongst themselves."

Subpopulation analyses are often worth carrying out because of the possibility

of heterogeneity across subpopulations in the size of the treatment effect. As with covariates, however, subpopulation analyses should be carried out when there is an underlying theoretical reason to do so, not because of a failed balance test. This is not to say that there is anything wrong with running subgroup analyses or more generally playing with one's data. These practices can certainly lead to conceptual advances, though what significance levels should be attached to data explorations is not clear. What is clear is that this kind of post-hoc stratification is not a logical or useful response to a failed balance test. Modulation of the treatment effect by subpopulation membership is not in any way related to the balance or lack of balance of the subpopulation over the random assignment.

Re-randomization

If balance tests are not useful toward informing covariate selection nor toward deciding on a post-stratification analysis, they might still be useful in signaling when to re-randomize. Re-randomizing in response to a failed balance test may or may not be an option, depending on the type of experiment. If subjects arrive one at a time for a laboratory experiment, and assignment to experimental condition is made on the spot, a failed balance test will be evident only at the end when it is too late to start the randomization afresh. If a given study allows randomization to be performed ahead of time *and* the researcher has access in advance to the set of variables on which balance is desired, then it is possible to check for balance before treatments are administered and to re-do the randomization if the balance test fails. In other words, if the researcher desires balance on, say, gender, race and party, then it is

possible to generate a random assignment, test for significant imbalance on these three variables, accept the randomization if all three are balanced, and if not, re-do the randomization, possibly more than once, until balance is achieved on these three variables.

Such a scheme of sampling with rejection is mathematically equivalent to generating a single randomization that is conditional on meeting certain criteria, such as exact or approximate equality of treatment and control groups on specified variables. Blocking, which is done in the design phase of an experiment, is the logical extreme of rejection sampling (Imai et al., 2008). Blocking is an effective way of increasing efficiency and "is almost always preferable whenever feasible" (Imai et al., 2008, p. 489). Blocking is not feasible on more than a few variables, but if one only requires approximate balance, then one might use a sampling scheme in which samples are rejected until the desired degree of balance across conditions is achieved (see Lock 2011).

It is indeed possible to benefit from such an approach, but there are a number of pitfalls as well. First, the reported significance level must be adjusted to take into account the precise rejection sampling scheme[3]. Secondly, one must often make do with crudely approximated test statistic distributions and p-values determined by simulations, because the mathematics of most rejection sampling schemes has not yet been worked out. Unfortunately, it is common practice in some subfields to re-randomize for balance without ever reporting that this has taken place. Significance levels are reported as if this were the one and only randomization. Because the procedures to

[3]This is true even if the particular randomization was not rejected! This is because p-values are a function of what might have happened as well as what did happen. This is one way in which frequentist statistics is at odds with the likelihood principle.

compute the correct significance levels for protocols with possible re-randomization are not contained in any statistical package, most experimentalists make no effort to take rejection into account when reporting significance levels. A promising avenue for methodological research is the extent to which surrogate computations from standard packages might be available.

Successful use of re-randomization supposes a pre-specified balance test. When the rejection for imbalance is done in an *ad hoc* manner, e.g., by deciding after the fact whether each unbalanced variable is important, it is not possible to resurrect a valid significance statement. Provided the protocol for re-randomization is well specified, there is still a question as to whether re-randomization is really the most powerful design. An alternative is to include the variables that were to be balanced as covariates from the start. Using covariates will often lead to less noise than rejection sampling (Permutt, 1990). No randomization scheme has been shown to improve on simply including the important covariates regardless of the outcome of a balance test. Most importantly, as a practical matter, relatively few experimental designs allow one to gather the necessary covariates from all subjects in advance and then check balance before executing the experiment.

Hazards of Responding to Balance Tests

Thus far, our case against balance tests is that there is no sound reason to do them, and no valid response once one has failed. But is there a reason not to do them? After all, when a manuscript referee asks for a balance test, it is expedient to comply, and

absent any harmful fallout, a researcher would be tempted to "just do it". Potential losses in efficiency due to inclusion of useless covariates are often quite small and the balance test is likely to be negative anyway, so the potential harm appears negligible.

There is, however, a price to pay, manifested in several ways. Leaving aside the question of which balance test to do (there is little agreement) and how many and which variables to test, there are several hazards. These hazards, which do not affect analyses based on straightforward comparisons of means, include 1) model mis-specification, 2) the perceived credibility of the results, and 3) more complex calculations of significance levels.

Model mis-specification is perhaps the most significant drawback of the use of covariates in general. Multiple regression and analysis of covariance are less robust against model mis-specification than is bivariate linear regression or analysis of variance. Stunningly, this fact appears to be overlooked by the majority of experimental researchers in the social sciences. To elaborate, in the experimental setting, bivariate linear regression (or analysis of variance) always produces a valid estimate of the treatment effect and a valid confidence test for nonzero treatment effect. The only necessary assumption is that the randomization was as advertised, that is, produced by a valid source of randomness. Multivariate models, on the other hand, involve explicit assumptions about dependence between variables. Each variable is assumed to contribute to the dependent variable in a linear manner, and the effects of different covariates are assumed to be additive. One can add extra variables, such as squares or products, that correct for nonlinearities or non-additive interactions, but the required assumptions of linearity and additivity persist in the enlarged set of variables. When

these assumptions are violated, estimates produced by multivariate models are biased and inferences and confidence statements become invalid. (A brief example is noted by Freedman 2008, Example 5.)

Using covariates with experimental data is tempting because this may produce a more accurate estimate of a treatment effect or a more sensitive test for differing means. Indeed, its use when one believes a handful of covariates will have strong and more or less linear effects on the dependent variable often allows identification of effects that would otherwise remain hidden. But these statistical models may also be invalidated by problems that cannot arise in a simpler model. Green (2009) argues that the biases and inaccuracies are often small. There is some evidence that a simple multivariate model involving one treatment and one control condition, with a single covariate and data that are not obviously bimodal, will not be off by too much. The reader is left to judge whether an analysis that is likely just a little wrong is better than one that is demonstrably right. To date nothing is known about the extent of risk due to mis-specification for the most common approach to analyzing political science experiments, namely large multiple regression equations involving many covariates, combined with complex, multi-factor experimental treatments.

The second potential hazard of "just doing it" is the lack of transparency in the presentation and interpretation of findings. In a simple linear regression, it is quite clear what is being estimated and why the estimate is what it is; the only independent variables are the ones representing experimental conditions. In a multiple regression, one estimates the net effect of treatment minus a linear function of the covariates, this function itself depending on an estimate of some coefficients in an equation assumed

to be linear. While the estimate produced by this process could possibly be more accurate, our relative unease with the practice rests on a very real narrowing of the validity of the model. Balance tests, to the extent that they result in models with extra assumptions, work against transparency and credibility, not in favor of them.

Ironically, many reviewers respond to models that include more covariates as if they were inherently more robust than models without covariates. This is demonstrably not the case. This mistake on the part of reviewers stems from a misunderstanding of how experiments work. The sheer fact that covariates in experiments are often referred to as "controls" reveals the nature of this misunderstanding. Whereas more control variables in an observational analysis might strengthen confidence that an association is not spurious, in experiments the presence of covariates should lead to greater scrutiny.

A second issue concerning transparency occurs when the model is altered based on analyses involving the dependent variable. Statistical inference is never valid conditional on post-treatment measures, meaning that one cannot alter the model based on any analyses of the dependent variable. A researcher cannot, for example, decide to include a covariate because it is a good predictor of the dependent variable in the present data set. Limited inference may still be available in this case, but it is very weak and the rigorous framework for such inference is in its infancy (Berk et al., 2010).

Because this is well known, analyses conditioned on the dependent variable are never presented as such. However, when an article is submitted for publication, it is not usually possible to tell precisely how the model was chosen. Some studies, such as

FDA drug trials, require an advance specification of design and analysis. This practice is not the norm in social science research. Therefore, to avoid the appearance of so-called data snooping, it is incumbent on the researcher to choose a model that is readily identified as the natural model for the dependent variable in question. Covariates should be selected because previous research or theoretical reasoning indicates an expected relation with the dependent variable. Short of advance specification, this is the surest way to avoid the appearance of having chosen the model based on post-treatment data.

A related problem with credibility occurs when multiple analyses of the experimental hypothesis are performed or might have been performed. Running many analyses, each with a different selection of covariates, and then reporting the most significant one is a form of choosing the model based on post-treatment measures. In the worst case, running k different analyses distorts the p-value by a factor of k. This would be true if the analyses were all independent. Because they are not independent, the actual factor is much less than k. Because of this, or perhaps because the actual correction to the p-value is unknown, the practice of suppressing inconsequential analyses and using the single-analysis p-value as if it were correct is commonplace. This practice always errs in the direction of over-reporting significance. Solutions that have been advocated include automated model selection, a partition of the data into a half that informs model selection and a half on which the final analysis is performed, and adherence to parsimonious models that do not have the complexity to be the result of a fishing expedition, at least not a large one (see, e.g., Tsiatis et al. 2007).

The final issue concerning inaccurate reporting of significance is subtle. Significance levels for sequential analyses are not equal to the significance level for the resulting simple analysis. For example, suppose the researcher decides to run an analysis with or without a covariate X depending on the distribution of that variable across treatment and control groups. If p_1 is the p-value for the analysis for the simple test that always excludes X and p_2 is the p-value for the test that always includes X as a covariate, then the significance level of the two-stage analysis – that is, first running the balance test, then running the analysis with or without the covariate, as dictated by the balance test – is not given by p_1 if X is excluded and p_2 if X is included. In fact, as demonstrated for a single-covariate multivariate normal model by Permutt (1990), the resulting significance can be less than either p_1 or p_2. Statistical software is not set up to compute significance in this kind of sequential analysis. This leads to errors in confidence levels, again in the direction of inflated significance.

Compromises have been proposed such as reporting the result of a balance test but not using the result to alter the analysis. But then it is extermely puzzling for the reader to understand why it is included. If it is not relevant to the model choice, the credibility of the findings or to the efficiency of the model, then why do it?

When a researcher initiates balance testing, he or she should understand the purpose of doing so and be prepared to live with the consequences. If a balance test results in doubt as to whether randomization was carried out properly, then the problem cannot be fixed by adding covariates to the analysis. At best some inference might be salvaged by re-analyzing the data as an observational study.

In practice, it is clear from what usually happens after a failed balance test that

the randomization procedure itself is not being deemed faulty by the author or the reviewers. Once an analysis has gone to peer review, we have seldom if ever seen a failed balance test lead to the response that the data will no longer be treated as experimental data, as would have to be the case if the randomization were believed to be flawed. Either the cases were randomly assigned or they were not; there is no middle ground.

Although the central point of this paper is statistical rather than historical, upon reaching our conclusions we found ourselves puzzled as to the origins of this practice. If the inclusion of variables as covariates based on failed randomization tests is not useful toward increasing the credibility of the findings nor the efficiency of the analysis, except in the contexts we have outlined, then a natural question to ask is why this has become such standard practice. We can find no one source that explicitly recommends this practice or documents its utility, thus we can only speculate as to its likely origins. To be clear, a number of sources appear to advocate balance tests for randomized experiments, but either they do so for general reasons (e.g., asserting greater credibility without justification), or upon closer inspection the recommendation is limited to situations falling outside the scope of the fully randomized experiments considered here.

One possibility is that it stems from a lack of faith in or thorough understanding of probability theory. A related possibility is confusion between frequentist and Bayesian paradigms. It seems intuitive to adopt the (Bayesian) likelihood principle, using a second look at the data to refine (frequentist) statements of significance. Attempts have been made to blend these two frameworks: first (as a frequentist) check whether

the findings are significant at say a $p < 0.05$ level, and second (as a Bayesian) if they are, bolster our confidence by assessing further the likelihood that the confounding random event of probability less than 0.05 has, in fact, occurred. Unfortunately, to date, no rigorous statistical practice along these lines improves upon traditional significance computations.

A second possibility is that this practice results from mistakenly applying methods for observational analyses to experimental results. Researchers new to experimental analysis often try to extend methods from a more familiar setting, namely that of observational survey data. These two methods require different analysis practices. Nonetheless, authors of experimental studies are often encouraged to run multivariate regressions including "control" variables such as a laundry list of demographic characteristics. The very notion of a "control" variable makes no sense in the context of an experiment.

The rise of large-N population-based survey experiments further complicates matters, as researchers have what is essentially hybrid data to analyze, often including hundreds of potential variables on which experimental conditions might be unbalanced. Such approaches encourage further confusion regarding best practices for analysis. In practice, most such studies should be analyzed as experiments (see Mutz 2011).

Yet another possibility is that the popularity of this practice comes from the field experimental literature in which treatment cannot always be controlled as well as it can be in purely experimental studies. As we describe when setting the parameters for our argument, balance tests may be useful if characteristics of individuals influence

their likelihood of experiencing treatment, or when differential attrition may occur based on individual characteristics.

Whatever the origins of this practice, researchers today overuse balance tests to achieve ends that these tests are not capable of accomplishing. The conclusion for our purposes is that the integrity of one's findings is increased by choosing covariates (or randomization schemes, stratification of population, etc.) according to a scheme that is specified in advance and parsimonious. Throwing in more covariates because they "can't hurt", aside from losing efficiency to a small degree, sacrifices integrity in the eyes of readers and reviewers, as well as intrinsic accuracy and validity.

References

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

Altman, D. (1985). Comparability of randomised groups. *J. Royal Statist. Soc., ser. D*, 34:125–136.

Begg, C. (1990). Significance tests of covariate imbalance in clinical trials. *Controlled Clinical Trials*, 11:223–225.

Berk, R., Brown, L., and Zhao, L. (2010). Statistical inference after model selection. *J. Quant. Criminol.*, 26:217–236.

Bowers, J. (2009). Making effects manifest in randomized experiments. *Preprint*.

Feldstein, M. (1973). Multicollinearity and the MSE of alternative estimators. *Econometrica*, 41:337–346.

Franklin, C. (1991). Efficient estimation in experiments. *Pol. Methodologist*, 4:13–15.

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

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

Green, D. (July, 2009). Regression adjustments to experimental data: do David Freedman's concerns apply to Political Science? *Paper presented to the 26th annual meeting of the Society for Political Methodology.*

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

Holt, D. and Smith, T. (1979). Post stratification. *J. Royal. Stat. Soc. Ser. A*, 142:33–46.

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.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *J. Royal. Statist. Soc., ser. A*, 171, part 2:481–502.

Little, R. J. A. (1993). Post-stratification: A modeler's perspective. *J. Amer. Stat. Assoc.*, 88:1001–1012.

Lock, K. (2011). *Rerandomization to improve covariate balance in randomized experiments*. PhD thesis, Harvard University.

Mutz, D. C. (2011). *Population-based Survey Experiments*. Princeton University Press, Princeton, NJ.

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

Pocock, S., Assmann, S., Enos, L., and Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:2917–2930.

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

Thye, S. (2007). Logic and philosophical foundations of experimental research in the social sciences. In *Laboratory Experiments in the Social Sciences*, pages 57–86. Academic Press, Burlington, MA.

Tsiatis, A., Davidian, M., Zhang, M., and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, 27:4658–4677.