Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee that Prepared the Reporting Guidelines

Alan S. Gerber, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling and D. Sunshine Hillygus

Journal of Experimental Political Science / Volume 2 / Issue 02 / December 2015, pp 216 - 229
DOI: 10.1017/XPS.2015.20, Published online: 12 January 2016

Journal of Experimental Political Science **2** (2015) 216–229
doi:10.1017/XPS.2015.20

Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee that Prepared the Reporting Guidelines

Alan S. Gerber*, Kevin Arceneaux†, Cheryl Boudreau‡, Conor Dowling§ and D. Sunshine Hillygus¶

INTRODUCTION

We welcome the comments on our committee's Reporting Guidelines (2014, *Journal of Experimental Political Science* 7 1(1): 81–98) from Diana Mutz and Robin Pemantle, as well as the opportunity to clarify our recommendations. We appreciate the points they raise and share their goal of encouraging a better understanding of experimental methods. Nonetheless, we are not in complete agreement with their proposed revisions to our recommendations.

Mutz and Pemantle (henceforth, MP) discuss and critique a broad range of common experimental practices. In our relatively brief reply, we focus on the narrow question of the merits of MP's recommendations for revising the Reporting Guidelines. MP's essay discusses three topics with implications for the Reporting Guidelines—manipulation checks, response rates, and pretreatment balance tables—and concludes with four proposed changes (Mutz and Pemantle 2015, 192–215):

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

*Department of Political Science, Yale University, New Haven, CT, USA, e-mail: alan.gerber@yale.edu
†Department of Political Science, Temple University, Philadelphia, PA, USA
‡Department of Political Science, University of California, Davis, Davis, CA, USA
§Department of Political Science, University of Mississippi, University, MS, USA
¶Department of Political Science, Duke University, Durham, NC, USA

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

Before we discuss our reaction to these proposed revisions, we want to emphasize our interpretation of our committee's charge. Our goal was to provide a set of reporting guidelines that would facilitate transparency in experimental research, by which we mean that readers should have a very clear view of what the researcher did and found. With this simple but crucial goal in mind, we prepared reporting guidelines for the American Political Science Association's Experimental Research Section to assist researchers in reporting the information that reviewers and readers would most want to see reported. We emphasize two aspects of our task: (1) that the guidelines are recommendations for clear and thorough reporting of the research procedures used and (2) the guidelines are meant for the community of readers and reviewers. We did not intend, nor do we see it as part of our charge, to offer advice on how researchers *should* design or analyze experiments. Although we have our opinions, we were not appointed to serve in the role of statistics and design authority to the section, and we leave these assessments of how experiments ought to be done to the broader community of scholars. In this same spirit, in our reply we make various statistical arguments to explain the basis for conventional views on such issues as the reporting of balance tables, but these rationales should be viewed as our particular take on things. Further, MP discuss the possible relationship between the Reporting Guidelines and statistical practices they criticize. However, the link, if any, between what appears in the Reporting Guidelines and specific statistical practices, such as those related to model specification, is entirely speculative, and pointing out and correcting allegedly flawed statistical practices is a job for methodologists, instructors, and reviewers. Our position is that if there are sound reasons for the readers to want to see an item, the item should be reported. Regarding the evaluation and improvement of statistical practice, we invite other researchers interested in the important issues MP raise to engage with MP further.

To succinctly summarize our response: Of the four recommendations that MP make, only two—manipulation checks and response rates—clearly relate to the Reporting Guidelines. Regarding manipulation checks, we do not agree with MP's suggestion that the Reporting Guidelines should explicitly recommend that researchers perform manipulation checks. Although MP offer valid reasons for why scholars should consider incorporating manipulation checks into their experimental designs, this recommendation does not fit with the overarching goal of the guidelines, which is to encourage transparency. Consequently, the guidelines request that researchers report the experiments they conduct, rather than recommend a specific research strategy or research priority. If researchers embed a manipulation check

into their design, they should report the results, as the current guidelines already state in Section D. We view it as beyond the scope of the Reporting Guidelines to make recommendations about best practices in experimental design.

Regarding response rates, we do not agree with MP's suggestion that the guidelines should only recommend the reporting of response rates for a narrow class of cases. Many researchers find the response rate to be useful information and there are sound reasons why it is viewed as informative.

MP also present arguments criticizing how political scientists discuss balance tables and randomization testing, as well as many cautions regarding model specification. However, the items that MP label as recommendations 3 and 4 do not actually propose any clear changes to the Reporting Guidelines. Nevertheless, because the Reporting Guidelines do ask researchers to provide balance tables, we will offer further discussion as to why we think providing a balance table is a good idea.

Although we are not in full agreement with MP's recommendations, their comments did prompt us to revisit and modify some features of the Reporting Guidelines. At the end of this response, we provide a link to a revised version of the Reporting Guidelines that incorporates several modifications, available as online supplementary material, as well as a new document, a checklist of reporting items, which summarizes the longer Reporting Guidelines.

WHY THE REPORTING GUIDELINES SHOULD NOT INCLUDE A RECOMMENDATION THAT RESEARCHERS PERFORM MANIPULATION CHECKS

MP provide an informative discussion of manipulation checks and why they ought to be encouraged. For instance, they note that the finding of no treatment effect does not necessarily show that the researcher's theory is wrong. Rather, it could be due to a failure of the treatment to move the latent variable that the theory posits is responsible for the causal effect. This is a useful caution when interpreting null results. A check that the treatment did have its intended effect on the pathway of interest would improve many experiments. Such an investigation may be thought of as its own experiment, potentially valuable as a stand-alone investigation, in which the latent variable is the outcome variable.

That said, the purpose of the Reporting Guidelines is to help ensure that researchers accurately disclose what they did, and we think it is the wrong vehicle to instruct researchers regarding how they ought to design their experiments or allocate their time and resources. If a manipulation check is performed, it should be fully reported. The current Reporting Guidelines remind the researcher to report manipulation checks for lab experiments (where they are probably most common) and also instruct researchers that this recommendation applies to other experiments as well if it is relevant. Consonant with the role of the Reporting Guidelines, and

prompted by MP's discussion, we revised the Reporting Guidelines to make it more clear that information regarding any manipulation checks that were performed should be reported for all types of experimental research.

In sum, MP make a good case for why reviewers and editors should be attentive to the question of whether an intervention did, as intended, affect a latent variable. However, it is up to reviewers and editors to determine whether the lack of a manipulation check is crucial to their assessment of the research. It is reasonably left to the discretion of reviewers, editors, and readers whether a study is publishable or incomplete until the issue of the mechanism is investigated to varying degrees, and whether it is recommended that the next dollar spent in a research program would be better allocated to a manipulation check, a replication, an assessment of the robustness of the treatment effect to variations in experimental context, or some other use.

WHY THE REPORTING GUIDELINES SHOULD REQUEST THAT RESEARCHERS REPORT RESPONSE RATES

The Reporting Guidelines specify that "If there is a survey: Provide response rate and how it was calculated" (95). MP recommend dropping the reporting of response rates for all cases except those in which the researchers claim that the subjects are random probability samples of some larger population. MP raise a number of valid points that lead us to offer a modest modification to this reporting item, and we thank them for raising this important issue. Because we believe that response rates are usually easy to calculate and are sufficiently informative to be worth reporting generally, we now recommend that researchers report response rates for any experiment (not simply for ones labeled "survey experiments") if the relevant information is available.

MP offer several objections to reporting response rates. They note that it may not be possible to calculate response rates. That is, for many common sources of subjects for survey experiments (e.g., Amazon.com's Mechanical Turk [MTurk]), it is unclear how to calculate a response rate even if one wanted to do so. Further, the response rate is uninformative as "even if such a figure could be calculated, it would have no bearing on the quality of the study" (Mutz and Pemantle 2015, 197). Finally, MP argue that there is no reason to report response rates if the researcher chooses to view the subjects as a convenience sample since there is no effort to generalize from the subset of subjects to the set of individuals the subset belongs to. They state: "If there are no claims to representativeness being made by the authors, we see no reason to require response rates" (Mutz and Pemantle 2015, 198).

MP are correct when they note that there are times when a standard response rate cannot be calculated.[1] For example, we typically do not know how many

[1]In its simplest form, a response rate—sometimes called a cooperation rate or participation rate—is calculated as the number of subjects who participate in a study divided by the number of subjects invited

individuals viewed a Craigslist or MTurk invitation. However, response metrics can be calculated for many convenience samples. For example, it is often straightforward to calculate a response rate when using a college student sample recruited from a research pool or course enrollments. There are also now commonly-used response metrics for online surveys—including non-probability online surveys (Callegaro and DiSogra 2008).[2] Although they are not standard response rates, these cooperation or participation rates for non-probability online surveys offer an informative metric for comparing across studies since they can reflect "the respondent's interest in the survey and/or the ability of the survey company to maximize cooperation" (Callegaro and DiSogra 2008, 1026).[3]

Why is it informative to provide a metric of cooperation or participation? (Rather than the more familiar term in survey work "cooperation rate," we use the broader term "participation rate" here because the reasons to report response rates typically apply more generally to any situation in which some non-random subset of those invited to participate agrees to do so). We see at least two reasons. First, researchers will sometimes have knowledge about typical participation rates in various experimental contexts, and this knowledge can lead to a more informed assessment of research. An extreme example of this point is the recent event involving the article published by LaCour and Green (2014) and subsequently retracted, in which the high participation rate reported in the article was an important clue to another group of scholars attempting similar research that the published study was problematic (Broockman et al. 2015).

Second, and more basic, participation rates inform the reader about the quantity being estimated by the experiment (that is, the estimand). Reporting participation rates improves our ability to understand what has been estimated and to compare estimates across studies. To pick a simple example, suppose that an Internet survey firm invites N potential subjects to participate according to their standard procedures, and a proportion P agree to participate; these subjects are then randomly assigned to treatment or control groups. It is useful to know the participation rate because the estimand for this experiment is the average causal effect among those who agree to participate, which we can write as T(P, X), where P is a participation rate and X are context conditions. This quantity is not necessarily equal to T(100%, X), the average causal effect for the population of N individuals invited to participate. Moreover, there is typically no reason to expect that all researchers and readers are willing to assume that T(P,X) = T(X) for all P. It is, therefore, informative to report P.

to participate. The American Association of Public Opinion Research has standardized formulas that vary depending on the treatment of factors like break-offs, noncontacts, and unknown eligibility.

[2]For example, the 2010 Cooperative Congressional Election Study, administered using the nonprobability panel of YouGov/Polimetrix, invited 196,235 email addresses to participate. Of those, 9,262 were deemed ineligible, 79,723 did not respond, 27,155 were partial interviews, and 75,450 were completed interviews. Although we cannot calculate a standard response rate, we can say the study had a 40.3% cooperation rate.

[3]In a study comparing eight different online panels, Yeager et al. (2011) found cooperation rates ranging from 2.4% to 51%.

Knowing P for a particular experiment will be valuable if the reader has some belief about how variation in P affects T or if he or she wants to compare results across experiments (in which each experiment has a particular value for P). Reporting participation rates also contributes to the basic methodological transparency needed for conducting replication studies. An apparent failure to replicate a finding across studies could reflect differential participation rates rather than other issues. More generally, for the participation rate to be useful, there is no need to assert a general relationship between participation rates and the "quality" of the study, as MP imply.

Researchers sometimes explicitly discuss the possibility that the treatment effect for non-participants will differ from that of participants. As an example, a study in which politically interested subjects participate at a disproportionately higher rate might find smaller persuasion effects compared to what would be found in the general electorate. Barabas et al. (2015) explain that "the factors that influence whether subjects participate in an experiment may alter treatment effects. Theorized effects may not manifest and effects not theorized (but observed) could attenuate or magnify relationships that researchers set out to test" (6). More generally, an understanding of any differences between those who participate and those who do not helps reviewers and readers gauge the extent to which findings are likely to apply in different populations or contexts. For example, if a study recruited 30 students (presumably those most in need of extra credit) from an invited 700 enrolled in an introductory psychology class, we might question if the same treatment effect would be observed for a sample less in need of extra credit. Given the minimal burden associated with reporting participation rates, it seems well worth the effort.

Finally, MP recognize the value of response or participation rate reporting when the goal is to generalize to a larger population, but make an exception when it is the author's intent that such generalization not be performed. We worry that this approach may be giving too much importance to the perspective of the author. Perhaps the author is really only interested in the particular group of individuals who respond to an invitation to participate; that does not mean that the reader will not want to draw general (or different) lessons from the research. If a participation rate will assist the reader, then there is a justification for providing it regardless of the researcher's own interpretation. In practice, researchers often do not clearly state what level of generalization they are claiming. In any event, whatever the researcher says, it is rarely the case that the lessons we learn from an experiment are supposed to apply to just the specific set of subjects involved. Perhaps the researcher is truly satisfied with a convenience sample because he believes that the treatment effect is uniform with respect to participation levels; there is no reason to expect that the reader will also be satisfied with this assumption.

When the author does intend to generalize and draw conclusions about a larger population, MP advise: "If the authors are making claims about accurately representing some larger population, then it would make sense to ask for a demographic comparison of their sample to the population in question" (198). Although this is not inconsistent with our recommendation to report response

rates, it is important to understand the limitations of this advice. Even if the respondents are demographically identical to the population from which they are drawn, this does not imply that the subset of the population that steps forward to participate has the same treatment effect as the population. There is considerable evidence that non-probability samples can differ in consequential ways—e.g., levels of political interest, knowledge, engagement—even if they look demographically similar (Callegaro et al. 2014; Chang and Krosnick 2009). More generally, the participators are a non-random subset of the population, and they may be biased samples even if the observables are similar to (or even identical to) the population averages. Participation rates are often useful to know, since as the rate of non-response rises, the bounds on possible bias rise as well (Manski 2007).

To conclude, we modified the Reporting Guidelines in light of MP's informative critique. MP correctly point out that it is not always possible to calculate a response rate for some surveys (e.g., MTurk samples). MP also note the often blurry line between survey methods and lab and field experiments. In light of these points, we revised the Reporting Guidelines to recommend that researchers report a response rate or other participation metric, and how it is calculated, for any type of experiment for which it is possible to do so. The Reporting Guidelines and checklist at the end of this essay reflect this change.

WHY THE REPORTING GUIDELINES SHOULD REQUEST THAT RESEARCHERS REPORT BALANCE TABLES

The Reporting Guidelines ask researchers to report averages and standard deviations for pre-treatment variables:

> "... provide a table (in text or appendix) showing baseline means and standard deviations for demographic characteristics and other pretreatment measures ..." (95).

To provide some reassurance that our request for balance tables will not make our section appear "less methodologically sophisticated than is desirable" (Mutz and Pemantle 2015, 213)[4], here is the similar reporting recommendation in the 2010 CONSORT reporting guidelines (reporting item 15), a set of widely adopted reporting standards for experimental research developed by a large group of medical researchers and statisticians and revised over the course of two decades:

> "A table showing baseline demographic and clinical characteristics for each group." (Schulz et al. 2010)

We believe a sufficient justification to request balance tables (tables of baseline means and standard deviations for each experimental group) is that the community

[4] "Evidence of widespread misunderstandings of experimental methods is plentiful throughout our major journals, even among top scholars in the discipline." Unless these are addressed "the discipline as a whole will be seen as less methodologically sophisticated than is desirable." (Mutz and Pemantle 2015, 213)

of researchers and readers finds such information useful and informative. Indeed, we are unaware of any substantial constituency that believes balance tables should go unreported or be withheld.

Moreover, as we briefly explained in the document that accompanied the Reporting Guidelines, there are good reasons to be interested in the information contained in these tables.

First, we argue that balance tables can be useful in detecting errors in randomization. We understood that this was a point of agreement with MP. However, MP write:

> The Report somewhat mischaracterizes our argument in saying that we agree "that formal tests or their rough ocular equivalents may be useful to detect errors in the implementation of randomization." The important points are (1) that such tests are not *necessary* in order to detect randomization problems; and (2) that they are not, in and of themselves, sufficient evidence of a randomization problem. (203, italics in original)

It was not our intent to mischaracterize MP's position, and upon closer reading, we believe this to be a misunderstanding. Regarding the first of MP's clarifications, we did not claim that a randomization test is "necessary" to detect randomization problems. There are other paths to detect randomization errors, such as direct examination of the researcher's code or a report by the researcher that he or she has made a mistake. But, we maintain that examination of the data for notable imbalances or other unusual patterns is a useful method to detect errors. We also do not claim that randomization tests are, in and of themselves, generally sufficient evidence of a randomization problem. Rather, as the observed imbalance is more and more unlikely under the randomization plan the researchers describe, readers will update their assessment of the odds some error was made.

We revise our statement to account for MP's concerns: "Although neither necessary nor sufficient to detect randomization errors, formal tests or their rough ocular equivalents may be useful to detect errors in the implementation of randomization."

Even after accounting for MP's qualifications, this provides a clear and common sense justification for both balance tables and even for some form of formal balance test (we will return to this question later in our reply).

Balance Tables Help Detect Errors

In the brief discussion that accompanied our presentation of the Reporting Guidelines we remarked that there are a variety of reasons why errors in the implementation of randomization might occur. Regarding the source of observable imbalances, we wrote:

> "Detectable imbalances can be produced in several ways (other than chance). These include, but are not limited to, mistakes in the randomization coding, failure to account for blocking

> or other nuances in the experimental design, mismatch between the level of assignment and the level of statistical analysis (. . .), or sample attrition." (92)

To this, MP respond:

> It is worth considering these additional rationales individually. Mistakes in coding variables do indeed occur with regularity, but why should they be more likely to occur with randomization variables than with the coding of other variables? Failure to account for blocking is already addressed elsewhere in the requirements where authors are required to describe whether and how their sample was blocked, as well as how they accomplished the random assignment process. Likewise, the description already must include mention of the unit of analysis that was randomized, so if the authors then analyze the data at a different unit of analysis, this will be evident. (204)

We respond to these arguments in order. MP concede that coding errors are common but afflict variables generally, not just the randomization. Fortunately, we have a way to detect coding errors in the randomization. As it is common for outside firms (e.g., YouGov) or organizations to implement the randomization in a study, it seems especially important for researchers to attempt to detect any errors in the process. Regarding the possibility of redundancy, yes, researchers are requested to describe how they did their randomization, but experience teaches that they may fail to do so accurately or completely.

MP continue and concede the value of examining balance when attrition is present:

> The one scenario in which balance testing does make sense is when experimental studies take place over time, thus raising the possibility of differential sample attrition due to treatment. (204)

They explain:

> Assuming a control condition is present, it sets an expectation for acceptable attrition levels. And if there is differential attrition across experimental conditions, then it makes perfect sense to conduct balance tests on pretreatment variables among post-test participants. If the post-attrition distribution of pre-treatment measures across treatment groups is distinguishable from the random pre-treatment distribution, then the experiment is clearly confounded. (204)

We are pleased to find common ground on the need for balance tables here, but to avoid any misunderstanding, one issue regarding attrition merits a small clarification. Similar to the threat of biased estimation that arises when there is non-participation, the subset of subjects that remains after attrition is typically a non-random subset of the subjects. Thus, even if the observables are the same as before attrition, or even if the attrition rate in the treatment group matches that in the control group, there still may be unobservable differences between the subjects who do not drop out and those who were randomly assigned to participate and this may produce bias. It is, however, as MP point out, reassuring to see that the observables are balanced after attrition.

One final note on the use of balance tables for error detection: The recommendation that researchers provide balance tables may lead to error discovery and correction in the course of preparing the tables. If so, much of the work done by this reporting recommendation will go uncredited.

Balance Tables Provide Useful Information to Readers

The second rationale we provided for reporting balance tables is that researchers may wish to engage in conditional analysis of a common sense sort:

> [T]here may be other uses of summary statistics for covariates for each of the experimental groups. For instance, if there is imbalance, whether statistically significant or not, in a pretreatment variable that is thought by a reader to be highly predictive of the outcome, and this variable is not satisfactorily controlled for, the reader may want to use the baseline sample statistics to informally adjust the reported treatment effect estimates to account for this difference. (92–93)

To be concrete, suppose that a research paper reported a voter turnout experiment in which the treatment group had higher turnout in past elections than the control group. Some readers will want to adjust their assessment of the magnitude of the treatment effect downward. MP take issue with the mention of statistical significance in the passage quoted above, but our reference to significance is clearly intended to highlight that p-values are not the issue here. Rather, the core concern which motivates much of the work that takes note of baseline differences in observables is that sometimes by chance the randomization results in imbalance in a variable that the researcher and/or reader believes predicts the outcome and thereby represents a realization of the randomization that will tend to yield an inaccurate treatment effect estimate. This is not a question of the statistical significance of the observed difference in outcomes, but rather one of obtaining from the experiment a fair measure of the magnitude of the treatment effect. Saying that the researcher should have stratified on the variable in question or taken some other design approach evades the issue. "Solving" this potential concern by not investigating whether there is a difference at the baseline, ignoring such information, and withholding tables containing such information, does not seem satisfactory unless we are very confident that there will be many experiments on the question at hand and the issues in any particular trial will wash out. Unfortunately, experimental replications remain quite rare in our discipline, underscoring the value of balance tables.

The informal adjustment of the treatment effect estimate described above is, we believe, fairly common practice. What formal steps should be taken in response to discovering a difference in some important variable across groups after the treatment has been applied is a matter on which opinions differ. To provide modeling and design advice is well beyond our committee's role, but there are some references that we have, as individuals, found to be helpful. These references should not in any way be taken to be an endorsement of particular articles or practices by the Experimental Research Section or as the recommended practices by the committee charged with preparing the Reporting Guidelines. Rather, they are provided to

present a range of perspectives offered by scholars who have reflected on this issue. Altman (2005) argues that the unadjusted analysis should be viewed as the primary one, but that it is arguably advisable to carry out an adjusted analysis (in Altman, a suggested adjustment is to include the baseline measures that would have been thought to be prognostic ex ante in a regression model).[5] He states that if the adjusted and unadjusted analyses show different results, "the existence of such a discrepancy must cast some doubt on the veracity of the overall (unadjusted) result (196)." Note that he implies there is value to presenting both sets of results and not just providing the reader with "the best" data analysis, as determined by the researcher (contra Mutz and Pemantle 2015, 212). Donald Rubin also addresses the issue of biased treatment effect estimates and advocates correcting for post-treatment imbalances through propensity score methods. He states that the goal of such analysis is that "we should be convinced that gold standard answers would not be materially altered had the trial been either a randomized block that ensured balance on prognostically important covariates or a re-randomized design that avoided such chance imbalances, rather than a completely randomized trial" (2008, 1352). He cautions that this post-treatment design phase should be completed before examination of the outcome data. For further discussion of these issues, including simulation evidence on the value of controlling for variables that exhibit baseline imbalances, see Bruhn and McKenzie (2009, 225–230). Some of the difficulties associated with adjusting for imbalances ex-post, such as the possibility of data-mining, will be reduced by the adoption of pre-analysis plans, though issues will still remain if the baseline measures the researcher prioritizes exclude variables that others think ought to be included.

MP raise further issues regarding the recommendation for reporting balance tables. They are concerned that researchers will find the terminology we use confusing and will have difficulty deciding which and how many variables to include in the balance table. We think that this concern is unfounded and that researchers in political science and other disciplines are capable of navigating this difficulty. However, in their discussion of the possible confusion caused by our directions regarding which variables to include in a balance table, MP write, "at other times it appears they are concerned with those variables not included in the model, such as demographic and other available pre-treatment measures . . . But if those variables are not central to the outcome of interest, it is unclear why balance on those variables is important." (Mutz and Pemantle 2015, 200)

There are two reasons to take a different view regarding the examination of variables that are thought to not be predictive of the outcome. First, there may be scholarly disagreement as to which variables are of central importance to the outcome and it is useful to allow the reader to form an independent judgment. Again, we emphasize that the guidelines are for the reader as well as the researcher.

[5]Altman (2005) recommends trying to mimic what would have been included in a model if prognostic variables were to be included, not just those shown to be imbalanced. To avoid investigator bias, "an independent source could be used to identify such variables" (5).

Second, and more critical, highly unlikely differences in group means on "irrelevant" variables may indicate that there is either a mistake in the random assignment process or that the random assignment process has not been correctly described.

The Role of Formal Statistical Tests

Turning briefly to the question of formal statistical tests of balance ("randomization tests"), we note that the Reporting Guidelines do not request the reporting of such tests, and so this portion of MP's reply is moot from the standpoint of the guidelines. That said, as a method for error correction, we do not believe formal balance tests warrant complete derision. The Bayesian logic that links observation of a baseline anomaly (usually the observation of large mean differences between treatment and control groups) to a greater chance of experimental error is easy to reconstruct. The reader begins with the prior odds of some mistake or misreporting, the balance table provides data, the reader assesses the likelihood of the observed data given random assignment versus its appearance if there is some error, and the reader updates his or her views of the chance error has occurred. If the balance table shows, for example, extreme differences that are highly unlikely by chance, this will be informative unless the observed data is equally likely in the event of error or if the prior beliefs of error are literally zero. Conducting a formal Bayesian analysis would require assumptions about the distribution of the data under the various possible errors. However, it is intuitive that very extreme values of difference of means, such as might occur one time in a million, are much more likely in the event of error than no error, and therefore might produce substantial updating in the direction of error having occurred. That said, the use of the p-value from a test of difference of means as a rough proxy for something proportional to the likelihood ratio entails some slippage and is not the same as a careful analysis of the possible errors and the observable consequences of these. Further, as MP's discussion maintains, use of the 0.05 level as a magnitude of special concern (or as a trigger for model specification) does seem an inheritance from other applications in which the 0.05 level is treated as something of importance.

Turning finally to the specific recommendation made by MP regarding balance tables (we do not address the recommendation on randomization tests, as the Reporting Guidelines do not call for such tests), we see that it is not inconsistent with our current recommendation. They conclude that "If tables of pretreatment means and standard errors are to be required, provide a justification for them" (213), and they provide examples of justifications they do not consider suitable such as "(a) Confirmation of "successful" randomization, (b) Supporting the validity of causal inference, and (c) Evidence of the robustness of inference" (213). We believe that inclusion of the tables is supported both by the mission of the guidelines to request the information readers want and by the sound reasons for such interest. As there are ample sound justifications for reporting balance tables, we leave it to the authors to choose which justifications, if any, they want to provide, and for the referees and editors to make appropriate judgments about the manuscripts researchers submit for review.

CONCLUSION

MP make some important and provocative arguments about experimental research. Although a full discussion of all of the points they raise is beyond the scope of our reply and beyond the role of our committee, we hope that our response has addressed some of the concerns MP raise. In the initial report accompanying the Reporting Guidelines we stated that we would be happy to revisit the guidelines from time to time to improve them and to adjust them to better reflect the current practices and concerns of the researchers and readers in our section and beyond. Please share with us your experience with the guidelines and any ideas you have for how they can be made more useful and informative.

SUPPLEMENTARY MATERIALS

For supplementary material for this article, please visit http://dx.doi.org/10.1017/XPS.2015.20.

REFERENCES

Altman, Doug. 2005. Adjustment for Covariate Imbalance. In *Encyclopedia of Biostatistics*, eds. P. Armitage and T. Colton. Chichester, UK: John Wiley.

Barabas, Jason, Jennifer Jerit, and Carlos Paez. 2015. "Representative Experiments: Using Registration-Based Sampling to Explore the Generalizability of Causal Effects." Presented at the 2015 Annual Meeting of the American Association of Public Opinion Researchers, Hollywood, FL.

Broockman, David, J. Kalla, and P. Aronow. 2015. "Irregularities in LaCour (2014)." Stanford University. http://stanford.edu/~dbroock/broockman_kall_aronow_lg_irregularities.pdf.

Bruhn, Miriam and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1(4): 200–32.

Callegaro, Mario and Charles DiSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72(5): 1008–1032.

Callegaro, M., R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, eds. 2014. *Online Panel Research: A Data Quality Perspective*. West Sussex, UK: John Wiley and Sons.

Chang, L. and J. A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing vs. the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4): 641–678.

LaCour, Michael J. and Donald P. Green. 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346(6215): 1366–9.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Mutz, Diana C. and Robin Pemantle. 2015. "Standard for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2(2): 192–215.

Rubin, Donald B. 2008. "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *The Journal of the American Statistical Association* 103(484): 1350–3.

Schulz, Kenneth F., Douglas G. Altman, and David Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials." *Annals of Internal Medicine* 152(11): 726–32.

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75(4): 709–47.

A Checklist of Reporting Items for Experimental Research

	Items to Report:
	Eligibility and exclusion criteria for participants
	Details of recruitment and selection of participants, including incentives and any firms used
	Type of experiment (lab, survey, field), mode, location, and dates conducted
	Response rate or other participation metric (and how calculated), when possible
	Details of randomization procedure
	Baseline means and standard deviations for demographics and other pretreatment measures by experimental group
	Whether blinding took place and how it was accomplished
	Description of the treatment(s), as well as description of control group
	Details of experiment: Its duration, number of participants, within- versus between-subjects design, piggybacking/ordering/repetition of treatments, use of deception, use of incentives
	Evidence treatment delivered as intended, if available
	Definitions of outcome measures and covariates as well as noting if level of analysis differs from level of randomization
	Identification of analyses specified ex ante versus ex post exploratory analyses
	Information in CONSORT participant flow diagram
	Sample means and standard deviations for outcome variables using intent-to-treat analysis
	Patterns of missing data, attrition, and methods of addressing these issues, if missing data and/or attrition are present
	Description of weighting procedures, if used
	IRB approval, preregistration, source of funding, conflict of interest
	Availability of replication materials and dataset