Journal of Experimental Political Science **4** (2017) 161–172
doi:10.1017/XPS.2017.2

Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments

Annie Franco*, Neil Malhotra†, Gabor Simonovits‡ and L. J. Zigerell§

Abstract

Weighting techniques are employed to generalize results from survey experiments to populations of theoretical and substantive interest. Although weighting is often viewed as a second-order methodological issue, these adjustment methods invoke untestable assumptions about the nature of sample selection and potential heterogeneity in the treatment effect. Therefore, although weighting is a useful technique in estimating population quantities, it can introduce bias and also be used as a researcher degree of freedom. We review survey experiments published in three major journals from 2000–2015 and find that there are no standard operating procedures for weighting survey experiments. We argue that all survey experiments should report the sample average treatment effect (SATE). Researchers seeking to generalize to a broader population can weight to estimate the population average treatment effect (PATE), but should discuss the construction and application of weights in a detailed and transparent manner given the possibility that weighting can introduce bias.

Keywords: Survey experiment, weighting, external validity, representativeness, transparency, SATE, PATE

Experiments have emerged as an important tool for studying political questions. Population-based survey experiments in particular allow researchers to test causal relationships that generalize to well-defined populations (Mutz 2011). The earliest of such studies were conducted on probability samples and administered via telephone (e.g., Sniderman et al. 1991) or the Internet (e.g., Clinton and Lapinski 2004). However, researchers have increasingly relied on samples from platforms such as the Cooperative Congressional Election Study (Vavreck and Rivers 2008) or Amazon's Mechanical Turk (Berinsky et al. 2012). Convenience samples may be less representative than those recruited using more traditional techniques because

*Department of Political Science, Stanford University, Stanford, CA, USA, e-mail: annniefranco@gmail.com
†Graduate School of Business, Stanford University, Stanford, CA, USA, e-mail: neilm@stanford.edu
‡Department of Politics, New York University, New York, NY, USA, e-mail: simonovits@nyu.edu
§Department of Politics and Government, Illinois State University, Normal, IL, USA, e-mail: ljzigerell@ilstu.edu

the sampling frames may contain higher coverage error.[1] Although their use has no effect on the experimenter's ability to correctly estimate the average treatment effect for the sample (sample average treatment effect (SATE)), it does raise questions about the ability of a survey experiment to provide an unbiased estimate of the average treatment effect for the corresponding population of interest (population average treatment effect (PATE)).[2]

Unrepresentativeness of survey samples caused by systematic non-response or self-selection into surveys is commonly addressed through various weighting methods.[3] The core idea of weighting techniques is to use information about the differences between the sample and the population of interest in order to estimate population quantities via adjustment of sample quantities. However, all weighting methods are based on explicit or implicit assumptions about the selection process from the population to the sample. As a result, estimates based on weighted data have desirable properties such as unbiasedness only if the assumptions underlying the weighting procedure are satisfied.

In the context of a survey experiment, the theoretical justification for the use of weights—and in fact for the use of expensive probability samples as opposed to cheaper convenience samples—is the possibility of heterogeneous treatment effects. If the treatment has the same effect on all respondents, then the SATE is an unbiased estimate of the PATE for *any* sample (e.g., Miratrix et al. 2013). Under this assumption, there is no reason to weight the data and also no reason to use more costly samples. If the treatment has differing effects across respondents, then the extent to which the SATE differs from the PATE will depend on the composition of the sample. Under certain assumptions, weighted data can yield an unbiased estimate of the PATE, but if these (untestable) assumptions fail, there is *no guarantee* that weighted estimates are better than unweighted estimates.[4]

There are legitimate reasons for applying weighting techniques in the context of a survey experiment, and there are also reasons for not using them. Unfortunately, as it will be very clear in the following section, the use of weighting methods in published work employing survey experiments is haphazard. Some articles

[1]There is, of course, substantial heterogeneity among non-probability samples. For example, YouGov/Polimetrix employs techniques such as sample matching to yield samples that approximate those obtained from probability sampling on some observable demographic characteristics. Researchers who use Mechanical Turk data often employ no strategies for making the composition of their samples more representative.

[2]Even though the quality of non-probability samples can be high (e.g., Ansolabehere and Shaffner 2014), results often substantially differ between probability samples and non-probability samples (e.g., Malhotra and Krosnick 2007; Yeager et al. 2011). Further, even probability samples have high rates of non-response that are likely non-random (Brick and Williams 2013).

[3]The arguments made in this paper are general and pertain to *any* weighting method such as post-stratification, inverse-probability weighting, or raking. To simplify the exposition we use the word "weighting" to refer to all such methods.

[4]See Online Appendix A for a technical discussion, as well as Miratrix et al. (2013, 2014) and Hartman et al. (2015).

report and discuss only weighted results, while others present only unweighted results. More importantly, most published articles fail to justify this methodological choice (e.g., simply stating in a short footnote that weights were applied). Because reviewers and editors do not seem to require authors to justify the choice of weighting methodology, researchers may cherry-pick estimates based on substantive or statistical significance.[5] Thus, in current practice weighting is a researcher degree of freedom akin to the selective reporting of outcome variables, experimental conditions, and model specifications (Franco et al. 2015; Simmons et al. 2011).

As we discuss below, the estimation of the SATE is straightforward, and we recommend that all studies employing survey experiments report this estimand as a matter of standard practice. Estimating the PATE is more complicated. In the presence of a correlation between survey non-response and individual-level treatment effects, adjusting the SATE using survey weights can help to reduce the bias of the PATE. At the same time, it may fail to mitigate all bias, and depending on the extent to which the assumptions behind the weighting method are satisfied, could also introduce additional biases. It is for this reason that we recommend that researchers reporting weighted results justify their use and be transparent about how weights were constructed and applied. An analogy can be drawn between our recommendations and standard practice in field experiments, which regularly report intent-to-treat effects even when other estimands are the primary research focus (e.g., the average treatment effect on the treated).

We first present our review of weighting practices in the literature, which indicates a lack of standard operating procedures for weighting survey experiments. We then provide a brief, non-technical review of the statistical literature on weighting and discuss the pros and cons of these adjustment techniques. We conclude by recommending best practices for the use of weights in survey experiments, with

[5]We provide two examples here of how weighting choices can affect inferences about both statistical and substantive significance. Although it is unclear whether the unweighted or weighted estimates better estimate the PATE, the key point here is that weighting has the potential to be used as a researcher degree of freedom. For instance, Harbridge and Malhotra (2011) examined the effect of providing information about partisan conflict on confidence in the U.S. Congress. The unweighted analyses showed that the bipartisan treatment increased confidence in Congress by 3.8% of the response scale (0.163 standard deviations, $p = 0.010$, two-tailed test). However, when the Polimetrix-provided weights are applied, the treatment effect is only 1.5% of the scale and does not reach statistical significance (0.064 standard deviations, $p = 0.415$, two-tailed test). Baker (2015) reported experimental evidence that non-black Americans were more likely to support in-kind donations to persons in the African country of Guyana than to persons in the Eastern European country of Armenia, but did not find a statistical difference in support for unconditional cash donations to these countries. The key regression model term for the difference in support for in-kind donations had a standardized coefficient of 0.260 ($p = 0.031$, two-tailed) in the weighted analysis when YouGov/Polimetrix-provided weights were applied, but a standardized coefficient of only 0.095 ($p = 0.317$, two-tailed) in the unweighted analysis. These examples illustrate how the application of weights can sometimes produce substantially smaller estimates (Harbridge and Malhotra 2011) and sometimes produce substantially larger estimates (Baker 2015) compared to estimates from unweighted analyses.

the hope that the discipline moves to a more standardized procedure of reporting results. One can disagree with our specific recommendations, but the goal of this article is to begin a dialogue such that political scientists address the issue of weighting more systematically. Indeed, the recent article by the Standards Committee of the Experimental Research Section of the American Political Science Association (Reporting Guidelines for Experimental Research) published in the *Journal of Experimental Political Science* includes a single line on weighting: "For survey experiments: Describe in detail any weighting procedures that are used" (Gerber et al. 2014, 98). This paper builds on and extends this guideline.

HOW DO POLITICAL SCIENTISTS EMPLOY WEIGHTS IN POPULATION-BASED SURVEY EXPERIMENTS?

Our review of the use of weights in political science survey experiments encompasses the three leading, general-interest political science journals: *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*. We conducted Google Scholar searches for each journal to locate all articles from 2000 to 2015 that used data from four commonly used online data sources for population-based survey experiments: (1) Knowledge Networks (now known as GfK Custom Research) employs probability sampling methods such as random digit dialing and address-based sampling to obtain representative samples; (2) YouGov/Polimetrix does not build its panel via probability sampling but employs model-based techniques such as sample matching to approximate population marginals; (3) Survey Sampling International (SSI) also does not employ probability sampling but allows researchers to set response quotas; and (4) Amazon's Mechanical Turk (MTurk) is an online platform that allows people to take surveys for money.[6,7] These four data sources were chosen because of their popularity and because researchers using these samples often seek to make inferences about population quantities. Google Scholar search terms and data collection procedures can be found in Online Appendix B.

After removing observational studies and false positives (e.g., articles referencing the Amazon River) from the search results, our final sample contained 113 unique studies in 85 published articles. Then two authors independently coded each article to determine whether and how the article reported handling survey weights. We first coded whether the articles mentioned weighting at all. Then, among the articles that mentioned weighting, we coded whether weighted results, unweighted results,

[6]The order in which the firms are mentioned corresponds to costliness of data collection, with 1,000-person samples from Knowledge Networks costing tens of thousands of dollars and Mechanical Turk studies of the same size running in the hundreds of dollars. Along with the raw data, GfK, and YouGov/Polimetrix provide researchers with post-stratification weights which, when applied, match the sample to population benchmarks on key demographics. SSI and MTurk do not provide weights, but researchers can calculate weights on their own.

[7]Search date: July 1, 2015 (Updated: April 1, 2016, to include the entire 2015 calendar year).

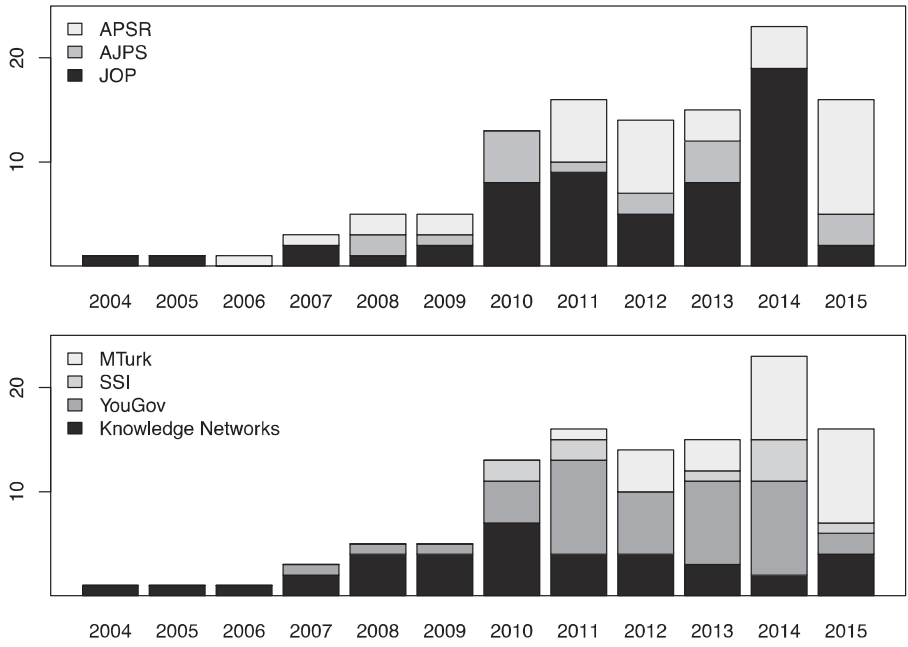

Figure 1

Trends in the Use of Online Survey Pools

Note: Columns depict the counts of articles by year for each journal (top panel) and subject pool (bottom panel). See the text and Online Appendix B for details about the sample of studies.

or both sets of results were reported for each study in the article. While some papers present additional results in online appendices, we only considered such results as "reported" if they were explicitly mentioned in the main text of the article. The agreement rate across the full set of coded observations was 92%; for the nine cases in which two authors disagreed, all four authors discussed the coding as a group and agreed upon a decision.

Trends in the use of these four samples are shown in Figure 1. The figure reveals a shift over time from traditional, more expensive online data sources used for survey experiments (Knowledge Networks/GfK, YouGov/Polimetrix) toward newer, cheaper alternatives (SSI, MTurk). Of the 45 studies published between 2004 and 2012, 24 used Knowledge Networks/GfK data. In contrast, of the 29 studies published in 2014 and 2015, only six used Knowledge Networks/GfK data, while 17 used data from MTurk.

The results of our review of the literature appear in Table 1. Across all studies, over three-quarters did not mention weighting at all. Among the 24 studies that discussed weighting, 13 reported weighted results but not unweighted results, 3 reported unweighted results but not weighted results, and 8 reported both weighted and unweighted results. For articles that did not specify weighting procedures,

Table 1
Treatment of Weighting in Political Science Survey Experiments

	KN	YouGov	SSI	MTurk	Total
Weighting not mentioned	70.3%	73.2%	100.0%	92.0%	78.8%
Only unweighted	2.7	2.4	0.0	4.0	2.7
Only weighted	16.2	17.1	0.0	0.0	11.5
Both	10.8	7.3	0.0	4.0	7.1
Number of studies	37	41	10	25	113

presumably many or all of the reported estimates are unweighted, but we cannot be sure. Clearly, the discussion of post-hoc weighting in the leading political science journals has been both rare and inconsistent.

Table 1 also presents the distributions of weighting practices across survey firms. Studies using SSI and MTurk samples almost never discussed weighting, presumably because weights are typically not provided by SSI to researchers and would need to be constructed from scratch for MTurk studies. On the other hand, while studies that use Knowledge Networks and YouGov samples also rarely discuss weighting, when they do, they often report weighted estimates only. Because these survey firms provide weights, it seems reasonable to conclude that articles using these samples and not discussing weighting are reporting unweighted estimates, but this is merely an assumption.

Given the reporting inconsistencies in Table 1, we present a practical guide for how researchers should deal with weighting in hopes of starting a discussion for what standard operating procedures should be for survey experimentalists. To provide some methodological background for our recommendations, we first provide a non-technical summary of the advantages and possible drawbacks of weighting techniques.

SATE VS. PATE

In order to infer the PATE from a treatment effect estimated in a given sample, one of two assumptions needs to be satisfied: (1) constant treatment effect across respondents (i.e., no treatment effect heterogeneity); or (2) random sampling of the population. Satisfaction of either of these two assumptions guarantees that the estimated treatment effect in the sample is an unbiased estimator of the PATE (Cole and Stuart 2010; Imai et al. 2008; see Online Appendix A).[8] Under constant treatment effects, *any* sample can be used to estimate the PATE. On the other hand, under random sampling the distribution of treatment effects in the sample is, in expectation, the same as in the population.

[8]In general, to make our discussion as broadly accessible as possible, our discussion here will remain non-technical. We provide a more rigorous treatment of weighting in Online Appendix A.

The SATE is no longer an unbiased estimate of the PATE when the probability of selection into the sample is correlated with the treatment effect (Bethlehem 1988; Cole and Stuart 2010; see also Online Appendix A). As our survey of the literature has shown, most current research does not use samples that could be plausibly considered random, and with the sharp decline of response rates, even probability samples cannot be considered truly random.[9] This is especially problematic because individual-level characteristics that are known to influence selection into surveys are also plausible moderators of a host of treatments employed in survey experiments. Weighting methods attempt to compensate for this potential source of bias.

The shared foundation of different weighting methods is the idea that, even if the sampling probability differs *across* subgroups, sampling can be assumed to be random *within* subgroups based on observable covariates (mostly demographics). If this missing-at-random (MAR) assumption holds, an estimator which weights strata-specific treatment effects by strata-specific inverse response probabilities is unbiased (e.g., Kalton and Maligalig 1991; Little and Rubin 2002; see also Online Appendix A). Different approaches to the two key issues of how to define strata (within which sampling probabilities are assumed to be equal) and how to calculate response probability in each stratum have given rise to a large number of different weighting methods.[10]

While the promise of weighting methods is to allow researchers to estimate the PATE even in the face of heterogeneous treatment effects and non-random sampling, the required assumptions are rather strong.[11] In particular, weighted estimates are no longer unbiased estimates of the PATE when there exists *any* unobserved, individual-level factor that is correlated with both the treatment effect and the sampling probability conditional on observables (Bethlehem 1988; Cole and Stuart 2010; see also Online Appendix A).

For instance, if the effect of a treatment is stronger for those with higher interest in politics and these persons are also more likely to self-select into a study, then one would need to weight on political interest in order to recover the PATE. Note, however, that because political interest is usually not observable in the population of non-respondents, one cannot use it to construct weights. While in practice

[9]Survey respondents can systematically differ from the population to which they belong for reasons related to survey design, the coverage of the sampling frame, or differential propensity to participate in a given survey (Brick and Kalton 1996).

[10]We give a brief overview of these methods in Online Appendix C and refer interested readers to read Kalton and Flores-Cervantes (2003) and Brick (2013).

[11]When data on the joint distribution of treatment and outcomes in the population are available, researchers can check whether some of these identifying assumptions hold for the PATE or for the PATE on the treated (the PATT) (Hartman et al. 2015). Hartman et al. (2015) suggest using equivalence tests to compare the weighted mean outcomes of each experimental group to the mean outcomes in the population. In contrast to medical studies, however, it is often unfeasible to conduct these tests in survey experiments because both treatment status and outcomes of interest are unobserved among unsampled individuals.

this issue can be "solved" by assuming that political interest is ignorable as a moderator of the treatment effect, or a determinant of sample selection *conditional on observables*, such assumptions are similarly as strong as the ones that motivate the use of experiments to begin with.[12]

Weighting methods also come with some practical problems. First, weighting procedures applied to the entire sample (as opposed to within treatment groups) can lead to covariate imbalance across experimental conditions. This can happen because although weights are distributed identically across treatment groups *in expectation,* there is no guarantee for this in individual samples. This can be particularly problematic when samples are small and some respondents receive very large weights, since in such cases estimates can be very sensitive to individual cases.[13]

Second, while more fine-grained weights are desirable because they make the assumption of equal selection probability within cells more plausible, they also lead to increased variability in the survey weights and, in turn, to a loss of precision. Weighting also complicates estimation of the sampling variance of estimated treatment effects (Gelman 2007), especially when the "population" frequencies used to weight strata are themselves estimated (Cochran 1977; Shin 2012; see also Online Appendix A).[14]

In sum, unweighted estimates are always unbiased estimates of the SATE, and given one of two assumptions (no treatment effect heterogeneity; random sampling) are unbiased estimates of the PATE. Weighted estimates, on the other hand, may not be unbiased estimates of the PATE, and applying weights may even introduce bias in finite samples.

Despite these drawbacks, weighting is a useful strategy because it can reduce bias in estimating the PATE from a survey sample even if that bias is not totally eliminated. In this sense, this methodological problem is no different than many we seek to tackle in the social sciences where we rely on assumptions. Yet, in order to properly move from SATE to PATE, researchers must apply weights carefully and make arguments that they are accounting for factors related to self-selection into the sample and/or treatment effect heterogeneity.

[12]The placebo tests recommended in Hartman et al. (2015) are sensitive to this heterogeneity, and can be used as a validation check whenever the placebo tests are also feasible (i.e., population data is available).
[13]Constructing a separate set of weights for each experimental group is preferable (see Hartman et al. 2015), although this creates additional practical issues for researchers, such as selecting relevant weighting variables and choosing a weighting method that takes advantage of all the sample and population information available to them (see Online Appendix C for a brief overview).
[14]Given that classical standard errors will underestimate the variance of weighted estimates, researchers should consult the literature on their post-hoc weighting method of choice for guidance on parametric adjustments and non-parametric alternatives (e.g., bootstrapping) suitable for various sampling designs and weighting scenarios.

DISCUSSION

The survey experiment is a powerful tool for identifying causal effects in political science, but the generalizability of experimental findings depends crucially on the population studied. Our review of survey experiments in the three leading political science journals using the four most prevalent online subject pools suggests that many researchers have not fully appreciated the distinction between the PATE and the SATE. While the SATE can always be estimated without bias, it is not necessarily informative about population parameters of interest. On the other hand, using methods to recover population parameters from experiments conducted on non-random samples involves making often untestable assumptions about either treatment effect heterogeneity or the process of self-selection into surveys.

These assumptions are problematic as they involve unobserved characteristics of individuals both in and out of the sample. Weighted analyses attempting to estimate the PATE can thus potentially fall prey to the very same issues so prevalent in observational research and that motivate the use of experiments in the first place. In particular, weighting experimental data to obtain the PATE can actually introduce bias if survey non-response is not properly modeled and is correlated with treatment effect heterogeneity. In the context of political science survey experiments, this is fairly likely given that many of the same variables that often predict survey response (e.g., cognitive skills, political interest) are also often moderators of political treatments (e.g., Kim et al. 2015; Xenos and Becker 2009).

Much of the discussion of weighting procedures among methodologists in political science and elsewhere creates the impression that weighting is primarily an issue of statistical methodology—that is, estimation and inference. This is partially true; advances in weighting methods can contribute to a better understanding and mitigation of problems arising from non-random selection into surveys. At the same time, given the practical limits on how much we can learn about individuals who simply *never* opt into surveys and whose politically relevant covariates remain unobserved, survey researchers should remain cautious of how much their data can tell them about population quantities.

Six Recommendations

1. Researchers should explicitly state whether they seek to estimate causal effects that generalize to a specific population (i.e., if their quantity of interest is a PATE), and whether they are reporting unweighted or weighted analyses.
2. Researchers should always report the estimate of the SATE.
3. If researchers interpret an unweighted experimental finding as a PATE, they should justify this by providing evidence that either (i) the treatment effect is constant across subgroups, or (ii) the sample is a random sample of the population of interest, with regards to measured and unmeasured variables that would plausibly moderate the treatment.

4. If researchers interpret a weighted experimental finding as a PATE, then they should be transparent about how the weights were constructed and applied.
5. Researchers using convenience samples should consider constructing weights based on some of the available demographic data for which there is sufficient variance. If a sample does not vary on observables that plausibly moderate a treatment effect, such as when the sampling frame for a study excludes some demographic groups, researchers should discuss how this limits the generalizability of their findings and/or redefine their target population.
6. Given that weighting is a researcher degree of freedom, we recommend that the full list of demographic characteristics and benchmark values used to construct the weights be reported. For studies using pre-analysis plans in advance of collecting and analyzing data (see Casey et al. 2012), weighting methodology should also be specified before data collection.[15]

Readers may disagree with these specific recommendations. The goal here is to begin a dialogue on how experimental political scientists should deal with survey weighting. We have demonstrated problems with the status quo. If the discipline adopts standard operating procedures with respect to the use of weights in survey experiments, inferential learning will be substantially improved.

SUPPLEMENTARY MATERIALS

For supplementary material for this article, please visit Cambridge Journals Online: https://doi.org/10.1017/XPS.2017.2.

REFERENCES

Ansolabehere, S. and B. F. Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22(3): 285–303.

Baker, A. 2015. "Race, Paternalism, and Foreign Aid: Evidence from U.S. Public Opinion." *American Political Science Review* 109(1): 93–109.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351–368.

Bethlehem, J. G. 1988. "Reduction of Nonresponse Bias through Regression Estimation." *Journal of Official Statistics* 4(3): 251–260.

Brick, J. M. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29(3): 329–353.

Brick, J. M. and D. Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *The Annals of the American Academy of Political and Social Science* 645(1): 36–59.

[15]While discussion of pre-analysis plans has generally focused on specifying treatment conditions, outcome variables, variable codings, and analyses, less attention has been paid to weighting.

Brick, J. M. and G. Kalton. 1996. "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5(3): 215–238.
Casey, K., R. Glennerster, and E. Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127(4): 1755–1812.
Clinton, J. D. and J. S. Lapinski. 2004. "'Targeted' Advertising and Voter Turnout: An Experimental Study of the 2000 Presidential Election." *Journal of Politics* 66(1): 69–96.
Cochran, W. G. 1977. *Sampling Techniques*. New York: John Wiley & Sons.
Cole, S. R. and E. A. Stuart. 2010. "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial." *American Journal of Epidemiology* 172(1): 107–115.
Franco, A., N. Malhotra, and G. Simonovits. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23(2): 306–312.
Gelman, A. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22(2): 153–164.
Gerber, A., K. Arceneaux, C. Boudreau, C. Dowling, S. Hillygus, T. Palfrey, D. R. Biggers, and D. J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1(1): 81–98.
Harbridge, L. and N. Malhotra. 2011. "Electoral Incentives and Partisan Conflict in Congress: Evidence from Survey Experiments." *American Journal of Political Science* 55(3): 494–510.
Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon. 2015. "From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects." *Journal of the Royal Statistical Society, Series A*. (forthcoming). doi: 10, 1111.
Imai, K., G. King, and E. A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2): 481–502.
Kalton, G. and D. S. Maligalig. 1991. A Comparison of Methods of Weighting Adjustment for Nonresponse, *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 409–428.
Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2): 81–97.
Kim, N., J. Krosnick, and D. Casasanto. 2015. "Moderators of Candidate Name-Order Effects in Elections: An Experiment." *Political Psychology* 36(5): 525–42.
Little, R. J. A. and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
Malhotra, N. and J. A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis* 15(3): 286–323.
Miratrix, L. W., J. S. Sekhon, and A. G. Theodoridis. 2014. "Why You Should (Usually) Post-Stratify on Sample Weights in Survey Experiments." *Paper presented at the Annual Meeting of the Society for Political Methodology*. Athens, GA.
Miratrix, L. W., J. S. Sekhon, and B. Yu. 2013. "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2): 369–396.
Mutz, D. C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

Shin, H. 2012. "A Cautionary Note on Post-Stratification Adjustment." Paper presented at the Section on Survey Research Methods, Joint Statistical meeting (JSM). San Diego, CA.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11): 1359–1366.

Sniderman, P. M., R. A. Brody, and P. E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. Cambridge: Cambridge University Press.

Vavreck, L. and D. Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion and Parties* 18(4): 355–66.

Xenos, M. A., and A. B. Becker. 2009. "Moments of Zen: Effects of The Daily Show on Information Seeking and Political Learning." *Political Communication* 26(3): 317–332.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability Samples and Non-Probability Samples." *Public Opinion Quarterly* 75(4): 709–747.