Worth Weighting? How to Think About and Use Weights in Survey Experiments

Luke W. Miratrix[1], Jasjeet S. Sekhon[2],
Alexander G. Theodoridis[3] and Luis F. Campos[4]

[1] *Graduate School of Education, Harvard University, Cambridge, MA 02138, USA*
[2] *Department of Political Science and Statistics, UC Berkeley, Berkeley, CA 94720, USA. Email: sekhon@berkeley.edu*
[3] *Department of Political Science, UC Merced, Merced, CA 95340, USA*
[4] *Department of Statistics, Harvard University, Cambridge, MA 02138, USA*

Abstract

The popularity of online surveys has increased the prominence of using sampling weights to enhance claims of representativeness. Yet, much uncertainty remains regarding how these weights should be employed in survey experiment analysis: should they be used? If so, which estimators are preferred? We offer practical advice, rooted in the Neyman–Rubin model, for researchers working with survey experimental data. We examine simple, efficient estimators, and give formulas for their biases and variances. We provide simulations that examine these estimators as well as real examples from experiments administered online through YouGov. We find that for examining the existence of population treatment effects using high-quality, broadly representative samples recruited by top online survey firms, sample quantities, which do not rely on weights, are often sufficient. We found that sample average treatment effect (SATE) estimates did not appear to differ substantially from their weighted counterparts, and they avoided the substantial loss of statistical power that accompanies weighting. When precise estimates of population average treatment effects (PATE) are essential, we analytically show poststratifying on survey weights and/or covariates highly correlated with outcomes to be a conservative choice. While we show substantial gains in simulations, we find limited evidence of them in practice.

Keywords: survey design, survey experiments, causal inference, generalizability

1 Introduction

Population-based survey experiments have become increasingly common in political science in recent decades (Gaines, Kuklinski, and Quirk 2007; Mutz 2011; Sniderman 2011). However, practical advice remains limited in the literature and uncertainty persists among scholars regarding the role of weights that capture differing probabilities of eventual inclusion across units in the analysis of survey experiments (Franco *et al.* 2017). Should they be used or ignored? If they are to be used, which estimators are to be preferred? As Mutz (2011, 113–120) notes,

> "there has been no systematic treatment of this topic to date, and some scholars have used weights while others have not . . . the practice of weighting was developed as a survey research tool—that is, for use in observational settings. The use of experimental

Political Analysis (2018)
vol. 26:275–291
DOI: 10.1017/pan.2018.1

Published
25 May 2018

Corresponding author
Jasjeet S. Sekhon

Edited by
R. Michael Alvarez

Authors' note: For research support and funding of the experimental studies analyzed here, Theodoridis thanks the University of California, Merced, the National Science Foundation Awards #1430505, #1225750 and #0924191, the Empirical Implications of Theoretical Models program and the Integrative Graduate Education and Research Traineeship program, and the Berkeley Institute of Governmental Studies and its Mike Synar Graduate Research Fellowship; Miratrix thanks the Institute of Education Sciences, U.S. Department of Education, through grant R305D150040; and Sekhon thanks Office of Naval Research (ONR) grants N00014-15-1-2367 and N00014-17-1-2176. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, ONR, or any of the organizations mentioned above. All studies were approved or deemed exempt by the appropriate institutional research ethics committees (University of California, Berkeley or University of California, Merced). Please send comments to lmiratrix@stat.harvard.edu, sekhon@berkeley.edu, atheodoridis@ucmerced.edu, or luiscampos@g.harvard.edu. Replication materials are publicly available, see Miratrix *et al.* (2017).

> methodology with representative samples is not yet sufficiently common for the analogous issue to have been explored in the statistical literature”

We seek to fill this void with a systematic evaluation of using weights, based on sound statistical principles rooted in the Neyman–Rubin model, to obtain practical advice for scholars seeking to make the best possible decisions when using (or electing not to use) weights in their analysis of survey experiments. We explore the topic through a combination of mathematical analysis, simulation, and examination of real data.

Taken together, these explorations lead to the conclusion that, for scholars examining population treatment effects using the high-quality, broadly representative samples recruited and delivered by top online survey firms, sample quantities, which do not rely on weights, are often sufficient. Sample average treatment effect (SATE) estimates tend not to differ substantially from their weighted counterparts, and they avoid the statistical power loss that accompanies weighting. When precise estimates of population average treatment effects (PATE) are essential, we find that a “double-Hàjek” weighted estimator is a very straightforward and reliable option in many cases. We also analytically show that poststratifying on survey weights and/or covariates highly correlated with the outcome is a conservative choice for precision improvement, because it is unlikely to do harm and could be quite beneficial in certain circumstances.

The greater prevalence of online surveys has gone hand-in-hand with the boom in survey experiments. Firms such as YouGov (formerly Polimetrix) and Knowledge Networks (now owned by GfK) provide researchers platforms through which to run experiments. The firms offer representative samples generated through extensive panel recruitment efforts and sophisticated sample matching and weighting procedures. By reducing or eliminating costs, subsidized, grant-based and collective programs such as Time-Sharing Experiments for the Social Sciences (TESS), the Cooperative Congressional Election Study (CCES), and Cooperative Campaign Analysis Project (CCAP) have further facilitated researchers’ access to time on high-end online surveys. Other firms and platforms, such as Survey Sampling International, Google Consumer Surveys (Santoso, Stein, and Stevenson 2016), and Amazon’s Mechanical Turk (Berinsky, Huber, and Lenz 2012), offer even less costly access to large and diverse convenience samples on which researchers can also conduct survey experiments. Researchers using these sometimes generate their own weights to improve representativeness. However, because we view population inferences with such convenience samples as rather tenuous, our primary interest is in methods for analysis of data from sources, such as YouGov and Knowledge Networks, that actively recruit subjects and provide the researcher with weights.

Survey experiments are a two-step process where a sample is first obtained from a parent population, and then that sample is randomized into different treatment arms. The sample selection and treatment assignment processes are generally independent of each other. Sampling procedures have changed in recent years because of increasing rates of nonresponse and new technologies. As a result, weights can vary substantially across units, with some units having only a small probability of being in the sample. In contrast, the treatment assignment mechanisms are usually simple and relatively balanced, rendering the SATE straightforward to estimate. Estimating the PATE, however, is less so because these estimates need to incorporate the weights, which introduces additional variance as well as a host of complexities.

In this work we assume the weights are known, and further assume that they incorporate both sampling probabilities and nonresponse. In particular, if there is nonresponse, and the nonresponse is correctly modeled as a function of some set of covariates, the overall weight would then be the product of being included in the sample and of responding conditional on that inclusion. We use *weight* rather than *sampling weight* to indicate this more general view. In fact, for our primary type of data targeted by this work, typically the weights are calculated by the

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

survey firms to represent the relative chances that a newly arrived recruit would get selected into the survey; as volunteering is in part self-selection, the nonresponse is built into the final weight calculations automatically. We believe the findings based on these assumptions are nevertheless informative, but we also discuss the additional complications of weight uncertainty in the body of this paper.

Overall, we encourage researchers choosing between these approaches to first give serious thought to the types of inferences they will make. Do they simply wish to establish the presence or absence of an effect in a given population? If so, the SATE may suffice. Or do they hope to measure the magnitude of an effect that may not already be documented? In this case, the scholar should possibly consider her options for weighted estimators.

In Section 2, we overview general survey methodology. In Section 3, we formally consider survey experiments and relate them to the SATE. We formally define the PATE and some estimators of it in Section 4, where we also discuss weights and uncertainty in weights in more detail, and we introduce a poststratification estimator in Section 4.3. We then investigate the performance of these estimators through simulation studies in Section 5, and analyze trends and features of real survey experimental data collected through YouGov in Section 6. We conclude with an extended discussion, providing some advice and high-level pointers to applied practitioners.

2 Surveys and Survey Experiments through the Lens of Potential Outcomes

We formalize surveys and survey experiments in terms of the Neyman–Rubin model of potential outcomes (Splawa-Neyman, Dabrowska, and Speed 1990). Assume we have a population of N units indexed as $i = 1, \dots N$. We take a sample from this population using a sample selection mechanism, and we then randomly assign treatment in this sample using a treatment assignment mechanism. Both mechanisms will be formally defined in subsequent sections. Each unit i in the population has a pair of values, $(y_i(0), y_i(1))$, called its potential outcomes. Let $y_i(1) \in \mathbb{R}$ be unit i's outcome if it were treated, and $y_i(0)$ its outcome if it were not. For each selected unit, we observe either $y_i(1)$ or $y_i(0)$ depending on whether we treat it or not. For any unselected unit, we observe neither.

We make the usual no-interference assumption that implies that treatment assignment for any particular unit has no impact on the potential outcomes of any other unit. This assumption is natural in survey experiments. The treatment effect Δ_i for unit i is then the difference in potential outcomes, $\Delta_i \equiv y_i(1) - y_i(0)$. These individual treatment effects are deterministic, pretreatment quantities.

Let $\mathcal{S}$ be our sample of n units. Then the SATE is the mean treatment effect over the sample:

$$\tau_{\mathcal{S}} = \frac{1}{n} \sum_{i \in \mathcal{S}} \Delta_i = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i(1) - \frac{1}{n} \sum_{i \in \mathcal{S}} y_i(0). \tag{1}$$

This is a parameter for the sample at hand, but is random in its own right if we view the sample as a draw from the larger population. By comparison, a parameter of interest in the population is the PATE defined as

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \Delta_i = \frac{1}{N} \sum_{i=1}^{N} y_i(1) - \frac{1}{N} \sum_{i=1}^{N} y_i(0).$$

In general $\tau_{\mathcal{S}} \neq \tau$, and if the sampling scheme is not simple (e.g., some types of units are more likely to be selected), then potentially $\mathbb{E}[\tau_{\mathcal{S}}] \neq \tau$.

We discuss some results concerning the sample selection mechanism in the next section. After that, we will combine the sample selection with the treatment assignment process.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

2.1 Simple surveys (no experiments)

Let S_i be a dummy variable indicating selection of unit i into the sample, with $S_i = 1$ if unit i is in the sample, and 0 if not. Let $\mathcal{S}$ be $(S_1, \ldots, S_N)$, the vector of selections. In a slight abuse of notation, let $\mathcal{S}$ also denote the random sample. Thus, for example, $i \in \mathcal{S}$ would mean unit i was selected into sample $\mathcal{S}$. Finally let the overall *selection probability* or *sampling probability* for unit i be

$$\pi_i \equiv \mathbf{P}\{S_i = 1\} = \mathbb{E}[S_i],$$

which is the probability of unit i being included in the sample. The more the π_i vary, the more the sample could be unrepresentative of the population. We assume $\pi_i > 0$ for all i, meaning every unit has some chance of being selected into $\mathcal{S}$. The π_i depend, among other things, on the desired size of sample $\mathbb{E}[n]$. We assume the π_i are fixed and known and incorporate nonresponse; we discuss uncertainty in them in Section 4.2.

Consider the case where we have no treatment and we see $y_i \equiv y_i(0)$ for any selected unit. Our task is to estimate the mean of the population, $\mu = (1/N)\sum_{i=1}^{N} y_i$. Estimating the mean of a population under a sampling framework has a long, rich history. We base our work on two estimators from that history here.

Let $\bar{\pi} = (1/N)\sum \pi_i$ be the average selection probability in the population and $n = \sum S_i$ be the realized sample size for sample $\mathcal{S}$, with $\mathbb{E}[n] = N\bar{\pi}$. Then let $w_i = \bar{\pi}/\pi_i$ be the *weight*. These weights w_i are relative to a baseline of 1, which eases interpretability due to removing dependence on n. A weight of 1 means the unit stands for itself, a weight of 2 means the unit "counts" as 2 units, a weight of 0.5 means units of this type tend to be overrepresented and so this unit counts as half, and so forth. The total weight of our sample is then

$$Z \equiv \sum_{i=1}^{N} \frac{\bar{\pi}}{\pi_i} S_i = \sum_{i=1}^{N} w_i S_i.$$

Z is random, but $\mathbb{E}[Z] = \mathbb{E}[n]$.

The Horvitz–Thompson estimator (Horvitz and Thompson 1952), an inverse probability weighting estimator, is then

$$\hat{y}_{HT} = \frac{1}{\mathbb{E}[n]} \sum_{i=1}^{N} \frac{\bar{\pi}}{\pi_i} S_i y_i = \frac{1}{\mathbb{E}[Z]} \sum_{i=1}^{N} w_i S_i y_i.$$

Although unbiased, the Horvitz–Thompson estimator is well known to be highly variable. This variability comes from the weights; if you randomly get too many rare units in the sample, the inverse of their weights will inflate $\hat{y}_{HT}$, even if all y_i are the same. We are not controlling for the realized size of the sample. This is reparable by normalizing by the realized weight of the sample rather than the expected.

This gives the Hàjek estimator, which is the usual weighted average of the selected units, and which likely reflects the approach used by most scholars:

$$\hat{y}_{H} = \frac{1}{Z} \sum_{i=1}^{N} w_i S_i y_i.$$

The Hàjek estimator is not unbiased, but it often has smaller MSE than Horvitz–Thompson (Hàjek 1958; Särndal, Swensson, and Wretman 2003). The bias, however, will tend to be negligible, as shown by the following lemma:

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

LEMMA 2.1 (A variation on Result 6.34 of Cochran (1977)). *Under a Poisson selection scheme, i.e. units sampled independently with individual probability π_i, the bias of the Hàjek estimator is $O(1/\mathbb{E}[n])$. In particular, the bias can be approximated as*

$$\mathbb{E}[\hat{y}_H] - \mu \doteq -\frac{1}{\mathbb{E}[n]}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)\frac{\bar{\pi}}{\pi_i}\right) = -\frac{1}{\mathbb{E}[n]}\mathrm{Cov}[y_i, w_i].$$

See Appendix C for proof. The above shows that, for a fixed population, the bias decreases rapidly as sample size increases. If we sample with equal probability or if the outcomes are constant, the bias is 0. However, if the covariance between the weights and y_i is large, the bias could potentially be large also. In particular, the covariance will be large if rare units (those with small π_i) systematically tend to be outliers with large $y_i - \mu$ because, as weights are nonnegative inverses of the π_i, their distribution can feature a long right tail that drives the covariance.

3 Survey Experiments and SATE

Survey experiments are surveys with an additional treatment assigned at random to all selected units. Independent of S_i let T_i be a treatment assignment, with $T_i = 1$ if unit i is treated, 0 otherwise. The most natural such assignment mechanism for our context is Bernoulli assignment, where each responding unit i is treated independently with probability p for some p. Another common mechanism is the classic complete randomization, when a np-sized simple random sample of the n units is treated. Regardless, we assume randomization is a separate process from selection. In particular, we assume that randomization does not depend on the weights.

If our interest is in the SATE, then a natural estimator is Neyman's difference-in-means estimator of

$$\hat{\tau}_{\mathrm{SATE}} = \frac{1}{n_1}\sum_{i=1}^{n} T_i y_i - \frac{1}{n - n_1}\sum_{i=1}^{n}(1 - T_i) y_i, \tag{2}$$

with n_1 the (possibly random) number of treated units (see Splawa-Neyman, Dabrowska, and Speed 1990).

This estimator is essentially unbiased for the SATE ($\mathbb{E}[\hat{\tau}_{\mathrm{SATE}}|\mathcal{S}] = \tau_{\mathcal{S}}$), but unfortunately, the SATE is not generally the same as the PATE and $\mathbb{E}[\tau_{\mathcal{S}}] \neq \tau$ in general. The bias, for fixed n, is

$$\begin{aligned} \mathrm{bias}(\hat{\tau}_{\mathrm{SATE}}) = \mathbb{E}[\hat{\tau}_{\mathrm{SATE}}] - \tau &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\pi_i}{\bar{\pi}} - 1\right)\Delta_i \\ &= \mathrm{Cov}\left(\frac{\pi_i}{\bar{\pi}}, \Delta_i\right). \end{aligned} \tag{3}$$

See Appendix C for derivation. As units with higher π_i will be more likely to be selected into $\mathcal{S}$, the estimator will be biased toward the treatment effect of these units. Equation (3) shows an important fact: if the treatment impacts are not correlated with the weights, then there will be no bias. In particular, if the selection probabilities are all the same, or there is no treatment effect heterogeneity, then the bias of $\hat{\tau}_{\mathrm{SATE}}$ for estimating the PATE will be 0.

The variance of $\hat{\tau}_{\mathrm{SATE}}$, conditional on the sample $\mathcal{S}$, is well known, but we include it here as we use it extensively.

THEOREM 3.1. *Let sample $\mathcal{S}$ be randomly assigned to treatment and control with $\mathbb{E}[T_i] = p$ for all i with either a complete randomization or Bernoulli assignment mechanism. The unadjusted*

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

simple-difference estimator $\hat{\tau}_{\text{SATE}}$ is unbiased[1] *for the SATE, i.e.* $\mathbb{E}[\hat{\tau}_{\text{SATE}}|\mathcal{S}] = \tau_{\mathcal{S}}$. *Its variance is*

$$\begin{aligned}\text{Var}[\hat{\tau}_{\text{SATE}}|\mathcal{S}] &= \frac{1}{n}[(\beta_1 + 1)\sigma^2_{\mathcal{S}}(1) + (\beta_0 + 1)\sigma^2_{\mathcal{S}}(0) + 2\gamma_{\mathcal{S}}] \qquad (4)\\ &= \frac{1}{n}[\beta_1\sigma^2_{\mathcal{S}}(1) + \beta_0\sigma^2_{\mathcal{S}}(0) - \sigma^2_{\mathcal{S}}(\Delta)], \qquad (5)\end{aligned}$$

where $\sigma^2_{\mathcal{S}}(z)$ *and* $\sigma^2_{\mathcal{S}}(\Delta)$ *are the variances of the individual potential outcomes and treatment effects for the sample, and* $\beta_\ell = \mathbb{E}[n/n_\ell]$ *are the expectations (across randomizations) of the inverses of the proportion of units in the two treatment arms.*

If n_1 is fixed, such as with a completely randomized experiment, then $\beta_1 = 1/p$, $\beta_0 = 1/(1-p)$ and the above simplifies to Neyman's result of

$$\text{Var}[\hat{\tau}_{\text{SATE}}|\mathcal{S}] = \frac{1}{n}\left[\frac{1}{p}\sigma^2_{\mathcal{S}}(1) + \frac{1}{1-p}\sigma^2_{\mathcal{S}}(0) - \sigma^2_{\mathcal{S}}(\Delta)\right].$$

For Bernoulli assignment, the β_ℓ are complicated because of the expectation of the random denominator, and there are mild technical issues because the estimator is undefined when, for example, $n_1 = 0$. One approach is to use $n/\mathbb{E}[n_1] = 1/p$ as an approximation for β_1. This approximation is quite good; the bias is of high order for the same reasons as the bias for the Hàjek estimator (see Lemma 2.1). Furthermore, the undefined issue is of small concern as the chance of $n_1 = 0$ is exponentially small; giving the estimator an arbitrary value (e.g., 0) if this rare event occurs, introduces only a small bias. An alternate approach is to condition on the number of units treated: set $p = n_1/n$ and use Neyman's results. Conditioning is a reasonable choice that we prefer. It leads to a more accurate (and more readable) formula. For details, including formal definitions and the derivations, see Miratrix, Sekhon, and Yu (2013).

It is important to underscore that any SATE analysis on the sample, given a truly modeled treatment assignment mechanism, is valid. That is, such an analysis is estimating a true treatment effect parameter, the SATE. If $\mathbb{E}[\tau_{\mathcal{S}}] = \tau$, then any SATE analysis will be correct for PATE as well (although estimates of uncertainty may be too low if they do not account for variability in $\tau_{\mathcal{S}}$). In particular, if there is a constant treatment effect, then $\tau_{\mathcal{S}} = \tau$ for any sample, and the SATE will be the PATE, and all uncertainty estimates for the SATE will be the same as for the PATE.[2] But a constant treatment effect is a large assumption.

4 Estimating the PATE

Imagine we had both potential outcomes $y_i(1)$, $y_i(0)$ for all the sampled $i \in \mathcal{S}$. These would give us exact knowledge of the SATE, and we could also use this information, coupled with the weights, to estimate the PATE. In particular, with knowledge of the $y_i(\ell)$ we have a sample of treatment effects:

$$\Delta_i = y_i(1) - y_i(0) \quad \text{for } i \in \mathcal{S}.$$

We can use these Δ_i to estimate the PATE, τ, with, for example, a Hàjek estimator:

$$v_{\mathcal{S}} = \frac{1}{Z}\sum S_i\frac{\bar{\pi}}{\pi_i}\Delta_i = \frac{1}{Z}\sum S_i\frac{\bar{\pi}}{\pi_i}y_i(1) - \frac{1}{Z}\sum S_i\frac{\bar{\pi}}{\pi_i}y_i(0). \qquad (6)$$

[1] Nearly unbiased, that is. Under randomizations where the estimator could be undefined (e.g., there is a chance of all units getting assigned to treatment, such as with Bernoulli assignment where n_1 is random and $\mathbf{P}\{n_1 = 0\} > 0$ or $\mathbf{P}\{n_0 = 0\} > 0$), this unbiasedness is conditional on the event of the estimator being defined. Because this probability is generally exponentially small the bias is as well, however. See Miratrix, Sekhon, and Yu (2013) for further discussion.

[2] The estimated uncertainty will, however, depend on the sample $\mathcal{S}$. For example, if $\mathcal{S}$ happens to have widely varying units, $\hat{\tau}_{\text{SATE}}$ will have high variance and the sample-dependent SATE SE estimate should generally reflect that by being large to give correct coverage for $\tau_{\mathcal{S}}$. Now, as this is true for any sample, the overall *process* will have correct coverage.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

This oracle estimator $\nu_{\mathcal{S}}$ is slightly biased, but the bias is small, giving $\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}] \approx \tau$. If we wanted an unbiased estimator, we could use a Horvitz–Thompson estimator by replacing Z with $\mathbb{E}[n] = N\bar{\pi}$, the expected sample size.

Unfortunately, we do not, for a given sample $\mathcal{S}$, observe $\nu_{\mathcal{S}}$. We can, however, estimate it given the randomization and partially observed potential outcomes. Estimating the PATE is now implicitly a two-step process: estimate the sample-dependent $\nu_{\mathcal{S}}$, which in turn estimates the population parameter τ. Under this view, we have two concerns. First, we have to accurately estimate $\nu_{\mathcal{S}}$, using all the tools available, by simple randomized experiments such as adjustment methods or, if we can control the randomization, blocking. Second, we have to focus on a sample parameter $\nu_{\mathcal{S}}$, that is itself a good estimator of τ. See Appendix A for further discussion.

4.1 Estimating $\nu_{\mathcal{S}}$

Equation (6) shows that our estimator is the difference in weighted means of our treatment potential outcomes and our control potential outcomes. This immediately motivates estimating these means with the units randomized to each arm of our study, as with the following "double-Hàjek" estimator

$$\hat{\tau}_{hh} = \frac{1}{Z_1}\sum_{i=1}^{N} S_i T_i \frac{\bar{\pi}}{\pi_i} y_i(1) - \frac{1}{Z_0}\sum_{i=1}^{N} S_i(1-T_i)\frac{\bar{\pi}}{\pi_i} y_i(0) \tag{7}$$

with

$$Z_1 = \sum_{i=1}^{N} S_i T_i \frac{\bar{\pi}}{\pi_i} \quad \text{and} \quad Z_0 = \sum_{i=1}^{N} S_i(1-T_i)\frac{\bar{\pi}}{\pi_i}.$$

The Z_ℓ are the total sample masses in each treatment arm. $\mathbb{E}[Z_1] = pN\bar{\pi} = \mathbb{E}[n_1]$, the expected number of units that will land in treatment (similarly for control).

This estimator is two separate Hàjek estimators, one for the mean treatment outcome and one for the mean control. Each estimator adjusts for the total mass selected into that condition. This difference of weighted means is the one naturally seen in the field. It corresponds to the weighted OLS estimate from regressing the observed outcomes Y^{obs} on the treatment indicators T with weights w_i. This equivalence is shown in Appendix C.

Because this is a Hàjek estimator, there is bias for $\hat{\tau}_{hh}$ in the randomization step as well as the selection step because the Z_ℓ depend on the realized randomization. Again, this bias is small, which means the expected value of our actual estimator, conditional on the sample, is approximately $\nu_{\mathcal{S}}$, our Hàjek "estimator" of the population τ: $\mathbb{E}[\hat{\tau}_{hh}|\mathcal{S}] \approx \nu_{\mathcal{S}}$. (For unbiased versions, see Appendix A.)

We can obtain approximate results for the population variance of $\hat{\tau}_{hh}$ if we view the entire selection-and-assignment process as drawing two samples from a larger population. We ignore the finite-sample issues of no unit being able to appear in both treatment arms (i.e., we assume a large population) and use approximate formula based on sampling theory. For a Poisson selection scheme and Bernoulli assignment mechanism we then have:

Theorem 4.1. *The approximate variance (AV) of $\hat{\tau}_{hh}$ is*

$$AV(\hat{\tau}_{hh}) \approx \frac{1}{p\mathbb{E}[n]}\frac{1}{N}\sum_{j=1}^{N} w_j (y_j(1)-\mu(1))^2 + \frac{1}{(1-p)\mathbb{E}[n]}\frac{1}{N}\sum_{j=1}^{N} w_j (y_j(0)-\mu(0))^2,$$

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

with $\mu(z) = (1/N)\sum_{i=1}^{N} y_i(z)$. This formula assumes the π_j are small; see Appendix C for a more exact form. This variance can be estimated by

$$\widehat{V}(\hat{\tau}_{sd}) = \frac{1}{Z_1^2}\sum_{j=1}^{N} S_i T_i w_j^2 (y_j(1) - \hat{\mu}(1))^2 + \frac{1}{Z_0^2}\sum_{j=1}^{N} S_i(1 - T_i) w_j^2 (y_j(0) - \hat{\mu}(0))^2,$$

where $\hat{\mu}(1) = (1/Z_1)\sum_{i=1}^{N} S_i T_i w_i y_i(1)$ and $\hat{\mu}(0) = (1/Z_0)\sum_{i=1}^{N} S_i(1 - T_i) w_i y_i(0)$.

See Appendix C for the derivation, which also gives more general formulas that can be adapted for other selection mechanisms. For related work and similar derivations, see Wood (2008) and Aronow and Middleton (2013).

4.2 Uncertain and misspecified weights

Following the survey sampling literature, this paper assumes the weights are exact, correct, and known. They are considered to be the total chance of selection into the sample. In particular, again following standard practice, the π_i are the product of any original sampling weights and any nonresponse weights, given a classic sampling context (Groves *et al.* 2009). By contrast, for surveys such as YouGov the nonresponse is built in, as the recruited panels are in effect self-selected, so we get the overall weights (which they call propensity or case weights) directly. Our results are regarding these total weights.

Of course, especially when considering nonresponse, weights are not known but instead estimated using a model and, ideally, a rich set of covariates. This raises two concerns. The first is if the weights are systematically inaccurate due to some selection mechanism that has not been correctly captured. In this case, as the weights are independent of the assignment mechanism, the SATE estimates are still valid and unbiased. The PATE estimates, however, can be arbitrarily biased, and this bias is not necessarily detectable. For example, if only those susceptible to treatment join the study, the PATE estimate will be too high, and there may be no measured covariate that allows for detection of this.

The second concern is whether there is additional uncertainty that needs to be accounted for, given the estimated weights, when doing inference for the PATE. There is, although we believe this uncertainty can often be much smaller than the uncertainty in the randomized experiment itself.[3] While this uncertainty could be taken into account, much of the literature does not tend to do so. Interestingly, it is not obvious whether estimating the weights given the sample could actually *improve* PATE precision similar to using estimated propensity scores instead of known propensity scores—see, for example, Hirano, Imbens, and Ridder (2000). We leave this as an area for future investigation.

For further thoughts on concerns regarding uncertainty in the weights, we point to the literature on generalizing randomized trials to wider populations, such as discussed in Hartman *et al.* (2015) and Imai, King, and Stuart (2008). Here, the approach is generally to estimate units' propensity for inclusion into the experiment, and then weight units by these quantities in order to estimate population characteristics. These propensities of inclusion are usually estimated by borrowing from the propensity score literature for observational studies (Cole and Stuart 2010). One nice aspect of this approach is it provides diagnostics in the form of a placebo test. In particular, the characteristics of the reweighted control group of the randomized experiment should match the characteristics of the target population of interest (see Stuart *et al.* (2010) for a discussion).

Relatedly, O'Muircheartaigh and Hedges (2014) and Tipton (2013) propose poststratified estimators, stratifying on these estimated weights. In their case, however, they also have the

[3] Consider that the estimated weights are usually calibrating the full sample to a larger known population, while the uncertainty of the experiment is of the difference in two subsamples, which will tend to have about four times the variance, at least.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

population proportions of the strata as given, which allows for simpler variance expressions and arguably less sensitivity to error in the weights themselves. Furthermore, they do not incorporate the unit-level weights once they stratify. Tipton (2013) investigates the associated bias-variance trade-offs due to stratification, and gives advice as to when stratification will be effective.

Generalization assumes we know the assignment mechanism, but not necessarily the sampling mechanism. There is some work on the reverse case, with estimated propensity scores of treatment and known weights, see DuGoff, Schuler, and Stuart (2013). Here the final propensity weights are also treated as fixed for inference.

4.3 Poststratification to improve precision

One can improve the precision of an experiment by adjusting for covariates. For an examination of this under the potential outcome framework, see, for example, Lin (2013). We use poststratification for this adjustment. In poststratification, treatment effects are first estimated within each of a series of specified strata of the data and then averaged together with weights proportional to strata size (Miratrix, Sekhon, and Yu 2013).

We use poststratification because it relies on very weak modeling assumptions and naturally connects with the weighting involved in estimating the PATE. See Appendix B for the overall framework and associated estimators. Other estimators that rely on regression and other forms of modeling are also possible; see Zheng and Little (2003) or, more recently, Si, Pillai, and Gelman (2015). For poststratification, the more the mean potential outcomes vary between strata, the greater the gain in precision. And given that it is precisely when the weights and outcomes are correlated that we must worry about the weights, poststratifying on them is a natural choice. Such stratification is easy to implement: simply build K strata prerandomization (but not necessarily presampling) by, e.g., taking the K-weighted quantiles of the $1/\pi_i$ as the strata.

When the units are divided into K quantiles by survey weight, the cut points of those quantiles depend on the realized weights of the sample. Because this is still prerandomization, this does not impact the validity of the variance and variance-estimation formulas of the SATE estimate of $\tau_{\mathcal{S}}$. It does, however, make generating appropriate population variance formulas difficult. Given this, we propose using the bootstrap, incorporating the variable definition of strata to take this stage being sample-dependent into account. Bootstrap is natural in that for survey experiments we are pulling units from a large population, and so simulating independent draws is reasonable. While a technical analysis of this approach is beyond the scope of this paper, we discuss some particulars of implementation in Appendix B. Reassuringly, our simulation studies in the next section show excellent coverage rates.

5 Simulation Studies

We here present a series of simulation studies to assess the relative performance of the respective estimators. We also assess the performance of the bootstrap estimates of the standard errors.

Our simulation studies are as follows: we generate a large population of size $N = 10,000$ with the two potential outcomes and a selection probability for each unit. Using this fixed population, we repeatedly take a sample and run a subsequent experiment, recording the treatment effect estimates for the different estimators. In particular, we first select a sample of size n, sampling without replacement but with probabilities of selection inversely proportional to the weights.[4] Once we have obtained the final sample, we randomly assign treatment and estimate the treatment effect. After doing this 10,000 times we estimate the overall mean, variance, and MSE of the different estimators to compare their performance to the PATE. We also calculate bootstrap

[4] We ignore a mild technical issue of the π_i not being exactly proportional to the weights due to not sampling with replacement.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

Table 1. Simulations A & B. Performance of different estimators as estimators for the PATE for (A) a heterogeneous treatment effect scenario with $\tau = 32.59$ and (B) a constant treatment effect of $\tau = 30$. For each estimator, we have, from left to right, its expected value, bias, standard error, root mean squared error, average bootstrap SE estimate, and coverage across 10,000 trials.

	Estimator	Mean	Bias	SE	RMSE	boot SE	Coverage
A							
1	$\tau_{\mathcal{S}}$	40.36	7.77	1.35	7.89		
2	$\nu_{\mathcal{S}}$	32.58	0.00	1.84	1.84		
3	$\hat{\tau}_{\text{SATE}}$	40.37	7.78	3.14	8.39	3.12	30%
4	$\hat{\tau}_{hh}$	32.62	0.03	3.91	3.91	3.79	95%
5	$\hat{\tau}_{ps}$	32.60	0.01	2.67	2.67	2.69	95%
B							
1	$\tau_{\mathcal{S}}$	30.00	0.00	0.00	0.00		
2	$\nu_{\mathcal{S}}$	30.00	0.00	0.00	0.00		
3	$\hat{\tau}_{\text{SATE}}$	30.01	0.01	2.58	2.58	2.58	95%
4	$\hat{\tau}_{hh}$	30.03	0.03	3.35	3.35	3.31	95%
5	$\hat{\tau}_{ps}$	30.02	0.02	3.32	3.32	3.29	95%

standard error estimates for all the estimators using the case-wise bootstrap scheme discussed in Appendix B.

5.1 Simulation A

Our first simulation is for a population with a heterogeneous treatment effect that varies in connection to the weight. See Appendix D for some simple plots showing the structure of the population and a single sample. Our treatment effect, outcomes, and sampling probabilities are all strongly related. We then took samples of a specified size from this fixed population, and examined the performance of our estimators as estimators for the PATE.

Results for $n = 500$ are on Table 1. Other sample sizes such as $n = 100$, not shown, are substantively the same. The first two lines of the table show the performance of the two "oracle" estimators $\tau_{\mathcal{S}}$ (Equation (1)) and $\nu_{\mathcal{S}}$ (Equation (6)), which we could use if all of the potential outcomes were known. For $\tau_{\mathcal{S}}$ there is bias because the treatment effect of a sample is not generally the same as the treatment effect of the population. The Hàjek approach of $\nu_{\mathcal{S}}$, second line, is therefore superior despite the larger SE. Line 3 is the simple estimate of the SATE from Equation (2). Because it is estimating $\tau_{\mathcal{S}}$, it has the same bias as line 1, but because it only uses observed outcomes, the SE is larger. Line 4 uses the "double-Hàjek" estimator shown in Equation (7). This estimator is targeting $\nu_{\mathcal{S}}$, reducing bias, but has a larger SE relative to line 3 due to the fact that we are incorporating weights. Line 5 is the poststratified "double Hàjek" given in Appendix B. Units were stratified by their survey weight, with $K = 7$ equally sized (by weight) strata. For this scenario, poststratifying helps, as illustrated by the smaller SE and RMSE, compared to $\hat{\tau}_{hh}$.

An inspection of the coverage rates reflects what we have already discussed: the estimate $\hat{\tau}_{\text{SATE}}$ does not target the PATE while the other two sample estimates, $\hat{\tau}_{hh}$ and $\hat{\tau}_{ps}$, do. Therefore, it has terrible coverage. Furthermore, the latter two estimates give correct coverage, which is reflective of the bootstrap SE estimates hitting their mark.

5.2 Simulation B

As a second simulation we kept the original structure between $Y(0)$ and w, but set a constant treatment effect of 30 for all units. Results are on the bottom half of Table 1. Here, $\tau_{\mathcal{S}} = \tau$ for any sample $\mathcal{S}$, so there is no error in either estimate with known potential outcomes (lines 1–2). This also means that $\hat{\tau}_{\text{SATE}}$ is a valid estimate of the PATE and this is reflected in the lack of bias and

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

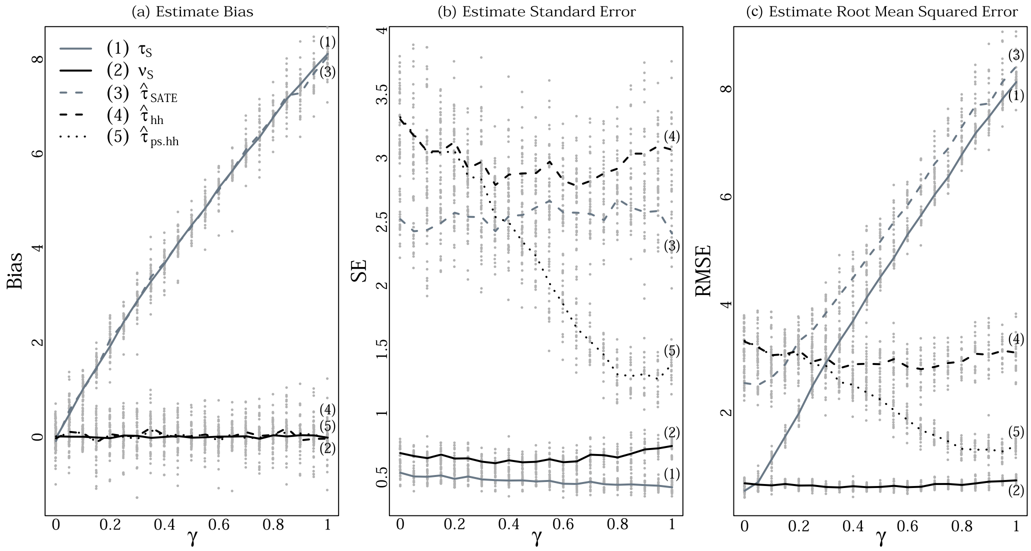

Figure 1. (a) Bias, (b) standard error, and (c) root mean squared error of estimates when selection probability is increasingly related to the potential outcomes (Simulation C). The horizontal axis, γ, varies the strength of the relationship between outcome and weight. Gray are SATE-targeting, black PATE-targeting. Solid are oracle estimators using all potential outcomes of the sample, dashed are actual estimators. The thicker lines are averages over the 20 simulated populations in light gray dots.

nominal coverage rate (line 3). The increase of SE of the weighted and poststratified estimators (lines 4–5) reflects the use of weights when they are in fact unnecessary. Overall, the SATE estimate is the best, as expected in this situation.

5.3 Simulation C

In our final simulation, we systematically varied the relationship between selection probability and outcomes while maintaining the same marginal distributions in order to examine the benefits poststratification.

In our data generating process (DGP) we first generate a bivariate normal pair of latent variables with correlation γ, and then generate the weights as a function of the first variable and the outcomes as a function of the second. Then, by varying γ we can vary the strength of the relationship between outcome and weight. (See Appendix D for the particulars.) When $\gamma = 1$, which corresponds to Simulation A, w_i and $(Y_i(0), Y_i(1))$ have a very strong relationship and we benefit greatly from poststratification. Conversely when $\gamma = 0$, w_i and $(Y_i(0), Y_i(1))$ are unrelated and there will be no such benefit.

For each γ we generated 20 populations, conducting a simulation study within each population. We then averaged the results and plotted the averages against γ on Figure 1. The solid lines give the performance of the oracle estimators $\tau_{\mathcal{S}}$ and $\nu_{\mathcal{S}}$, and the nonsolid lines are the estimators. The gray lines are estimators that do not incorporate the weights, and the black lines are estimators that do. The light gray points show the individual population simulation studies; they vary due to the variation in the finite populations.

We first see that, because both the double Hàjek and its poststratified version are targeting $\nu_{\mathcal{S}}$, which in turn estimates the PATE, they remain unbiased regardless of the latent correlation. On the other hand, the SATE and its estimator, $\hat{\tau}_{\text{SATE}}$, are affected. The bias continually increases as the relationship between weight and treatment effect increases.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

As expected, the SE of the estimators that do not use weights, $\tau_{\mathcal{S}}$ and $\hat{\tau}_{\text{SATE}}$, stay the same regardless of γ because the marginal distributions of the outcomes are the same across γ. The estimators that only use weights to adjust for sampling differences, $\nu_{\mathcal{S}}$ and $\hat{\tau}_{hh}$, also remain the same, although their SEs are larger than for $\tau_{\mathcal{S}}$ and $\hat{\tau}_{\text{SATE}}$ because of incorporating the weights. We pay for unbiasedness with greater variability. The poststratified estimator $\hat{\tau}_{ps}$, however, sees continual precision gains as the weights are increasingly predictive of treatment effect. For low γ, it has roughly the same uncertainty as $\hat{\tau}_{hh}$, but is soon the most precise of all (nonoracle) estimators.

These conclusions are tied together in the rightmost panel of Figure 1, showing the RMSE, which gives the combined impact of bias and variance on performance. As γ increases, the RMSE of $\hat{\tau}_{\text{SATE}}$ steadily climbs due to bias, eventually being the worst at $\gamma = 0.2$. Meanwhile, the poststratified estimator that exploits weights, $\hat{\tau}_{ps}$, performs better and better. Overall, if weights are important then (1) the bias terms can be too large to be ignored, and (2) there is something to be gained by adjusting the estimates of treatment effects with those weights beyond simple reweighting. Otherwise, SATE estimators are superior, as incorporating weights can be costly.

6 Real Data Application

To better understand the overall trade-offs involved in using weighted estimators of PATE versus simply estimating the SATE on actual survey experiments, we analyzed a set of survey experiments embedded in 7 separate surveys fielded by us though YouGov over the course of 5 years. Studies appeared in two modules of the 2010 CCES, one module each of the 2012 and 2014 CCES, a survey of Virginia voters run prior to the 2013 gubernatorial election in that state, and two other national YouGov surveys. Each survey had *post hoc* weights assigned by YouGov through that firm's standard procedure. Across these surveys were 18 separate assignments of respondents to binary treatments. In several of these cases, multiple outcomes were measured, producing 46 randomization/outcome combinations.[5] All of the studies examined were conducted in the United States and focused on topics related to partisan political behavior.[6] As such, we make our comparisons of SATE and PATE by looking at Democratic and Republican respondents separately. This is because treatment effects for such studies are generally highly heterogeneous by respondent party identification. Our set of 46 randomization/outcome combinations produce 92 experiments (half among Democrats and half among Republicans). Sample sizes for the experiments analyzed range from 145 to 504.[7] Weights varied substantially in these samples, ranging from near 0 to around 8, when normalized to 1 across the sample (standard deviation of 1.04). Sixty five of them (70.7%) showed SATEs that were significantly different from zero. However, once the weights provided by YouGov were taken into account to estimate the PATE (via the double-Hàjek estimate) only 52 experiments (56.5%) had significant effects.

Our first finding is that incorporating weights substantially increased the standard errors. Figure 2 shows a 32.1% average increase in standard error estimates of $\hat{\tau}_{hh}$ over $\hat{\tau}_{\text{SATE}}$ across experiments.

We next examined whether there is evidence of some experiments having a PATE substantially different from the SATE. To do this, we calculated bootstrap estimates of the standard error for the

[5] In the analyses presented here, we use all of these outcomes. We elected to do this because more outcomes provide more opportunity for divergence between SATE and PATE, and thus provide a more conservative test of our conclusion that such divergence is rare. Also, selecting outcomes would represent an added researcher degree of freedom, which we sought to avoid. In the interest of transparency, we present, in Appendix E, the results of our examination when only the primary outcome for each randomization is used. Our findings remain the same.

[6] Some of the studies featured random assignment of campaign advertisements shown to respondents with ad tone and partisan source varied. Several of the studies presented respondents with vignettes or news stories describing candidates or groups of voters with characteristics such as party label and gender randomized. Another set of studies asked respondents to evaluate artwork when told (or not told) that the art was produced by Presidents Bush and Obama. More details regarding the specific studies used can be found in Table 1 in the supplementary materials.

[7] These survey experiments along with the code to replicate the following analyses are publicly available, see Miratrix *et al.* 2017.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

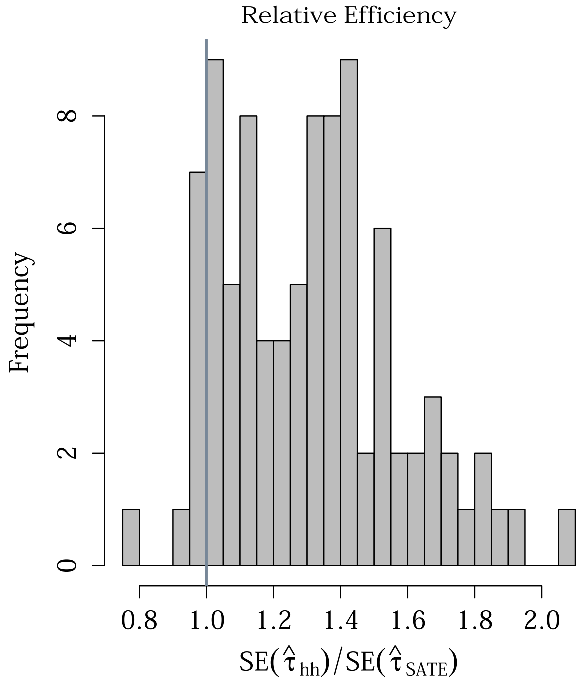

Figure 2. Relative efficiency of $\hat{\tau}_{hh}$ versus $\hat{\tau}_{\mathrm{SATE}}$ of the estimates for the 92 experiments.

difference in the estimators, and calculated a standardized difference in estimates of $\hat{\delta} = (\hat{\tau}_{\mathrm{SATE}} - \hat{\tau}_{hh})/\widehat{SE}$. If there were no difference between the SATE and the PATE, the $\hat{\delta}$s should be roughly distributed as a standard Normal. First, the average of these $\hat{\delta}$ is –0.115, giving no evidence for any systematic trend of the PATE being larger or smaller than the SATE. Second, when we compared our 92$\hat{\delta}$ values to the standard normal with a qq-plot (Figure 3), we find excellent fit. While there is a somewhat suggestive tail departing from the expected line, the bulk of the experiments closely follow the standard normal distribution, suggesting that the SATE and PATE were generally quite similar relative to their estimation uncertainty. A test using q-statistics (modeled after Weiss *et al.* 2017) failed to reject the null of no differences across the experiments ($p > 0.99$); especially considering the possible correlation of outcomes would make this test anticonservative, we have no evidence that the PATE and SATE estimates differ (see Appendix E for further investigation of this). An FDR test also showed no experiments with a significant difference.

Finally, we consider whether poststratification on weights improved precision. Generally, it did not: the estimated SEs of $\hat{\tau}_{ps}$ are very similar to those for $\hat{\tau}_{hh}$, with an average *increase* of about 0.6%. Further examination offers a hint as to why poststratification did not yield benefits: the weights generated by YouGov for these samples do not correlate meaningfully with the outcomes of interest. In no case did the magnitude of the correlation between weights and outcome exceed 0.23. To further explore potential benefits of poststratification we examine the effects of adjusting for a covariate, respondent party identification, on the full experiments not separated by party ID. Relative to no stratification, if we poststratify on party ID, weighting the strata by the total sampled mass in both treatment and control, our PATE estimate shows an average reduction of 1.4% in estimated SEs across experiments with participants of both major parties. If we poststratify on both party ID and the weights, we see an average estimated SEs reduction of 1%. These reductions would in no way make up for the larger SEs from attempting to estimate the PATE.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

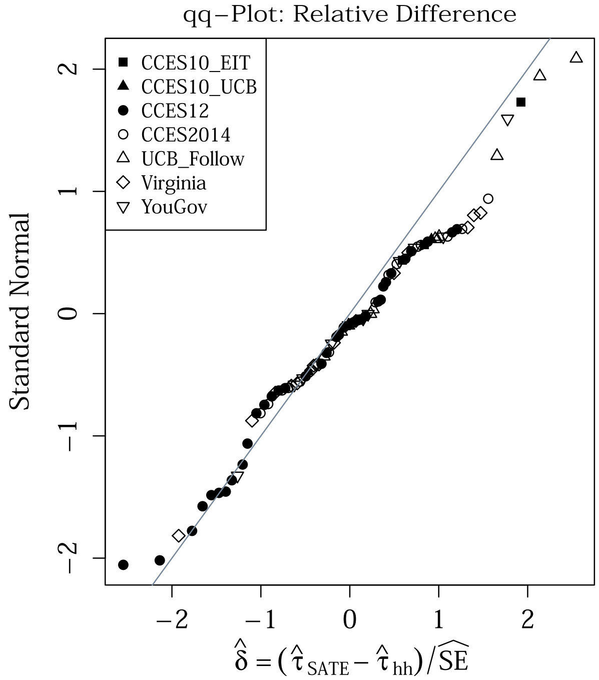

Figure 3. Quantile–quantile plot of the *relative standardized differences* $\hat{\delta}$ of the estimates for the 92 experiments grouped into the larger surveys they are part of.

While this does not imply that scholars need not consider poststratifying on weights, it does show that outcomes of interest in political science studies are not necessarily going to be correlated with these weights. This makes clear the importance of researchers understanding, and reporting, the process used to generate weights and being aware of the covariates with which those weights are likely to be highly correlated (for online surveys, such a list would often includes certain racial and education-level categories).

6.1 Discussion

Overall, it appears that in this context and for these experiments, the survey weights significantly increase uncertainty, and that there is little evidence that the RMSE (which includes the SATE–PATE bias) for estimating the PATE is improved by estimators that include these weights. Furthermore, the weights are not predictive enough of outcome to help the poststratified estimator. With regard to poststratification, we note that in practice the analysis of any particular experiment would likely be improved by poststratifying on known covariates predictive of outcome rather than naïvely on the weights.

In understanding these findings, it is useful to consider the ways in which data from these leading online survey firms (in this case, YouGov) may differ from more convenience-based online samples. Even unweighted, datasets from these firms tend to be more representative. This is because they often engage in extensive panel recruitment and retention efforts and assign subjects from their panels to client samples through mechanisms such as block randomization. As a result, the unweighted data are often largely representative of the overall population along many relevant dimensions. Relatedly, firms may use a clean-up matching step, such as the one employed by YouGov, where they downsample their data to generate more uniform weights (Rivers 2006).

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

This will likely increase the heterogeneity of the final sample, which could decrease precision.[8] We recommend that researchers request the original, preweighted data, in order to work with a larger and more homogeneous sample. For the SATE the gains are immediate. For the PATE, one might generate weights for the full sample by extrapolating from the weights assigned in the trimmed sample or by contracting with the survey firm to obtain weights for this full unmatched sample. Then, by poststratifying on the weights, the researcher can take advantage of the additional units to increase precision in some strata without increasing variability in the others. For both SATE and PATE estimation, power would be improved.

7 Discussion and Practical Guidance

We investigate incorporating weights in survey experiments under the potential outcomes framework. We focus on two styles of estimator, those that incorporate these weights to take any selection mechanisms into account, and those that ignore weights and instead focus on estimating the SATE. We primarily find that incorporating weights, even when they are exactly known, substantially decreases precision. Because of this, researchers are faced with a trade-off: more powerful estimates for the SATE, or more uncertain estimates of the PATE. We conclude with several observations that should inform how one navigates this trade-off.

The PATE can only be different from the SATE when two things hold: (1) there is meaningful variation in the treatment impact, and (2) that variation is correlated with the weights (see Equation (3)). Moreover, the random assignment of treatment protects inference for the weighted estimator, even if the weights are incorrect or known only approximately: because the randomization of units into treatment is independent of the (possibly incorrect) weights, any inference conditional on the sample and the weights is a valid inference. When PATE is the estimand, we are estimating the treatment effect for a hypothetical population defined by the weights and sample, even if it does not correspond to the actual population. For example, if we find a treatment effect in our weighted sample, we know the treatment does have an effect for at least some units. See Hartman *et al.* (2015) for a discussion of this issue in the case of evaluating the external validity of an experiment.

It is important to compare the PATE and SATE estimates. A meaningful discrepancy between them is a signal to look for treatment effect heterogeneity and a flag that weight misspecification could be a real concern. If the estimates do not differ, however, and there is no other evidence of heterogeneity, then extrapolation is less of a concern—and furthermore the SATE is probably a sufficient estimate for the PATE. Of course, with misspecified weights if there is heterogeneity associated with being selected into the experiment that is not captured by the covariates, then PATE estimation can be undetectably biased. For more on assessing heterogeneity, see Ding, Feller, and Miratrix (2018).

[8] Consider a standard scenario wherein a researcher purchases a sample of 1000 respondents. To generate these data, the survey firm might recruit 1400 respondents, all of whom participate in the study. Two datasets result from this. The first contains all 1400 respondents. The second is a trimmed version, where the firm drops 400 of the most overrepresented respondents (which is tantamount to assigning these respondents a weight of 0). This second set, which comes with weights assigned to each observation, is what many scholars analyze. Some firms will, upon request, also provide the full dataset, but these data do not generally include weights, as the process for generating these weights is combined with the procedure for trimming down the larger dataset by matching it to some frame based upon population characteristics. The weights will be less extreme than they would have been had the entire original sample been included, and the trimmed sample will be more heterogeneous, as many similar observations will be purged. This will make it more difficult to estimate its SATE compared to the full set (do note the SATEs could differ). Furthermore, poststratification shows that estimators that include weights for the trimmed set will also be less variable than for the same estimators on the full dataset (assuming weights could be obtained), even though the trimmed dataset weights will be less variable. Consider a case with two classes of respondents, reluctant and eager, equally represented in the population. The trimmed sample will have fewer eager respondents. Then, compared to the full dataset, we will have a less precise estimate of the eager respondents in the trimmed dataset. The precision for the reluctant respondents would be the same. Overall, our combined estimate will be, therefore, less precise.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

Interestingly, our examination of real survey data found no strong connection between the weights and outcomes. The SATE and PATE estimators tended to be similar. Based on this, we have several general pieces of practical guidance: (1) When analyzing survey experiments using high-quality, broadly representative samples, such as those recruited and provided by firms like YouGov and Knowledge Networks, SATE estimates will generally be sufficient for most purposes. (2) If a particular research question calls for estimates of the PATE, a "double-Hàjek" estimator is probably the most straightforward (and a defensible) approach, unless weights are highly correlated with the outcomes variables. (3) If weights are strongly correlated with a study's outcome(s) of interest, poststratification on the weights with bootstrap standard errors can help offset the cost of including weights for those seeking to draw population inferences.

This motivates a two-stage approach. First, focus on the SATE using the entire, unweighted sample and determine whether the treatment had impact. This will generally be the most powerful strategy for detecting an effect, as the weights, being set aside, will not inflate uncertainty estimates. Then, once a treatment effect is established, work on how to generalize it to the population. This second stage is an assessment of the magnitude of an effect in the population once an effect on at least some members of the population has been established. First estimate the PATE with the weights, and then compare it to the SATE estimate. If they differ, then consider working to explain any treatment effect heterogeneity with covariates, and think carefully about weight quality. Regardless, ensure that all analyses preserve the original strength of the assignment mechanism; the weights do not need to jeopardize valid assessment of the presence of causal effects. Part of preserving valid statistical inference would be to commit to a particular procedure before analyzing the given dataset. A preanalysis plan or sample splitting would help prevent a fishing expedition to find treatment effects.

Acknowledgements

For helpful and careful comments, we would like to thank Henry Brady, Devin Caughey, Christopher Fariss, Erin Hartman, Steve Nicholson, Liz Stuart, Chelsea Zhang, and participants of the ACIC 2015 and 2014 Society of Political Methodology Summer Meetings. We would also like to thank Guillaume Basse for his insights into the connection between weighted linear regression and the double-Hàjek estimator. We also thank the valuable feedback and commentary received from two anonymous reviewers and the Editor, who pushed us to clarify and refine our findings.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2018.1.

References

Aronow, P. M., and J. A. Middleton. 2013. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference* 1(1):135–154.

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.

Cochran, W. G. 1977. *Sampling techniques*. 3rd edn. New York: John Wiley and 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.

Ding, P., A. Feller, and L. Miratrix. 2018. Decomposing treatment effect variation. *Journal of the American Statistical Association*, doi:10.1080/01621459.2017.1407322.

DuGoff, E. H., M. Schuler, and E. A. Stuart. 2013. Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research* 49(1):284–303.

Franco, A., G. Simonovits, N. Malhotra, and L. J. Zigerell. 2017. Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science* 4(2):161–172.

Gaines, B. J., J. H. Kuklinski, and P. J. Quirk. 2007. The logic of the survey experiment reexamined. *Political Analysis* 15(1):1–20.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press

Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Hàjek, J. 1958. On the theory of ratio estimates. *Applied Mathematics* 3:384–398.

Hartman, E., R. Grieve, R. Ramsahai, and J. Sekhon. 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 3:757–778.

Hirano, K., G. Imbens, and G. Ridder. 2000. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260):663–685.

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.

Lin, W. 2013. Agnostic notes on regression adjustments to experimental data: reexamining freedman's critique. *The Annals of Applied Statistics* 7(1):295–318.

Miratrix, L. W., J. S. Sekhon, A. G. Theodoridis, and L. F. Campos. 2017. Replication data for: Worth weighting? How to think about and use weights in survey experiments, https://doi.org/10.7910/DVN/52UGJT, Harvard Dataverse, V1.

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* 75(2):369–396.

Mutz, D. C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

O'Muircheartaigh, C., and L. V. Hedges. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63(2):195–210.

Rivers, D. 2006. Sample matching: representative sampling from internet panels. *Polimetrix White Paper Series*.

Santoso, L. P., R. Stein, and R. Stevenson. 2016. Survey experiments with google consumer surveys: promise and pitfalls for academic research in social science. *Political Analysis* 24(3):356–373.

Särndal, C.-E., B. Swensson, and J. Wretman. 2003. *Model assisted survey sampling*. New York: Springer-Verlag.

Si, Y., N. S. Pillai, and A. Gelman. 2015. Bayesian nonparametric weighted sampling inference. *Bayesian Analysis* 10(3):605–625.

Sniderman, P. M. 2011. The logic and design of the survey experiment. In *Cambridge handbook of experimental political science*, ed. J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. Cambridge: Cambridge University Press, p. 102.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5(4):465–472.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2010. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society (Series A)* 174(3):1–18.

Tipton, E. 2013. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38(3):239–266.

Weiss, M. J., H. S. Bloom, N. Verbitsky-Savitz, H. Gupta, A. E. Vigil, and D. N. Cullinan. 2017. How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness* 10(4):843–876.

Wood, J. 2008. On the covariance between related Horvitz–Thompson estimators. *Journal of Official Statistics* 24(1):53–78.

Zheng, H., and R. J. A. Little. 2003. Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics* 19(2):99–117.

https://doi.org/10.1017/pan.2018.1 Published online by Cambridge University Press