# Supplementary Materials for: Worth Weighting? How to Think About and Use Weights in Survey Experiments

Luke W. Miratrix
Harvard University

Jasjeet S. Sekhon
University of California, Berkeley

Alexander G. Theodoridis
University of California, Merced

Luis F. Campos
Harvard University

# Appendix A: A general class of estimators

When estimating the PATE, our overall estimation error is a combination of our error due to the randomized experiment for estimating $\nu_S$ and the difference between our survey-sampling estimate $\nu_S$ and the PATE $\tau$. We can break this error down for any estimator $\hat{\tau}_*$ of $\nu_S$. First, given $\hat{\tau}_*$, we have $\mathbb{E}[\hat{\tau}_*|\mathcal{S}] = \nu_S + b_S$, with $b_S$ being a bias term. Then

$$\begin{aligned} \text{MSE}[\hat{\tau}_*] &= \mathbb{E}\left[(\hat{\tau}_* - \tau)^2\right] \\ &= \mathbb{E}\left[(\hat{\tau}_* - \nu_S)^2\right] + \mathbb{E}\left[(\nu_S - \tau)^2\right] + 2\,\mathbb{E}_S\left[b_S(\nu_S - \tau)\right] \qquad (1) \\ &= \mathbb{E}_\mathcal{S}[\text{MSE}[\hat{\tau}_*|\mathcal{S}]] + \text{MSE}[\nu_S] + 2\,\mathbb{E}_\mathcal{S}\left[b_S(\nu_S - \tau)\right] \end{aligned}$$

Given a choice of $\nu_S$, the first term is the expected MSE of the estimator for estimating $\nu_S$ when we consider all possible *randomizations* of treatment assignment on the given sample $\mathcal{S}$. The second term is the MSE of $\nu_S$ as an estimator for $\tau$ across all samples. The third term is a cross-bias term; it depends on how the bias of a sample is correlated with the error of its $\nu_S$. We generally assume it is small and ignore it. This gives a rough formula for the overall mean square error of

$$\text{MSE}[\hat{\tau}_{hh}] \approx \mathbb{E}_\mathcal{S}[\text{MSE}[\hat{\tau}_{hh}|\mathcal{S}]] + \text{MSE}[\nu_S]\,. \qquad (2)$$

The first term will tend to be a function of the randomization method used and sample-dependent parameters such as $\sigma^2_S(1)$, $\sigma^2_S(0)$, $\sigma^2_S(\Delta)$, and, importantly, the choice of estimator $\hat{\tau}_*$. For a given choice of $\nu_S$, if we reduce this inner term, we reduce the expectation and therefore increase the overall precision of the estimator for PATE. We reduce this term with better estimators, e.g., ones that exploit covariates; this is the goal of post-stratification.

The sampling scheme and choice of $\nu_S$ governs the second term. If we reduce it by changing $\nu_S$, we increase precision. The main way to do this is to sample better, e.g., move closer to equal probability sampling. No estimation strategy can reduce this term.

**Alternate estimators.** Given the above, our primary“double-Hàjek” estimator $\hat{\tau}_{hh}$ can be viewed as doubly biased: the expected value across randomizations is approximately $\nu_{\mathcal{S}}$, and the expected value of $\nu_{\mathcal{S}}$ is approximately $\tau$. We could instead use Horvitz-Thompson style estimators at either or both levels to remove these biases. In particular, if we select an estimator that is unbiased at the randomization level, i.e. $\mathbb{E}[\tau_* | \mathcal{S}] = \nu_{\mathcal{S}}$, then we have

$$\text{MSE}[\hat{\tau}_*] = \mathbb{E}_{\mathcal{S}}[\text{Var}[\hat{\tau}_* | \mathcal{S}]] + \text{Var}_{\mathcal{S}}[\nu_{\mathcal{S}}] + (\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}] - \tau)^2$$

One such estimator is the “single-Hàjek” estimator of

$$\hat{\tau}_h = \frac{1}{Zp} \sum_{i=1}^{N} S_i T_i w_i y_i(1) - \frac{1}{Z(1-p)} \sum_{i=1}^{N} S_i (1 - T_i) w_i y_i(0).$$

This estimator is tied to double-Hàjek by $\mathbb{E}[Z_1 | \mathcal{S}] = pZ$ and $\mathbb{E}[Z_0 | \mathcal{S}] = (1-p)Z$. It is a Horvitz-Thompson estimator with respect to the randomization for the two parts of our estimand $\nu_{\mathcal{S}}$. Interestingly, this estimator has the same asymptotic variance expression found in Theorem 4.1 as $\hat{\tau}_{hh}$.

Finally, if $\mathbb{E}_{\mathcal{S}}[\nu_{\mathcal{S}}] = \tau$ we have

$$\text{MSE}[\hat{\tau}_*] = \mathbb{E}_{\mathcal{S}}[\text{Var}[\hat{\tau}_* | \mathcal{S}]] + \text{Var}_{\mathcal{S}}[\nu_{\mathcal{S}}] \, .$$

For fixed $n$, we have such an estimator as

$$\hat{\tau}_{sd} = \frac{1}{n_1} \sum_{i \in \mathcal{S}} T_i w_i y_i(1) - \frac{1}{n - n_1} \sum_{i \in \mathcal{S}} (1 - T_i) w_i y_i(0). \tag{3}$$

This estimator generally pays a large price for unbiasedness with high variance.

# Appendix B: Post-Stratification for PATE in Survey Experiments

Post-stratification is motivated by viewing PATE estimation as a two step process. In particular, estimators $\nu_S$ that have higher precision will give overall gains. Say we had a categorical covariate $b$ associated with our outcomes. We can then express our overall estimand $\tau$ as:

$$\tau = \sum_{k=1}^{K} \frac{N_k}{N} \left( \frac{1}{N_k} \sum_{i:b_i=k} (y_i(1) - y_i(0)) \right) = \sum_{k=1}^{K} f_k \tau_k$$

with $N_k$ being the number of units in the population in stratum $k$ and $f_k = N_k/N$ being the proportion of the population in stratum $k$. We could then estimate the population $\tau_k$ with strata level estimators of

$$\nu_{Sk} = \frac{1}{Z_k} \sum_{i:b_i=k} w_i (y_i(1) - y_i(0)).$$

As before, we would then need to estimate these $\nu_{Sk}$.

This motivates a post-stratified estimator as a combination of estimates of population strata size estimates and population strata effect estimates:

$$\hat{\tau}_{ps} = \sum_{k=1}^{K} \hat{f}_k \hat{\tau}_k$$

where $\hat{f}_k = Z_k/Z$ estimates $f_k$, with the $Z$ being the total weight in the sample and the

$$Z_k = \sum_{i:b_i=k} w_i S_i \text{ for } k = 1, \ldots, K$$

being the total weights of the strata. These are not dependent on the randomization so

$$\mathbb{E}[\hat{\tau}_{ps} | \mathcal{S}] = \sum_{k=1}^{K} \hat{f}_k \, \mathbb{E}[\hat{\tau}_k | \mathcal{S}] \, .$$

If we had population knowledge we might actually know the $f_k$ and simply plug them in; this connects to the generalization of experiments. See, for example, Tipton (2013).

For the $\tau_k$ we have several options. Arguably the most natural is the double-Hàjek estimator of

$$\hat{\tau}_k = \frac{1}{Z_{k1}} \sum_{i:b_i=k} S_i T_i w_i y_i(1) - \frac{1}{Z_{k0}} \sum_{i:b_i=k} S_i (1-T_i) w_i y_i(0)$$

with $Z_{k1}$ being the total weight in the treatment group in stratum $k$, and similarly for the control. The $\hat{\tau}_k$ will have the usual bias from being Hàjek estimators. Here, however, this bias is of order $n_k$, not $n$ (see Lemma 2.1), and so could potentially be larger than one might expect.

Regardless, combining gives our final

$$\hat{\tau}_{ps} = \sum_{k=1}^{K} \frac{Z_k}{Z} \left( \frac{1}{Z_{k1}} \sum_{i:b_k=k} S_i T_i w_i y_i(1) - \frac{1}{Z_{k0}} \sum_{i:b_k=k} S_i (1-T_i) w_i y_i(0) \right). \tag{4}$$

If we want to avoid this bias, we could instead use a single-Hàjek estimator in each strata:

$$\hat{\tau}_k^{(h)} = \frac{n_k}{Z_k} \left( \frac{1}{n_{Tk}} \sum_{i:b_i=k} T_i w_i y_i(1) - \frac{1}{n_k - n_{Tk}} \sum_{i:b_i=k} (1-T_i) w_i y_i(0) \right).$$

For the single-Hàjek, we immediately have $\mathbb{E}\left[\hat{\tau}_k^{(h)} | \mathcal{S}\right] = \nu_{S_k}$, i.e., unbiasedness in the randomization step. This also causes the $Z_k$ to cancel. If the weights within strata are generally homogenous, the single-Hàjek will be essentially the same as the double. And if $b$ is built by stratifying on weights then we would indeed expect such homogeneity. Thus, with post-stratification, we can remove some bias for very little cost in variance.

## Variance Estimation

As discussed in the main text, the post-stratification step can be sample-dependent. For example, if 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 pre-randomization, this does not impact the validity of the variance and variance-estimation formulae of the SATE estimate of $\tau_S$. It does, however, make generating appropriate population variance formulae difficult. Furthermore, even if the strata are pre-defined, the formulae of Theorem 4.1 are actually for a linearized version of the ratio estimators, and as the strata are smaller than the overall sample, one might be concerned that these approximations would be not that good when applied to individual strata. This is why we propose the bootstrap.

Appropriate implementation of the bootstrap deserves some discussion. Bootstrap is a "by analogy" technique. To obtain the variability of an estimator we repeatedly simulate obtaining a sample from some population using our hypothesized sampling mechanism, randomizing it into treatment, and estimating the treatment effect using our estimator on that sample. We first, therefore, need to have a population to sample from. Our best estimate of this population is the sample weighted by the weights. We then take a size-$n$ i.i.d. sample from this population with probability proportional to the inverse of these weights. The treatment assignment being Bernoulli means we take a case-wise bootstrap, bootstrapping the original treatment assignment along with the outcome. This avoids any need to impute any missing potential outcomes.

The up-weighting and subsequent weighted sampling steps collapse to generating a bootstrap sample by taking a classic with-replacement unweighted sample (i.e., a case-wise bootstrap) from the original sample of the triples $(Y_i^{obs}, Z_i, w_i)$.

# Appendix C: Derivations

In the following we derive the bias of the Hàjek estimator, show that it is small, and derive the bias of $\tau_{SATE}$ as an estimator for the PATE. After this we show how a weighted OLS regression can be used in practice to estimate the double-Hàjek. Finally, we derive properties of the unstratified PATE estimators.

## Bias of the Hàjek Estimator

The proof of Lemma 2.1, that the bias of a Hàjek estimator is $O(1/\mathbb{E}[n])$, follows a similar strategy to the proof of Result 6.34 in Cochran (1977). That result is of the bias of a general ratio estimator for a fixed sample size under simple random sampling. We adapt this result to the Hàjek estimator (also a ratio estimator) under independent Poisson random sampling with variable sample size. A fixed sample size correction is possible, but is not needed for our purposes.

We extend the notation described in Section 2.1. Denote $Z_y = \sum_{i=1}^{N} \frac{\bar{\pi}}{\pi_i} S_i y_i$ so that we can write $\hat{y}_H = \frac{Z_y}{Z}$. The expected values of both the numerator and denominator are

$$\begin{aligned} \mathbb{E}[Z_y] &= N\bar{\pi}\mu, \qquad (5) \\ \mathbb{E}[Z] &= N\bar{\pi}. \end{aligned}$$

These results alone should motivate why the Hàjek estimator should be approximately unbiased, but let us be a bit more rigorous. By first manipulating the difference of the estimator and its target and then applying the first order Taylor approximation, $(1 + A)^{-1} \doteq (1 - A)$, we can get the approximate difference.

$$\begin{aligned} \hat{y}_H - \mu = \frac{Z_y}{Z} - \mu = \frac{Z_y - \mu Z}{Z} &= (Z_y - \mu Z)\frac{1}{Z} \\ &= (Z_y - \mu Z)\frac{1}{N\bar{\pi}}\frac{N\bar{\pi}}{Z} = (Z_y - \mu Z)\frac{1}{N\bar{\pi}}\left(\frac{Z}{N\bar{\pi}}\right)^{-1} \\ &= (Z_y - \mu Z)\frac{1}{N\bar{\pi}}\left(\frac{N\bar{\pi} + (Z - N\bar{\pi})}{N\bar{\pi}}\right)^{-1} \\ &= (Z_y - \mu Z)\frac{1}{N\bar{\pi}}\left(1 + \frac{Z - N\bar{\pi}}{N\bar{\pi}}\right)^{-1} \\ &\doteq (Z_y - \mu Z)\frac{1}{N\bar{\pi}}\left(1 - \frac{Z - N\bar{\pi}}{N\bar{\pi}}\right) \end{aligned}$$

Taking expectations and noting that $\mathbb{E}[Z_y - \mu Z] = 0$ by Equation 5 leads to the approximate bias:

$$\begin{aligned}\mathbb{E}[\hat{y}_H] - \mu &\doteq \, - \frac{1}{(N\bar{\pi})^2}\,\mathbb{E}\left[(Z_y - \mu Z)(Z - N\bar{\pi})\right] \\ &= -\frac{1}{(N\bar{\pi})^2}\left(\mathbb{E}[Z_y Z] - N\bar{\pi}\,\mathbb{E}\left[Z_y\right] + N\bar{\pi}\mu\,\mathbb{E}\left[Z\right] - \mu\,\mathbb{E}\left[Z^2\right]\right).\end{aligned} \tag{6}$$

These expanded terms can be calculated individually for our estimator using properties of variance and covariance.

$$\begin{aligned}\mathbb{E}[Z_y Z] &= Cov(Z_y, Z) + \mathbb{E}[Z_y]\,\mathbb{E}\left[Z\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\bar{\pi}^2}{\pi_i \pi_j} y_i\, Cov(S_i, S_j) + (N\bar{\pi}\mu)(N\bar{\pi}) \\ &= \sum_{i=1}^{N}\frac{\bar{\pi}^2}{\pi_i^2} y_i\, Var(S_i) + N^2\bar{\pi}^2\mu \\ &= \bar{\pi}^2\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i} y_i + N^2\bar{\pi}^2\mu \\ &= \bar{\pi}^2\sum_{i=1}^{N}\frac{y_i}{\pi_i} - N\bar{\pi}^2\mu + N^2\bar{\pi}^2\mu\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbb{E}[Z^2] &= Var(Z) + \mathbb{E}[Z]^2 \\ &= \sum_{i=1}^{N}\frac{\bar{\pi}^2}{\pi_i^2} var(S_i) + N^2\bar{\pi}^2 \\ &= \bar{\pi}^2\sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right) + N^2\bar{\pi}^2 \\ &= \bar{\pi}^2\sum_{i=1}^{N}\frac{1}{\pi_i} - N\bar{\pi}^2 + N^2\bar{\pi}^2\end{aligned} \tag{8}$$

Finally, substitute Equations 5, 7 and 8 into Equation 6 and simplify:

$$
\begin{aligned}
\mathbb{E}[\hat{y}_H] - \mu \doteq -\frac{1}{(N\bar{\pi})^2}\Bigg(&\bar{\pi}^2 \sum_{i=1}^{N} \frac{y_i}{\pi_i} - N\bar{\pi}^2\mu + N^2\bar{\pi}^2\mu \\
&- N^2\bar{\pi}^2\mu + N^2\bar{\pi}^2\mu \\
&- \mu\bar{\pi}^2 \sum_{i=1}^{N} \frac{1}{\pi_i} + N\bar{\pi}^2\mu + N^2\bar{\pi}^2\Bigg) \\
= -\frac{1}{N\bar{\pi}}&\left(\bar{\pi}\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i} - \mu\bar{\pi}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}\right) \\
= -\frac{1}{\mathbb{E}[n]}&\left(\bar{\pi}\frac{1}{N}\sum_{i=1}^{N}\frac{y_i - \mu}{\pi_i}\right) \\
= -\frac{1}{\mathbb{E}[n]}&\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)\frac{\bar{\pi}}{\pi_i}\right).
\end{aligned}
$$

We finally use the relation

$$
\mathrm{Cov}[A, B] = \mathbb{E}\big[(A - \bar{A})(B - \bar{B})\big] = \mathbb{E}\big[(A - \bar{A})B\big] - \mathbb{E}\big[(A - \bar{A})\bar{B}\big] = \mathbb{E}\big[(A - \bar{A})B\big]
$$

to get our final covariance formulation.

We have ignored a mild technical issue of an undefined estimator with probability $\mathbf{P}\{Z = 0\}$. For the Poisson selection scheme, with the $S_i$ independent, $\mathbf{P}\{Z = 0\} = \prod(1 - \pi_i)$ which will be exponentially small in $n$. Letting the estimator be defined as 0 under this circumstance gives a bounded, exponentially small term far less in magnitude than other bias terms.

## Bias of the SATE for the PATE

To see that $\hat{\tau}_{SATE}$ (or $\tau_{\mathcal{S}}$) is a biased estimate for PATE, assume fixed sample size $n$ to obtain:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{SATE}] = \mathbb{E}_{\mathcal{S}}[\mathbb{E}[\hat{\tau}_{SATE}|\mathcal{S}]] = \mathbb{E}_{\mathcal{S}}[\tau_{\mathcal{S}}] = \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{N} S_i(y_i(1) - y_i(0))\right] \\ = \frac{1}{N}\sum_{i=1}^{N}\frac{N\pi_i}{n}\left(y_i(1) - y_i(0)\right).\end{aligned}$$

For a random sample size, there is an additional, but negligible, a bias term. We can see that the above is a first order approximation of the overall bias by replacing $\mathbb{E}_{\mathcal{S}}\left[{S_i}/{n}\right]$ with ${\mathbb{E}_{\mathcal{S}}[S_i]}/{\mathbb{E}_{\mathcal{S}}[n]}$. The difference in these terms is of order $1/n$, as with our bias lemma.

## The double-Hàjek as weighted OLS

In Section 4.1 we introduced the "double-Hàjek" estimator. Here we show that this estimate is equivalent to a weighted OLS where the weights are $w_i = \frac{\bar{\pi}}{\pi_i}$ and we regress on the treatment indicator. In other words we fit the model

$$y_i = \alpha + \tau T_i + \varepsilon_i$$

with weights $w_i$. The weighted OLS estimates $\hat{\alpha}$ and $\hat{\tau}$ are the solutions to the normal equations:

$$\sum_{i \in S} w_i(y_i - \hat{\alpha} - \hat{\tau}T_i) = 0, \qquad (9)$$

$$\sum_{i \in S} w_i T_i(y_i - \hat{\alpha} - \hat{\tau}T_i) = 0. \qquad (10)$$

These are obtained by taking derivatives with respect to $\alpha$ and $\tau$ of the weighted sum of squares, $\sum_{i \in S} w_i(y_i - \alpha - \tau T_i)^2$, and setting them to $0$. Grouping by treatment indicators, we get the following:

$$\sum_{i:T_i=1} w_i(y_i - \hat{\alpha} - \hat{\tau}) + \sum_{i:T_i=0} w_i(y_i - \hat{\alpha}) = 0,$$
$$\sum_{i:T_i=1} w_i(y_i - \hat{\alpha} - \hat{\tau}) = 0.$$

Taking the difference of these equations implies that

$$\hat{\alpha} = \frac{\sum_{i:T_i=0} w_i y_i}{\sum_{i:T_i=0} w_i}.$$

To make the connection to the "double-Hàjek" estimate, denote $Z_0 = \sum_{i:T_i=0} w_i$ and $Z_1 = \sum_{i:T_i=1} w_i$, as before. If we distribute the summation in the second normal equation (Equation 10), we get

$$\sum_{i:T_i=1} w_i y_i - \hat{\alpha} Z_1 - \hat{\tau} Z_1 = 0$$
$$\sum_{i:T_i=1} w_i y_i - \frac{Z_1}{Z_0} \sum_{i:T_i=0} w_i y_i - \hat{\tau} Z_1 = 0$$
$$\hat{\tau} = \frac{1}{Z_1} \sum_{i:T_i=1} w_i y_i - \frac{1}{Z_0} \sum_{i:T_i=0} w_i y_i$$

Written in the most general sense and replacing the weights, we get back our "double-Hàjek" estimate.

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

Hence one way of calculating $\hat{\tau}_{hh}$ is by fitting a weighted OLS regression onto the treatment indicator and inspecting the coefficients.

## Properties of $\hat{\tau}_{hh}$

Our estimator can be expressed as

$$\begin{aligned}\hat{\tau}_{hh} &= \frac{1}{Z_1}\sum_{i=1}^{N} S_i T_i \frac{\bar{\pi}}{\pi_i} y_i(1) - \frac{1}{Z_0}\sum_{i=1}^{N} S_i(1-T_i)\frac{\bar{\pi}}{\pi_i} y_i(0) \\ &= \hat{\mu}(1) - \hat{\mu}(0).\end{aligned}$$

For the expectation of $\hat{\tau}_{hh}$, we have

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{hh}|\mathcal{S}] &\approx \mathbb{E}\left[\frac{1}{\mathbb{E}[Z_1|\mathcal{S}]}\sum_{i=1}^{N} S_i T_i \frac{\bar{\pi}}{\pi_i} y_i(1) - \frac{1}{\mathbb{E}[Z_0|\mathcal{S}]}\sum_{i=1}^{N} S_i(1-T_i)\frac{\bar{\pi}}{\pi_i} y_i(0)|\mathcal{S}\right] \\ &= \frac{1}{Z}\sum_{i=1}^{N} S_i \frac{\bar{\pi}}{\pi_i} y_i(1) - \frac{1}{Z}\sum_{i=1}^{N} S_i \frac{\bar{\pi}}{\pi_i} y_i(0) = \nu_{\mathcal{S}}\end{aligned}$$

For variance we use results and notation from Särndal, Swensson and Wretman (2003) to obtain approximate variance terms as follows. Define $\tilde{S}_i = S_i T_i$ as the event of unit $i$ being selected and also treated. We then have $\tilde{\pi}_i = \mathbb{E}\left[\tilde{S}_i\right] = p\pi_i$ and the probability that units $j$ and $k$ are both selected and treated is

$$\tilde{\pi}_{jk} = \mathbb{E}\left[\tilde{S}_j = 1 \text{ and } \tilde{S}_k = 1\right] = \mathbf{P}\{T_j = 1 \text{ and } T_k = 1 | S_j = 1, S_k = 1\}\, \pi_{jk}$$

For the treatment group specifically we have

$$\hat{\mu}(1) = \frac{\bar{\pi}\sum_{i=1}^{N} S_i T_i \frac{y_i(1)}{p\pi_i}}{\bar{\pi}\sum_{i=1}^{N} S_i T_i \frac{1}{p\pi_i}} = \frac{\sum_{i=1}^{N} S_i T_i \frac{y_i(1)}{p\pi_i}}{\sum_{i=1}^{N} S_i T_i \frac{r_i}{p\pi_i}} = \frac{\sum_{\tilde{S}} \check{y}_i}{\sum_{\tilde{S}} \check{r}_i} = \frac{\hat{t}_y}{\hat{t}_r}.$$

with $r_i = 1$. The check notation denotes a value divided by its probability of being included in the sample: $\check{a}_i = a_i/\pi_i$. The above is a classic ratio estimator with selection probabilities of $\tilde{\pi}_j$ for the ratio of

$$R = \frac{t_y}{t_r} = \frac{\sum_{i=1}^N y_i(1)}{\sum_{i=1}^N r_i} = \frac{\sum_{i=1}^N y_i(1)}{N} = \mu(1)$$

since $t_r = \sum_{i=1}^N r_i = N$.

The approximate variance of a ratio estimator (Särndal, Swensson and Wretman, 2003) is:

$$\begin{aligned} AV(\hat{\mu}(1)) &= \frac{1}{t_r^2} \sum_{j=1}^N \sum_{k=1}^N \tilde{\Delta}_{jk} \frac{y_j(1) - Rr_j}{\tilde{\pi}_j} \frac{y_k(1) - Rr_k}{\tilde{\pi}_k} \\ &= \frac{1}{N^2} \sum \sum \tilde{\Delta}_{jk} \frac{y_j(1) - \mu(1)}{p\pi_j} \frac{y_k(1) - \mu(1)}{p\pi_k} \\ &= \frac{1}{N^2} \sum \sum \frac{\tilde{\Delta}_{jk}}{p^2 \pi_j \pi_k} \left(y_j(1) - \mu(1)\right) \left(y_k(1) - \mu(1)\right) \end{aligned}$$

with

$$\tilde{\Delta}_{jk} \equiv \tilde{\pi}_{jk} - \tilde{\pi}_j \tilde{\pi}_k = \tilde{\pi}_{jk} - p^2 \pi_j \pi_k.$$

We can estimate this variance with a sum over the treatment group of

$$\widehat{V}(\hat{\mu}(1)) = \frac{1}{\widehat{N}^2} \sum_{j=1}^N \sum_{k=1}^N S_j T_j S_k T_k \frac{\tilde{\Delta}_{jk}}{\tilde{\pi}_{jk} p^2 \pi_j \pi_k} \left(y_j(1) - \hat{\mu}(1)\right) \left(y_k(1) - \hat{\mu}(1)\right)$$

with $\hat{\mu}(1) = \frac{1}{\widehat{N}} \sum_{i=1}^N S_i T_i \breve{y}_i(1)$ and $\widehat{N} = \sum S_i T_i / \pi_i p$.

**The Poisson-Bernoulli Model.** Under Poisson selection we have $\pi_{jk} = \pi_j \pi_k$ for $j \neq k$ (with $\pi_{jj} = \pi_j$). With Bernoulli assignment we have $\tilde{\pi}_{jk} = p^2 \pi_j \pi_k$ for $j \neq k$ (with $\tilde{\pi}_{jj} = p\pi_j$) giving

$\tilde{\Delta}_{jk} = 0$ for $j \neq k$ and $\tilde{\Delta}_{jj} = p\pi_j(1 - p\pi_j)$ for $j = k$. This gives

$$AV(\hat{\mu}(1)) = \frac{1}{N^2} \sum_{j=1}^{N} \frac{1 - p\pi_j}{p\pi_j} \left(y_j(1) - \mu(1)\right)^2$$

and

$$\widehat{V}(\hat{\mu}(1)) = \frac{1}{\widehat{N}^2} \sum_{j=1}^{N} S_j T_j \frac{1 - p\pi_j}{p^2\pi_j^2} \left(y_j(1) - \hat{\mu}(1)\right)^2 .$$

The above formula are problematic in that they depend on our $\pi_j$ rather than the weights $w_j = \bar{pi}/\pi_j)$. However, if we assume $N \gg n$ we can make progress. In particular, in this case, under mild regularity conditions on the sampling probabilities, we can assume $\pi_j \ll 1$ for all $j$. This means that $1 - p\pi_j \approx 1$. Couple this with $N\bar{\pi} = \mathbb{E}[n]$ to get a fairly tight upper bound on our two formula of

$$AV(\hat{\mu}(1)) \leq \frac{1}{p\,\mathbb{E}[n]} \frac{1}{N} \sum_{j=1}^{N} w_j \left(y_j(1) - \mu(1)\right)^2$$

and, using $\widehat{N} = Z_1/(\bar{\pi}p)$ with $Z_1 = \sum S_j T_j w_j$,

$$\begin{aligned}\widehat{V}(\hat{\mu}(1)) &= \frac{\bar{\pi}^2}{Z_1^2} \sum_{j=1}^{N} S_j T_j \frac{1 - p\pi_j}{\pi_j^2} \left(y_j(1) - \hat{\mu}(1)\right)^2 \\ &\leq \frac{1}{Z_1^2} \sum_{j=1}^{N} S_j T_j w_j^2 \left(y_j(1) - \hat{\mu}(1)\right)^2 .\end{aligned}$$

Finally, to get overall variance presented in Theorem 4.1 we first view the sample into the treatment arm as independent of the sample into the control arm, which is again motivated by the $N \gg n$ assumption. For the control arm, we then do the above derivation with $\tilde{S}_i = S_i(1 - T_i)$ and $\tilde{\pi}_i = (1 - p)\pi_i$. More lengthy derivations that account for the dependence structure will give higher-order terms which are in the end negligible. See Wood (2008) for an approach.

# Appendix D: The simulation's DGP

In this section we provide additional simulation details and explanations of some of the choices made throughout the simulations of Section 5. In all our simulations, the potential outcomes are simulated as nonlinear functions of the weights.

To generate our populations we use the following algorithm: let $\gamma \in [0,1]$ be a correlation measuring the strength of the relationship between the weights and outcomes. We then generate two latent parameters $(\varepsilon_i, \tilde{\varepsilon}_i)$ as a bivariate standard normal draw with correlation $\gamma$. (We do this by generating $\varepsilon_i \sim N(0,1)$, and $\tilde{\varepsilon} = \gamma\varepsilon_i + \sqrt{1-\gamma^2}\eta_i$, with $\eta_i \sim N(0,1)$.)

We then generate uniformly distributed weights on pre-specified interval $(a,b)$ by using the c.d.f. transformation:

$$w_i = a + b\,\Phi(\varepsilon_i),$$

where $\Phi$ is the standard normal c.d.f. We also generate shadow weights

$$\tilde{w}_i = a + b\,\Phi(\tilde{\varepsilon}_i),$$

also uniform, and with the same distribution as $w_i$.

Our potential outcomes are then a function of the shadow weights $\tilde{w}_i$:

$$Y_i(0) = 120 - 20\sqrt{\tilde{w}_i} + 5\epsilon_i$$
$$Y_i(1) = Y_i(0) + 10\sqrt{b - \tilde{w}_i}$$

with $\epsilon_i$ as independent Gaussian noise. The treatment potential outcomes are generated to give a non-linear heterogeneous treatment effect. When $\gamma = 1$, $\tilde{w}_i = w_i$, giving the strongest possible relationship between outcome and weight. Conversely when $\gamma = 0$ the weights are completely unrelated to the potential outcomes, so stratifying on them should not help improve estimation.

Once we have a population, we then sample inversely proportional to the weight $w_i$. For example, in Simulation A we take a fixed sample size of $n = 500$ (5% of the population). Our post-stratification estimator stratifies based on the weight $w_i$ to increase precision. The stratifying variable $b_i$ is defined in Section 4.3.

Simulation A has maximal covariance, with $\gamma = 1$. Figure 1 shows a subset of the population and a sample from this scenario to illustrate the structure of our DGP. Figure 1a shows the characteristics of the simulated population while Figure 1b shows how a weighted sample might look.

Overall, Figure 1 shows that the weight $w_i$ and potential outcome distributions differ in the sample and population. Furthermore, because the potential outcomes are related to the weights they are consequently related to the post-stratification levels $b_i$ in the sample.

For Simulation B we simply replace the formula for $Y_i(1)$ with a constant treatment effect of 30, so $Y_i(1) = Y_i(0) + 30$. We still have the sample general relationships between the sample and population, but as we see in Section 5 the estimators behave quite differently.

For Simulation C we varied $\gamma$, which controls the relationship between the weight and the potential outcomes. The top two right-most panels of Figure 1b show there is smaller variability within strata for $Y_i(0)$ and $Y_i(1)$ than if we consider the entire sample at once. As our weights become less predictive of outcome, this variability will increase. Our formulation, however, maintains the marginal distributions of $w_i$, $Y_i(0)$, and $Y_i(1)$ as $\gamma$ changes so that any benefits we see from post-stratification can only be attributed to the changing relationship.

# Appendix E: Further Details and Results of the Real Data Application

As mentioned in the main text, the 92 survey experiments analyzed in Section 6 were generated from 18 unique randomizations on 7 separate surveys. We split each randomization by subject party identification and considered multiple outcomes per treatment randomization. One might worry that the potential correlation of the multiple outcomes might be influencing the results, so we append here the results when considering only one unique outcome per randomization.

The 18 unique randomizations give rise to 36 survey experiments after splitting each randomization by subject party identification, considering only the larger Democratic and Republican leaning subgroups. 28 of them (78%) showed SATEs that were significantly different from zero. Once the weights were taken into account to estimate the PATE (via the double-Hàjek estimate) $25$ experiments ($69\%$) had significant effects. Even though more experiments showed significant PATE than SATE estimates, incorporating weights still increased standard errors: there was a $31.6\%$ average increase in variance of $\hat{\tau}_{hh}$ over $\hat{\tau}_{SATE}$ across experiments. The raw SE increases can be seen in Figure 2(a).

We further examined whether there is evidence of some experiments having a PATE substantially different from the SATE. We calculated the 36 $\hat{\delta}$ values and compared them to a standard normal with a qq-plot (Figure 2(b)). While visually there do seem to be some distributional departures from a standard Normal, a KolmogorovSmirnov test does not support this hypothesis (with a p-value of $0.14$). Furthermore, an FDR test also fails to find any experiments with significant differences. All of this suggests a general equivalence between the SATE and the PATE in this subset of experiments as well.

To explore whether post-stratification on weights improved precision, we compared the estimated SEs. The estimated SEs of $\hat{\tau}_{ps}$ are very similar to those for $\hat{\tau}_{hh}$, with an average increase of about $0.2\%$. Post-stratifying on party ID on the original 18 experiments led to modest variance reduction. Relative to no stratification, we see an average reduction of $2.6\%$ in variance across experiments with participants of both major parties. If we post-stratify on both party ID and the weights, we see an average reduction of $2.3\%$. These findings, similar to the main text, show that while post-stratification should help reduce the variance in theory the gains can be rather modest in practice.

| Survey | Year | Geography | Base Stimulus | Conditions | Outcome(s) | N Democrats | N Republicans |
|---|---|---|---|---|---|---|---|
| CCES Module 1 | 2010 | National | News Report | Democratic/Republican Candidate | Report Fair; Report Biased; Topic Important; Candidate Deserves Credit; Candidate Typical | 231 | 211 |
| CCES Module 1 | 2010 | National | Image | Democratic Donkey/Republican Elephant | Unemployment Rate Estimate | 190 | 219 |
| CCES Module 2 | 2010 | National | Image | Democratic Donkey/Republican Elephant | Unemployment Rate Estimate | 201 | 193 |
| YouGov Study | 2011 | National | News Report | Democratic/Republican Candidate | Report Fair; Report Biased; Topic Important; Candidate Deserves Credit; Candidate Typical | 442 | 372 |
| CCES | 2012 | National | Campaign Advertisement Video | Obama/Romney | Time watched; Repeat; Share; See More | 326 | 228 |
| CCES | 2012 | National | Campaign Advertisement Video | Negative/Positive | Time watched; Repeat; Share; See More | 326 | 228 |
| CCES | 2012 | National | Candidate Vignette | Democratic/Republican Label | Trait Ratings: Compassionate; Moral; Strong Leader; Really Cares; Knowledgeable; Greedy; Indecisive; Hard Working; Honest | 321 | 225 |
| CCES | 2012 | National | Voter Fraud Hypothetical | Democrats/Republicans | Would this group commit fraud? | 223 | 145 |
| Gubernatorial Election | 2013 | Virginia | Campaign Advertisement Video | McCauliffe/Cuccinelli | Time watched; Repeat; Share; See More | 454 | 350 |
| Gubernatorial Election | 2013 | Virginia | Campaign Advertisement Video | Negative/Positive | Time watched; Repeat; Share; See More | 454 | 350 |
| YouGov Study | 2013 | National | News Report | Democratic/Republican Candidate | Report Fair; Report Biased; Topic Important; Candidate Deserves Credit; Candidate Typical | 456 | 353 |
| CCES | 2014 | National | Candidate Conjoint 1 | Male/Female Candidate | Is candidate more likely a Democrat or Republican? | 504 | 330 |
| CCES | 2014 | National | Candidate Conjoint 2 | Male/Female Candidate | Is candidate more likely a Democrat or Republican? | 504 | 330 |
| CCES | 2014 | National | Candidate Conjoint 3 | Male/Female Candidate | Is candidate more likely a Democrat or Republican? | 504 | 330 |
| CCES | 2014 | National | Candidate Conjoint 4 | Male/Female Candidate | Is candidate more likely a Democrat or Republican? | 504 | 330 |
| CCES | 2014 | National | Painting by George W. Bush | Bush Revealed as Artist/Not Revealed | Rating of Painting Quality | 504 | 329 |
| CCES | 2014 | National | Sketch by Barack Obama | Obama Revealed as Artist/Not Revealed | Rating of Sketch Quality | 394 | 278 |
| CCES | 2014 | National | News Story about Stampede at July 4 Gathering | Democratic/Republican Event | In-Party Shame | 338 | 204 |

Table 1: Details for Studies Used

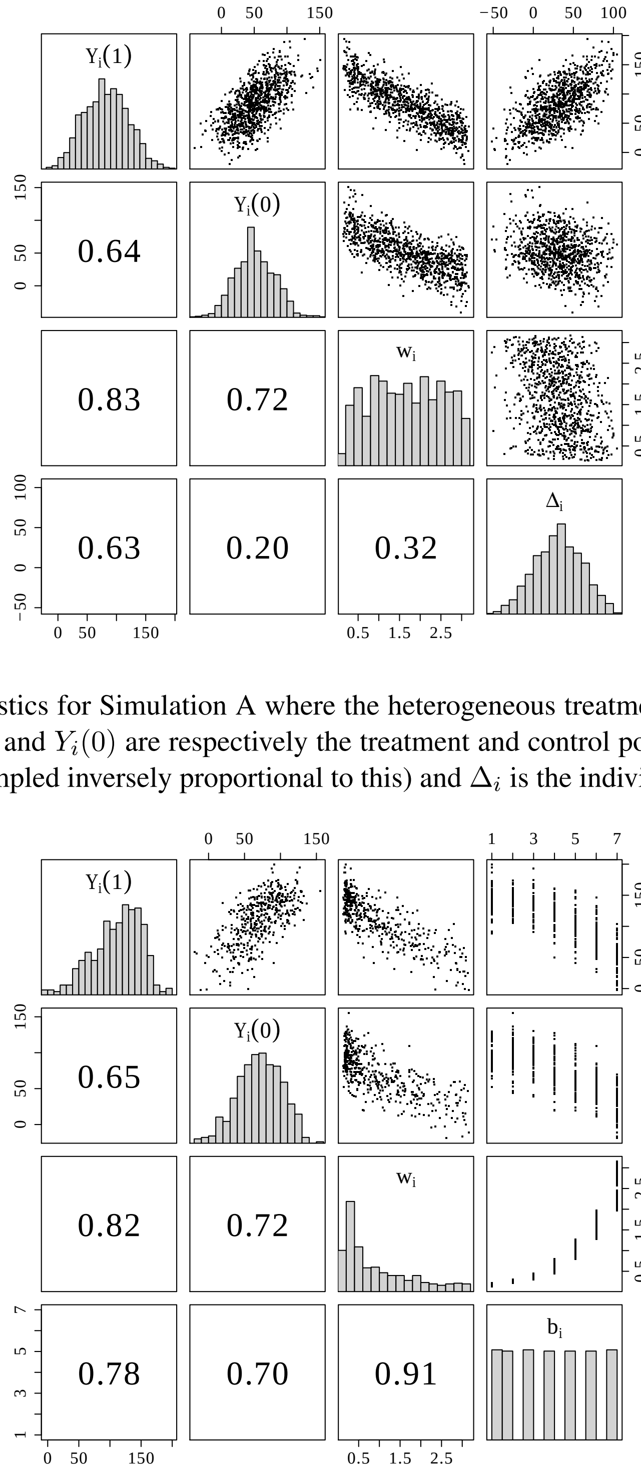
# References

Cochran, William G. 1977. *Sampling Techniques, 3rd Edition.* New York: John Wiley and Sons.

Särndal, Carl-Erik, Bengt Swensson and Jan Wretman. 2003. *Model assisted survey sampling*. Springer.

Tipton, Elizabeth. 2013. "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38(3):239–266.

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



(a) Population characteristics for Simulation A where the heterogeneous treatment effect varies in connection to the weight. $Y_i(1)$ and $Y_i(0)$ are respectively the treatment and control potential outcomes, $w_i$ is the unit weight (units are sampled inversely proportional to this) and $\Delta_i$ is the individual treatment effect.

(b) Characteristics of a sample from Simulation A. $b_i$ is the post-stratification generated on this particular sample.

Figure 1: Characteristics of the Population and a Sample from Simulation A

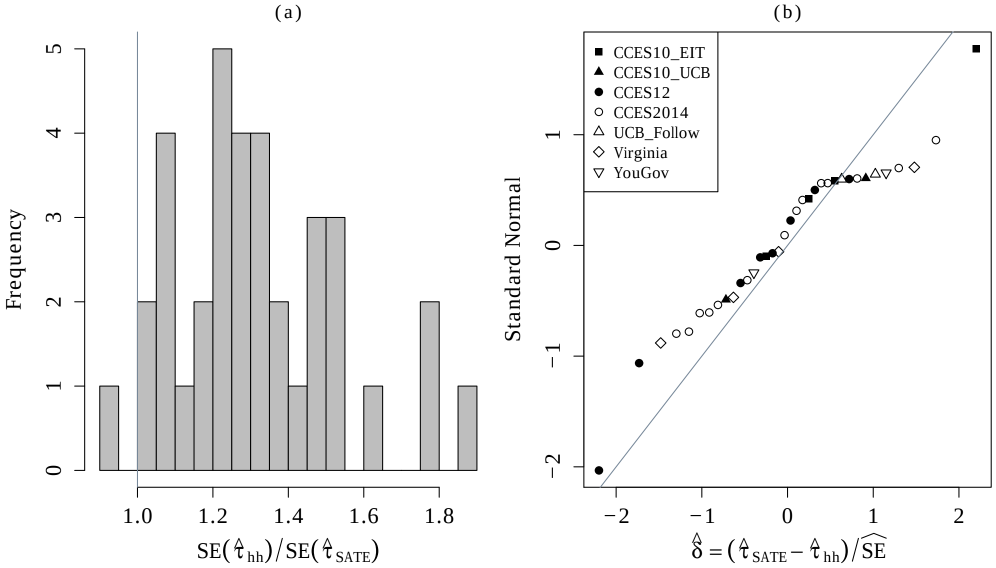


Figure 2: (a) Standardized efficiency of estimates of $\hat{\tau}_{hh}$ vs $\hat{\tau}_{SATE}$. (b) quantile-quantile comparison plot of the *relative difference* $\hat{\delta}$ of the estimates for the 36 experiments grouped by containing survey.