Online Appendix A: Generalizing from the SATE to the PATE

A General Framework

Consider the case in which we sample n units from a larger finite population of size N. S, R, and T are random variables that represent sample selection, survey response, and treatment assignment, respectively. Let $S_i = 1$ if unit i is in the sample, and $S_i = 0$ if it is not. Similarly, let $R_i = 1$ if unit i responds to the survey and $R_i = 0$ if it does not. Finally, let $T_i = 1$ if unit i receives an experimental treatment and $T_i = 0$ if it does not.

The potential outcomes variables Y_{it} represent the value of the outcome of interest when $T_i = t$ for $t = 0, 1$. We define the *treatment effect* for unit i as the difference in unit-level outcomes under treatment and control,

$$\tau_i \equiv Y_{i1} - Y_{i0}, \tag{1}$$

and the *population average treatment effect* $(PATE)$ as these individual treatment effects averaged over all units in the population,

$$\bar{\tau} \equiv \frac{1}{N} \sum_{i=1}^{N} \tau_i. \tag{2}$$

Not all sampled individuals will respond to a survey. Under a stochastic non-response model, we assume that units i have an unknown, nonzero probability of responding ϕ_i.[1] When i is selected at random from the population, it responds with probability ϕ_i and does not respond with probability $1 - \phi_i$. The response indicator R_i is observed for sampled units only, and the realized sample size is

$$n_r = \sum_{i=1}^{N} S_i R_i.$$

[1] Note that $\phi_i = 0$ corresponds to permanent exclusion of a unit from the sample. This noncoverage can arise due to sampling, i.e. units that share some characteristic(s) $\mathbf{x_i} = \mathbf{x}$ are never sampled, or due to selection, i.e. the units never opt-in to the survey. Weighting methods can reduce noncoverage error in population estimates when at least some units with characteristic(s) $\mathbf{x_i} = \mathbf{x}$ are sampled and respond to the survey, i.e. $P(R = 1|\mathbf{x_i} = \mathbf{x}) = \phi_i > 0$. Hence, all units in a target population must have a nonzero probability of being selected and responding in order to recover an unbiased estimate for that population. When this assumption is not satisfied, researchers should redefine the target population to that for which $0 < \phi_i \leq 1$.

Further, we only observe response under treatment and control, Y_{it}, for units that respond. We now define the *sample average treatment effect* ($SATE$) as the individual-level treatment effects averaged over units in the realized sample,

$$\bar{\tau}_r \equiv \frac{1}{n_r} \sum_{i \in \{R_i = 1\}}^{n_r} \tau_i. \tag{3}$$

In general, the expected value of this quantity is not equal to $\bar{\tau}$ (Bethlehem 1988). Rather, the expected value of the$SATE$ is a biased estimate of the $PATE$,

$$\mathrm{E}(\bar{\tau}_r) = \tilde{\tau},$$

where

$$\tilde{\tau} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi_i}{\bar{\phi}} \tau_i$$

and

$$\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i.$$

Bethlehem (1988) derives an expression for the bias in $\bar{\tau}_r$ as an estimate of $\bar{\tau}$, and shows that it is approximately equal to

$$\mathrm{B}(\hat{\bar{\tau}}) = \frac{\rho_{\phi,\tau} \sigma_\phi \sigma_\tau}{\bar{\phi}}, \tag{4}$$

where $\rho_{\phi,\tau}$ is the correlation between individual-level treatment effects and response probabilities, σ_ϕ is the standard deviation of the response probabilities ϕ_i, and σ_τ is the standard deviation of the treatment effects τ_i.

From (4), we can see that the $SATE$ is an unbiased estimate of the $PATE$ when any one of the following conditions hold:

1. When there is a constant treatment effect, i.e. $\tau_i = \tau \ \forall i$, then $\sigma_\tau = 0$ and the bias is zero.[2]

2. When the response probability is the same for all sampled units, i.e. $\phi_i = \phi \ \forall i$, or when there is *no* non-response, then $\sigma_\phi = 0$ and the bias is zero.

[2] This is what Schouten, Cobben, and Bethlehem (2009) term "strong representativeness," and is equivalent to the missing-completely-at-random (MCAR) assumption (Little and Rubin 2002).

3. When there is no correlation between the treatment effects and response probabilities, i.e. $\rho_{\phi,\tau} = 0$, the bias is also zero.

We can also see from the expression above that the bias increases as (a) the correlation between treatment effects and response propensities increases, and (b) response rates decrease, i.e. the mean of ϕ_i decreases. Note that because $\rho_{\phi,\tau}$ is independent of the sample size, increasing the sample size does *not* reduce non-response bias.

Weighting to Recover the PATE

Often researchers will have some information about non-respondents and/or the population, which they can use to correct for non-response bias. We can account for this auxiliary information by redefining

$$\phi_i = \phi(\mathbf{x}_i) = P(R_i = 1|\mathbf{x}_i = \mathbf{x}),$$

where $\mathbf{X}$ is the set of covariates known for the full sample or the entire population. Suppose we use this information to post-stratify the sample into classes, or strata $h = 1, 2, \dots, H$. The bias of the post-stratified estimator is now given by

$$\mathrm{B}(\hat{\hat{\tau}}) = \frac{1}{N} \sum_{h=1}^{H} \frac{\rho_{\phi_h,\tau_h} \sigma_{\phi_h} \sigma_{\tau_h}}{\bar{\phi}_h}. \tag{5}$$

This estimator is unbiased if the response probabilities within each stratum are constant, i.e. if units can be assumed to be a truly random sample of their corresponding class. If $\phi_h = \frac{1}{N_h} \sum\limits_{i \in h} \phi_{ih} = \phi \ \forall h$ then $\sigma_{\phi_h} = 0$ and the bias vanishes.[3] We can reduce non-response bias–even if we do not eliminate it completely–using weighting methods as long as the correlation between response and treatment effect is attenuated or the variability of response propensities is lower within weighting classes (Bethlehem 1988).

It is evident from the preceding discussion that the most effective variables to use when creating weighting classes are those that are correlated with both the outcome variables and the individual-level treatment effects (Kalton 1983; Kalton and Flores-Cervantes 2003). Unfortunately, it is often the case that hypothesized treatment effect moderators can only be measured in-sample (i.e. with surveys). Data on the distribution of these variables among non-respondents or the whole population is rarely available to researchers. In these situations, non-response bias in sample estimates will be reduced to the extent that the variables used to construct the weights are correlated with the treatment effect moderators, but some bias will remain.

[3] This is what Schouten, Cobben, and Bethlehem (2009) term “weak representativeness.”

It is also possible that researchers may unintentionally exacerbate imbalances due to selection on treatment effect moderators when adjusting a sample to other population characteristics, or even when information about marginal distributions is available but cell proportions or probabilities must be estimated (Shin, n.d.). Note that for the stratified SATE, i.e.

$$\hat{\tau}_{st} = \sum_{h=1}^{H} w_h \hat{\tau}_h,$$

where w_h are the estimated cell weights with corresponding W_h true cell weights, the bias is $\left[\sum_{h=1}^{H} (w_h - W_h)\bar{\tau}_h\right]^2$ (Cochran 1977). The further the estimated weights deviate from the true weights, the larger the bias, irrespective of sample size.

References

Bethlehem, Jelke G. 1988. "Reduction of nonresponse bias through regression estimation." *Journal of Official Statistics* 4(3):251–260.

Cochran, William. 1977. *Sampling techniques.*

Kalton, Graham. 1983. *Introduction to survey sampling.* Vol. 35 Sage.

Kalton, Graham and Ismael Flores-Cervantes. 2003. "Weighting methods." *Journal of Official Statistics* 19(2):81.

Rubin, Donald B and Roderick JA Little. 2002. *Statistical analysis with missing data.*

Schouten, Barry, Fannie Cobben, Jelke Bethlehem et al. 2009. "Indicators for the representativeness of survey response." *Survey Methodology* 35(1):101–113.

Shin, Hee-Choon. N.d. "A Cautionary Note on Post-stratification Adjustment." . Forthcoming.

Online Appendix B: Data Collection Procedures

The survey of the literature was limited to the three leading, general-interest political science journals: *American Political Science Review*, *American Journal of Political Science*, and *The Journal of Politics*. Google Scholar searches for each of these journals were conducted to locate all articles that used data from four commonly used online data sources for population-based survey experiments: Knowledge Networks (now known as GfK Custom Research), YouGov/Polimetrix, Survey Sampling International (SSI), and Amazon's Mechanical Turk (MTurk).

Four searches were conducted, with the following search terms:

- "Knowledge Networks" OR "GfK" OR "KN" OR "TESS" OR "Time-sharing Experiments"
- "YouGov" OR "Polimetrix" OR "CCES" OR "Cooperative Congressional Election Study" OR "CCAP" OR "Cooperative Campaign Analysis Project"
- "SSI" OR "Survey Sampling International" OR "Survey Sampling Inc." OR "Survey Sampling, Incorporated"
- "Mechanical Turk" OR "MTurk" OR "Turk" OR "Amazon" OR "Amazon's"

Restrictions for the searches were made in the Google Scholar fields for “Return articles dated between” and “Return articles published in,” with respective restrictions of 2000 and 2015 and the name of the journal. For the initial coding, the search date was July 1, 2015; the full set of data from 2015 was subsequently obtained and coded on April 1, 2016.

Each article returned in the search was assessed for whether the article was a false positive. The first round of false positive removals was for the journal. For example, the search for

“Journal of Politics” returned journals such as the *Journal of Politics & International Affairs* and the *Australian Journal of Politics and History*. Subsequent rounds of false positive removals were conducted for articles in which a search term had an alternate meaning (e.g., Amazon rainforest, Supplemental Security Income).

For each non-false positive article, the following elements were independently coded for each survey experiment in the article by two of the four authors:

1. journal name
2. article name
3. author name(s)
4. year published
5. journal volume
6. journal number
7. survey company
8. a dichotomous variable for whether the article had a survey experiment
9. a dichotomous variable for whether the article indicated the weighting used (weighted, unweighted, or both)
10. for articles that indicated the weighting used, a trichotomous variable to describe unweighted results (0=no unweighted results or values are reported, 1=unweighted values are reported, and 2=unweighted values are not reported but the article indicates that results are the same with and without weights)
11. for articles that indicated the weighting used, a trichotomous variable to describe weighted results (0=no weighted results or values are reported, 1=weighted values are reported, and 2=weighted values are not reported but the article indicates that results are

the same with and without weights)

12. direct quote excerpts from the article that described the weighting process

For searches for whether the article indicated that weighting was employed, descriptions of weighting were searched for in each article with the search terms “weight” and “stratification.” For item 9 above, results reported in online appendices were considered to have been indicated only if the results were mentioned in the main text of the article. Finally, we removed all observational studies that did not include a survey experiment.

The agreement rate across the full set of codings was 92%; for the 9 cases where two authors disagreed, all four authors discussed each case as a group and agreed upon a coding.

Online Appendix C: Weighting Methods

Weight construction usually proceeds in separate stages to address different sources of bias: sampling, non-coverage, and non-response (Kalton and Kasprzyk 1986; Brick and Kalton 1996). Base weights represent the probability of inclusion in the sample, typically calculated as inverse-sampling probabilities. These tend to be fixed for the population and do not depend on the final composition of the sample. Weights based on selection probabilities are sufficient for producing unbiased population estimates from a truly random sample (e.g., Horvitz and Thompson 1952). However, estimates will be biased if there is non-response correlated with the outcome variable (e.g., Bethlehem 1988; Cole and Stuart 2010).

Thus, the second stage of weight construction adjusts the base weights to reflect the characteristics of the full sample, including non-respondents. The last stage of weight construction accounts for incomplete coverage of population subgroups. The weights are once again adjusted, now to conform to known population totals. The choice of procedure to adjust for non-coverage or non-response depends on the extent to which the distribution of relevant characteristics is known for non-respondents or for the population. We discuss three commonly used techniques—post-stratification, raking, and inverse-propensity weighting—but many others exist (see Brick and Kalton 1996; Kalton and Flores-Cervantes 2003).

Post-stratification. In post-stratification, the sample is divided into strata h and units in the strata are weighted such that the sum of the weights in each stratum equals the population totals for that stratum (Kish 1965; Holt 1979; Little 1993). The post-stratified estimator for the SATE is a weighted average of the stratum-specific SATE (Miratrix et al. 2013). Post-stratification requires knowing unconditional cell probabilities or population counts for the characteristics being adjusted, and can fail if there are deep interactions with few observations in a cell. Units in cells with a small number of observations will have large weights which may yield unstable estimates, thought trimming can help in this regard (Kalton and Flores-Cervantes 2003). Post-stratification can also fail when the population and sample characteristics are measured differently and there is no one-to-one correspondence between strata and cell probabilities.

Raking. An alternative to post-stratification is raking (Deming and Stephan, 1940; Deville, Sarndal, and Sautory, 1993; Ireland and Kullback, 1968; Oh and Scheuren, 1983). Raking is used when the unconditional cell probabilities are unknown but the marginal distributions are known. This method can also be useful in cases where cell sizes are small. Researchers can simply collapse across categories or cells. For instance, surveys that adjust for non-response and weight to population demographics (e.g., from the Census) usually do not know the joint distribution of the interacted categories. Indeed, raking is often the method of choice of survey firms that conduct probability samples and provide researchers with pre-computed survey weights.

Raking assumes mutual independence (e.g., no interactions) between the classification variables, which has the advantage of leading to less variability in weight estimates. However, this assumption is often implausible and the estimated cell probabilities can differ substantially from the true population parameters (Shin, n.d.). If raking weights are treated as ground truth and subsequently used in post-stratification adjustment, then the raked mean will be biased regardless of sample size and its variance will underestimate the true error (Shin, n.d.; Cochran, 1977).

Inverse-propensity score weights. Both raking and post-stratification have the disadvantage of requiring discrete or coarsened auxiliary variables. Propensity score methods, in contrast, can incorporate continuous predictors and high-dimensional interaction terms to estimate the probability of survey response (see Rosenbaum and Rubin, 1983). Conditional on the propensity score, the distribution of covariates between respondents and non-respondents is equal. Any model that estimates the predicted probability of non-response can be used to calculate the propensity score, such as generalized regression (e.g., logistic regression), regression trees, random forests, neural networks, etc. One major disadvantage with IPSW methods is that it cannot ensure that the sample marginal joint distributions match those of the population (see e.g. Hazlett, n.d.).

Other methods. Post-stratification, raking, weight class estimators, and generalized regression estimators are special cases of the calibration estimator, which forces the sum of the adjusted weights to equal the population totals for each auxiliary variable used in estimation while minimizing the distance between the adjusted and unadjusted (inverse-sampling probability) weights (see Brick, 2013; Lumley et al., 2011). Many of the aforementioned methods can also be combined, such as using post-stratification for cells with large sample sizes and another method such as raking for cells with small sample sizes. More generally, researchers can combine methods to automatically build strata, such as matching (Stuart, 2010; Diamond and Sekhon, 2013; Sekhon and Grieve, 2012), with weighting or balancing methods to adjust strata characteristics to the observed population characteristics.

References

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

Brick, J.M. 2013. Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of*

Official Statistics 29(3): 329-353.

Brick, J. M. and Kalton, G. 1996. Handling Missing Data in Survey Research. *Statistical Methods in Medical Research* 5 (3): 215-238.

Cochran, W. G. 2007. *Sampling Techniques*. New York: John Wiley & Sons.

Cole, S. R., and Stuart, E. A. 2010. Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology* 172 (1): 107-115.

Deming, W. E., & Stephan, F. F. 1940. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *The Annals of Mathematical Statistics* 11 (4): 427-444.

Deville, J. C., Särndal, C. E., & Sautory, O. 1993. Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* 88 (423): 1013-1020.

Diamond, A., & Sekhon, J. S. 2013. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics* 95 (3): 932-945.

Gelman, A. 2007. Struggles with Survey Weighting and Regression Modeling. *Statistical Science* 22 (2): 153–164.

Hazlett, C. 2015. Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects. SSRN 2746753.

Holt, D., & Smith, T. M. F. 1979. Post Stratification. *Journal of the Royal Statistical Society. Series A (General)* 142 (1): 33-46.

Horvitz, D. G., & Thompson, D. J. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47 (260): 663-685.

Ireland, C. T., & Kullback, S. 1968. Contingency Tables with Given Marginals. *Biometrika* 55(1): 179-188.

Kalton, G., & Flores-Cervantes, I. 2003. Weighting Methods. *Journal of Official Statistics* 19 (2): 81-97.

Kalton, G., & Kasprzyk, D. 1986. *Treatment of Missing Survey Data*. Department of Biostatistics, University of Michigan.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Little, R. J. 1993. Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association* 88(423): 1001-1012.

Lumley, T., Shaw, P. A., & Dai, J. Y. 2011. Connections between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. *International Statistical Review* 79 (2): 200-220.

Miratrix, L. W., Sekhon, J. S., and Yu, B. 2013. Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (2): 369-396.

Oh, H. L., & Scheuren, F. J. 1983. Weighting Adjustment for Unit Nonresponse. *Incomplete Data in Sample Surveys*. New York: Academic Press.

Sekhon, J. S., & Grieve, R. D. 2012. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics* 21 (6): 695-714.

Shin, H. N.d. A Cautionary Note on Post-Stratification Adjustment. Working paper.

Stuart, E. A. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* 25(1): 1-21.