Generalizing Randomized Trial Findings to a Target Population using Complex Survey Population Data

Benjamin Ackerman, Catherine R. Lesko & Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Background: Randomized trials are considered the gold standard for estimating causal effects. Trial findings are often used to inform education policy and program implementation. While trials have strong internal validity by design, they may suffer from poor external validity, or generalizability, to the target population of interest (Bell et al., 2016; Cook, 2014; Tipton, 2014). Formally, the sample average treatment effect estimated in the trial (SATE) will be a biased estimate of the population average treatment effect (PATE; Olsen et al., 2013) if there are treatment effect moderators whose distributions differ between the trial and target population. Statistical methods have been developed to improve generalizability by combining trials and population data, and weighting the trial to resemble the population on baseline covariates. The implementation of these methods requires the identification of a dataset for the target population of interest, one that contains individual-level data on all relevant treatment effect modifiers in the trial. While data availability and quality make this challenging to do (Stuart and Rhodes, 2017), in practice, large nationally representative surveys collected by government agencies are often good sources of information on policy-relevant populations. However, these surveys often have complex sampling designs, with weights assigned to each participant relating them to the target population, and yet there is currently no best practice for incorporating survey weights when generalizing trial findings to a complex survey. Failing to account for these can result in estimates that are generalized to the complex survey *sample*, and not to the true target population of interest.

Purpose: We propose and investigate an approach to incorporate survey weights when generalizing trial findings to a population using population data from a complex survey sample. We examine the performance of this method (and the consequences of ignoring the complex survey weights) in simulations and then apply the methods to generalize findings from PREMIER, a lifestyle intervention trial, to a target population from NHANES. The work highlights the importance in properly accounting for the complex survey sampling design when generalizing trial findings to a population represented by a complex survey sample.

Methods: When a trial is not representative of the target population of interest, a current common approach to estimating the PATE is to transport the effect in the trial to a survey by weighting the trial to look more like the survey on pre-treatment characteristics. The first step to doing so is to specify a model of sample membership for the trial vs. the survey conditional on the set of covariates that influence sample membership and moderate treatment effect. This is similar to fitting a propensity score model of treatment assignment in non-experimental studies. The sample membership model is then used to predict the probability of trial participation, ê(X), and to construct weights equal to the inverse odds of êX) for trial participants, and 0 for survey participants. The PATE is then estimated as the mean difference of outcomes under treatment and control in the trial, weighted by the inverse odds of sample membership.

However, when complex survey data are used as the population data, this transported estimate is a biased estimate of the PATE because it ignores the survey's weights relating the survey to the target population. In order to account for this, we propose a two-stage weighting approach, where we first *weight the sample membership model*, such that survey participants are weighted by the inverse probability of survey selection (using the complex survey weights), and trial participants are given a weight of 1. For example, if a survey participant has a probability of survey selection of 0.02, the corresponding weight of 1/0.02 = 50 suggests that the individual in the survey should count for 50 people in the population when fitting the sample membership model. Weighting the survey participants in the sample membership model in this way enables us to better compare the trial demographics to the target population, and not just to the survey sample.

Simulation Results: Figure 1 shows the bias of estimating the PATE using the naïve trial estimator (red), the transportability estimator ignoring survey weights (green), and the transportability estimator using the survey weights to weight the sample membership model (blue). As the trial differs more greatly from the target population (moving down the rows), the naïve trial estimate becomes increasingly biased as expected. As the survey differs more greatly from the target population (moving from left to right on the x axis), the transportability estimate ignoring the survey weights becomes increasingly biased, and eventually more biased than the naïve trial estimate. On the other hand, the transported estimate that uses the survey weights to fit a weighted sample membership model is uniformly less biased than the other estimators across all scenarios. Note also that incorporating the survey weights appears to protect the transported estimate from becoming more biased as the survey becomes less representative.

Conclusion: When transporting trial findings to a population dataset that come from a complex survey, it is crucial to incorporate the survey weights in order to estimate the PATE. Our work has shown that fitting a sample membership model weighted by survey weights can only improve upon our ability to draw population-level inferences from RCTs, and that failing to do so may actually result in *more* biased estimates. Given that complex survey data often come ready for use with a variable containing the necessary survey weights, implementing this approach does not require specifying any additional models other than those needed for the standard transportability weighting methods. Our two-stage weighting method will ultimately allow researchers to draw more accurate population inferences from trials, and therefore better leverage information from trials when formulating education policies.

References:

Bell, S.H., Olsen, R.B., Orr, L.L., and Stuart, E.A. (2016). Estimates of external validity bias when impact evaluations select sites non-randomly. *Educational Evaluation and Policy Analysis* 38(2): 318-335.

Cook, T. (2014). Generalizing causal knowledge in the policy sciences: External validity as a task of both multi-attribute representation and multi-attribute extrapolation. *Journal of Policy Analysis and Management* 527-536. DOI: 10.1002/pam.

Olsen, R., Bell, S., Orr, L., and Stuart, E.A. (2013). External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of Policy Analysis and Management* 32(1): 107-121. NIHMS 382967

Stuart, E. A., & Rhodes, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation review*, *41*(4), 357-388.

Tipton, E. (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics, 39*(6): 478 – 501.

Figures:

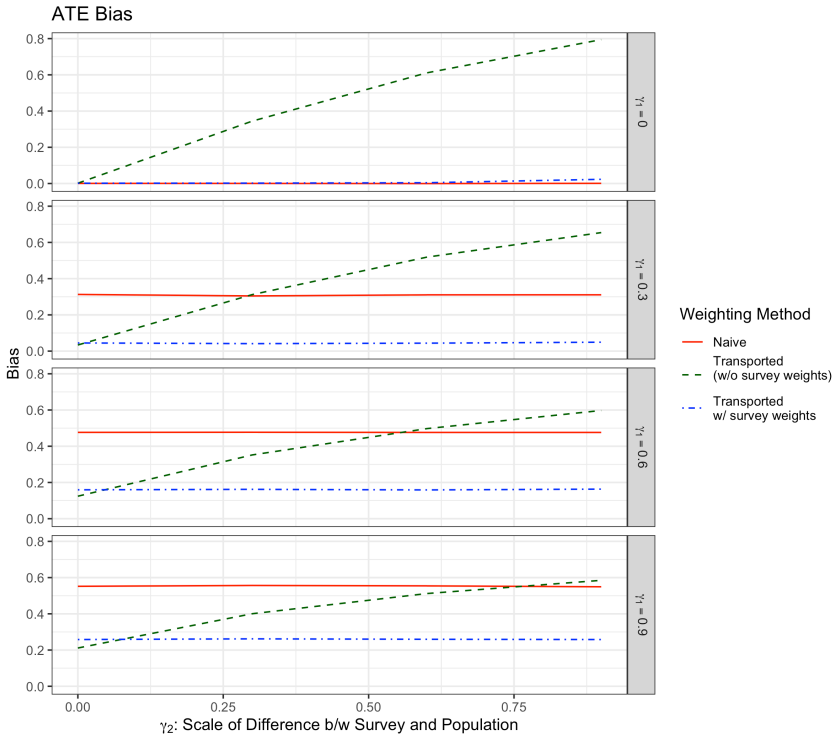

Figure 1: Bias of estimating the PATE by weighting method. Each column represents a different scenario of missing a variable used to calculate survey weights in the analytic survey dataset. From top to bottom row, the γ1 "scale" parameter for how much the trial differs from the population by the Xs increases. The different line types and colors represent the different weighting approaches: Naïve trial estimate (red solid), transported estimate ignoring the survey weights (green dash) and transported estimate using the survey weights (blue dotted dash).