Making Generalizations from Experimental Findings: An Introduction to New Methods

Elizabeth Tipton
Department of Human Development
Teachers College, Columbia University

Robert B. Olsen
George Washington Institute of Public Policy
George Washington University

June 2017

STRUCTURED ABSTRACT

Background

School-based experiments are increasingly common in education research. Ideally, the results of these experiments are used to make evidence-based policy decisions for students. However, it is difficult to make generalizations from RCTs because the types of schools included in RCTs are often based on convenience.

Purpose

This paper provides an overview of new statistical methods for improving generalizations from experiments in education. These methods have all been developed in the past 5 – 10 years and include methods for selecting sites into an experiment; for assessing how representative the sample in an experiment is of one or more target populations; and for estimating population average treatment effects with less bias.

Main argument

Selecting samples based on convenience alone can lead to biased impact estimates for the populations of policy interest. This bias can be reduced by identifying the target population, using stratified site selection methods to choose the sample, and using regression-based and propensity score methods to adjust statistically for any differences between the sample and the population. The degree of similarity between a sample and population can be assessed quantitatively as well.

Conclusions

The methods described here provide researchers with a way forward for improving the generalizability of findings from experiments in education. The implementation of these methods is facilitated by the availability of existing data for describing the population of interest and the availability of tools helpful for implementing at least some of these methods.

Key Words: *Generalizability; External Validity; Sampling; Propensity Scores.*

INTRODUCTION

Randomized controlled trials (RCTs) in education are conducted to estimate the causal impacts of educational interventions, programs, and policies on student outcomes. Evidence from RCTs are often used to inform education policy decisions, such as whether to adopt an educational intervention or cancel an education program. However, if the impact of an intervention varies across students or schools, the results from the RCT may not always "generalize" from the sample in the RCT to the population of students and schools that would be affected by these decisions.

In recent years, this *causal generalizability* problem (Shadish, Cook, & Campbell, 2002) has become increasingly addressed in the education, social welfare, and medical communities through the development of a variety of new statistical methods. This paper presents a summary of these methods, providing researchers new to this area with an overview of methods and approaches. Some of these methods can be used when the sample is selected; other methods involve improved analysis techniques or post hoc assessments of the likely generalizability of the study findings.

PROBLEMS GENERALIZING FROM RCTS

In an RCT, randomization to treatment ensures that the *average treatment effect* (*ATE*) estimated is the *causal* effect in the *sample*. However, researchers are rarely interested in restricting the generalizations from an RCT to the sample, instead aiming to predict the ATE in a target population of policy importance. For example, this target population might include all elementary schools in a particular school district, state, or region, while the RCT sample might include only schools in one or two districts chosen based on convenience.

It is important to begin this discussion by noting that generalizability is straightforward when treatment impacts are constant. In this case, the impact estimated in an RCT in *any* sample would lead to an unbiased estimate of the impact in *any* population. This constant treatment impact assumption is strong, however, particularly given growing evidence that for at least some interventions, impacts vary substantially across schools (Weiss et al, forthcoming).

If we begin with the assumption that treatment impacts vary, it becomes clear that the generalization problem in RCTs arises when the schools selected into the RCT were *not randomly selected*. Studies that include both random assignment and random sampling are exceedingly rare in education and social welfare, accounting for fewer than 3% of all RCTs (Olsen et al., 2013). Instead, the school districts and schools taking part in

education RCTs are chosen based on convenience. Recent reviews indicate, for example, that the schools taking part in RCTs funded by the Institute of Education Sciences are typically larger (in terms of student enrollments), in larger school districts, and include fewer rural and Title I schools than many important target populations (Stuart et al., 2017; Fellers, 2016; Tipton et al, 2016).

The generalizability of results from RCTs is thus problematic when this selection process inadvertently favors sites in which the impacts of the intervention are larger or smaller on average than that in the population (Olsen et al., 2013). In aggregate, the result is an ATE estimate that is much larger or smaller than the ATE in the target population. For example, in one evaluation, this bias was found to be on the order of 0.10, which is of similar magnitude to bias due to treatment selection in observational studies (Bell et al., 2016).

In any particular study, this bias is a function of the degree of variation in treatment impacts, the probability that different sites in the population will end up in the RCT, and the correlation between these treatment impacts and the probability of selection (Olsen et al., 2013). When selection is not random, these probabilities can be estimated based upon observed characteristics of the sites in the population and RCT (Stuart et al, 2011; Tipton, 2013), and – as we will show in later sections of this paper – these estimated *sampling propensities* (Tipton, 2013) can provide a means to getting a less biased estimate of the population ATE.

SELECTING A TARGET POPULATION

The methods provided in this review all begin with a common first principle: It is impossible to discuss generalization without specifying *to whom*. The results of an RCT may generalize well to schools in Colorado but not so well to schools in Wyoming. For this reason, the first step towards making generalizations is to clearly define one or more *target populations*.

Ideally, the choice of a target population is determined by the policy decisions that the study would inform. In RCTs of federal education programs, the target population could be all students that participate in the program since they are affected by policy decisions regarding the program. In other cases, the RCT may be designed to inform local decisions, such as district decisions about whether to adopt a specific math curriculum. For this type of RCT, the target population could be defined to include all districts that could potentially choose to implement the intervention.

There is a second principle as well: Every study has a target population, though some populations might be more broadly or narrowly defined than others. A broad target population – perhaps found in an effectiveness trial – might include all public, Title I elementary schools serving 3rd graders throughout the United States. In comparison, a narrow target population – perhaps found in an efficacy trial – might focus on public, Title I elementary schools that are in need of improvement in the Des Moines Public School District.

DATA REQUIRED FOR GENERALIZATION

Generalizing to the chosen target population requires data. The target population must be enumerated and include data on the key variables that are believed to moderate the impact of the intervention. Fortunately, in the United States, data on schools and districts – the level of sampling – are widely available and can be used for this purpose, including[1]:

- **The Common Core of Data (CCD).** Available for download from the National Center for Education Statistics (https://nces.ed.gov/ccd/index.asp), the CCD provides an annual census of public schools and school districts in the United States, which can be used as a sampling frame when selecting schools or districts for RCTs. The CCD includes data that can be used to construct several possible impact moderators, including measures of school and district size, per pupil funding, dropout rates, and student composition with respect to poverty (i.e., eligibility for free or reduced price lunch), disability, English language learner status, race and ethnicity.

- **Stanford Education Data Archive (SEDA).** Available for download from the Stanford Center for Education Policy Analysis (https://cepa.stanford.edu/seda/overview), SEDA provides estimates of average student achievement in mathematics and reading for school districts across the country.[2] The data also include measures of district and neighborhood racial and socioeconomic composition, school and neighborhood racial and socioeconomic segregation patterns, and other features of the schooling system.

[1] Outside of K-12, population data is not always so readily available, though often population frames can be created by combining data across various sources (for a Pre-K study, see Stuart & Rhodes, in press).
[2] SEDA's project director, Dr. Sean Reardon, reported at a pre-conference workshop of the 2017 spring conference for the Society for Research on Educational Effectiveness (SREE) that SEDA is planning to release achievement data for individual schools to complement the data already available on individual districts.

- **State-specific data sources.** Individual states typically maintain online data about its schools and districts for accountability reporting or research purposes. For example, Texas makes rich performance data available for its schools through the Texas Academic Performance Reports (TAPR) (http://tea.texas.gov/perfreport/tapr/index.html), while California make similar data available through its DataQuest system (http://www.cde.ca.gov/ds/sd/cb/dataquest.asp).

While these are promising, a note of caution is in order. Data from these sources will only improve generalizations to the extent that the variation in impacts in the population can be explained by the variables that can be constructed from these data (this is referred to in the literature as a *sampling ignorability assumption*; see Stuart et al, 2011; Tipton, 2013). However, knowledge of the predictors of treatment effect variation is currently limited and an area of growing research (e.g., Ding, Feller & Miratrix, 2015; Weiss et al., in press). For these reasons, we focus on the methods provided here as avenues for potentially *reducing* bias in estimates of population ATEs, since eliminating it is probably infeasible in most studies.

SITE SELECTION FOR IMPROVED GENERALIZATION

In the ideal, questions of the generalizability of findings from an RCT would be raised at the *beginning* of the study in the study design phase. In doing so, the discussion of the appropriate target population for the study becomes tied to the overall goals for the RCT. Once the target population is defined, the goal then is to develop a strategy to select a sample of schools and students that is like the population on the set of characteristics that might moderate the impacts of the intervention (Olsen & Orr, 2016; Tipton, 2014a; Tipton et al, 2014). In practice, this process involves two steps, detailed below.

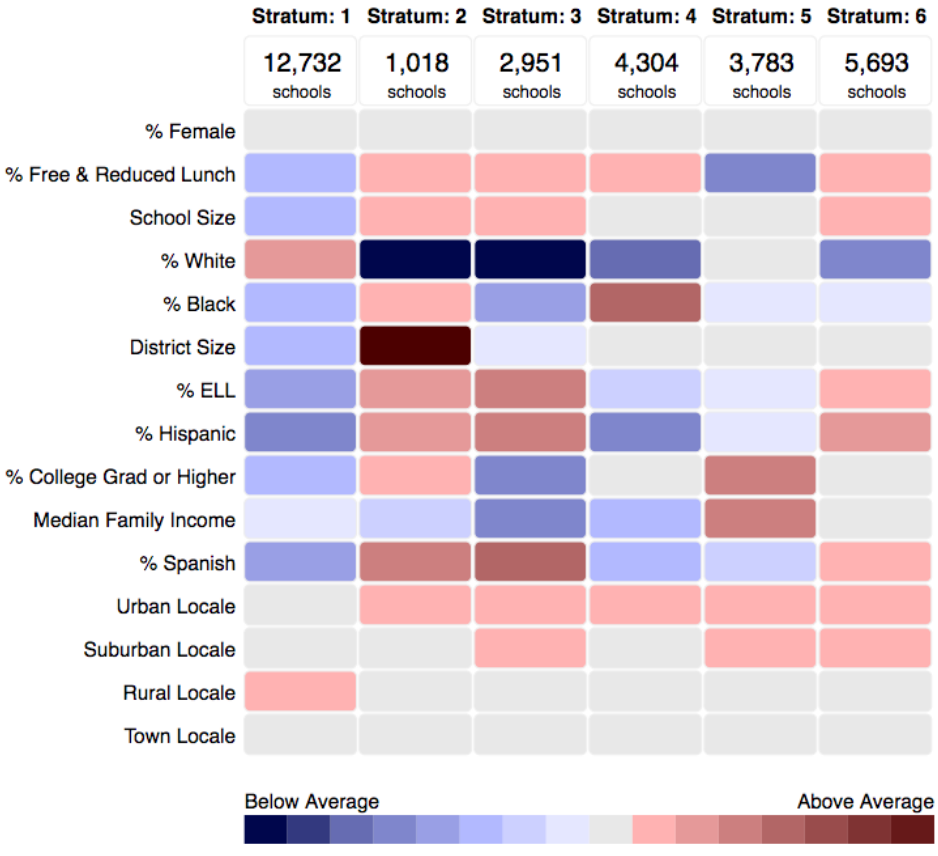
Stratifying the Population

Stratification is an important tool in selecting a sample to match a target population. Here the strata are defined in relation to moderators of the treatment impact (Olsen & Orr, 2016; Tipton, Yeager, Schneider & Iachan, in press). Stratifying the target population is straightforward when there are only a few variables that are categorical. However, the number of potential moderators can lead to an impractically large number of strata. For example, if an RCT identifies five possible moderators with two categories each, this would create 2^5 or 32 different strata. In this example, it would be impossible to include some schools from each stratum if the target number of schools is less than the total number of strata (32).

To create a smaller number of strata, researchers have developed several approaches. One approach is to simply combine multiple strata into a single stratum based on theory or pragmatism. Another approach is to use some form of dimension reduction, like cluster analysis or propensity score analysis (Tipton, 2014; Tipton et al, 2014). These methods are straightforward to implement in most statistical software, as well as in a free webtool (www.thegeneralizer.org; Tipton & Miller, 2016).

Dividing the population into many strata may reduce the bias—but it also complicates site recruitment. Every additional stratum in the design adds an additional resource constraint, requiring researchers to continue recruiting schools in more difficult strata, even when recruitment in other strata proves easy. Therefore, researchers need to strike a balance between bias reduction and ease of implementation. In cluster-randomized designs in education experiments, between 4 – 6 strata is often a good compromise.

Figure 1: Example Heat Map Comparing Strata
Source: The Generalizer



An additional benefit of stratification is that these strata can provide *descriptive* information on the target population. Figure 1 illustrates the division of a population into 6 strata. This figure indicates, for example, that the first stratum includes the largest proportion of schools and that these schools are smaller and include primarily rural, white students in communities that are not highly educated.

Selecting a Stratified Sample of Schools

The first step in selecting a stratified sample of schools is to decide how many schools to select from each stratum. Since the goal is to select a sample for the RCT that matches the target population on all potential moderators, proportional allocation is ideal. In this scheme, if 40% of the population is in Stratum 1, then in the RCT, 40% of the sample should also be in Stratum 1 (Tipton, 2014). In practice, this may not always be feasible, however, since recruitment in some strata may be easier than in others. A minimal goal, therefore, is to recruit enough schools in each stratum to estimate a stratum specific ATE, and to adjust for differences in composition between the sample and population using a post-stratification estimator (Tipton, 2013; O'Muircheartaigh & Hedges, 2014). This minimal goal eliminates *under-coverage* – which occurs when some portion of the target population is not at all represented in the RCT (aka *coverage error*; Tipton, 2013; 2014a) – thus avoiding a situation in which it is impossible to estimate the population ATE without heroic assumptions.

The second step in selecting a stratified sample of schools is to decide how to select schools from each stratum. Ideally, schools in the same stratum would have the same values of the treatment effect moderators—thus acting as replicates of one another—so it would not matter how the schools were selected. In practice, however, schools will vary within each stratum, either because some moderators are continuous or because the number of total strata is restricted to improve implementation.

Within each stratum, there are three possible selection methods to ensure a good match between the sample and the population on the distributions of moderator variables:

1. **Systematic site selection.** Researchers select sites that are as similar as possible to the average site in each stratum, based upon a "distance" measure[3]. Schools

[3] Calculating the distance requires an approach to weighting the different variables used for stratification. Researchers could choose equal weighting, larger weights for factors believed to be more important impact moderators, or larger weights for factors that are measured more precisely.

closer to the stratum-average school would be prioritized for recruiting (Tipton, 2014a; Tipton et al, 2014).[4]

2. **Random site selection.** Researchers select sites randomly from each stratum. When there are no refusals, random site selection ensures that there are no systematic differences between the sample selected and the population on both observed and *unobserved* characteristics (Olsen & Orr, 2016).

3. **Compromise selection.** Given resource constraints, researchers often begin studies with at least a handful of sites already agreeing to take part in the study. In this approach, these schools are first located in the strata. Then either strategy (1) or (2) is used to recruit the remainder of the sample.

With any of these approaches, researchers are encouraged to track data on which schools were selected, if they agreed or refused to take part in the study, and reasons for refusing. This information can then be used to better understand sources of sample selection bias (e.g., Tipton et al, 2016).

ASSESSMENT OF GENERALIZABILITY

Once the sample has been selected – before or after the RCT is complete – researchers can *assess* the likely generalizability of the study findings to one or more target populations. These methods all focus on quantifying the degree of *similarity* between the sample in the RCT and a target population on the set of potential moderators.

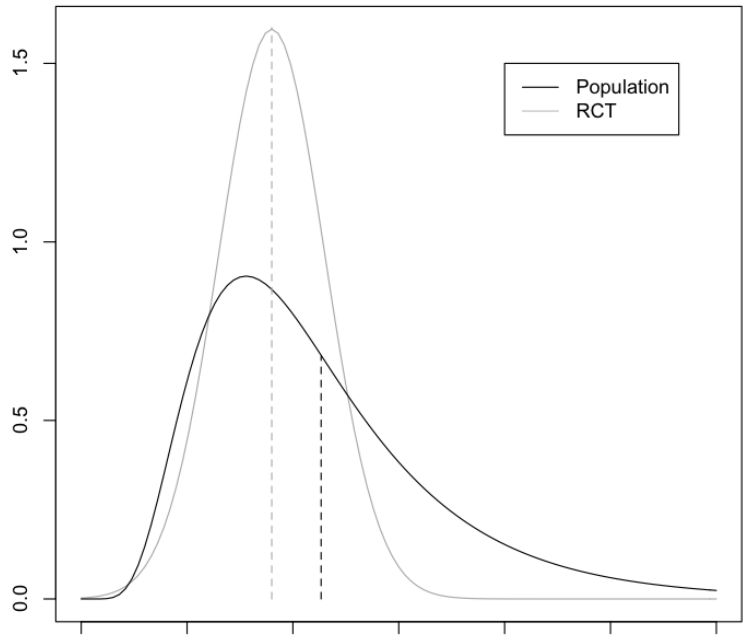
A difficulty in assessing similarity between the sample and a target population is that there are many potential moderators to consider. One approach to reduce the dimensionality of this problem is to use *propensity score* methods. Foundational research on propensity score methods has shown that if two groups are well matched on the propensity score, they will also be well matched on the variables—in this case, treatment effect moderators—that were included in the propensity score model (Rosenbaum and Rubin, 1983).

The *sampling propensity* is the probability that a site from the target population would take part in the RCT. These can be estimated using the observed covariates that potentially moderate the treatment impact and a wide variety of methods, from logistic

[4] Tipton (2014a) shows that even when a large share of selected sites refuse to participate in the study, systematic site selection—with systematic replacement of sites that refuse—can yield a sample that much more closely resembles the population that standard approaches to site selection.

regression to regression trees and neural networks (see Stuart, 2010 for a review). Once estimated, the similarity between the distribution of these sampling propensities in the RCT and target population can be compared; see Figure 2 for an example.

Figure 2: Distributions of Propensity Scores in Population and RCT
Note: In the figure, vertical dashed lines indicate the average value in each group.



As Figure 2 shows, the distributions in the two groups differ. To date, the following measures have been proposed as methods for summarizing these differences:

1. **Coverage.** This measure provides the proportion of the population that is represented by the RCT. When certain types of schools are not represented in the RCT, it will be difficult—maybe impossible—to generalize from the sample to the full population using post-hoc adjustments. Importantly, under-coverage is common even when samples are randomly selected (see Tipton, Hallberg, Hedges & Chan, in press). In the example in Figure 2, about 83% of the population is represented by the RCT (with under-coverage occurring in the long-tail of the population distribution).

2. **Standardized mean difference (SMD)**. This can be calculated on the propensity score scale or on their logits (see Stuart et al., 2011), providing the degree of difference *on average* between the distributions. This metric is standard in the propensity score literature. When this SMD is larger than 0.25 it indicates that

regression adjustments may not be warranted. In the example in Figure 2, the SMD is -0.386, indicating that inferences from the sample to population adjusted using regression would involve extrapolations.

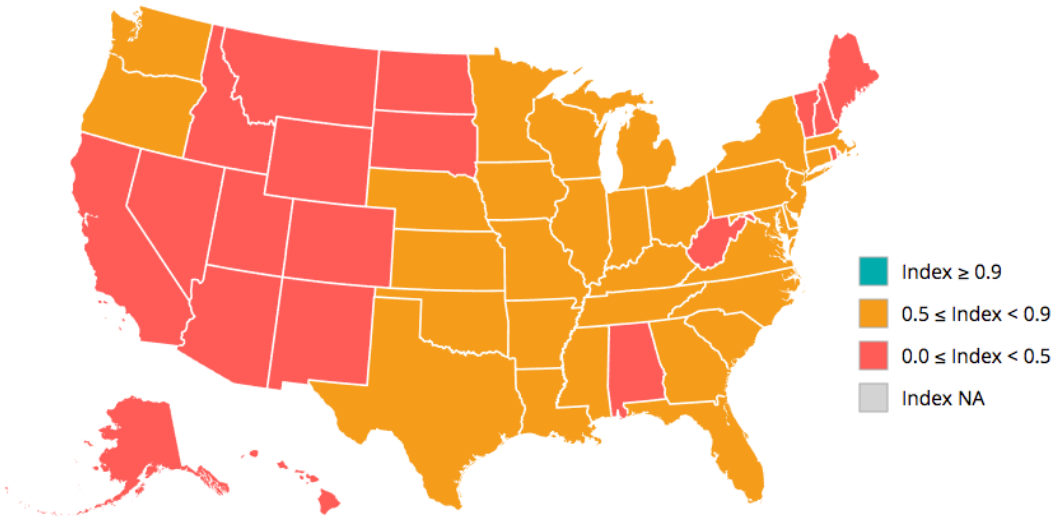
3. **Generalizability index.** This index summarizes the overall degree of distributional similarity (Tipton, 2014b). The index takes values between 0 and 1, with a 1 indicating that the sample is perfectly matched to the target population on the observed treatment effect moderators. The index is a function of both coverage and the SMD, as well as the proportion of the sample represented by the population (Tipton, 2014b). In cluster randomized trials, values greater than about 0.90 indicate that the sample is about as similar to the target population on the moderators as a random sample of the same size. Additionally, values smaller than 0.5 indicate that reweighting of the type described in the next section will be largely unsuccessful. This approach is implemented in the free webtool mentioned earlier (www.thegeneralizer.org; Tipton & Miller, 2016). In the example in Figure 2, if the RCT included a sample of 40 schools, the index value is approximately 0.85, indicating that while different from a random sample, differences between the sample and population could be adjusted for using the methods in the next section.

These measures of assessment can be particularly useful when there are multiple target populations. For example, in Figure 3, we provide a map in which the degree of similarity between an RCT and the populations of each of the 50 states is provided, offering information on where generalizations are most warranted (e.g., the southeast) and where they are not (e.g., the west).

The application of this index to studies in education suggests that the problem of generalizability is more than just an academic consideration. A reanalysis of data from two large-scale randomized trials in education calculated a generalizability index of .61 for one study and .57 for the other study. Indices this far below 1 indicate that study findings do not directly generalize to the population without any statistical adjustments, but that statistical adjustments will substantially increase the standard errors of the impact estimates (Tipton et al., 2016). Early results from an ongoing study suggest that this problem is prevalent among RCTs in education (Fellers, 2016).

In some situations, outcome data are also available for the target population – for example, if the outcome in the RCT is a student's score on a state-mandated achievement test and data on these scores are available for all schools in the state. In this case, researchers can also test how accurately they can predict the average outcome in the target population from the average outcome in the *control* condition in the RCT using the propensity score methods described above (see Stuart et al., 2011).

Figure 3: Map indicating similarity between an RCT sample and populations of each of the 50 states
Source: The Generalizer



IMPACT ANALYSIS FOR IMPROVED GENERALIZATION

In many situations, the sample in the RCT and the target population will not be perfectly aligned, indicating that the standard estimator of the ATE in the RCT will be biased for the target population ATE. A less biased estimate of the target population ATE would therefore consider compositional differences between the RCT and population on potential moderators. In this section, we discuss two propensity score-based estimators that serve this purpose.

Poststratification / Subclassification

One approach to reducing the site selection bias is to post-stratify the sample. Poststratification is much like stratification, but it is used to reweight the sample that was included—*post* site selection—to match the sample to the population (see Lohr, 1999, pp. 114-15). However, when the number of potential moderators is large, the number of strata becomes unwieldy, as stratification when selecting sites.

The propensity score methods described earlier – and useful for dimension reduction – can be used for poststratification: The target population can be divided into k strata based on the estimated sampling propensity. Within each of these j = 1,…,k strata, the sample from the RCT is located, and from these units, an estimate d_j of the stratum ATE

is calculated, as well as a standard error, SE(d_j). The population ATE and standard error can then be estimated using,

$$\widehat{PATE} = \sum_{j=1}^{k} w_j d_j$$

$$SE\left(\widehat{PATE}\right) = \sqrt{\sum_{j=1}^{k} w_j^2 \mathrm{SE}(d_j)^2}$$

where w_j are the population proportions—which usually equal 1/*k*, since the sample is usually divided into equal-sized groups.[5] When there is large under-coverage, the post-stratification approach can reduce bias, but typically does not eliminate it (e.g., Tipton, 2013; Tipton, Hallberg, Hedges, & Chan, in press). Furthermore, when there are large differences in the distributions of sampling propensity scores – as evidenced by a low generalizability index value – this approach can result in a much larger standard error for the population ATE.

Inverse-probability weighting

In the inverse-probability weighting (IPW) approach, the estimated sampling propensities are used to reweight the RCT to be more compositionally like the target population (Stuart et al, 2011). This approach is like Horvitz-Thompson estimators in survey sampling (Lohr, 1999) and can be written,

$$\widehat{PATE} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} d_{ij}/p_{ij}$$

where p_{ij} is the estimated sampling propensity in the i^{th} school in the j^{th} stratum. In theory, this inverse-probability weighting approach is a version of the post-stratification estimator with many, very small strata. In some cases there are very small sampling propensities (p_{ij}), and these can result in extreme weights, greatly affecting the standard error of the overall estimator.

In practice, the best estimator is the one that results in the greatest similarity (i.e., balance) between the sample and target population on the set of moderators under study. The process of determining the best estimator is therefore iterative (see Stuart, 2010). In head-to-head comparisons, the evidence suggests that one method does not always outperform the other, and as such both estimators are recommended in practice (Tipton, Hallberg, Hedges, & Chan, in press).

Other Methods

In addition to the propensity score based methods given above, other methods are also under development and testing. The simplest of these involves modeling the treatment impacts based upon a set of covariates, and then predicting the PATE using the population averages for each of the covariates. Kern, Stuart, Hill, and Green (2016) investigate properties of this linear regression approach, as well as several other more

[5] Equal-population strata (i.e., each contain 1/kth of the population) lead to the greatest bias reductions though other methods for creating strata (e.g., full matching) hold promise (Tipton, 2013).

flexible methods using a series of simulation studies. At the other end, Chan (2017) provides a bounding approach that requires different assumptions and may be particularly useful when there is under-coverage.

DISCUSSION

The knowledge, tools, and methods for conducting RCTs in education have increased dramatically over the past 30 years. As policy-makers at all levels begin to use results from these studies to make policy decisions, it becomes especially important to know *where*, *when*, and *for whom* results from an RCT apply. As this paper has shown, there is now a growing body of data, research methods, statistics, and tools available for researchers to design, assess, and estimate treatment impacts with generalizability in mind.

In using these tools, researchers should keep in mind the following guiding principles:

- **Defining the target population is a critical first step in conducting an RCT.** The target population identifies the group of schools or students about which the study hopes to learn.

- **Established methods exist for learning about the intervention's impacts in this population.** To improve the likelihood that RCT findings generalize to the target population, researchers can select the sample to be as similar to the target population as possible using stratified sampling techniques described earlier; they can also correct for any remaining differences using propensity score and other methods.

- **Completed RCTs can be assessed for their generalizability to the target population.** Tools described in this paper allow researchers to assess how well the RCT findings will generalize to the target population, giving researchers a sense of the likely bias in making these generalizations and the reduction in precision when applying propensity score methods for reducing the bias.

Lastly, we are under no illusions that the methods described in this paper "solve" the problem of generalizability in RCTs. Relatively little is known about the factors that moderate the impacts of educational interventions. But the methods described here rely on researchers' almost certainly imperfect knowledge of those moderators—and data that captures them—to improve generalizations from RCT findings.

REFERENCES

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

Chan, W. (2017). Partially Identified Treatment Effects for Generalizability. Forthcoming in *Journal of Research on Educational Effectiveness*.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society Series B*, *78*(3), 655-671.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103-127.

Lohr, S. (2009). Sampling: design and analysis. Nelson Education.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. Journal of Policy Analysis and Management, 32(1), 107-121.

Olsen, R. B., & Orr, L. L. (2016). On the “where” of social experiments: Selecting more representative samples to inform policy. New Directions for Evaluation, 2016(152), 61-71.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 41-55.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2), 369-386.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. Journal of Research on Educational Effectiveness, 10(1), 168-206.

Stuart, E. A., & Rhodes, A. Generalizing Treatment Effect Estimates From Sample to Population A Case Study in the Difficulties of Finding Sufficient Data. Forthcoming in *Evaluation Review*.

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

Tipton, E. (2014a). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review, 37*(2), 109-139.

Tipton, E. (2014b). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics, 39*(6), 478-501.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, 9(sup1), 209-228.

Tipton, E., Hallberg, K., Hedges, L.V., & Chan, W. Implications of small samples for generalization: Adjustments and rules of thumb. Forthcoming in a special issue on external validity at *Evaluation Review.*

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness, 7*(1), 114-135.

Tipton, E. & Miller, K. (2016) The Generalizer: A webtool for improving the generalizability of results from experiments. Available at http://www.thegeneralizer.org

Tipton, E., Yeager, D., Schneider, B., & Iachan, R. Designing Probability Samples to Identify Sources of Treatment Effect Heterogeneity. *Book chapter under review*.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials. Forthcoming in *Journal of Research on Educational Effectiveness*.