A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions

Elizabeth Tipton[1] and Robert B. Olsen[2]

School-based evaluations of interventions are increasingly common in education research. Ideally, the results of these evaluations are used to make evidence-based policy decisions for students. However, it is difficult to make generalizations from these evaluations because the types of schools included in the studies are typically not selected randomly from a target population. This paper provides an overview of statistical methods for improving generalizations from intervention research in education. These are presented as a series of steps aimed at improving research design—particularly recruitment—as well as methods for assessing and summarizing generalizability and estimating treatment impacts for clearly defined target populations.

Keywords: educational policy; evaluation; experimental design; experimental research; external validity; generalizability; multisite studies; policy; program evaluation; propensity scores; research methodology; sampling; statistics

In education research, evaluations are routinely conducted to estimate the causal impacts of education interventions, programs, and policies on student outcomes. The gold standard for evaluating causality is the randomized controlled trial (RCT), though some quasi-experimental designs—such as regression discontinuity,[1] instrumental variables, and matching—also offer valid approaches to causality under certain assumptions (see Shadish, Cook, & Campbell, 2002; What Works Clearinghouse, 2017). Evidence from these evaluations is often used to inform education policy decisions, such as whether to adopt a curriculum or cancel an education program. However, if the impact of an intervention varies across students or schools, the results from the study may not "generalize" from the sample in the evaluation to the population of students and schools that would be affected by these decisions.

In recent years, this *causal generalization* problem (Shadish et al., 2002) has become increasingly addressed in the education, social welfare, and medical communities through the development of a variety of statistical methods. This paper presents a summary of these methods, providing researchers new to this area with an overview of methods and approaches. We divide the process of implementing these methods in an evaluation into five steps, following the process of study design through analysis: (1) selecting a target population, (2) gathering data on this target population, (3) recruiting with this target population in mind, (4) assessing generalizability, and (5) addressing any mismatch between the sample and population.

Throughout, we focus on both introducing methods and practical guidelines for their implementation. To situate this work, we focus on cases in which schools are first recruited to a study and then either schools, teachers, classrooms, or students are randomized to treatment conditions. These methods apply more generally, however, including situations in which school districts, teachers, or students are directly recruited and in which instead of random assignment, some other quasi-experimental approach is used instead.

It is our hope that researchers planning evaluations of educational interventions—particularly RCTs—will use this paper as a primer, leading to important conversations about generalization early in the study design process. By addressing generalizability in the design phase instead of at the end, researchers have a greater ability to make thoughtful decisions about generalizability, thus extending the scope and impact of their research findings.[2]

[1]Columbia University, New York, NY
[2]George Washington University, Washington, DC

Educational Researcher, Vol. 47 No. 8, pp. 516–524
DOI: 10.3102/0013189X18781522

Problems Generalizing From Evaluations

In an RCT, randomization to treatment ensures that the *average treatment effect* (ATE) estimated is the *causal* effect in the sample.[3] Similarly, quasi-experimental designs also provide approaches for unbiased estimation of the *sample* ATE under certain assumptions. However, researchers are rarely interested in understanding if an intervention improves outcomes only in the sample—instead, they aim to predict the ATE in some population of scientific or policy importance. For example, this population might be vaguely defined as "all kindergarten at-risk kids" or more specifically defined as all elementary schools in a particular school district, state, or region, while the sample in the evaluation might include only schools in one or two districts.

It is important to begin this discussion by noting that generalizability is straightforward under the strong—and often implausible—assumption that treatment impacts are constant. In this case, the effect of the intervention is the *same* for every unit in the population—for example, students, classrooms, and schools. As a result, it does not matter which sample is included in the evaluation since the impact estimated in *any* sample would lead to an unbiased estimate of the impact in the population. The assumption that treatment effects are constant is strong, however, and not empirically based; in fact, there is growing evidence that for at least some interventions, impacts vary substantially across schools (e.g., Weiss et al., 2017).

If we begin instead with the assumption that treatment impacts vary, it becomes clear that the generalization problem arises when the schools recruited into an evaluation differ in important ways from the population of interest. One approach that addresses this problem is to both randomly sample schools into the study and randomly assign units to treatment. Unfortunately, these double-random designs are exceedingly rare in education and social welfare, accounting for fewer than 3% of all RCTs (Olsen, Orr, Bell, & Stuart, 2013). Instead, the school districts and schools taking part in education evaluations are typically chosen nonrandomly based on factors such as convenience, concerns with implementation fidelity, assumed responsiveness to the intervention, and cost. 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 (Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016).

The generalizability of results from educational evaluations is thus problematic when this selection process inadvertently favors sites in which the impacts of the intervention are larger or smaller than that in the population (Olsen et al., 2013), resulting in an ATE estimate that is much larger or smaller than the ATE in the target population. For example, in one RCT, this bias from nonrandom selection was found to be on the order of 0.10 standard deviations, which is of similar magnitude to bias due to treatment selection in observational studies (Bell, Olsen, Orr, & Stuart, 2016). Overall, this suggests that researchers planning intervention studies would be wise to focus on developing study designs that minimize the *total* bias (Olsen et al., 2013), where

$$\text{total bias} = \text{internal bias} + \text{external bias}.$$

Here, by *internal bias*, we are referring to the usual threats to internal validity—for example, treatment selection, attrition—and by *external bias*, we are referring to threats to external validity—for example, differences between sample and population unit treatment effects.

This focus on minimizing total bias differs markedly from standard evaluation methods in psychology, health, and social welfare, where research design has focused nearly entirely on reducing internal bias. We argue, however, that given findings from this new literature on the extent of external bias, the profound differences between the types of schools and students taking part in evaluations compared to the population, and the increasing calls to use results from evaluations in decision making in schools (e.g., ESSA), the time has come for a change.

Guidelines for Generalizing Impact Evaluation Results

Having established that external bias can be as much a concern as internal bias in evaluations of interventions, in the remainder of this review paper we provide a primer on methods for reducing this bias in future studies. This primer builds on a literature on sample selection, generalizability assessment, and treatment effect estimation that spans the fields of education, psychology, epidemiology, and medicine. Throughout, we provide guidance on both methods and tools as well as practical considerations.

In general, our approach assumes that the best strategy for improved generalizability is via improved research design. However, this is not a requirement per se as the methods that will be described for assessing generalizability and estimating population ATEs can be implemented post hoc. As we will show, however, these analysis methods work best in tandem with improved design.

Step 1: Select a Target Population

All methods regarding improved generalizability begin with a common first principle: It is impossible to discuss generalization without specifying *to whom*. The results of an intervention study may generalize well to schools in Colorado but not so well to schools in Wyoming. For this reason, the first step toward 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 study may be designed to inform local decisions, such as district decisions about whether to adopt a specific math curriculum. For this type of evaluation, the target population could be defined to include all districts that could potentially choose to implement the intervention.

There is also a second principle: 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 third 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.

Importantly, in many studies, there is a tension between the desire for a broad target population and the resources required to study such a broad population. For example, a broad target population may have a larger degree of variation between schools (resulting in a large intraclass correlation), which requires a larger sample size to detect if there is a non-zero average effect in the population (i.e., statistical power). Similarly, a broad target population may require greater resources for travel—for example, in order to include both large and small school districts in both urban and rural areas. For these reasons, a narrow population may be preferred for many studies, though this is only helpful if the population is accurately defined to include schools where the intervention has the largest effect. The difficulty here, of course, is that this is not always possible to know a priori. Our position is not that one is better than the other, only that researchers must make these assumptions and tradeoffs clear when proposing and reporting their work.

Step 2: Gather Data Required for Generalization

Generalizing to the chosen target population requires data. While population data on individual students is typically not available, in the United States, data on schools and school districts—the level of recruitment—are widely available and can be used for this purpose, including[4]:

- **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. The CCD includes 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.[5] 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.
- **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).

The ideal population data not only enumerate the population units but also include key variables. Since external validity bias arises when the units in a study have lower or higher treatment impacts than the rest of the population, the key variables of focus here are those that are believed to moderate the impact of the intervention. Unfortunately, there is little evidence on the factors that moderate the impact of most educational interventions.

When the factors that moderate treatment effects are unknown, we recommend focusing on variables that have been shown to be related to the outcome of interest since the impact is simply the difference between outcomes in the treated and untreated conditions, and variables related to the outcome may also be related to impact. Such variables include race, socioeconomic status, prior test scores, and other aggregated student demographics. We also recommend considering variables that prior research has found to be associated with the inclusion of schools or districts in evaluations. For example, Stuart et al. (2017) found that school districts that participate in randomized trials were more likely to be larger, more urban, and more disadvantaged than the average school district; Tipton et al. (2016) and Fellers (2017) found similar results for schools. Gathering data on these factors is an important first step to selecting a representative sample on these dimensions—or at least attempting to correct for the ways in which the sample is not representative.

Finally, 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, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013). This requires researchers' almost certainly imperfect knowledge of those moderators—and data that capture them—to improve generalizations from RCT findings. Sensitivity analyses—conducted in the analysis phase—provide one approach to determining if this ignorability condition has been met (Nguyen, Ebnesajjad, Cole, & Stuart, 2017). Given these concerns, 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.

Step 3: Recruit With Generalization in Mind

Once the target population is defined, the goal, then, is to develop a strategy to recruit 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 in the following.

Stratifying the Population

Stratification is an important tool for recruitment. Here the strata are defined in relation to moderators of the treatment

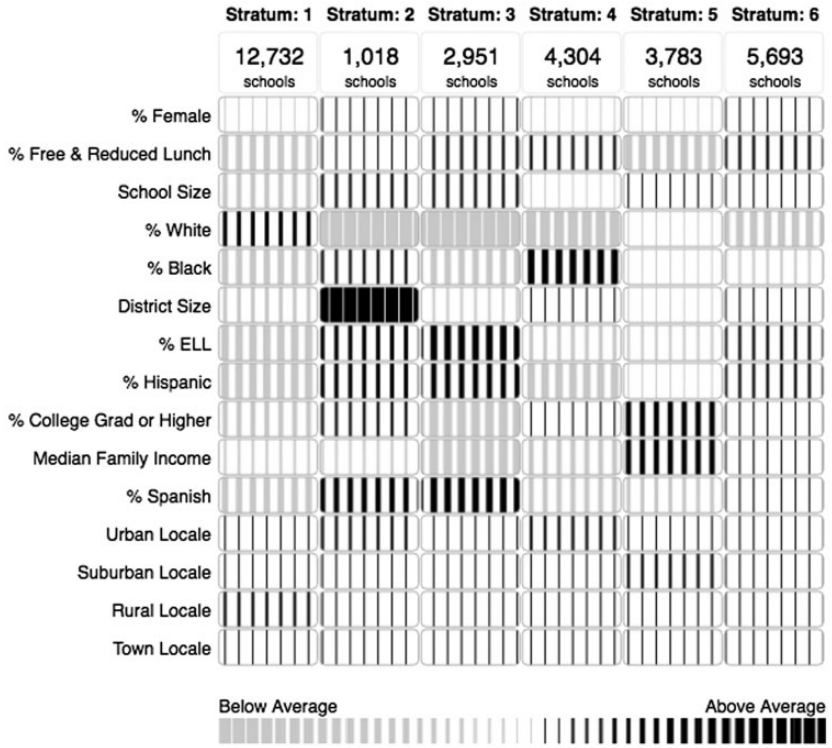

FIGURE 1. *Example heat map comparing strata.*
Source. The Generalizer.

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 evaluation 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, 2014b; 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, defining between four and six strata is often a good compromise.

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 six 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.

Developing Stratified Recruitment Goals

The first step in recruiting a stratified sample of schools is to decide how many schools to recruit from each stratum. Since the goal is to recruit 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 evaluation, 40% of the sample should also be in Stratum 1 (Tipton, 2014b). 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 adjust for differences in composition between the sample and population using a post-stratification estimator (see Step 5; O'Muircheartaigh & Hedges, 2014; Tipton, 2013). This minimal goal eliminates *under-coverage*—which occurs when some portion of the target population is not at all represented in the study (aka *coverage error*, see Step 4; Tipton, 2013, 2014b)—thus avoiding a situation in which it is impossible to estimate the population ATE without heroic assumptions.

The second step in recruiting a stratified sample of schools is to decide how to recruit schools from within 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

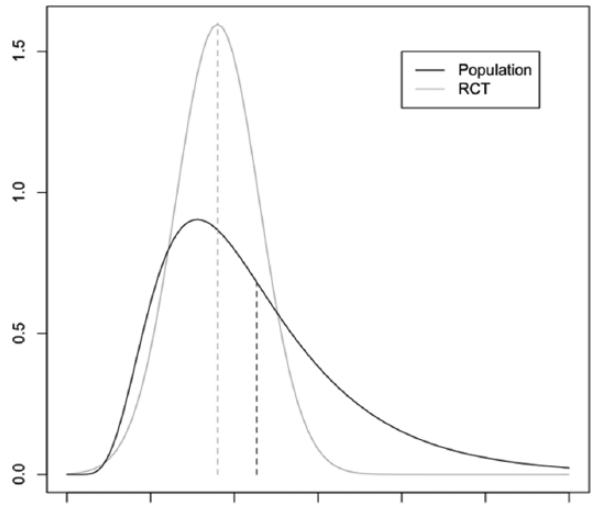

FIGURE 2. *Distributions of propensity scores in population and randomized controlled trial.*
Note. In the figure, vertical dashed lines indicate the average value in each group.

stratum, either because some moderators are continuous or 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 schools that are as similar as possible to the average school in each stratum based on a "distance" measure.[6] Schools closer to the stratum-average school would be prioritized for recruitment (Tipton, 2014b; Tipton et al., 2014).[7]
2. **Random site selection.** Researchers select schools randomly from each stratum. When there are no refusals, random 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 schools 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 recruited, 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).

Step 4: Assess Generalizability and Iterate

Quantifying Generalizability

Once the final sample has been recruited—before or after the study 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 evaluation 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. Of course, the sample and target population can be compared for each of these moderators, one variable at a time, using *t* tests or chi-square tests. For a single comparison that combines these moderators, researchers can 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 & Rubin, 1983).

The *sampling propensity* is the probability that a school from the target population would be in a sample in the evaluation given a set of covariates. These can be estimated using the observed covariates that potentially moderate the treatment impact and a wide variety of methods, from logistic regression to regression trees and neural networks (for a review, see Stuart, 2010). Once estimated, the similarity between the distribution of these sampling propensities in the sample and target population can be compared; see Figure 2 for an example.

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 evaluation. When certain types of schools are not represented in the evaluation, it will be difficult—maybe impossible—to generalize from the sample to the population using post hoc adjustments. Importantly, under-coverage is common even when samples are randomly selected (see Tipton, Hallberg, Hedges, & Chan, 2017). 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 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 absolute 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, 2014a). The index takes values between 0 and 1, with 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. In studies with around 40 schools, values greater than about 0.90 indicate that the sample is about as similar to

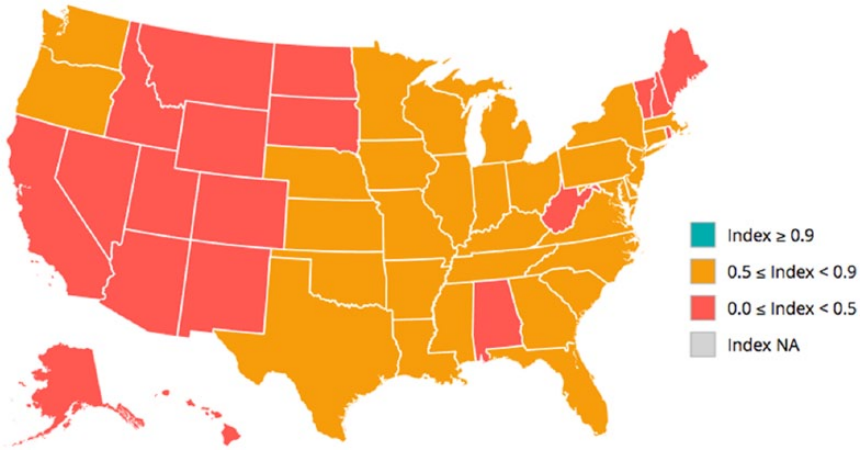

FIGURE 3. *Map indicating similarity between a randomized controlled trial sample and populations of each of the 50 states. Source.* The Generalizer.

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 study 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 illustrating the degree of similarity between the sample taking part in an RCT and the populations of schools in each of the 50 states, offering information on where generalizations are most warranted (e.g., the Southeast) and where they are not (e.g., the West). (A color version of Figure 3 is available in the online article.)

Finally, 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 previously (see Stuart et al., 2011).

A Note on Difficult Generalizations

If recruitment was difficult or finished before concerns with generalizability were raised, it is likely that the resulting sample will differ from the target population, sometimes in large ways. For example, a reanalysis of data from two large-scale randomized trials in education calculated a generalizability index of 0.61 for one study and 0.57 for the other study (Tipton et al., 2016). 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. Early results from an ongoing study suggest that this problem is common among RCTs in education (Fellers, 2017).

In these situations, it can be useful to *define the population to which generalizations are possible*, thus increasing similarity between the sample and *some* target population. Determining the best subpopulation, however, is not straightforward and is typically iterative. To do so, the first step is to determine which covariate values differ largely between the sample and population. For categorical covariates, this is straightforward. For example, if the sample does not include any rural schools but the population does, redefining the population to exclude rural schools could help. For continuous covariates, this is less straightforward: Generally, the most problematic covariates are those that not only differ on average but are also less variable and have a shorter range in the sample compared to the population. In this case, redefining the population to include only schools with fewer than the maximum observed in the sample (e.g., 500 students) might help.

Tipton et al. (2016) provides a case study that illustrates this process in action. In their example study with a generalizability index of 0.57, using two inclusion criteria, they could define a new target population with a generalizability index of 0.80. This new population accounted for about 33% of the original target population of interest. While perhaps not ideal given the study's initial goals, given the difficulties of recruitment, this approach led to a realistic accounting of where the study results are most and least credible for making curricular decisions.

Finally, it is important for researchers to be as transparent as possible about any changes to the target population that occur in their studies. Just as researchers commonly report details on attrition after random assignment, it is important for researchers to report details on both the originally defined target population as well as the subpopulation where generalizations are stronger based on actual recruitment.

Step 5: Address Mismatch Between the Sample and Population

In most practical cases, the sample in the evaluation and the target population will not be perfectly aligned, indicating that

the standard estimator of the ATE will be biased for the target population ATE. This section describes three different approaches to reducing the bias when estimating the ATE for the population: (1) propensity score methods, (2) model-based approaches, and (3) bounding.

Propensity Score Methods

Propensity score methods can be used either to identify and reweight entire subclasses of schools in the sample or to reweight individual schools in the sample to reduce the differences between the sample and population on potential moderator variables.

Propensity score subclassification. The propensity score methods described earlier can be used to post-stratify or subclassify the sample. The target population can be divided into k strata based on the estimated sampling propensity. Within each of these $j = 1, \ldots, k$ strata, the sample from the evaluation is located, and from these units (e.g., schools), 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,

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

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

where w_j are the population proportions—which usually equal $1/k$ since the sample is usually divided into equal-sized groups.[8] When there is large under-coverage, post-stratification can reduce bias but typically does not eliminate it (e.g., Tipton, 2013; Tipton et al., 2017). In general, using regression adjustment within strata can help reduce this bias (see Stuart, 2010), though this can result in extrapolations. Overall, when there are large differences in the distributions of sampling propensity scores—as evidenced by a low generalizability index value—subclassification (with or without regression adjustment) can result in a much larger standard error for the population ATE.

Inverse-probability weighting. Inverse-probability weighting (IPW) uses the estimated sampling propensities to reweight the schools in the sample to be more compositionally like the target population (Stuart et al., 2011). This approach is like Horvitz-Thompson estimators in survey sampling (Horvitz & Thompson, 1952; Lohr, 1999) and, in the case of a cluster randomized trial, can be written,

$$\widehat{PATE} = \sum_{i=1}^{N} \frac{W_i Z_i Y_i}{p_i q_i} - \sum_{i=1}^{N} \frac{W_i (1 - Z_i) Y_i}{p_i (1 - q_i)}$$

$$SE\left(\widehat{PATE}\right) = \sqrt{\sum_{i=1}^{n_T} Y_i^2 \left(\frac{1 - p_i q_i}{p_i q_i}\right) - \sum_{i=1}^{n_C} Y_i^2 \left(\frac{1 - p_i (1 - q_i)}{p_i (1 - q_i)}\right)}.$$

Here, W_i indicates if school i (for $i = 1, \ldots, N$ in the population) is in the sample, Z_i indicates if the school has been assigned to the treatment condition, and for the ith school, p_i is the estimated sampling probability, q_i is the probability of being assigned to treatment, and Y_i is the outcome. In theory, this IPW approach is a version of the post-stratification estimator with many, very small strata. In some instances, very small sampling propensities (p_i) can result in extreme weights, which inflates the standard error of the overall estimator. This problem can be addressed by trimming the most extreme weights (Lee, Lessler, & Stuart, 2011).

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. 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 et al., 2017).

Model-Based Approaches

An alternative approach to addressing differences between the sample and the population involves regression modeling. This approach builds on the regression models that researchers use to estimate the ATE. When those models are enhanced to include interactions between the treatment and potential impact moderators, they can be used to predict the impact of the intervention for any combination of the moderator variables. Therefore, if the average values of those moderators in the population are known, researchers can insert these values in the estimated regression model to predict the average impact in the population.

Standard regression models may yield poor predictions if the linearity and additivity assumptions are not satisfied or the model omits important treatment-by-moderator interactions. To address this problem, researchers can use algorithms like Bayesian additive regression trees (BART; Chipman, George, & McCulloch, 2007, 2010). BART relaxes key assumptions of standard regression models, automates the process of selecting interaction terms, and can be used to create Monte Carlo estimates of the posterior distribution of the PATE (for more details, see e.g., Kern, Stuart, Hill, & Green, 2016).

Bounding Approaches

Finally, another set of methods is that of Chan (2017), which provides a bounding approach to population treatment effect estimation from RCTs. Instead of providing a point-estimate for the ATE, this approach provides an interval estimate. These interval estimates are a function of the proportion of a target population in an RCT, the treatment effect in the RCT, and assumptions regarding the minimum and maximum possible treatment impact. The most general of these estimators requires no assumptions but can result in a very wide range. Other estimators in this class add assumptions regarding the treatment effect outside of the RCT. These approaches may be particularly useful when propensity score methods are not possible, such as when there is under-coverage or covariate information is unavailable or sparse in a population.

Discussion

As we have highlighted throughout this paper, the knowledge, tools, and methods for conducting evaluations of interventions in education have increased dramatically over the past 30 years.

As policymakers 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 evaluation apply. Our goal in this paper has been to provide education researchers with a broad overview of the methods and tools available for addressing concerns with generalizability in evaluations of educational interventions and programs. Our hope is that moving forward, researchers will take seriously these concerns with generalizability and take advantage of this growing base of tools available for design, assessment, and estimation.

It is important to understand what can reasonably be accomplished using the tools described in this paper. The impacts of educational interventions may vary for a host of reasons that challenge our ability to generalize, including variation in implementation of the intervention, take-up rate when participation is voluntary, dose of the intervention received by participating students, fidelity of implementation to the program model, services that students would otherwise receive without the intervention, characteristics of participating students, and characteristics of the context or setting (see Weiss, Bloom, & Brock, 2014). The tools described in this paper are designed to address variation in factors for which data are typically available (e.g., the characteristics of students and schools); they cannot help us address variation in factors for which data are typically unavailable outside the evaluation sample (e.g., fidelity of implementation).

Furthermore, the success of these methods for generalizing depends heavily on our knowledge of the factors that influence the impact of the intervention and data that capture these factors. When the factors that moderate the impact are unknown or unmeasured and the sample is selected nonrandomly—either by design or due to self-selection into the study—no statistical methods can help us generalize evaluation findings from one set of students and schools to another. In practice, with helpful but imperfect theory and data, the methods described in this paper should facilitate generalizations that are imperfect but still better than we could make without them.

NOTES

Tipton would like to acknowledge funding from the Institute of Education Sciences, Award Number R305D170024 and from the Spencer Foundation, Award Number 201500057. Tipton completed this paper while faculty at Teachers College, Columbia University. She is now faculty at Northwestern University. Olsen would like to acknowledge funding from the Institute of Education Sciences, Award Number R305D150003.

[1]Importantly, not all of these methods estimate the same parameter. For example, in regression discontinuity (RD), the effect estimated is the "local" average treatment effect near the cutoff. Generalization to a target population from an RD design thus also requires generalizing over cutpoints on the assignment variable, which is beyond the scope of this paper.

[2]For those interested in making generalizations from a study that is already completed, Step 3 can be omitted.

[3]This is the intent-to-treat effect when there is noncompliance.

[4]Outside of K–12, population data are 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, 2017; for a community college study, see Tipton & Matlen, 2018).

[5]Stanford Education Data Archive's (SEDA) 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.

[6]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.

[7]Tipton (2014b) shows that even when a large share of selected sites refuses 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.

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

ORCID ID

Elizabeth Tipton https://orcid.org/0000-0001-5608-1282

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. *Journal of Research on Educational Effectiveness*, *10*(3), 646–669.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian ensemble learning. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19 (NIPS 2006)* (pp. 265–272). Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-19-2006

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Fellers, L. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 10256121)

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.

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.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS One*, *6*(3), e18174.

Lohr, S. (1999). *Sampling: Design and analysis*. Boston, MA: Duxbury Press.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, *11*(1), 225–247.

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.

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.

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*, *70*(1), 41–55.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1.

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., 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., & 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. (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). 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. (2014b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109–139.

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*(Suppl 1), 209–228.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472–505.

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., & Matlen, B. (2018). *The development and implementation of a strategic recruitment plan for generalizability: A case study*. Working paper.

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

Tipton, E., Yeager, D., Schneider, B., & Iachan, R. (In press). Designing probability samples to identify sources of treatment effect heterogeneity. In P. J. Lavrakas (Ed.), *Experimental methods in survey research: Techniques that combine random sampling with random assignment*. New York, NY: Wiley.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

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

What Works Clearinghouse. (2017). *What Works Clearinghouse: Standards handbook (Version 4.0)*. Washington, DC: U.S. Department of Education.

AUTHORS

ELIZABETH TIPTON, PhD, is an associate professor of statistics at Northwestern University, 2006 Sheridan Road, Evanston, IL 60208; *tipton@northwestern.edu*. Her research focuses on the development of methods for improving the generalizability of causal effects in randomized trials and meta-analysis.

ROBERT B. OLSEN, PhD, is an associate director at Westat, 1600 Research Boulevard, Rockville, MD 20850; *robolsen@westat.com*; and a research professor at the George Washington Institute of Public Policy, The George Washington University, Media and Public Affairs Building, 805 21st Street NW, Washington DC 20052; *robolsen@gwu.edu*. His research focuses on the impacts of educational interventions and federal programs and research methods for impact evaluation.

Manuscript received June 20, 2017
Revisions received January 30, 2018, and April 24, 2018
Accepted May 4, 2018