Article

Evaluation Review
1-34

DOI: 10.1177/0193841X16655665
erx.sagepub.com

Implications of Small Samples for Generalization: Adjustments and Rules of Thumb

Elizabeth Tipton[1], Kelly Hallberg[2], Larry V. Hedges[3], and Wendy Chan[4]

Abstract

Background: Policy makers and researchers are frequently interested in understanding how effective a particular intervention may be for a specific population. One approach is to assess the degree of similarity between the sample in an experiment and the population. Another approach is to combine information from the experiment and the population to estimate the population average treatment effect (PATE). **Method:** Several methods for assessing the similarity between a sample and population currently exist as well as methods estimating the PATE. In this article, we investigate properties of six of these methods and statistics in the small sample sizes

[1] Department of Human Development, Teachers College, Columbia University, New York, NY, USA
[2] University of Chicago – Urban Labs, Chicago, IL, USA
[3] Institute for Policy Research, Northwestern University, Evanston, IL, USA
[4] Quantitative Methods Division, Graduate School of Education, University of Pennsylvania, Philadelphia, PA USA

Corresponding Author:
Elizabeth Tipton, Department of Human Development, Teachers College, Columbia University, 525 W. 120th St., Box 118, New York, NY 10027, USA.
Email: tipton@tc.columbia.edu

common in education research (i.e., 10–70 sites), evaluating the utility of rules of thumb developed from observational studies in the generalization case. **Result:** In small random samples, large differences between the sample and population can arise simply by chance and many of the statistics commonly used in generalization are a function of both sample size and the number of covariates being compared. The rules of thumb developed in observational studies (which are commonly applied in generalization) are much too conservative given the small sample sizes found in generalization. **Conclusion:** This article implies that sharp inferences to large populations from small experiments are difficult even with probability sampling. Features of random samples should be kept in mind when evaluating the extent to which results from experiments conducted on nonrandom samples might generalize.

Keywords
education, content area, methodological development

Policy makers are frequently interested in understanding how effective a particular intervention may be for a specific (and often broad) population. Evidence-based policy making requires that the policy makers should inform these decisions using results from well-designed evaluations. In many fields, particularly education and social welfare, the ideal form of these evaluations is a large-scale randomized experiment. The fact that sites or units within sites are randomly assigned to different interventions (or a control group) allows the causal impact of an intervention to be assessed without selection bias, which often compromises other evaluation designs. Making generalizations of this causal effect to an inference population is difficult, however, since, as recent research highlights, sites in large-scale experiments are rarely randomly sampled (Olsen, Orr, Bell, & Stuart, 2013).

Recently, statisticians and evaluators have begun developing methods for improving generalizations from randomized experiments based on nonprobability samples (Orr, 2015). One strand of this research focuses on the assessment of generalizability (Olsen et al., 2013; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2014a), providing policy makers with tools to gauge how similar a sample from an experiment is to a policy relevant inference population. This research highlights that the results of an experiment may generalize well to one population but not another. In many situations, the experimental sample and population differ in composition on a few

important covariates and another strand of research has focused on the development of bias-reduced estimators. The general approach of these methods is to reweight the sample using poststratification (O'Muircheartaigh & Hedges, 2014; Tipton, 2013) or inverse probability weighting (IPW; Stuart, Bradshaw, & Leaf, 2014; Stuart et al., 2011).

Propensity score methodologies are at the core of both these assessment and reweighting approaches. Propensity scores were developed for removing bias resulting from nonrandom treatment assignment in observational studies (Rosenbaum & Rubin, 1983) and have since been applied to address other missing data problems (e.g., Little, 1986). The literature on propensity scores in observational studies is extensive, providing a standard suite of methods for assessing covariate balance (e.g., Stuart, 2010), for selecting covariates (e.g., Steiner, Cook, Shadish, & Clark, 2010), and for judging the performance of reweighting methods (e.g., Guo & Fraser, 2014; Rosenbaum & Rubin, 1984). Along with standard methods, formal and informal guidelines have been developed about how to assess the adequacy of these methods in controlling the bias.

However, these rules of thumb were developed in the context of the kinds of problems and data found in observational studies. For example, observational studies typically include moderate-sized treatment groups (greater than 100 units) and pools of eligible cases from which control groups are composed that are 2 or more times as large as the treated group.

In problems of generalization from randomized trials, propensity scores are used to compare the sample of units in an experiment to the set of units in one or more inference populations. In many fields, particularly in education and social welfare, a cluster sampling design is employed; in this design, the primary sampling unit is that of an aggregate—for example, schools, school districts, or counties—and within these units, all individuals are included in the study. This means that the typical generalization study includes only a small number of sampling units—typically from 10 to 70—while the inference population often includes over 100 times as many units. The smallest of these experimental sample sizes (e.g., 10–20 units) are found in multisite randomized experiments (i.e., experiments with randomized block designs, where the blocks are treated as random), wherein units in each site are randomized to treatment conditions, while somewhat larger sample sizes (e.g., 30–70 units) are common in cluster randomized experiments, wherein whole sites are randomized to conditions. This *small sample size* problem means the application of propensity score methods to support generalization from an experiment presents challenges not commonly addressed in the quasi-experimental literature (e.g., Tipton, 2013). Since

these features are different than those of the typical observational study, little is known about how well these post hoc propensity score–based assessments and adjustments may perform in supporting generalizations from typical experiments.

A starting point for understanding methods for generalization under nonrandom samples is to understand how well these methods would work under the theoretically ideal condition, random sampling. Random sampling is typically considered ideal since it eliminates the bias that results when sites volunteer or select into an experiment. For example, to assess how well propensity score matching could be expected to improve the covariate balance, the covariate balance that would be expected under random sampling is a useful benchmark. The goal of this article is to understand properties of statistical methods that can be used to assess and support generalization from experiments under random samples and to use these properties to develop rules of thumb that can be used to help researchers and policy makers when assessing the generalizability of an experiment to a population.

We begin by exploring four indices commonly used for assessing similarity between a sample and population—the absolute standardized mean difference (ASMD; for individual covariates), the standardized mean difference (SMD) of the logits (commonly used to assess generalizability; see Stuart et al., 2011), a measure of undercoverage (Tipton, 2013), and the generalization index (Tipton, 2014a). We then turn our focus to two reweighting estimators—poststratification and IPW—that are commonly used to estimate the population effects. In order to develop rules of thumb, we ask: What values of these statistics might we expect to find in a random sample? We answer this using both theoretical approximations and a simulation study based on data from the Common Core of Data (a commonly used population frame). In order to make the problem and application of the findings clear, we situate this article around an experiment recently conducted in Indiana, where the focus is on generalizing the treatment effect to a population of schools in the state.

Framework for Generalization

Propensity Scores

We begin by briefly reviewing the sampling propensity score framework used in the generalization literature. In order to understand how these approaches work as well as the assumptions needed for generalization, it

is useful to adapt the potential outcomes framework to the context of generalizability. In the typical case, we have a sample, S, that contains n units (these could be aggregates, such as schools or clinics or individuals) and an inference population, P, that includes N units. We assume that the experiment was conducted in the sample and the sample was not randomly drawn from the population. The experiment provides an estimate of the sample average treatment effect (SATE), but the effect of substantive interest is the population average treatment effect (PATE). To see why these two quantities may differ, for each unit in the population and the experimental sample, let $W = 1$ if the unit receives the treatment and $W = 0$ otherwise. Further, let $Z = 1$ if the unit is in the experimental sample and $Z = 0$ otherwise. Each unit is assumed to have two potential outcomes $Y(0) = Y(W = 0)$ under the control condition and $Y(1) = Y(W = 1)$ under the treatment condition, such that the treatment effect for each unit in the sample or the population is $\Delta = Y(1) - Y(0)$. However, for each unit, we observe only one of the two potential outcomes. Because of randomization in the experimental sample, the experiment assures that potential outcomes are missing completely at random. So an unbiased SATE, $\tau^S = E[\Delta|Z = 1]$, can be estimated using the difference in mean outcomes in the treatment and control groups. However, the PATE, $\tau^P = \Pr(Z = 1)\, E[\Delta|Z = 1] + (1 - \Pr(Z = 1))\, E[\Delta|Z = 0]$. The PATE is equivalent to the SATE only when $E[\Delta|Z = 1] = E[\Delta|Z = 0]$. This is the case when the experimental sample is randomly drawn from the population, when $\Pr(Z = 1) = 1$, or if the sample selection and treatment effect heterogeneity are independent (Imai, King, & Stuart, 2008; Rubin, 1974). Otherwise the SATE is considered to be a naïve estimate of the PATE and in most cases biased.

At its core, generalization is a missing data problem. In other missing data contexts—for example, observational studies, missing data, and survey nonresponse—propensity score methods have been used to account and adjust for bias induced from missingness. In generalization, these standard propensity score methods have been adapted and used to develop tools for both assessing generalizability and for producing a bias-reduced estimate of the PATE (e.g., O'Muircheartaigh & Hedges, 2014; Stuart et al., 2011; Tipton, 2013). This approach requires that the researcher has access to a set of covariates, $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ for each unit in the sample and the population; in education, common population frames include the Common Core of Data (National Center for Education Statistics) and state longitudinal data systems.

The sampling propensity score is defined as $s(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$ and is typically estimated using a logistic regression model, such as:

$$\text{Log}\Big(s(\mathbf{X})/[1 - s(\mathbf{X})]\Big) = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j,$$

where $\mathbf{X}$ is an $N \times p$ design matrix containing covariate information on both units in the sample and population and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. Based on this model, estimated propensity scores and their logits are defined for each unit, $i = 1, \ldots, N$, in the population (and thus sample) as:

$$s_i = 1/\Big(1 + \exp(-L_i)\Big) \text{ and } L_i = \text{logit}\Big(s(\mathbf{X}_i)\Big) = \mathbf{x}_i\mathbf{b},$$

where $\mathbf{x}_i$ is a $1 \times p$ vector containing information on features of unit i and $\mathbf{b}$ is a $p \times 1$ vector of estimated logistic regression coefficients. Typically, these estimated logits are used to assess generalizability (by comparing their distributions in the sample and population) and to conduct post hoc adjustments in order to estimate the PATE.

Assumptions

Random sampling provides a model free basis for generalization. Propensity score–based methods for generalization require three assumptions to ensure their validity. First, the stable unit treatment value assumption (SUTVA) must hold for all units in the experiment (Rubin, 1978, 1980, 1990) and in the population (Tipton, 2013). As in the traditional experimental literature, this condition requires that there is no interference between treatment and control units in the experiment and there can be only one version of the treatment. As such, each unit's potential outcomes are determined only by whether or not they receive treatment and not the treatment assigned to any other unit. In the case of generalization, however, the SUTVA also extends to the sample selection process as well. That is, in the generalization context, the SUTVA also requires no interference between units in the experiment and those not included in the experiment. Notably, this requires that potential outcomes cannot be a function of the proportion of units that take part in the experiment.

Second, generalization using propensity score methods requires strongly ignorable treatment assignment in the experiment. That is:

$$[Y(1), Y(0)] \perp W|Z = 1, s(\mathbf{X}) \text{ and } 0 < \Pr(W = 1|Z = 1, s(\mathbf{X})) < 1.$$

This means that within the experiment, the probability of receiving the treatment must be nonzero for all units and treatment assignment must not be confounded with potential outcomes. This assumption is generally met through random assignment, but it could also be met in an observational study if all covariates affecting both treatment selection and outcomes were controlled for.

Finally, generalization using propensity score methods requires a strongly ignorable sample selection given the propensity score:

$$\Delta = [Y(1) - Y(0)] \perp Z \,|s(\mathbf{X}) \text{ and } 0 < s(\mathbf{X}) \leq 1.$$

This condition requires that $\mathbf{X}$ includes all covariates that differ in the experimental sample and the population and are related to treatment effects. That is, $\mathbf{X}$ must include all characteristics that explain the treatment effect heterogeneity that differ between the experimental sample and the population. This condition also requires that there is sufficient overlap between the experimental sample and the population on $s(\mathbf{X})$, a condition that can occasionally be challenging to meet when the experimental sample is small (as we will discuss below in the Results section, "Simulation Study 2"). When these conditions have been met, the estimate produced using post hoc adjustments—providing sufficient similarity is achieved—produces an unbiased estimate of the PATE.

Indiana Example

To make this more concrete, we situate our investigation of generalization from experiments with small samples in an example. The data we examine are drawn from a cluster randomized trial (Konstantopoulos, Miller, & Van der Ploeg, 2013) that was designed to study the effect of Indiana's benchmark assessment system on student achievement in mathematics and English Language Arts based on annual Indiana Statewide Testing for Educational Progress Plus scores. Fifty-six K-8 schools volunteered to implement the system in the 2009–2010 school year. The process of selection into the study primarily reflected the principal's interest in implementing the new benchmark assessment system. A sample of study school principals were asked why they had decided to apply, and their responses suggest a wide variety of reasons, none having to do with student performance levels or concerns about how these levels were changing in their schools. Some principals saw the intervention as an opportunity to take advantage of free resources from the state, others cited a preexisting interest

in data-driven decision-making, yet others mentioned knowing other schools that had implemented the program in the earlier pilot test.

Of the volunteering schools, 34 were randomly assigned to the state's benchmark assessment system while 22 served as controls. Students in the treatment schools were regularly assessed using formative assessments aligned with the state test and teachers received feedback on their performance intended to guide instruction. The size of the sample in this randomized controlled trial (RCT) is typical of many cluster RCTs conducted in education and is thus a fitting example for the purposes of this article.

Given the process by which schools volunteered to be in the study, the experimental schools were clearly not a randomly sampled subset of schools in the state. However, an important policy question the study sought to answer was: What would be the effect of implementing the benchmark assessment system statewide (rather than just those schools that volunteered to be randomly assigned)?

Using the generalization methods described in the previous section, the first step to answering this question was to precisely define the population of interest. We began by defining the population of interest as all "regular" schools serving students in Kindergarten through an 8th grade in the state of Indiana in the 2009–2010 school year; based on data from the Common Core of Data, this resulted in an inference population of 1,514 schools.[1] Figure 1 shows the locations of the experimental and nonexperimental schools in the state.

After defining the population, we selected a set of covariates (that were measured on the entire population of schools) that we believed to explain variability in school average treatment effects (ATEs). At present, there is limited empirical evidence about what covariates are related to treatment effect heterogeneity, so covariates were selected based on substantive theory about what might account for treatment effect heterogeneity. For example, the effect of the benchmark assessment system might vary based on a school's past performance, the demographic makeup of the school, or the student–teacher ratio in the school. We drew on data on schools in Indiana (and thus the experiment) from both the national Common Core of Data and from the Indiana Department of Education. Overall, we selected 14 covariates, including the aggregate demographic measures (percentage of male, percentage of White, percentage of English language learners, percentage of special education students, and percentage of students who qualify for free or reduced price lunch), school features (whether the school qualified for Title I and school-wide Title I, school size, teacher–student ratio, attendance, and the size of the county the school was located in), and most

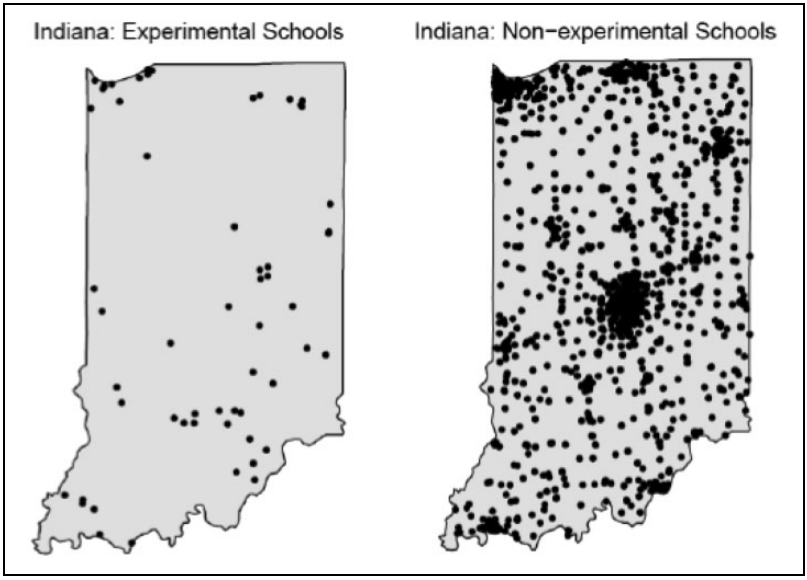

Figure 1. Experimental and nonexperimental schools in Indiana.

importantly, a pretest measure (past academic performance in both math and English language). The pretest measure used here is particularly important, since it aligns exactly with the outcome measure used in the experiment.

In the next section, we use this inference population data frame (of 1,514 schools and their associated covariates) to investigate properties of small random samples. In particular, we focus on small sample properties of measures of covariate balance, generalizability assessment, and the effectiveness of reweighting measures. After developing these properties (based on random samples), we return to the experimental schools and use the intuition and rules of thumb to both assess how generalizable the experimental sample is to the Indiana population and the extent to which a less-biased treatment effect estimate can be created using existing methodology.

Generalizability Statistics

Recall that for each unit in the sample and population, we can estimate the propensity score logit ($L_i = \mathbf{x}_i\mathbf{b}$) using logistic regression. Since propensity scores (and thus their logits) are balancing scores (Rosenbaum & Rubin,

1983), we know that, conditional on the value of L_i—when the assumptions have been met—the marginal distributions of the covariates in **X** are identical. This means that the question of similarity between a sample and population on the multivariate covariates in **X** can be reduced to a question of similarity on the univariate logits instead. Thus, the questions of assessment and the PATE estimation hinge on understanding the degree of similarity between the distribution of L_i in the sample and population.

In this section, we begin by introducing four statistics commonly used for assessment of covariate balance. Each of these statistics (or their combination) can be used to answer if a sample is "similar enough" to a population (on a set of observable covariates) to warrant generalization. The question of similar enough requires a rule of thumb or threshold for each statistic; to date, the literature has answered this question by relying on rules of thumb generated from applications in observational studies (where treatment and control are compared). As we argued in the introduction, the application of propensity score matching to bias reduction in observational studies is similar to its application in generalization from experiments, but the conditions under which it is applied are different in important ways. This raises the question of whether the guidelines used in applications to reducing bias in observational studies are applicable to applications in generalization. We therefore focus here on developing properties based on the performance of these statistics in applications to generalization.

The goal of propensity score matching in generalization from experiments is to obtain covariate balance that is sufficient for generalization to the inference population. Random samples are the "gold standard" for generalization; therefore, we use the degree of covariate balance that is likely to be obtained in random samples as the criterion of whether covariate balance is sufficient to support generalization. We begin by deriving the sampling distribution of each statistic in simple cases; in the next section, we compare rules of thumb based on these theoretical values to empirical values based on a simulation study with real data.

After discussing measures of assessment, we then turn to methods for reweighting. While many methods are available in the more general propensity score literature, in research on generalization, two PATE estimators are most common: Poststratification (i.e., subclassification; see O'Muircheartaigh & Hedges, 2014) and IPW (see Stuart et al., 2011). Our focus here is on highlighting possible problems that can arise with each estimator; we then compare these estimators in a broad variety of conditions using real data in the next section.

Assessment

Statistics for assessing generalizability. When assessing generalizability, three global measures and one covariate-level measure are often used. We begin here with the covariate-level measure—the ASMD defined for each covariate X_j ($j = 1, \ldots, p$) as:

$$|d_j| = \left| \frac{\bar{X}_{jS} - \mu_{jP}}{\sigma_{jP}} \right|, \tag{1}$$

where $\bar{X}_{jS}$ is the mean in the sample, μ_{jP} is the population mean, and σ_{jP} is the population standard deviation. This is calculated for each of the p covariates included in the propensity score. We focus here on the absolute value since, following the literature, we are rarely interested in direction but instead in magnitude. When assessing similarity, researchers are often interested in both covariate by covariate comparisons (each value of $|d_j|$) and aggregates of these (e.g., $|d| = \sum_{j=1}^{p} \Sigma |d_j|/p$).

The next three measures are global and focus not on the individual covariates but instead on the propensity score logits. Let L_s and L_p be the average logits in the sample and population, respectively, and s_P be the standard deviation of the logits in the population. Then similar to the ASMD for covariates, the SMD for the logits provides a global summary (Stuart et al., 2011),

$$L = \frac{L_S - L_P}{s_P}. \tag{2}$$

Assuming the logits are normally distributed in the sample and population, the logit SMD (L) thus provides a measure of the distance between the two distributions. Like the ASMD for the individual covariates, the logit SMD is commonly used as a measure of assessment (i.e., balance) in observational studies (see Stuart, 2010).

A third measure is the generalizability index, B (Tipton, 2014a), which provides a comparison of the densities of the logits in the sample and population. Unlike the logit SMD, theory for this index does not hinge on the distribution of the logits being normal. This index ranges from 0 to 1, with a 1 indicating that the sample is identical to the population (the highest degree of generalization). B is defined as:

$$B = \sum_{j=1}^{k} \sqrt{w_{p_j} w_{s_j}}. \tag{3}$$

In order to calculate B, the distributions of logits are divided into $j = 1, \ldots, k$ bins, each containing the proportions w_{p_j} and w_{s_j} of the population and sample, respectively. Tipton suggests using Silverman's rule (1986) to define the bins, where the bin width is defined as $[4\,\nu^5/(3r)]^{1/5}$, where ν and r are the variance and size of the logits in the sample or population, respectively.

The fourth statistic commonly used is a measure of the degree of overlap of the distributions of logits in the sample and population. Tipton (2013) shows that when there is insufficient overlap between experimental sample and inference population, it reduces the number of strata that can be created for poststratification and the amount of bias reduction expected. Put another way, this overlap problem indicates that there is undercoverage—parts of the population without similar units in the sample—making generalization difficult (since it entails extrapolation). Undercoverage is typically defined as:

$$O = 1 - \int_{\min s_S(\mathbf{x})}^{\max s_S(\mathbf{x})} s_P(\mathbf{x})\, dp, \tag{4}$$

which is simply the proportion of the distribution of the propensity scores in the population that fall outside the support of the propensity score distribution in the sample. Following the observational studies literature, it is typical to define the *min* and *max* of the propensity score distributions in the sample based on the 5th and 95th percentile.

Each of these four statistics can then be used to determine if the sample is similar enough to the population (on observables) to warrant generalization of the experimental findings. For the ASMD and the logit SMD, researchers have borrowed rules of thumb generated in observational studies, with similarity achieved when the values are smaller than 0.25 (Rubin, 2001) or 0.10 (e.g., Normand et al., 2001). For the generalizability index, Tipton (2014a) shows, based upon a simulation study, that $B > .90$ corresponds to being "like a random sample" (very high generalizability) and values of $B < .50$ indicating that generalization is unwarranted (low generalizability). Finally, for the measure of undercoverage, no thresholds have been proposed; however, when undercoverage is high, the literature shows that post hoc adjustments may not perform well, making it difficult to generalize (e.g., Tipton, 2013).

Properties of assessment statistics. Given that these statistics are commonly used to assess similarity between a sample and population, an important

question is how these statistics would perform in random samples of the size commonly found in field experiments. While properties of the ASMD are straightforward (as we will show next), in order to develop properties of the three global measures, we begin by approximating the distribution of the logits, L_i, recall that in the logistic regression model, for units, $i = 1, \ldots, N$, and covariates, $j = 1, \ldots, p$, we have logits:

$$L_i = b_0 + \sum_{j=1}^{p} b_j X_{ij},$$

where the b_j are the logistic regression coefficient estimates. We can approximate the distribution of these logits as:

$$L_i \sim N\left(0, \sum_{j=1}^{p} d_j^2\right) \text{in the population} \tag{5}$$

and

$$L_i \sim N\left(\sum_{j=1}^{p} d_j^2, \sum_{j=1}^{p} d_j^2\right) \text{in the sample,}$$

where $\sum_{j=1}^{p} d_j^2$ is the sum of the squared individual SMD (d_j; Equation 1) for each of the p covariates in the propensity score model.

In order to make these approximations, we rely on three assumptions. First, we assume that the covariates in the model are independent (i.e., $\rho_{ij} = 0$ for all $j = 1, \ldots, p$). While this is clearly unrealistic, given the goal of meeting the sampling ignorability condition (wherein the goal is to explain *all* variation in treatment impacts), we would expect in practice for the average correlation to be relatively small. Second, we calculate the b_j using ordinary least squares (OLS; which approximates the first Newton–Raphson iteration in logistic regression), which gives closed form solutions. Third, we assume that the covariates X_{ij} are standardized to have mean 0 and variance 1 in the population. Each of these assumptions seems to be particularly strong, but as the results that follow indicate, approximations based on them perform very close to the behavior found in real data.

ASMD. Deriving the sampling distribution of the ASMD, $|d_j|$, is the simplest, since it is a function of the mean, $\bar{X}_{jS}$, that is:

$$d_j \sim N(0, 1/n)$$

and thus the ASMD follows a half-normal distribution,

$$|d_j| \sim \text{HN}\left(\sqrt{2/(\pi n)},\ 1/n\right). \tag{6}$$

Importantly, unlike the SMD, where $E(d_j) = 0$, for the half-normal distribution, the average value is a function of sample size, that is $E(|d_j|) = \sqrt{2/(\pi n)}$ (see, e.g., Johnson & Kotz, 1971). As we discuss in the Results section, this means that in small samples, we are likely to find sometimes large imbalances simply by chance.

Logit SMD. Applying the distribution of L_i developed above (Equation 5) and the requisite assumptions to the definition of L (Equation 2), it can be shown that, in general, L can therefore be written as:

$$L = \frac{\sum_{j=1}^{p} b_j d_j}{\sqrt{\sum_{j=1}^{p} b_j^2 + \sum_{j=1}^{p}\sum_{k=1}^{p} b_j b_k \rho_{jk}}} = \frac{\sum_{j=1}^{p} b_j d_j}{\sqrt{\sum_{j=1}^{p} b_j^2}} = \sqrt{\sum_{j=1}^{p} d_j^2}$$

$$= \sqrt{\frac{\sum_{j=1}^{p} (\sqrt{n} d_j)^2}{n}} \sim \sqrt{\frac{\chi_p^2}{n}}, \tag{7}$$

where in the first equality, ρ_{jk} is the correlation between covariates X_j and X_k in the population and the second equality is based upon the assumption of independence. The third equality comes from the use of OLS to estimate the coefficients (wherein $b_j = d_j$), and the final result comes from the fact that under random sampling, $d_j\sqrt{n}$, is a standard normal variable. The result is that under random sampling, nL^2 follows a χ^2 distribution with p degrees of freedom. Using properties of the χ^2 distribution, this means that $E(L) = \sqrt{p/n}$, which is a function of both the number of covariates, p, and the experimental sample size, n.

Generalizability index. Following the logic above, if we assume that the distribution of the logits in the sample and population are both normal (a more simple case than likely be found in practice), then we can use the analytic form of the B index (see Tipton, 2014a) in order to develop the sampling distribution, as:

$$B = \exp(-L^2/8) = \exp[-\chi_p^2/(8n)]. \tag{8}$$

Thus, $(-8n)\log(B)$ follows a χ^2 distribution with p degrees of freedom. Again, this means that following the properties of the χ^2 distribution, $E(B) \approx \exp[-(1/8)(p/n)]$.

Undercoverage. Finally, undercoverage is ubiquitous in generalization examples (e.g., Stuart et al., 2011; Tipton, 2013) and, following protocols developed in observational studies, often leads to a redefinition of the inference population. In order to investigate this, we again return to the definition of undercoverage (Equation 4), substituting assumptions regarding the distributions of logits (Equation 5). Since both distributions are normal, this results in:

$$\begin{aligned} O &= \Pr(L_i < \min S \,|P) + \Pr(L_i > \max S \,|P) \\ &= \Phi(\min S / \sqrt{\textstyle\sum_{j=1}^{p} d_j^2}) + 1 - \Phi(\max S / \sqrt{\textstyle\sum_{j=1}^{p} d_j^2}), \end{aligned}$$

where $\min S$ and $\max S$ are the $100b$-percentile and $100(1-b)$-percentile observed values of L_i in the sample, respectively (where typically $b = .05$). While it was possible to find the distribution of functions of both L and B it is not as easy to approximate the distribution of O, since it is a function of the distribution of the order statistics (which take a complex form). However, it is possible to approximate the expected value of the O as follows:

$$\begin{aligned} E(O) \approx 1 - \Phi\Big\{\sqrt{(p/n)} + \Phi^{-1}\Big(\Big(nr(1-b) - a\Big)\Big/(n - 2a + 1)\Big)\Big\} \\ + \Phi\Big\{\sqrt{(p/n)} + \Phi^{-1}\Big(\Big(r(1+bn) - a\Big)\Big/(n - 2a + 1)\Big)\Big\}, \end{aligned} \tag{9}$$

where $a = 3/8$, p is the number of covariates, n is the sample size in the experiment, b is the percentile used when defining overlap, and $\Phi(.)$ is the cumulative density function of the standard normal distribution. Here we include r as a correction factor that we develop empirically to minimize the distance between the simulated results and this theoretical approximation. The overall form of this result is found by applying Equations 5 and 7 to an approximation of the order statistics of the normal distribution by Blom (1958).

Rules of thumb. In this article, we argue that rules of thumb based on the observational study context may be inapplicable to the context of generalization from the experiments. Here we derive more appropriate guidelines

for the context of generalization based on critical values of the sampling distributions derived here (under random sampling). Importantly, this means that critical values of the ASMD can be based upon the normal distribution, those for the logit SMD and the generalizability index can be based on the χ^2 distribution, and for the expected degree of undercoverage based on the formula provided. In the remainder of this article, we use the 95th percentile critical value to develop these rules of thumb. Importantly, we focus here on the development of rules of thumb that take into account sample size since, as we will show later, features of probability samples—the benchmark for generalizability—differ markedly in small samples.

Reweighting

Reweighting estimators. The previous four measures allow for the degree of similarity between a sample and population to be summarized and assessed. When these statistics indicate differences, then the naive estimator of the SATE may be a biased estimator of the PATE. In order to reduce this bias, reweighting methods are typically employed. The two most common are poststratification (O'Muircheartaigh & Hedges, 2014; Tipton, 2013) and IPW (Stuart et al., 2014; Stuart et al., 2011).

To implement the poststratification approach, the sample and population cases are ordered by L_i and divided into k equal-population proportion strata. The poststratification estimator of the PATE is defined as:

$$\text{PATE}_{\text{strat}} = \sum_{j=1}^{k} w_{p_j}(\bar{Y}_{T_j} - \bar{Y}_{C_j}), \tag{10}$$

where for stratum j, $\bar{Y}_{T_j}$ is the sample mean for the units in the treatment group ($W = 1$), $\bar{Y}_{C_j}$ is the sample mean for the units in the control group ($W = 0$), N_j is the number of population cases, and $w_{p_j} = N_j/N$ is the proportion of the population in stratum j. The variance of the estimate can be written as:

$$V(\text{PATE}_{\text{strat}}) = \sum_{j=1}^{k} w_{p_j}^2 V(\bar{Y}_{T_j} - \bar{Y}_{C_j}), \tag{11}$$

which is a function of the squared population weights, w_{p_j}, and the variance of the stratum-specific treatment effects, $\bar{Y}_{T_j} - \bar{Y}_{C_j}$.

Alternatively, the IPW approach calculates the mean difference in the treatment and control cases, giving each case the weight that is the inverse

of the probability of that case being selected into the sample, $1/s(\mathbf{X})$. This approach can be seen as an extension of the Horvitz–Thompson estimator (Cole & Hernan, 2008). Using weighted least squares (e.g., Stuart et al., 2014), the PATE estimator can be defined as:

$$\text{PATE}_{\text{IPW}} = \sum_{i=1}^{n_T} w^*_{T_i} Y_{T_i} - \sum_{i=1}^{n_C} w^*_{C_i} Y_{C_i}, \tag{12}$$

where Y_{T_i} and Y_{C_i} are the outcomes for the ith treatment and control cases, respectively, in the experiment and $w^*_{j_i} = [1/s_i(\mathbf{X})]/[\sum_{i=1}^{n_j} 1/s_i(\mathbf{X})]$ for $j = \{C, T\}$are the normed inverse probability weights. The variance of the IPW estimator calculated using weighted least squares can be defined as:

$$V(\text{PATE}_{\text{IPW}}) = \sigma^2 \left(\sum_{i=1}^{n_T} (w^*_{T_i})^2 + \sum_{i=1}^{n_C} (w^*_{C_i})^2 \right), \tag{13}$$

where σ^2 is the variance of the outcomes. Note that for consistency with the poststratification variance formula, this variance estimator conditions on the observed propensity score estimates.

In practice, a researcher might try the poststratification and IPW approaches and reassess similarity between the sample and population after each method (but before calculating the PATE). The final estimator is thus selected, so that the two groups are most similar to one another, both on a global measure (e.g., logit SMD and average-ASMD) and on individual covariates (ASMD), with a particular focus on increasing similarity on covariates that are thought to best explain variation in treatment impacts.

Problems with estimators of PATE. Poststratification and IPW are commonly used in the observational studies literature to reduce bias with research indicating that they tend to perform well (e.g., five strata reduces bias 90%; Rosenbaum & Rubin, 1984). In applications in generalization, however, both estimators can have problems. Poststratification, while typically viewed as the simplest to implement (and preferred by experimenters who are used to blocking), can often result in either empty strata or a limitation on the number of possible strata. For example, Tipton (2013) found that at most four equal-population probability strata could be used in one example (the SimCalc experiment), and later, we will show that in the Indiana example, only three strata are possible. In comparison, IPW, while often preferred for its relation to the Horowitz–Thompson estimator in sampling, often results in extreme weights, giving disproportionate weight to a single (or small group) of units and resulting in a large standard error. Both of

these problems are common in problems of generalization because of the "long-tail" problem that results from probabilities being pushed close to 0 (because n/N is typically very small). In the next section, we investigate these potential problems—that is, limited number of strata and extreme weights—making head-to-head comparisons of poststratification and IPW in a variety of simulated conditions.

Simulation Studies

Simulation Study 1: Measures of Assessment

In the previous section, we derived sampling distributions (under random sampling) for four measures used to assess similarity between a sample and population and based upon these, provided methods for developing rules of thumb based on critical values of these distributions. In this section, we compare these critical values to those derived empirically using the Indiana population frame. The goal is to determine how well these theoretically derived sampling distributions perform in "real" data, particularly given that their derivations were based upon approximations and assumptions (e.g., normality and independence).

For the simulations, we use the population frame from Indiana, with $N =$ 1,514 schools. Focusing on this population frame is useful in two regards. First, it allows us to compare the theoretically derived and empirical sampling distributions of the statistics in "real" data. The Indiana data are of a subset of the Common Core of Data—which is a population frame commonly used when making generalizations in education research. As such, it includes a variety of covariate types—continuous (normal and nonnormal), categorical, and dichotomous—and, importantly, these covariates are correlated (whereas the derivations assume they are not). Second, it allows us to easily link the results of this study to the Indiana experiment example (revisited in the next section).

In the simulation, we vary both the sample size ($n = 10$, 20, 30, 50, and 70) and the number of covariates in the propensity score model ($p = 7$, 14, and 26). We focus on both n and p since in the previous section, we found that both parameters played an important role in the sampling distributions. We include very small sample sizes ($n = 10$ and 20) since these are often found in multisite designs (where treatment is randomized within units) as well as small and moderate sample sizes ($n = 30$, 50, and 70) that are more common in cluster randomized designs. When $p = 14$, we use the 14 covariates defined in the previous section, which were selected to meet the

ignorability condition. For $p = 7$, we selected random subsets of these covariates. For $p = 26$, we added 12 covariates to the $p = 14$ model, most of which we did not believe were necessary for sampling ignorability, but which we believed were likely to explain variation in the outcomes; for example, we also included pretest scores for previous years (highly correlated with the immediate pretest scores already in the model) as well as in other subject areas.

For each combination of parameters, we selected 20,000 random samples using the statistical program **R** (R Core Team, 2014). In each of the iterations, we compared the sample to the population by estimating propensity scores using logistic regression. In the results shown here, we standardized the covariates based on the population standard deviation; the results are similar when using unstandardized covariates. We then used the distribution of the covariates and logits to compare measures of assessment.

Simulation Study 2: Estimation Methods

In order to compare the degree of similarity between the sample and population using the naïve estimator, poststratification, and IPW, we conducted a second simulation study. Similar to Study 1, we varied the values of both n and p and generated 20,000 random samples for each. In order to investigate properties of poststratification, in each iteration, we recorded the maximum number of equal-population sized strata that could be created (subject to the requirement that there were at least two units in each stratum, thus allowing estimation of an ATE and within-stratum variance). In order to investigate properties of IPW, in each of the iterations, we also calculated the inverse propensity weights. These weights were then compared to the expected weight N/n, where N is the population size (1,514 in this case) and n is the sample size, and divided the weights into three categories: ''2–3 times'', ''4–5 times,'' and ''at least 6 times'' the expected weight for the given sample size.

For those replications with at most five maximum strata, the iterations were then cross-classified based upon the number of maximum strata (3, 4, 5) and the maximum weight (2–3, 4–5, $\geq$6). We focus on this subset (ruling out those with potential for more than five strata and/or maximum weights less than twice the expected) since these are cases in which adjustment is needed. In essence, we are thus randomly generating samples that are purposely *not* similar to the population. By randomly generating these differences, we hope to investigate a wider variety of structures than using a researcher generated selection model.

Table 1. Properties of the Absolute Standardized Mean Difference.

	Expected Value of |d|		95th percentile of |d|	
N	Simulated (Range)	Theoretical	Simulated (Range)	Theoretical
10	.24–.26	.25	.59–.65	.62
20	.17–.18	.18	.40–.44	.44
30	.14–.15	.15	.33–.36	.36
50	.11–.11	.11	.27–.28	.28
70	.09–.10	.10	.23–.24	.23

Note. The simulated range is over the 26 covariates under study, which include a broad range of covariate types.

With these identified cases, we compared the performance of the IPW estimator and the weighted (stratified) estimator using three criteria: average ASMD, ASMD of the worst covariate, and the average rank of the estimators among the 14 covariates. Under average ASMD, an estimator was considered superior if the average of the ASMDs among the 14 covariates was smallest. For the second criterion, we focused on the covariate in each sample that had the largest ASMD and identified the estimator for which the value of ASMD was smallest. For the third criterion, the three estimators (IPW, post stratification, and unweighted/naïve) were ranked using the ASMD for each covariate. Finally, we calculated both relative and absolute measures for each; for relative measures, we collected whether the IPW or poststratified estimator performed better, while for the absolute measure, we compared the biases remaining after adjustment.

Results

Simulation Study 1

ASMD. Table 1 provides some properties of the ASMD, indicating a few trends important for generalization. First, in multisite trials ($n = 10$ and 20), we can expect to see large imbalances simply by chance. For example, with $n = 10$, on average, we would expect the ASMD to be nearly 0.25, with 5% having values greater than 0.62. In the typical cluster randomized studies ($n = 30$, 50, and 70), the average values of the ASMD are still larger than the more strenuous 0.10 threshold used for the assessment of balance in observational studies. Using the 0.25 rule commonly provided in observational studies would also be too conservative, for $n = 30$, because 5% of values from a random sample are larger than 0.36. Finally, the simulation results,

which included a variety of covariate types and distributions, indicated that the theoretical results were a good approximation to the sampling distribution.

Logit SMD. Rules of thumb based on the 95th percentile of the χ^2 distribution, as well as results based on simulations are given in Table 2. As the first six columns of Table 2 indicate, both the average and 95th percentiles for L based on the χ^2 approximation were very similar to the values found in the simulation study. For $n > 10$, the expected values and 95th percentiles were typically within 0.05 units (a relative bias of about 10–15%); the differences were larger for $n = 10$, where the theoretical estimates were larger.

Furthermore, two trends are obvious. First, as the number of covariates in the model (p) increases, more and more large values of L are found simply by chance. For example, when $n = 70$, with $p = 26$ covariates, the average value of L is 0.61 and the 95th percentile is 0.75, both of which are considerably larger than the conventional 0.25 (or 0.10) rules of thumb. Second, the sampling distribution of L is strongly driven by n. When $p = 14$, the expected value of L falls from 0.84 to 0.45 as n increases from 20 to 70, while the 95th percentile falls from 1.09 to 0.58. Altogether, this indicates that even in random samples—which are ideal for generalization—we would expect large differences in the logit SMD just by chance.

Generalizability index. The next six columns after the "logit SMD" of Table 2 indicate that like the logit SMD, in multisite trials ($n = 10$ and 20), large differences can occur simply by chance, particularly when the number of covariates used in the model (p) is large. For example, when $n = 10$ and $p = 14$, the 5% critical value is smaller than 0.70, whereas when $n = 20$, the value climbs to 0.86. The values are even smaller when p is much larger than n; when $p = 26$, the values range from 0.59 to 0.77 (for $n = 10$ and 20, respectively). These small values arise because, like the other statistics studied here, the distribution of the generalizability index depends on the ratio p/n. This suggests that the rules of thumb developed in Tipton (2014a)—that is, that the sample is considered "like" a random sample when the index is larger than 0.90—is best applied when $p \leq n$; when $p > n$, and particularly when n is small, greater leniency is applied.

Undercoverage. As the last six columns of Table 2 show, undercoverage is far more common by chance than one might expect in random samples. Not surprisingly, in very small studies ($n = 10$), these values can be quite large, even with as few as seven covariates (42% undercoverage), and are even

Table 2. Properties of the Logit SMD (L), Undercoverage (O), and Generalizability Index (B) Under Random Samples.

		Logit SMD (L)						Generalizability Index (B)						Nonoverlap Proportion (O)		
		Expected Value			95th percentile			Expected Value			95th percentile			Expected Value[a]		
p	*n*	Sim	Theo	Bias	Sim	Theo	Bias	Sim	Theo	Bias	Sim	Theo	Bias	Sim	Theo	Bias
7	10	0.75	0.84	.09	1.14	1.19	.05	.90	.92	.02	.80	.82	.02	.42	.42	.00
	20	0.55	0.59	.04	0.82	0.84	.02	.95	.96	.01	.90	.90	.00	.29	.29	.00
	30	0.45	0.48	.03	0.67	0.68	.01	.96	.97	.01	.93	.94	.01	.24	.24	.00
	50	0.35	0.37	.02	0.52	0.53	.01	.98	.98	.00	.96	.96	.00	.19	.20	−.01
	70	0.30	0.32	.02	0.43	0.45	.01	.98	.99	.01	.97	.97	.00	.17	.19	.02
14	10	0.99	1.18	.19	1.39	1.54	.15	.83	.84	.01	.68	.72	.04	.56	.54	−.02
	20	0.77	0.84	.06	1.04	1.09	.05	.91	.92	.01	.86	.85	−.01	.38	.36	−.02
	30	0.64	0.68	.04	0.86	0.89	.03	.94	.94	.00	.90	.90	.00	.31	.29	−.02
	50	0.50	0.53	.03	0.67	0.69	.02	.96	.97	.01	.94	.94	.00	.24	.23	−.01
	70	0.42	0.45	.02	0.56	0.58	.02	.97	.98	.01	.96	.95	−.01	.21	.21	.00
26	10	1.27	1.61	.34	1.74	1.97	.23	.73	.72	−.01	.57	.59	.02	.71	.70	−.01
	20	1.05	1.14	.09	1.32	1.39	.07	.86	.85	−.01	.77	.77	.00	.49	.46	−.03
	30	0.88	0.93	.05	1.09	1.14	.05	.90	.90	.00	.85	.84	−.01	.39	.36	−.03
	50	0.69	0.72	.03	0.85	0.88	.03	.94	.94	.00	.91	.90	−.01	.30	.28	−.02
	70	0.58	0.61	.03	0.72	0.75	.03	.96	.95	−.01	.93	.93	.00	.26	.24	−.02

Note. Sim = simulated value; theo = theoretical value; SMD = standardized mean difference.
[a]Expected value is based on theoretical with $a = .375$ and $b = .075$.

Table 3. Maximum Possible Strata and Largest Weights in Random Samples.

			Range of Extreme Weights		
n	Strata	How Often	2–3 Times	4–5 Times	6+ Times
20	3	.61	.70	.22	.08
	4	.28	.75	.20	.06
	5	.08	.79	.17	.04
30	3	.21	.72	.20	.07
	4	.30	.76	.18	.06
	5	.28	.80	.16	.04
50	3	.01	.75	.17	.08
	4	.07	.79	.15	.05
	5	.16	.84	.13	.03
70	3	.00	.83	.17	.00
	4	.01	.82	.15	.03
	5	.04	.84	.12	.04

Note. "How often" gives the proportion out of the 20,000 replications that the maximum number of strata (strata) occurred for a particular sample size (*n*). For each sample size (*n*) and maximum number of strata (strata), the three extreme weights columns give the proportion of random samples falling within each category. These results are from a model with $p = 14$ covariates.

larger as the number of covariates increases. Even in more moderate-sized cluster randomized studies, these indicate that it is not uncommon for as much as 30% of the population to be without similar sample units. Furthermore, as the last column of the table indicates, the theoretical approximation developed in Equation 9 is close to that found in the simulation results; these results are based on $b = .05$ (i.e., overlap defined at the 5th percentile and 95th percentile of the sample data) with a correction factor of $r = 1.5$.

Simulation Study 2

Distributions of strata and weights. One question is how often the maximum number of strata is limited in random samples. As Table 3 indicates, in very small studies ($n = 10$ and 20) common in multisite trials, the majority of the time the maximum number of equal population strata was 3 (61% of the time). Even in more moderate sizes ($n = 30$), the maximum number of strata was 3 or 4 in just over 50% of the samples. For larger sizes ($n = 50$ and 70), five or more strata were possible most of the time (92% and 99%, respectively). This means that the typical rule of thumb (to use five strata) is often only possible in moderate to large random samples.

Second, Table 3 also indicates that extreme weights ($\geq$6 times the expected) occurred much more commonly by chance in small samples (about 8% of the time) than in large samples (about 1% of the time), though moderately large weights (4–5 times the expected) were much more common overall (a little more than 20% of the time when $n = 20$ vs. about 8% of the time for $n = 70$). Additionally, it is worth noting that weights that were at least 10 times as large as the expected weight were uncommon (about 2% of the time) in the simulation studies even in very small samples.

Head-to-head comparison. Results of the simulation comparing the naïve estimator, IPW, and poststratification are found in Figure 2. In each row, poststratification and IPW are compared relative to one another for the $n =$ 20, 50, and 70 cases; the first row focuses on a model with $p = 7$ covariates, while the second row focuses on the $p = 26$ covariate case (trends for $p =$ 14, while not included in the figure to save space, were typically in between these two values). In these figures, the size of the bars indicates the proportion of the time one estimator outperforms another. Additionally, in each column, three numbers indicate (from top to bottom), the average ASMD of the covariates for the poststratified, IPW, and unweighted models.

Figure 2 indicates several trends. First, for a given sample size, when there are no extreme weights (weights less than 6 times larger than expected), IPW typically outperforms poststratification. When there are extreme weights, however, poststratification is typically best. Second, as sample size increases, IPW outperforms both poststratification and the unweighted estimator in a larger share of cases. This highlights too that in small samples ($n = 10$ and 20), reweighting is not uniformly better than the original estimator. Third, as the number of covariates in the model increases, the IPW estimator outperforms the poststratification and original estimator, but less so compared to the models with fewer covariates. This is seen primarily in the larger samples ($n = 50$ and 70 schools). Finally, as the numbers in each column indicate, while reweighting approaches do typically reduce the ASMD, the percentage of reduction is typically not that large; importantly, these trends are similar for the other measures of balance indicated above.

Indiana Example Revisited

Having developed small sample properties and rules of thumb, we now return to the Indiana example to apply them as we attempt to generalize from the experimental sample of 56 schools to the population of 1,514

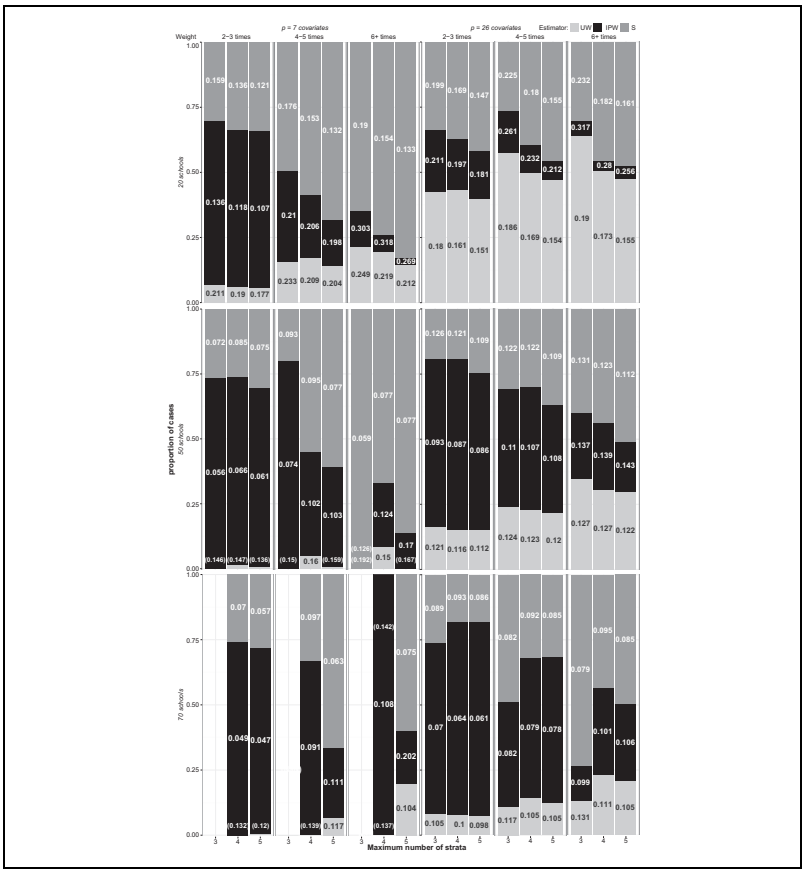

Figure 2. Head-to-head comparison of three different population average treatment effect estimates in terms of average standardized mean differences for the covariates.

schools. Having already defined the inference population and identified a set of 14 covariates, our next step is to estimate a propensity score model using logistic regression. The distribution of the propensity score logits in the experimental schools and those in the inference population can be seen in Figure 3, and the baseline balance statistics for the covariates follow in the first column of Table 4.

Figure 3 reveals that there is limited overlap between the distributions of logits in the two sets of schools. Put another way, there is an undercoverage

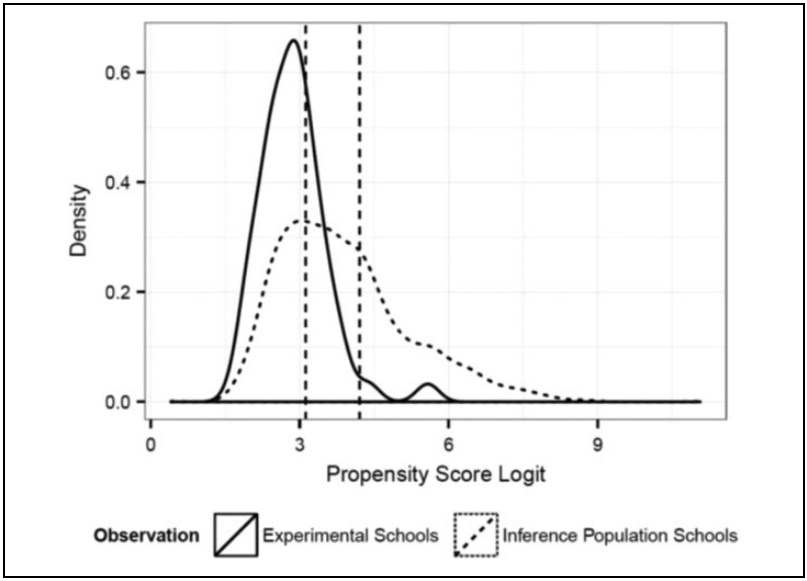

Figure 3. Distribution of logits in the experiment and population. In this figure, the vertical lines indicate the stratum "boundaries," which are the cutoff values in the propensity score logits that partitioned the distribution into three strata.

problem—approximately 14% of the 1,514 schools in the inference population have no or few experimental schools with similar propensity score logits. Furthermore, the SMD of the logits is 0.75, the maximum number of equal population strata possible is 3 (as depicted using the vertical lines), and the largest weight is 9.5 times larger than expected. As Table 4 indicates, of the 14 covariates, 6 have ASMD greater than 0.25 and 9 have ASMDs greater than 0.10.

Using the rules of thumb generated from observational studies, our initial evaluation of this example was that the study was quite different from the population of Indiana schools and would not generalize well. However, as our study of small sample properties indicates, these values and issues are much more common even in random samples than we initially anticipated. For example, the logit SMD of 0.75, while appearing large, is actually not that large given that the expected value in a random sample is approximately $\sqrt{(14/56)} = 0.50$. Similarly, by chance, we would expect the covariate ASMDs to be 0.10; following the 95th percentile of the normal distribution, only 5 of the 14 covariates have unusually large ASMD values

Table 4. ASMD Between Experimental Sample and Population.

	ASMD		
Covariates	Unweighted	IPW	Stratified
2008–2009 ELA	**.44**	.25	.17
2008–2009 Math	.03	.12	.09
2008–2009 Attendance	.07	**.28**	.25
2008–2009 Full-time staff	**.35**	.23	.16
2008–2009 Population of students	.27	.17	.17
2008–2009 Pupil–teacher ratio	.22	.19	.05
County population	**.44**	**.37**	.25
2008–2009 Title I status	.15	.22	.22
2008–2009 School-wide Title I status	.07	.17	.07
Proportion of male students	.01	.23	**.36**
Proportion of White students	**.29**	.18	.19
Proportion of special education students	.18	.18	**.38**
Proportion of free/reduced price students	.02	.11	.14
Proportion of limited English proficiency students	**.32**	**.46**	**.42**
Propensity score logits	**.75**	NA	.16

Note. Bold values indicate those larger than the 95th percentile for the ASMD under random samples. ASMD = absolute standardized mean difference; IPW = inverse probability weighting; ELA = English Language Arts; NA = not applicable.

(i.e., greater than 0.28; see Table 2). Even the overlap problems—which initially indicated a large degree of undercoverage—are more common than expected with the average random sample having even worse overlap (22% of the population undercovered; see Equation 9 with $p = 14$, $n = 56$, $b = .05$, $r = 1.5$). Overall, this suggests that the degree of mismatch between the sample and population found in this example is more common than we would have previously expected and that the sample is more "like" a random sample than is obvious at first blush. Importantly, of the previously defined rules of thumb studied here, only that for the generalizability index (calculated to be 0.90) effectively identifies this feature.

Furthermore, the results of the simulation study suggest that given this situation, with large maximum weights and a maximum of only three strata, that bias reduction will likely be better using poststratification (compared to IPW). However, in absolute terms, our simulation study also suggests that with only three strata, bias reduction may be limited, and that an unbiased estimator would be impossible. In Table 4, we include balance statistics for

both poststratification (with three strata) and IPW approaches. These results are mixed, but generally in line with our simulation results. Whereas before reweighting, the average ASMD was 0.20, for IPW, it was 0.22, and for poststratification, it was 0.21. Before reweighting, five of the covariates had ASMD greater than 0.28 (the 5th percentile upper bound, given in Table 2), whereas this was reduced to three for both IPW and poststratification. All three approaches had one covariate with an ASMD larger than 0.40, though the covariate differed in each approach. The results indicate that poststratification and IPW are rather similar and that while they offer some improvement over the unweighted estimator, the improvement is minimal.

Finally, we calculated the PATE for the population of 1,514 Indiana schools using both IPW and poststratification and compared it to the unweighted estimator. Given the generalizability index estimate of 0.90, Tipton (2014a) suggests that reweighting will not have large implications for standard errors. In comparison to the original ATE estimate of 0.20 (0.12) in the sample, using the IPW model results in an estimate of 0.19 (0.12), and the poststratification estimate is 0.26 (0.15), which indeed are not that much different.

While at first these results seem perplexing, the simulation results can help us to make sense of them. By allowing us to explicitly compare the experimental sample to an expected randomized sample, we conclude that the sample that selected into the experiment in this case was not so different (on observables) from what we would expect had 56 schools been drawn at random from the 1,514 schools in the inference population. This is a departure from what we might have concluded had we applied traditional rules of thumb from the observational studies literature. Further, the simulation results suggest that neither IPW nor poststratification is likely to yield a substantial improvement in bias reduction in this case, which is consistent with the empirical results we find.

Overall, in the example, this means that we would likely select the unweighted estimator as the best available estimate of the benchmark assessment system for K-8 schools in Indiana. Our confidence in this estimate would be strengthened by the fact that the experimental sample was not too different (on observables) from what would be expected of a random sample from the population in this case. However, it is important to note that the strength of this claim of generalizability depends not only on this assessment of similarity on observables, but also on the extent to which researchers believe that the observables included in this analysis are sufficient. Another way of saying this is that the requirement for similarity on observables is a *necessary* but not *sufficient* condition for generalizability.

Discussion and Conclusion

This article addresses the problems of generalization from experiments with small, nonprobability samples to well-defined inference populations. It is natural to approach this problem in terms of the extent to which the experimental sample matches the inference population on potentially confounding covariates (covariates that are related to treatment effects). However, this raises the issue of how to judge the adequacy of the match between the experimental sample and the inference population.

Because this matching problem is similar to that in matching treatment groups to comparison groups in observational studies, it is tempting to draw on practical experience from observational studies. But as this article highlights, there is a major difference between the matching problem in observational studies and in generalization from experiments. Observational studies typically involve large numbers of potential comparison group members, only a fraction of which need to be matched to treatment group members. With large samples, the benchmark then is a rule of thumb indicating bounds within which regression adjustments will perform well (Rubin, 2001). Experiments (particularly cluster randomized experiments) typically have a small number of units that need to be matched to all of the units in a large population. In this context, regression adjustments are rare, with the focus instead remaining on simply assessing overall similarity (a yes–no generalizability decision) or on adjusting through reweighting. These differences mean that practical experience and guidelines for ''goodness of matching'' developed in the observational study context may be less relevant to the context of generalization from experiments.

Probability sampling is the gold standard for generalizing from samples. Thus, we have emphasized the use of the performance of probability samples in making judgments of the adequacy of matching experimental samples to inference populations. The idea is to use the adequacy of matching that would be expected if the experiment had a probability sample to develop benchmarks of adequate matching. There is no reason to expect small experimental samples to match inference populations better than probability samples.

As it turns out, small probability samples exhibit larger degrees of mismatching and population undercoverage than might be expected in the observational study context. This degree of mismatch is particularly high when the number of covariates studied (p) is larger than the number of sites (n) in the experiment. The fact that the ratio p/n drives all the three of the global assessment statistics is important and can help us understand an

otherwise paradoxical result: That a study can have a sample size sufficient for estimating the average treatment impact but not for ensuring similarity between the sample and population. This is because in order to assess generalization, similarity must be compared on multiple covariates, thus inducing the same multiple-comparisons issues common in other areas of statistics.

Importantly, we are not saying that probability samples are biased. The sampling distribution of estimates takes the differences in values of confounding covariates into account. What we are saying is that the differences between the means of small probability samples on confounding covariates and the means of the inference population from which they are drawn are often surprisingly large. We may desire that these differences are small in nonprobability samples, but we should not expect them to be smaller than in probability samples. This article gives some insight about how small we can reasonably expect them to be and provides some theory about how to evaluate them.

One might argue that a traditional sample designer would use stratified random sampling to better match the inference population and reduce undercoverage. That is certainly true if sample design was considered in study planning. However, if sample design for generalization was a major consideration in study planning, stratification could also be used with nonprobability samples (see Tipton, 2014b; Tipton et al., 2014) and would similarly improve the match of the experimental sample to the inference population and reduce undercoverage. While planning for generalization was not a major objective in sample design, we would argue that the relevant probability-sampling model that should serve as a standard of comparison is simple random sampling, the only probability-sampling model that requires no a priori thinking about potential stratifiers.

In both cases, the promise of stratification to improve matching of sample to population depends on knowledge of covariates that are related to relevant variation (in this case, variation of treatment effects). Although variation in treatment effects is currently of great research interest, the amount of empirical evidence about treatment effect variation and its correlates is quite modest (see, e.g., Schochet, Puma, & Deke, 2014). Better empirical evidence about treatment effect variation and its correlates is badly needed. This limitation in current knowledge is true also regarding the relationship between bias in the covariates and that in treatment impacts. Ideally, rules of thumb would be based on this empirical relationship, indicating the degree to which different adjustment methods may reduce bias. However, even without solid empirical evidence, we might still

improve study design by careful thought about likely correlates of treatment effect variation and designing experimental samples to take those into account.

The approach we've taken throughout this article is to develop rules of thumb based upon features of random samples; importantly, these rules of thumb—based on the sampling distributions of the statistics under normality—take into account the sample size in the study. Our focus throughout has been on samples of size 10–70, however, and these rules of thumb may be less useful in experiments with very large numbers of sites. For example, in studies with more than 100 sites, it is easy to see that probability samples would perform quite well and even small SMD would be statistically significant. To some degree, these results may apply here nonetheless, since larger samples make it easier to adjust for more covariates, thus driving up the ratio p/n. In large samples, nonetheless, reweighting adjustments may perform quite well and regression adjustments may be feasible; thus this may be exactly the case in which the observational studies rules of thumb apply. In comparison, studies with very small numbers of sites—for example, those with 10 or fewer—may at first be heartened by these results, suggesting that generalizations can be easily made even when the sample and population are radically different. While this is certainly possible, the insufficient ability to detect variation in treatment impacts across sites—necessary for generalization—may provide an adequate limitation.

This article implies that sharp inferences to large populations from small experiments is difficult even with probability sampling. We have also suggested some methods for evaluating the adequacy of experimental samples to support generalizations. For example, we find that simply by chance, the SMD of the logits and the measure of overlap are both typically large, making standard rules of thumb for each somewhat poor measures for assessing generalizability; in contrast, in samples larger than 20, the rules of thumb proposed for the generalizability index give a much more straightforward assessment of similarity (by taking into account the sample size and the distribution of the logits). We have also derived the sampling distribution of these statistics and shown that these distributions can be used to determine if a sample is sufficiently similar to a population to be "like" a random sample. Additionally, we find that when there are large differences, reweighting is not always very effective, particularly in small sample sizes, and that whether one method is more or less effective depends on both the degree of overlap and the size of the weights. We hope that these findings will help education scientists to better judge the adequacy of generalizations that may be desired in the evidence-based policy context.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Authors Hedges and Tipton received an NSF grant 1118978.

Note

1. In defining "regular" or typical schools, we excluded from the inference population all charter schools, schools whose populations were over 95% male, were over 95% English language learners or students with special education needs or served fewer than 100 students; this reduced the population size from 1,587 to 1,514 schools.

References

Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York, NY: Wiley.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*, 656–664.

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (Vol. 11). Thousand Oaks, CA: Sage.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, *171*, 481–502.

Johnson, N. L., & Kotz, S. (1971). *Distributions in statistics: Continuous univariate distributions* (Vol. 1, Wiley Series in Probability and Mathematical Statistics). New York, NY: Wiley.

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*, 481–499.

Little, R. J. A. (1986). Survey nonresponse adjustments. *International Statistical Review / Revue Internationale de Statistique*, *54*, 139–157

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*, 387–398.

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*, 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*, 195–210.

Orr, L. L. (2015). 2014 Rossi award lecture: Beyond internal validity. *Evaluation Review*, *39*, 167–178. doi:10.1177/0193841X15573659

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472–480.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). Boca Raton, FL: CRC.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2014). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*, 1–11.

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*, *2*, 369–386.

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? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics*, *39*, 478–501.

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

Tipton, E., Hedges, L. V., Vaden-Kiernan, M., Borman, G. D., 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*, 114–135.

Author Biographies

Elizabeth Tipton is an assistant professor of applied statistics at Teachers College, Columbia University. Her research focuses on issues of generalizability in the design and analysis of large-scale experiments and meta-analysis.

Kelly Hallberg is the managing director at the University of Chicago-Urban Labs and a senior research associate at the Harris School of Public Policy. She oversees a portfolio of applied research projects examining innovative approaches to reducing violence and improving the academic and life outcomes of urban youth.

Larry V. Hedges is the board of trustee professor of statistics, education and social policy, and psychology at Northwestern University. His research specialties include the design and analysis of social experiments, meta-analysis, and social policy analysis.

Wendy Chan is an assistant professor of education in the Graduate School of Education at the University of Pennsylvania. Her research interests include applications of partial identification and small area estimation methods to problems of external validity in the social sciences.