Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 3, pp. 239–266
DOI: 10.3102/1076998612441947

Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts

Elizabeth Tipton
Columbia University

As a result of the use of random assignment to treatment, randomized experiments typically have high internal validity. However, units are very rarely randomly selected from a well-defined population of interest into an experiment; this results in low external validity. Under nonrandom sampling, this means that the estimate of the sample average treatment effect calculated in the experiment can be a biased estimate of the population average treatment effect. This article explores the use of the propensity score subclassification estimator as a means for improving generalizations from experiments. It first lays out the assumptions necessary for generalizations, then investigates the amount of bias reduction and average variance inflation that is likely when compared to a conventional estimator. It concludes with a discussion of issues that arise when the population of interest is not well represented by the experiment, and an example.

Keywords: *experiment; program evaluation; external validity; causal inference; propensity scores*

While prized for their high internal validity, social experiments typically have low external validity (Shadish, Cook, & Campbell, 2002), making it difficult to generalize from the treatment effect estimated in the experiment to the effect expected in a larger population. The fact that within an experiment, units are randomly assigned to treatment conditions results in the experiment having high internal validity, while the fact that the units in the experiment are *not* randomly selected from a well-defined population leads to the experiment's low external validity. For example, this problem occurs when policymakers need to generalize from the results of an experiment conducted on a convenience sample of schools to the effect expected for the population of schools in a particular state.

The problem of generalization is rarely addressed in the experimental design literature (Pearl & Bareinboim, 2011; Shadish, 2010; for exceptions see Cochran & Cox, 1992; Kish, 1987). In practice, the process commonly employed is like the two-bridges analogy described by Cornfield and Tukey (1956). They argue that making generalizations to a population of interest from an experiment involves two steps or bridges. The first bridge of generalization is a statistical span from the experiment to some putative population "like" it, whereas the second bridge is a subject matter span to the population truly of interest; importantly, this second bridge is largely astatistical. In practice, the decision to generalize is often framed dichotomously: If the convenience sample seems "like" the population of interest, policymakers assume the experimental estimate is unbiased for the population, whereas if the sample is "unlike" the population they decide not to use the experimental estimate at all. Notably, the definition of "like" is usually not made clear.

Recently, the problem of external validity in experiments has begun to be addressed more formally, with the goal of making this second bridge in generalization a statistical bridge. Stuart, Cole, Bradshaw, and Leaf (2011) and Hedges and O'Muircheartaigh (2011) recently proposed methods for adapting propensity score methodology to assess and improve the generalizability of results from experiments. These methods are a modification to the propensity score methods commonly used in quasi experiments and observational studies to address treatment selection bias (Rosenbaum & Rubin, 1983). The general idea in both is to use propensity scores to reweight the experimental sample so that its composition is similar to that of a population of interest on a variety of key covariates; this idea is similar to that of standardization in demography (Kitagawa, 1964). Stuart et al. (2011) use three propensity score methods—inverse propensity weighting (IPW), full matching, and subclassification—to evaluate if an experimental sample is similar enough to a population for an effect to be generalized. Hedges and O'Muircheartaigh (2011) develop a subclassification estimator and measure the degree of mismatch between the population and experiment via the standard error of this estimator. Both provide examples using large-scale experiments conducted in schools.

The subclassification estimator is well suited for the causal generalization case for three reasons. First, this estimator is easy to understand, explain, and implement. This is both because the subclassification estimator is commonly used in observational studies and because the related concepts of blocking and stratification are foundational to the design of experiments and sampling. This ease of explanation is important since the purpose of this work is to provide policymakers and nonscientists with information on the effectiveness of an intervention in a policy-relevant population. Second, the subclassification estimator, while the coarsest of the propensity score matching methods, has been found to perform nearly as well in terms of bias reduction (BR) as other more complicated estimators (e.g., Rosenbaum & Rubin, 1984; Stuart, Cole, Bradshaw, & Leaf, 2011). This is because the subclassification estimator essentially smooths over the individual propensity scores. Finally, other methods such as inverse propensity weights (IPW) have been shown to perform poorly when the selection probabilities are small (Kang & Schafer, 2007), which occur commonly in the generalization case.

The current article extends the work of Stuart et al. (2011) and Hedges and O'Muircheartaigh (2011) in three ways. In Section 1, we provide a more formal treatment of the assumptions required for generalization, including a discussion of SUTVA. In Section 2, we define the subclassification estimator and develop benchmarks for bias reduction and variance inflation (as compared to a conventional estimator); these results are extensions to the analytic study presented in Cochran (1968), which was extended to the propensity score case by Rosenbaum and Rubin (1984). Finally, in Section 3, we address the important case in which the population is not well represented by the experimental sample, making estimation of an average treatment effect for the full population difficult. We argue that this situation is likely to occur in many generalization contexts and investigate the use of the subclassification estimator here. We conclude with an example based on a mathematics intervention aimed at middle school students in the state of Texas.

Finally, it is important to note that the propensity score method used here focuses exclusively on the process of generalizing from a sample to a population. In the Cronbach (1982) *UTOS* (i.e., *U*nits, *T*reatments, *O*utcomes, and *S*ettings) external validity framework, we focus here on generalizing from a sample of units, *u,* to the population of units, *U,* or to a new population of units, *U**. We do not focus on or address other problems in external validity—for example, generalizing across time, changes in the treatment, or general equilibrium issues (e.g., Cook, 1993; Schneider & McDonald, 2007). These issues are certainly important but are beyond the scope of the work presented here.

1. Introduction to the Problem of Generalization

1.1. Population and Estimands

Let $\mathcal{P}$ be a well-defined population composed of N units for which an estimate of the average treatment effect for a particular intervention is of interest. The statistical ideal for making these inferences would involve first drawing a sample $\mathcal{S}$ of n units randomly from these N population units and then to assign $n/2$ of these units randomly to the treatment condition and the remaining $n/2$ units to the control condition. This dual randomization process, however, rarely occurs in practice and is often infeasible (Bloom, 2005; Rubin, 1974; Shadish et al., 2002). Notably, even when this process is feasible, random sampling allows unbiased estimation only for this one well-defined population and does not solve the problem of generalizing to a new or different population (Cronbach's U^*). Following the typical case, we will assume throughout that the sample of units in the experiment has been selected nonrandomly from a population. In contrast, we assume that once the experimental sample is selected, units within the experiment are in fact randomly assigned to treatment.

The problems of generalization and sample selection bias can be situated in the potential outcomes framework. Let $W = 1$ if a unit is assigned to the treatment condition. Then assume that for each unit in the population there exist two potential outcomes, $Y(0) = Y(W = 0)$ and $Y(1) = Y(W = 1)$, where $Y(1)$ is the unit's potential outcome under treatment and $Y(0)$ is the unit's potential outcome under some specified alternative condition. These are referred to as potential outcomes since they are theoretical quantities. The Fundamental Problem of Causal Inference arises because both outcomes can never be observed for a particular unit (Holland, 1986). Instead, for units in the experiment at most one outcome $Y = W\,Y(1) + (1 - W)\,Y(0)$ is observed, which is either $Y(1)$ if the unit is assigned to the treatment or $Y(0)$ if the unit is not. Note that it is not necessary for $Y(0)$ or $Y(1)$ to be observed for units in the population that are not in the experiment, although when this information is available it can be a useful diagnostic (Stuart et al., 2011).

For each unit in the population and sample, we can define the potential treatment effect $\Delta = Y(1) - Y(0)$. Note that Δ is never observed for an individual. Let $Z = 1$ if a unit is in the experimental sample. Then based on this notation, we can define the following estimands for the sample average treatment effect (SATE) and the population average treatment effect (PATE):

1. (Sample): SATE $= \tau^S = E[\Delta|Z = 1]$;
2. (Population): PATE $= \tau^P = \Pr(Z = 1)\, E[\Delta|Z = 1] + (1 - \Pr(Z = 1))\, E[\Delta|Z = 0]$.

The average treatment effect expected in the population and experimental sample are identical if $E[\Delta|Z = 1] = E[\Delta|Z = 0]$—which occurs under random sampling or when the potential treatment effects are constant—if $\Pr(Z = 1) = 1$, or if sample selection and treatment effect heterogeneity are independent (Imai, King, & Stuart, 2008; Rubin, 1974). Finally, note that the conventional estimator of the average treatment effect is the simple difference in means,

$$T = \bar{Y}(1|Z = 1, W = 1) - \bar{Y}(0|Z = 1, W = 0) = \bar{Y}_T - \bar{Y}_C,$$

where $\bar{Y}_T$ is the mean outcome in the treatment group ($W = 1$) and $\bar{Y}_C$ is the mean outcome in the control group ($W = 0$) in the experiment. Importantly, this is an unbiased estimator of the SATE. However, this estimator (and more complicated versions in cluster randomized or multisite trials) is also commonly used to estimate the PATE. In the remainder of this article, we will refer to this as the conventional or naïve estimator of the PATE.

1.2. Data Setup, Propensity Scores, and Assumptions

The problem of estimating the PATE under nonrandom sample selection has been addressed by Hedges and O'Muircheartaigh (2011) and Stuart et al. (2011) via use of a propensity score. Propensity scores are commonly used in observational studies to match units that receive a treatment with units that did not on a variety of covariates (Rosenbaum & Rubin, 1983). Propensity scores are also used in survey sampling to model the probability that a unit does not respond to a survey (Little, 1986).

In generalization, propensity scores are used to match units in the experimental sample to units in the population of interest. This requires that the same covariate information is available in data sets for both the sample $\mathcal{S}$ and the population $\mathcal{P}$. Throughout this section, we focus on the case in which the population data set is a census, such as an administrative data system, and in which units in the experimental sample are locatable within this data set ($\mathcal{S} \subset \mathcal{P}$). Importantly, this census must include a variety of covariates ($\mathbf{X}$) for each unit. In Section 2.3, we address two extensions to this case, which occur when the sample $\mathcal{S}$ is not a subset of the population $\mathcal{P}$; we separate these for ease of explanation.

Definition 1.1: Sampling propensity score

Let $Z = 1$, if a unit is in the experimental sample. Then the *sampling propensity score* is defined as $s(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$, where we assume the independence of units,

$$\Pr(Z_1, \ldots, Z_N|\mathbf{X}) = \prod_{i=1}^{N} s(\mathbf{X})^{Z_i}(1 - s(\mathbf{X}))^{1-Z_i}.$$

An important property of the propensity score is that it is a balancing score (see Rosenbaum & Rubin, 1983, Theorem 1). A balancing score is a function of $\mathbf{X}$ such that the conditional distribution of $\mathbf{X}$, given the balancing score is the same for cases in which $Z = 1$ and $Z = 0$, or here, in the sample and population.

There are three assumptions needed in order to use propensity scores for generalization. Importantly, these same assumptions would be needed even under random sampling.

Assumption 1: Stable Unit Treatment Value Assumptions:

SUTVA (Rubin 1978, 1980, 1990) must be met for all units in the experiment as well as all units in the population of interest. This means there are two forms of SUTVA. The first of these, SUTVA ($\mathcal{S}$), is in relation to the treatment assignment process within the experiment, while the second, SUTVA ($\mathcal{P}$), is in relation to the sample selection process:

SUTVA ($\mathcal{S}$): Assume $Z_i = 1$. Let $W_i = 1$ if unit i is assigned to treatment. Let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes for unit i under the treatment and control conditions, respectively. For each unit i in the experiment $\mathcal{S}$ and all possible pairs of treatment assignment vectors $\mathbf{W} = (W_1, \ldots, W_n)$ and $\mathbf{W}' = (W_1', \ldots, W_n')$,

if $\mathbf{W}_i = \mathbf{W}_i'$ then $Y_i(\mathbf{W}) = Y_i(\mathbf{W}')$.

SUTVA ($\mathcal{P}$): Let $Z_i = 1$ if unit i is selected into the experiment. Let $\Delta_i(1) = Y_i(W = 1, Z = 1) - Y_i(W = 0, Z = 1)$ and $\Delta_i(0) = Y_i(W = 1, Z = 0) - Y_i(W = 0, Z = 0)$ be the potential treatment effects for unit i when i is selected into the experimental sample $\mathcal{S}$ or not. For each unit i in the combined experiment and population data frame and all possible pairs of sample selection vectors $\mathbf{Z} = (Z_1, \ldots, Z_N)$ and $\mathbf{Z}' = (Z_1', \ldots, Z_N')$,

if $\mathbf{Z}_i = \mathbf{Z}_i'$, then $\Delta_i(\mathbf{Z}) = \Delta_i(\mathbf{Z}')$.

Additionally, assume that $\Delta_i = \Delta_i(1) = \Delta_i(0)$ for all units in the population.

The fact that SUTVA must be met at two levels for generalizations from experiments is important. SUTVA ($\mathcal{S}$) is defined in relation to the treatment assignment process within the study and therefore has to do with potential outcomes. It requires that within the experiment, there is no interference of units and that there is only one version of the treatment. This may not be met in cases in which knowledge of treatment assignment has an effect on the potential outcomes. For example, SUTVA ($\mathcal{S}$) is not met in cases in which units not receiving the treatment are demoralized (Shadish et al., 2002), or in cases in which peer effects occur (Hong & Raudenbush, 2006).

In contrast to SUTVA ($\mathcal{S}$), SUTVA ($\mathcal{P}$) is defined in relation to the sample selection process and as a result has to do with the potential treatment effects. There are two parts to SUTVA ($\mathcal{P}$). First, it requires that there is no interference of units and that there is only one version of the treatment. Here, however, this interference is across units included (or not) in the experiment. This may not be met if the potential treatment effects are a function of the proportion of the population receiving the treatment. This is related to the problem of scale-up (Schneider & McDonald, 2007); a notable example is that of the Tennessee STAR class size experiment, which while showing large positive gains in the experiment, when implemented on a large scale in California led to teacher labor market shifts that dramatically reduced the impact of the treatment (Bohrnstedt & Stecher, 1999). The second part of SUTVA ($\mathcal{P}$) requires that $\Delta = \Delta(0) = \Delta(1)$ for all units in the population. This may not be met in cases in which being involved in an experiment has an effect on units' potential treatment effects. This would occur if units react positively or negatively to being in an experiment, performing better or worse than under "real" conditions, or if the conditions in the experiment differ markedly from what would occur in nonexperimental conditions. When this condition fails, the PATE becomes the average treatment effect *when all units in the population are in the experiment*, which may differ from the average treatment effect expected in general nonexperimental conditions.

Assumption 2: Strongly ignorable treatment assignment (in the experiment):

Let $Z = 1$ if a unit is in the sample, let $W = 1$ if the unit is assigned to the treatment condition, and let $s(\mathbf{X})$ be the sampling propensity score. Then the treatment assignment is strongly ignorable if

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

This condition is met by the random assignment process typically used in experiments, which is the focus here. It would also be met in an observational study in which $\mathbf{X}$ included all covariates affecting sample *or* treatment selection. This condition means that within the experiment, the probability of receiving either treatment condition must be nonzero for all units and the treatment assignment must not be confounded with the potential outcomes.

Assumption 3: Unconfounded sample selection (Stuart et al., 2011):
The sampling process and the unit treatment effects are conditionally independent, given the propensity score $s(\mathbf{X})$,

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

This condition says that $\mathbf{X}$ must include all covariates that both explain variation in the potential treatment effects and differ in distribution in the population and experiment. This is a smaller set of variables than those associated with the potential outcomes, which is the requirement for matching in observational studies. For example, in observational studies local matches are encouraged since many contextual variables are related to educational outcomes (Cook, Shadish, &Wong, 2008); however, so long as the potential treatment effects do not vary in relation to these contextual variables, they need not be accounted for in the sample selection case.

Proposition 1.1: Strongly ignorable sample selection

The sample selection is *strongly ignorable* given the propensity score $s(\mathbf{X})$, if

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

Proof: By extension of Rosenbaum and Rubin (1983) Theorem 3.

Importantly, note that $s(\mathbf{X}) = 1$ is possible; this occurs when all units with a particular covariate vector are in the experiment. This is not a problem since $\mathcal{S} \subset \mathcal{P}$. Additionally, when Assumption 2 also holds, this means that *both* the treatment assignment and sample selection processes are strongly ignorable. Importantly, sampling ignorability can fail under either of two conditions. The first of these requires that no variables that explain variation in the potential treatment effects and that affect sample selection have been omitted from $\mathbf{X}$. The second requires that there are no population units with $s(\mathbf{X}) = 0$; when this occurs, it means that there do not exist relevant comparison units in the experiment.

In many cases, strong ignorability is not met because there exist some units with $s(\mathbf{X}) = 0$. Therefore, in order for the strong ignorability assumption to be met, a subpopulation $\mathcal{P}_0 \subset \mathcal{P}$ must be found so that $0 < s(\mathbf{X}) \leq 1$ for all units in $\mathcal{P}_0$. This requires a reconceptualization of the population of interest; issues related to this are discussed in Section 3. Note that in the propensity score literature, when propensity scores are estimated, this distinction between $\mathcal{P}$ and $\mathcal{P}_0$ is known as the problem of common support or overlap (e.g., Dehejia & Wahba, 1999, 2002). From here forward, we refer to the subpopulation $\mathcal{P}_0$ with N_0 units as that which meets the selection probability requirement ($0 < s(\mathbf{X}) < 1$), and to the average treatment effect in this subpopulation as the P$_0$ATE, τ^{P0}. Note that it may be the case that $\mathcal{P}_0 \equiv \mathcal{P}$ but this is not required.

1.3. Special Cases

The discussion thus far has focused on the case in which the experimental sample is a subset of the population ($\mathcal{S} \subset \mathcal{P}$). The propensity score method introduced by Stuart et al. (2011), and Hedges and O'Muircheartaigh (2011) however, can also be applied more broadly to two additional cases.

The first case is that in which $\mathcal{S} \subset \mathcal{P}$, but the population data set available is not a census of $\mathcal{P}$. Instead, the data set is a probability survey ($\mathcal{V}$), where $\mathcal{V} \subset \mathcal{P}$. However, units in the sample are either not in the survey or not locatable, so $\mathcal{S} \not\subset \mathcal{V}$ (for a medical study example, see Cole & Stuart, 2010). The second case is that in which the sample $\mathcal{S} \not\subset \mathcal{P}$. For example, the experiment may have been conducted in Texas, while the population of interest is Florida. In this second case, it is particularly important that the set of covariates **X** required for Assumption 3 does not include an indicator of location. That is, the potential treatment effects cannot be a function of a unit living in Florida or Texas, once other important covariates are taken into account; if this is so, the generalization is unwarranted.

These two cases are different than the standard $\mathcal{S} \subset \mathcal{P}$ case in a couple of ways. These differences occur because in the combined data set, a unit is *either* in the population data set ($Z = 0$) or in the experimental data set ($Z = 1$), but not both. This means that when $s(\mathbf{X}) = 1$, there exists a unit in the experimental data set for which there is no relevant comparison unit in the population. In order for strong ignorability to be met, therefore, a subsample $\mathcal{S}_0 \subset \mathcal{S}$ must be found so that $0 < s(\mathbf{X}) < 1$ for all units in $\mathcal{S}_0$; this means dropping units from the experiment. Second, in these cases, the propensity score cannot be thought of as a true selection probability. However, since it is a balancing score, even when the propensity score has no intrinsic meaning it provides a method for matching units in the population of interest to their most relevant comparison cases in the experiment (for a similar use in the observational studies literature, see Zanutto, 2006).

2. Subclassification Estimator of the P$_0$ATE

Given strongly ignorable sample selection and treatment assignment processes and the sampling propensity score $s(\mathbf{X})$, a variety of estimators of the P$_0$ATE, τ^{P0}, could be developed. In this section, we focus on studying the properties of the subclassification estimator proposed by Hedges and O'Muircheartaigh (2011). Throughout we focus on a subpopulation $\mathcal{P}_0$ and a subsample $\mathcal{S}_0$ defined so that the two strong ignorability conditions are met. Note that $\mathcal{P}_0 \equiv \mathcal{P}$ and $\mathcal{S}_0 \equiv \mathcal{S}$ are possible.

Definition 2.1: General subclassification estimator of P$_0$ATE

Assume that there are n_0 units in the sample $\mathcal{S}_0$ and N_0 units in the population $\mathcal{P}_0$. Let

$$T_{\text{sub}} = \sum_{j=1}^{k} w_{p_j}\left(\bar{Y}_{T_j} - \bar{Y}_{C_j}\right),$$

where for stratum j, $\bar{Y}_{T_j}$ is the sample mean for units in the treatment group ($W = 1$), $\bar{Y}_{C_j}$ is the sample mean for units in the control group ($W = 0$), N_{0j} is the number of population cases and $w_{pj} = N_{0j}/N_0$ is the proportion of population cases in the stratum. Importantly, the variance of this estimator can be written

$$V(T_{\text{sub}}) = \sum_{j=1}^{k} w_{p_j}^2\left[V\left(\bar{Y}_{T_j} - \bar{Y}_{C_j}\right)\right],$$

which is a function of the squared population weights, w_{pj}, and the variances of the stratum-specific treatment effect estimates, $\bar{Y}_{T_j} - \bar{Y}_{C_j}$.

The remainder of this section will develop conditions under which this estimator can be compared with the conventional estimator, $T = \bar{Y}_T - \bar{Y}_C$, in terms of bias and average variance. Later, these simplifications will be used to investigate the amount of the reduction in bias and increase in average sampling variance that can be expected under particular numbers of strata. While the w_{pj} can be defined in any way, following Cochran (1968) we limit our focus here to cases in which the k strata are defined such that $w_{pj} = 1/k$ for strata $j = 1 \ldots k$; the benefit of this approach to stratification and weighting is that it is optimal in terms of bias reduction and ease of implementation.

2.1. Reduction in Bias

The most well-known result from Cochran's (1968) study is that using a subclassification estimator with five equal-weight strata reduces the bias (as compared to a conventional estimator) by approximately 90%. In this section, we investigate these properties in the sample selection bias case.

Proposition 2.1: Expected bias reduction via subclassification in generalization

Let $\tau(s) = E[\Delta|s(\mathbf{X}) = s]$ be the average potential treatment effect for units with $s(\mathbf{X}) = s$. Under Assumption 1, the bias in the conventional estimator T can be written,

$$B_I = \text{Bias}\left(T|\tau^{\text{P0}}\right) = \tau^{\text{S0}} - \tau^{\text{P0}} = (1 - \Pr(Z = 1))\{E[\tau(s)|Z = 1] - E[\tau(s)|Z = 0]\}.$$

This is the difference between the S$_0$ATE and the P$_0$ATE. In comparison, the bias in the subclassification estimator T_{sub} with k strata can be written

$$B_{\text{sub}} = \text{Bias}(T_{\text{sub}}|\tau^{P0}) = \sum_{j=1}^{k} E[\tau(s)|s \in I_j]\text{Pr}_{P_0}\left(s \in I_j\right) - \tau^{P_0}$$

$$= (1 - \text{Pr}(Z = 1)) \sum_{j=1}^{k} \{E[\tau(s)|Z = 1, s \in I_j] - E[\tau(s)|Z = 0, s \in I_j]\}\text{Pr}_{P_0}\left(s \in I_j\right),$$

where I_j is the set of $s(\mathbf{X})$ values in population $\mathcal{P}_0$ in stratum j. Now, assume that $\tau(s)$ is a linear function of $s(\mathbf{X})$. Then the proportion reduction in bias in the $\tau(s)$ scale following subclassification on $s(\mathbf{X})$ equals the reduction in bias in the $s(\mathbf{X})$ scale, where the bias reduction (BR) can be written,

$$\text{BR} = 100\left(1 - \frac{B_{\text{sub}}}{B_I}\right)\% = 100\left(1 - \frac{\sum_{j=1}^{k} w_{p_j}\left\{E[s(\mathbf{X})|Z = 1, s \in I_j] - E[s(\mathbf{X})|Z = 0, s \in I_j]\right\}}{E[s(\mathbf{X})|Z = 1] - E[s(\mathbf{X})|Z = 0]}\right).$$

Proof: Follows immediately from Rosenbaum and Rubin (1984) Theorem A1.

Proposition 2.1 provides a method to conceptualize bias reduction in relative terms. It says that under the simplest case—when the conditional potential treatment effects $\tau(s)$ are a linear function of the sampling propensity $s(\mathbf{X})$—we can approximate the reduction in bias by looking only at the bias in terms of the distributions of $s(\mathbf{X})$ in the population $\mathcal{P}_0$ and the experiment $\mathcal{S}_0$. In other, nonlinear cases, bias will also be reduced by subclassification, but the amount of the reduction in these cases will be more difficult to quantify, since it will depend on the functional form of $\tau(s)$. In general, it is assumed that by reducing bias in the propensity score $s(\mathbf{X})$—and thus the covariates $\mathbf{X}$—bias in the treatment effect estimator is also reduced.

2.2. Effect on Sampling Variance

In terms of sampling variance, Cochran (1968) shows that in observational studies, sampling variance is often reduced by use of a subclassification estimator. In order to investigate sampling variance in the generalization case, we focus here on the case in which the relationship between $\tau(s)$ and $s(\mathbf{X})$ is linear.

Proposition 2.2: Expected variance inflation via subclassification in generalization

Recall that $\tau(s) = E[\Delta|s(\mathbf{X}) = s]$ is the average potential treatment effect for units with $s(\mathbf{X}) = s$. Assume that for those units in the experiment S_0, $Y(0) = \mu_C + \beta_C s + \varepsilon$ and $Y(1) = \mu_T + \beta_T s + \varepsilon$, where $s = s(\mathbf{X})$, $E(\varepsilon|W = 0) = E(\varepsilon|W = 1) = 0$, and $V(\varepsilon|W = 0) = V(\varepsilon|W = 1) = \sigma^2$, and where $E(.)$ is the expectation and $V(.)$ is the variance. That is, assume $\tau(s) = \delta + \beta s$, where $\delta = \mu_T - \mu_C$ and $\beta = \beta_T - \beta_C$. Let ρ_{st} = Corr(s, $Y|W$ = 1) and ρ_{sc} = Corr(s, $Y|W$ = 0) where Corr(.) is the correlation. Then under these conditions, the average variance inflation of the subclassification estimator T_{sub} (relative to the conventional estimator, T) can be written

$$\text{EVIF}(T_{\text{sub}}) = A[1 - \rho^2_{s*}(1 - B/A)],$$

where for stratum j, w_{pj} is the proportion of the population $\mathcal{P}_0$, w_{sj} is the proportion of the sample $\mathcal{S}_0$, and σ^2_{sj} is the variation in the distribution of $s(\mathbf{X})$ in $\mathcal{S}_0$, and where σ^2_s is the total variation of $s(\boldsymbol{X})$ in the sample, $A = \sum_{j=1}^{k} w_{p_j}\left(\frac{w_{p_j}}{w_{s_j}}\right)$, $B = \sum_{j=1}^{k} w_{p_j}\left(\frac{w_{p_j}}{w_{s_j}}\right)\left(\frac{\sigma^2_{s_j}}{\sigma^2_s}\right)$, and $\rho^2_{s*} = (\rho^2_{st} + \rho^2_{sc})/2$ is the average value of the correlation.

Proof: See Appendix A.

The expected variance inflation factor (EVIF) here is used instead of the VIF, since the exact sampling variance is a function of the treatment allocation observed in a particular realization of the random assignment process. In order to get intuition about the EVIF, focus on the case that $\rho^2_{s*} = 0$, indicating that there is no relationship between $s(\mathbf{X})$ and $\tau(s)$. In this case, the EVIF reduces to A, which is a rough measure of the similarity of the distributions of $s(\mathbf{X})$ in the population and experiment. Clearly $A > 1$ when these distributions differ and A is very large when there exists at least one stratum j' such that $w_{sj'}$ is very small.

2.3. Theoretical Investigation and Rules of Thumb for P_0ATE

An important question is how the subclassification estimator compares to the naïve estimator. Assuming that the strong ignorability conditions have been met, the amount of bias reduction and variance inflation (EVIF) for a subclassification estimator will depend on the distributions of $s(\mathbf{X})$ in the population and the experimental sample. In order to investigate this analytically, we focus on the distribution of the propensity score logits, $f = f(s) = \log[s/(1-s)]$, which take values on the whole real line instead of the interval (0,1).

Cochran (1968) evaluated this same question in the treatment selection case for distributions that differ by a locational shift. In generalization, however, the distributions themselves may often differ; in many situations, the distribution in the population is skewed, whereas that in the experiment is symmetric. One reason for this is that the population often contains units with very small probabilities of being in the sample. These cases will be less common in the sample, and as a result its distribution will be more symmetric. Second, the distributions of many population variables are highly skewed themselves (e.g., income), whereas current methods for selecting samples tend to focus on the inclusion of average and modal units, not those at the full range of covariate values (e.g., Bloom, 2005; Cook, 1993). Note that by focusing only on the subpopulation $\mathcal{P}_0$ and subsample $\mathcal{S}_0$ that meet the overlap conditions for strong ignorability, these distributional differences are made less dramatic, though some differences may persist.

The goal of this analysis is to compare the subclassification estimator, T_{sub}, with the conventional estimator, T, in terms of bias and variance; here T_{sub} has $k = 2,3,4$, and 5 equally sized strata based on the population density of f. For comparison purposes, we include two types of distributional pairs, and these are listed in Table 1. This includes nine pairs in which the sample and population are both normally distributed but differ by a shift ("locational shift"), and six pairs in which the population distribution is skewed (chi-square) and the sample distribution is symmetric ("skewed"). Additionally, in order to simulate the process of restricting the analysis to the subpopulation $\mathcal{P}_0$, we analytically truncated the distribution of the original population $\mathcal{P}$ at $F_S{}^{-1}(.99)$, where F_S is the cumulative distribution function for the distribution of f in the experimental sample $\mathcal{S}$. Note that if other values such as $F_S{}^{-1}(.9)$ or $F_S{}^{-1}(.8)$ are used instead for truncation, the results for the "skewed" cases will be more in line with those from the "locational shift" cases, since most of the population skew is removed.

Finally, for each of these pairs, we evaluate the bias reduction and EVIF using the results developed in Sections 2.1 and 2.2. These comparisons are analytic, based on the probability densities of these distributions; when necessary, we evaluate the means and variance of truncated distributions analytically using the extrunc() and vartrunc() functions (Nadarajah & Kotz, 2006) in the statistical program R (R Development Core Team, 2011). The analytic results for these distributions are summarized in Table 2. We collapse the individual results for the distributions into results for the "locational shift" and "skewed" cases.

The results in Table 2 indicate that bias is greatly reduced by subclassification in both distributional cases. The results for the "locational shift" models are generally in line with those distributions studied by Cochran (1968); these indicate, for example, that in general with $k = 5$ strata, bias reduction is generally at least 94%. For the "skewed" models, bias reduction is often smaller, but still significant, and tends to vary more across the distributional pairs. For example, in this case $k = 5$ strata is associated with bias reduction in the 56–96%, with an average of 78%.

In terms of variance inflation (EVIF), Table 2 shows results for the conservative case ($\rho_{s*} = 0$) and the moderate case ($\rho_{s*} = 0.50$). Here we focus only on the conservative case, noting that variance inflation can be much smaller when there is a relationship between the propensity score distribution and the conditional treatment effects, $\tau(s)$. In general, with $k = 5$ strata these EVIFs range from 1.04 to 2.04 for the distributional pairs under study, with an average of 1.53 for the "locational shift" pairs and 1.18 for the "skewed" pairs. Larger EVIFs occur when distributions are very different, indicating that the variance of the subclassification estimator accounts for the degree to which the experiment and population differ (Hedges & O'Muircheartaigh, 2011). Finally, note that in situations in which a more accurate estimator (one with approximately less mean square error) is preferred, it may be desirable to choose $k = 3$ or $k = 4$ strata.

TABLE 1
Distributional Pairs Studied

Type of Distributional Pair	Sample	Population	Original Bias Mean Difference	 |SMD|
Locational shift	N (3, 1)	N (4, 1)	1	1.000
	N (3.5, 1)		0.5	0.500
	N (3.75, 1)		0.25	0.250
	N (3, .75)	N (4, 1)	2	1.131
	N (3.5, .5)		0.5	0.632
	N (3.75, .5)		0.25	0.316
Skewed	N (8, 2)	χ^2 (10)	2	0.577
	N (9, 2)		1	0.289
	N (9, 3)		1	0.263
	N (4, 2)	χ^2 (6)	2	0.707
	N (5, 2)		1	0.354
	N (5, 3)		1	0.309
	N (2, 2)	χ^2 (4)	2	0.816
	N (3, 2)		1	0.408
	N (3, 1)		1	0.471

Note: The SMD is standardized in relation to the population.

3. Estimating the PATE When $\mathcal{P}_0 \subset \mathcal{P}$

Up until this point, we have focused on estimation of the P$_0$ATE, where $\mathcal{P}_0 \subset \mathcal{P}$ is the subpopulation for which sampling is strongly ignorable. While restricting the focus to the subpopulation $\mathcal{P}_0$ is useful theoretically and statistically, however, doing so can make interpretation very difficult. Since this situation is likely to arise in many policy contexts, in this section we offer two different approaches to dealing with this problem.

3.1. Conceptualizing $\mathcal{P}_0$

In most situations, the initial policy relevant question is about the population $\mathcal{P}$, not $\mathcal{P}_0$. A question then is how to describe this new subpopulation $\mathcal{P}_0$ that meets the strongly ignorable sample selection criteria. Here we offer a few options.

One simple way to relate the sub- and full populations $\mathcal{P}_0$ and $\mathcal{P}$ is through the idea of coverage. In survey sampling the fact that $\mathcal{P}_0$ and $\mathcal{P}$ may differ is referred to as the problem of *coverage error* (Groves, Fowler, Couper, Lepkowski, & Singer, 2009). In sampling, coverage error occurs when the population of interest (the target population) and the sampling frame are not identical. For example, if the population of American households is of interest and sampling is conducted using random digit dialing on landlines, then the sampling frame excludes

TABLE 2
Bias Reduction (BR) and Variance Inflation (EVIF) by Number of Strata

				Bias Reduction (BR)				EVIF ($\rho_{s*} = 0$)				EVIF ($\rho_{s*} = .50$)			
Population of Interest	Type of Distributional Pair	Statistic	θ (%)	$k = 2$ (%)	$k = 3$ (%)	$k = 4$ (%)	$k = 5$ (%)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$\mathcal{P}_0$	Locational shift	Min	75	67	82	91	94	1.03	1.04	1.04	1.04	0.70	0.62	0.59	0.57
		Mean	87	70	84	93	96	1.39	1.46	1.52	1.53	0.91	0.86	0.85	0.83
		Max	98	72	88	95	98	1.76	1.92	2.02	2.04	1.12	1.12	1.13	1.11
	Skewed	Min	74	7	20	44	56	1.00	1.00	1.01	1.02	0.68	0.60	0.58	0.56
		Mean	84	45	60	72	78	1.04	1.10	1.14	1.18	0.71	0.66	0.65	0.65
		Max	94	78	96	94	96	1.13	1.28	1.42	1.54	0.76	0.77	0.82	0.86
$\mathcal{P}$	Locational shift	Min	*NA*	61	74	85	89	1.04	1.04	1.06	1.06	0.71	0.63	0.60	0.58
		Mean		62	76	85	89	1.69	2.37	3.43	4.59	1.08	1.37	1.90	2.50
		Max		64	77	86	90	3.02	4.43	7.36	12.03	1.83	2.51	4.06	6.52
	Skewed	Min	*NA*	22	31	46	53	1.01	1.02	1.05	1.08	0.69	0.61	0.60	0.59
		Mean		29	41	53	60	1.12	1.43	2.15	3.38	0.75	0.85	1.22	1.85
		Max		43	57	67	73	1.34	2.56	6.08	13.12	0.88	1.50	3.37	7.11

Note: $\rho_{s*} = [(\rho_{st}{}^2 + \rho_{sc}{}^2)/2]^{1/2}$, where $\rho_t = \mathrm{Corr}(Yt, \mathrm{logit}[s(x)])$ and $\rho_c = \mathrm{Corr}\ (Yc, \mathrm{logit}[s(x)])$ are the correlations between the logits of the propensity score distributions and Y_c and Y_t are the outcomes in the treatment and control groups. [b]When the bias for $\mathcal{P}_0$ is very small, no additional subclassification estimator is needed. [c]The bias reduction shown for the population $\mathcal{P}_0$ is in relation $\mathcal{P}_0$ not $\mathcal{P}$.

households in the population without landlines. This means that the average for the target population and the sampling frame are not identical, inducing *coverage bias*. In generalizations from experiments, a similar coverage problem occurs. One way to summarize the degree of this coverage error is through the following definition.

Definition 3.1: Population coverage percent

Let the population $\mathcal{P}$ be composed of N units, and the subpopulation $\mathcal{P}_0$ contain N_0 units. Note that $\mathcal{P}_0 \subset \mathcal{P}$ is defined such that $\mathcal{P}_0$ meets the strong ignorability condition given in Proposition 1.1. Then the *population coverage percent*, θ, is defined as

$$\theta = (N_0/N) \times 100\%,$$

which is the percentage of the population $\mathcal{P}$ that is in the subpopulation $\mathcal{P}_0$.

In practice, when sampling propensity scores are estimated, determining whether the strong ignorability requirement that $0 < s(\mathbf{X}) \leq 1$ is met will depend on the method used for determining distributional overlap. One method for determining overlap is to truncate the population distribution of the estimated propensities, $\hat{s}(\mathbf{X})$, at the maximum value of the estimated experimental sample $\hat{s}(\mathbf{X})$ values. In order to investigate typical values of θ, we calculated θ for each distributional pair studied in Table 1. In Table 2, we report the minimum, average, and maximum of these values in the column "θ" for the "locational shift" and "skewed" cases studied and for the subpopulation $\mathcal{P}_0$. We find θ values averaging between 84% and 87%, which means that the subpopulation $\mathcal{P}_0$ includes over 84% of the population of interest. These values vary by distributional pair, however, and range from as low as 74% to as high as 98%.

While the population coverage percent is a helpful metric for relating $\mathcal{P}_0$ to $\mathcal{P}$, used alone it does not offer any information on which units were actually included in $\mathcal{P}_0$ or how to conceptualize this new population. When the population $\mathcal{P}$ is small, those units included in $\mathcal{P}_0$ and $\mathcal{P}$ can be enumerated in a list; this may be useful, for example, if schools are the units of comparison and school leaders wish to know if their school was included in the generalization. Other methods that are helpful in this regard include mapping units included in $\mathcal{P}_0$ (vs. those excluded), reporting tables of univariate statistical comparisons (including the minimum, mean, and maximums), and modeling the inclusion in $\mathcal{P}_0$ via a statistical classification framework. Importantly, note that this last case often leads to the definition of a new propensity score—the propensity for a unit in $\mathcal{P}$ to also be in $\mathcal{P}_0$.

3.2. Estimating the PATE

In many situations, despite the fact that sampling ignorability cannot be met, an estimator of the PATE is required. For example, the policy question may be, "Should we roll out this treatment in all schools in Texas?"; in this case, providing a treatment effect estimate for *only* θ% of the population does not answer the question. Since this problem is likely to be common in generalization, we here provide properties of the subclassification estimator when strong ignorability fails. Researchers and policymakers can use these results to guide their decision to use the conventional estimator or the subclassification estimator developed in this article. Note that here the same subclassification estimator is used, but that the strata and weights are defined in relation to the distribution of $s(\mathbf{X})$ in the full population $\mathcal{P}$.

There are two main problems that arise when using the propensity score subclassification estimator when the sampling ignorability condition fails. Both of these problems result in bias. First, recall that for the subclassification estimator in general, bias reduction is nearly optimal when the strata are defined so that each stratum contains an equal portion of the population ($w_{pj} = 1/k$); conversely, bias reduction is far from optimal when the strata are defined in terms of the distribution in the experimental sample (Cochran, 1968) A problem that arises in the full population case is that in order to guarantee that each stratum has enough experimental units for the estimation of stratum average treatment effects, only a small number of strata can be used (e.g., 3 or 4). Since there are fewer strata, this means that given a set of covariates that meet Assumption 3, bias reduction is likely to be smaller than in the $\mathrm{P_0}$ATE case (where an estimator with more strata could be used).

Second, when the ignorability condition is not met, there are values of the covariates $\mathbf{X}$ in the population that are not in the experiment. For example, it could be that in the population, students' ages range from 5 to 18, where 10 is the average age, while in the experiment they only range from 5 to 13, with an average age of 8. By reweighting the experimental sample via a subclassification estimator, the average values of $\mathbf{X}$ are likely to be closer to that in the population (e.g., an average age of 9.2 vs. 10 in the population), though generally not as close as when focusing on the subpopulation $\mathcal{P}_0$. Furthermore, the amount of bias reduction in terms of the squares and cross-products of the covariates is often smaller. Both residual imbalance problems are largely driven by the stratum which contains units with $s(\mathbf{X}) = 0$, since the population and experimental sample stratum averages often differ the most here. Altogether this means that when estimating the PATE, when Assumptions 2 and 3 have been met, the subclassification estimator is generally less biased than the conventional estimator, although some bias remains.

Together these two issues indicate that when estimating the PATE, while the subclassification estimator generally exhibits less bias in the covariates than the conventional estimator, both are biased. In this sense, neither estimator is ideal. Despite this, there are two benefits to using the subclassification estimator instead of the conventional estimator here. First, it allows for the assumptions regarding covariates and their functional form to be clearly stated and explained and for the amount of bias reduction in each covariate to be clearly evaluated.

Second, it allows the sampling variance of the estimate of the PATE to take into account mismatches between the population and experiment. This second point is important—when the variance is large, it indicates that the experiment provides little evidence for evaluating the effect of the treatment on a particular population (Hedges & O'Muircheartaigh, 2011).

3.3. Theoretical Investigation and Rules of Thumb for PATE

In order to investigate the effect of using T_{sub} for the PATE (instead of the naïve estimator T), we calculated the amount of bias reduction and variance inflation for the distributional pairs shown in Table 1. These results are summarized in the lower half of Table 2. Note that for both the "locational shift" cases and the "skewed" cases, the amount of bias reduction is smaller than for the P$_0$ATE. For example, with $k = 5$ strata, when generalizations are restricted to $\mathcal{P}_0$, bias reduction is on average 96% for the "locational shift" cases and 78% in the "skewed" cases, whereas when the full population $\mathcal{P}$ is used these fall to 89% and 60%, respectively. Notably for the "skewed" cases, when the full population $\mathcal{P}$ is of interest, with $k = 3$ strata the bias reduction ranges from 31% to 57%, while for $k = 5$ cases it ranges from 53% to 73% for the distributions studied here. While these reductions in bias are not as large as in the $\mathcal{P}_0$ case, they are still nonnegligible. Any bias reduction is better than none.

Additionally, the sampling variance for the estimator is also affected by the shift from $\mathcal{P}_0$ to $\mathcal{P}$. As Table 2 illustrates, when the full population $\mathcal{P}$ is of interest, the sampling variances are likely to be much larger. For example, for $k = 5$ strata, the EVIFs average around 4.5 for locational shift models and 3.38 for the skewed cases (as compared to 1.53 and 1.18, respectively, when $\mathcal{P}_0$ is the focus). One reason these EVIFs are so much larger is that in the stratum containing cases in which $s(\mathbf{X}) = 0$, the ratio w_{pj}/w_{sj} tends to be very large. Importantly, the results for $k = 3$ are not nearly as dramatic, suggesting a clear bias-variance trade-off in the number of strata used.

4. Example

SimCalc is a mathematics software program aimed at middle school students, which teaches proportionality, linearity, and rates of change. In order to evaluate the effectiveness of SimCalc, SRI conducted two cluster-randomized trials in Texas: a pilot study, which included 19 schools, and a full study, which included 73 schools (Roschelle et al., 2010). Every effort was made to include schools representative of the state, but random sampling was not feasible or implemented. In this section, we present a reanalysis of the $n = 92$ schools in the SimCalc experiments; the goal is to reweight the experimental sample so that it is similar in composition to the population of Texas on a variety of important covariates. As we have shown, when the population and sample are balanced and the ignorability conditions have been met, the estimate of the school average treatment effect for the population of schools in Texas will be less biased than the conventional estimator.

TABLE 3
Comparison of Estimates for Estimating the School-Average Treatment Effect for SimCalc

		Weights				
Population	Estimator	w_s	w_p	Estimate	*SE*	95% CI
Any	T			1.443	0.142	[1.165, 1.721]
P_0ATE	Stratum 1	0.490	0.200	1.673	0.199	[1.283, 2.063]
	Stratum 2	0.275	0.200	1.199	0.276	[0.658, 1.741]
	Stratum 3	0.107	0.200	0.801	0.418	[−0.019, 1.621]
	Stratum 4	0.047	0.200	1.466	0.578	[0.333, 2.599]
	Stratum 5	0.081	0.200	2.009	0.510	[1.010, 3.008]
	T_{sub}			1.430	0.188	[1.061, 1.798]
PATE	Stratum 1	0.530	0.200	1.597	0.191	[1.222, 1.971]
	Stratum 2	0.248	0.200	1.368	0.290	[0.799, 1.936]
	Stratum 3	0.107	0.200	0.40	0.405	[−0.389, 1.198]
	Stratum 4	0.060	0.200	1.786	0.546	[0.716, 2.856]
	Stratum 5	0.054	0.200	2.107	0.610	[0.910, 3.303]
	T_{sub}			1.452	0.195	[1.069, 1.835]

Note: The effect size estimates are standardized in relation to the between school variation.

The population was enumerated and defined using the publically available state academic excellence indicator system (AEIS). The population $\mathcal{P}$ was defined to be the $N = 1{,}713$ Texas schools in the 2008–2009 academic year with seventh-grade classrooms that were not charter schools. The outcome of interest here is the school-average student gain scores on a test that focuses directly on the issues of proportionality, linearity, and rates of change. As Table 3 shows, the conventional estimate of the treatment effect for SimCalc is $T = 1.443$ (.142), $p < .001$, which indicates that schools using SimCalc had larger average-gain scores than those using business as usual. Note that this effect is standardized in relation to the between-school variance (Hedges, 2007). This metric was chosen since different versions of the test were used in the pilot and full studies, and since interest here is in school average gains.

4.1. Propensity Score Estimation

Twenty-six covariates were selected in the AEIS system for matching the schools in the experiment with those in the population. These included variables on student and teacher demographic composition, school structure, and prior year school average test scores. Note that 9 of these variables are teacher aggregates, 16 are student aggregates, and 1 is a school-level variable. Table 4 lists all of these covariates, as well as their means in the experiment and population.

TABLE 4
Covariates and Bias in Covariates in Three Models

	Conventional	P_0ATE ($\theta = 92\%$)					PATE		
				Bias After Subclassification				Bias After Subclassification	
Description	Experimental Sample ($\mathcal{S}$), $n = 92$	Population $\mathcal{P}_0$, $N = 1,581$	Original Bias |SMD|	|SMD|	% BR	Population $\mathcal{P}$, $N = 1,713$	Original Bias |SMD|	|SMD|	% BR
Teacher tenure (mean)	6.801	7.002	0.033	0.032	5	7.091	0.066	0.073	−10
Teacher experience (mean)	10.947	11.500	0.213	0.015	93	11.583	0.225	0.043	81
Teacher–student ratio	13.266	12.870	0.143	0.071	50	12.699	0.189	0.005	97
Teachers that are African American (%)	2.557	6.687	0.196	0.040	80	8.393	0.253	0.071	72
Teachers that are Hispanic (%)	21.571	14.874	0.317	0.029	91	14.721	0.322	0.021	93
Teachers in the school (total)	42.984	41.150	0.220	0.052	76	39.873	0.272	0.030	89
Teachers in first year of teaching $\mathcal{P}_0$	8.739	8.294	0.038	0.008	80	8.321	0.031	0.020	35
Teachers with 1–5 years experience (%)	28.744	28.176	0.149	0.022	85	28.012	0.155	0.012	93
Teachers with >20 years experience (%)	17.695	19.874	0.206	0.002	99	20.251	0.221	0.033	85
Students in disciplinary alternative education programs (%)	3.418	3.158	0.157	0.190	−22	3.102	0.172	0.216	−26
Seventh-grade retention (rate)	1.307	1.386	0.066	0.001	99	1.833	0.109	0.061	44
Students that are mobile (%)	14.804	16.358	0.190	0.172	9	19.228	0.281	0.255	9
Students in school that are in seventh grade (%)	34.995	32.361	0.280	0.038	86	31.206	0.354	0.136	62
Students in seventh grade (total)	224.247	199.478	0.267	0.038	86	190.399	0.326	0.008	97
Students that are African American (%)	5.114	10.035	0.247	0.060	76	11.792	0.310	0.102	67
Students that are Hispanic (%)	47.195	40.486	0.255	0.028	89	40.268	0.264	0.039	85
Students that are LEP (%)	9.440	7.698	0.235	0.006	97	7.538	0.253	0.006	98
Students that are economically disadvantaged (%)	52.084	52.796	0.031	0.090	−192	53.643	0.008	0.115	−1436
Students that are at risk (%)	40.607	41.457	0.023	0.065	−185	43.467	0.135	0.138	−2

(continued)

Table 4 (continued)

	Conventional	P_0ATE ($\theta = 92\%$)					PATE		
				Bias After Subclassification				Bias After Subclassification	
Description	Experimental Sample ($\mathcal{S}$), $n = 92$	Population $\mathcal{P}_0$, $N = 1{,}581$	Original Bias |SMD|	|SMD|	% BR	Population $\mathcal{P}$, $N = 1{,}713$	Original Bias |SMD|	|SMD|	% BR
Students proficient in seventh-grade reading (%)	86.000	86.196	0.024	0.049	−104	81.901	0.200	0.178	11
Students proficient in seventh-grade math (%)	75.562	76.703	0.074	0.086	−17	72.791	0.139	0.261	−88
Students proficient in grades 3–11 math (%)	75.014	76.223	0.107	0.100	6	73.599	0.085	0.251	−196
Students proficient in grades 3–11 all (%)	63.836	65.298	0.133	0.049	63	63.287	0.016	0.181	−1012
Students with commended performance, grades 3–11, math (%)	20.315	20.646	0.044	0.004	92	19.612	0.053	0.105	−99
Students with commended performance, grades 3–11, reading (%)	8.877	9.167	0.071	0.047	34	8.710	0.003	0.037	−1201
County of school is rural	0.315	0.335	0.117	0.119	−2	0.329	0.105	0.164	−57
Logits of propensity scores (mean)	2.625	3.085	0.764	0.035	95	4.116	0.561	0.174	69
Propensity scores (mean)	0.919	0.943	0.886	0.052	94	0.959	0.896	0.035	96
(.10<|SMD| < .20) (.20 < |SMD| < .30) (.30 < |SMD|)			(8) (8) (1)	(4) (0) (0)			(8) (7) (4)	(8) (4) (0)	

Note: [a]For both the P_0 ATE and the PATE, the subclassification estimator uses five strata. [b]The propensity score models are formulated so that the event being predicted is not being in the experiment. [c]Variable names in the Texas AEIS system are available upon request. [d]Negative values of % bias reduction (BR) indicate that bias in the variable is increased.

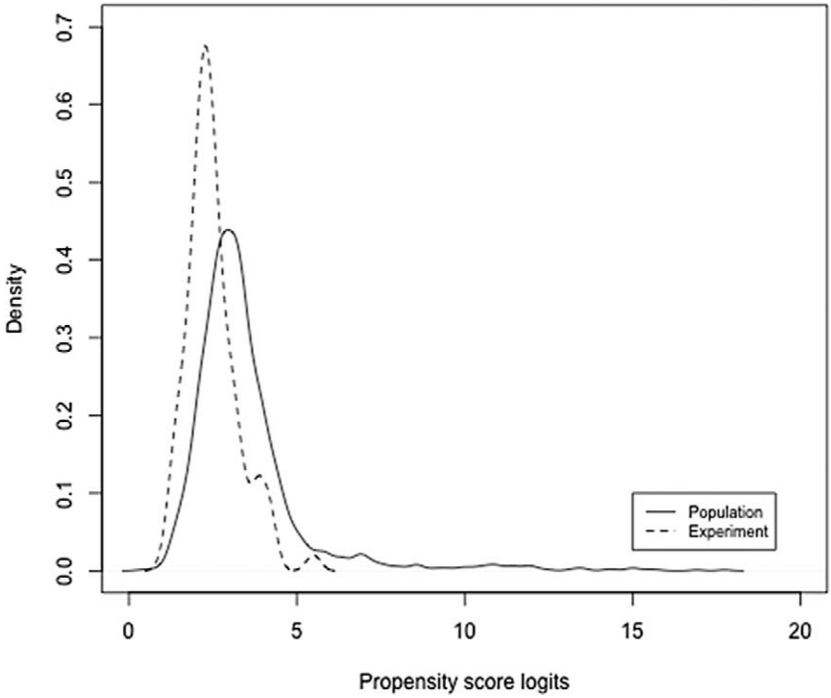

FIGURE 1. *Distribution of propensity score logits for population and experimental schools, when the predicated category is nonmembership in the experiment.*

We estimated the propensity scores $s(\mathbf{X})$, created strata, and evaluated balance using the package matchit (Ho, Imai, King, & Stuart, 2007) in the statistical program R, though this model could have been estimated with any standard statistical program. We estimated the propensity scores using the logistic regression model,

$$\log[s(\mathbf{X})/(1 - s(\mathbf{X}))] = \beta_0 + \beta_1 X_1 + \ldots + \beta_{26} X_{26}.$$

Note that in the matchit program, the event being predicted (the treatment) was nonmembership in the experiment; we did this to make the output simpler, since by analogy here the "population" is akin to the "treatment" group in observational studies. Figure 1 illustrates the distributions of the estimated logits in the experiment and population; note that these two distributions do not overlap completely, indicating that for some population units, $s(\mathbf{X}) = 0$. Based on the degree of overlap of these distributions, it was determined that $\theta = 92\%$ of the population units have similar experimental units available for subclassification. This means that the subpopulation meeting the sampling ignorability requirement,

$\mathcal{P}_0$, differs from the full population $\mathcal{P}$. Note that here $\mathcal{S}_0 \equiv \mathcal{S}$, and no experimental units were dropped; this is because $\mathcal{S} \subset \mathcal{P}$.

4.2. Assumptions Needed to Generalize

Before proceeding to estimate the $\mathrm{P_0ATE}$ and PATE, it is important to carefully discuss how well the three assumptions are met. Requirements and issues with meeting Assumption 1, SUTVA ($\mathcal{S}$) and Assumption 2, treatment assignment ignorability, have been discussed formally in other works focused on experiments (e.g., Shadish et al., 2002), so we will not focus our discussion on these. Clearly they would be violated if, for example, there was treatment contamination or compensatory rivalry, or if there was posttreatment assignment attrition or crossovers. Instead, we focus here on the two assumptions unique to the generalization case.

Meeting Assumption 1, SUTVA ($\mathcal{P}$) can be thought of as meeting a series of smaller assumptions. First, it requires that the potential treatment effect is not a function of being selected into the experiment. This would be violated if the schools in the experiment behaved differently (e.g., changed their curriculum in other ways than SimCalc) as a result of being in the study, and in ways that affected their associated treatment effects. Second, it requires that the version of SimCalc used in the experiment is the same version that would be used in the population. This would be violated, for example, if the version of SimCalc in the experiment—which depends on teacher training, computers, and software—is of a higher dosage than would occur outside the experiment. Finally, it requires that the potential treatment effects are not a function of the proportion of the population receiving the treatment. Problems of this type would occur if the SimCalc developers had difficulties providing adequate support when the program was rolled out on a large scale.

The unconfounded sample selection assumption (Assumption 3) requires that the 26 covariates included in the propensity score model include all of the covariates that explain variation in the potential treatment effects. Certainly, these 26 covariates cover a wide range of student, teacher, and school demographics. If the outcomes used here were also measured in all schools in the population, this information would be helpful for determining if this assumption is met (Stuart et al., 2011); unfortunately this information is not available. Additionally, it is important to note that this list does not include any proximal variables—those regarding school or student achievement and learning as it relates to the particular math concepts addressed by SimCalc—since these are not included in the AEIS system. In order for this assumption to be met, therefore, requires that either potential treatment effects do not vary in relation to these proximal measures or, if they do, that the effect of these proximal measures is accounted for by the other covariates in the model. If an important covariate has been left out of this model, then any propensity score based method will result in some degree of bias.

4.3. Estimating the P_0ATE

In order to estimate the P$_0$ATE, a subclassification estimator with $k = 5$ strata was used. The analysis with outcome data was conducted in the statistical program R using the zelig package (Imai, King, & Lau, 2006), though again any statistical package could have been used. The following regression model was estimated,

$$Y = \beta_1 S_1 + \beta_2 \mathrm{S}_2 + \ldots + \beta_5 \mathrm{S}_5 + \beta_6 \mathrm{S}_1 \cdot \mathrm{TRT} + \ldots + \beta_{10} \mathrm{S}_5 \cdot \mathrm{TRT} + \varepsilon,$$

where $S_1, \ldots, S_5$ are indicators for the five strata, TRT is an indicator for the treatment group, and it is assumed that $\varepsilon \sim N(0, \tau^2)$. By using this model, the stratum specific estimates of τ^2 could be pooled; this was important here since some strata had as few as three or four experimental units, leading to imprecise stratum specific variance estimates. The subclassification estimator used here and its associated variance can be written as

$$T_{\text{sub}} = \sum_{j=1}^{5} w_{p_j} \hat{\beta}_{j+5}, \text{ and } V(T_{\text{sub}}) = \sum_{j=1}^{5} w_{p_j}^2 V\left(\hat{\beta}_{j+5}\right).$$

Note that the weights and strata are defined here such that $w_{pj} = 1/5$ based on the distribution of the propensity scores $s(\mathbf{X})$ in the sub-population $\mathcal{P}_0$. As Table 4 shows, the combined estimate of the treatment effect for schools in $\mathcal{P}_0$ is $T_{\text{sub}} = 1.430$ (0.188), $p < .01$. This estimate is 0.93% smaller than T, while the standard error is 32% larger.

As Table 3 shows, using the subclassification estimator for the P$_0$ATE drastically reduces bias in the propensity scores and covariates. Using this estimator reduces bias by 94% in the propensity scores and 95% in their logits. Furthermore, in terms of specific covariates, only 4 of the 26 have absolute standardized mean differences greater than 0.10, as compared to 17 of the 26 covariates with the conventional estimator. Finally, note that this balance could be further improved by respecifying the propensity model to include higher order terms (squares and interactions).

4.4. Estimating the PATE

In order to estimate the PATE, a subclassification estimator with $k = 5$ strata was used. The same model and estimation process was used as for the P$_0$ATE. The main difference here is that the strata and weights were defined in relation to the distribution of the propensity score $s(\mathbf{X})$ in the *full* population $\mathcal{P}$ instead of the subpopulation $\mathcal{P}_0$. As Table 4 shows, here the combined estimate of the treatment effect for schools in $\mathcal{P}$ is $T_{\text{sub}} = 1.452$ (0.195), $p < .01$. This estimate is 0.64% larger than T, while the standard error is 38% larger. Importantly, note that the standard error is inflated more (38% vs. 32%) for estimation of the PATE than for the P$_0$ATE.

As Table 3 shows, the subclassification estimator reduces bias in the propensity score and the covariates; bias is reduced by 96% for the propensity scores and 68% for their logits. For both the conventional and the subclassification estimators, eight of the covariates have absolute standardized mean differences (|SMD|) between 0.10 and 0.20. However, for the subclassification estimator, only 4 covariates have |SMD|'s greater than 0.20; in comparison this is the case for 11 covariates with the conventional estimator. This indicates that the reweighted experimental sample is generally more similar to the population of interest than the original sample. However, since balance has not been achieved on all covariates, the decision of which estimator to use would depend on substantive knowledge regarding which covariates are more likely to explain variation in treatment effects (Hill, 2008).

5. Conclusions

This article highlights three important issues that arise when using propensity score methods, and the propensity score subclassification estimator in particular, to improve generalizations from experiments. The first issue is that without random sampling, any generalizations require some degree of modeling. When nonrandom sample selection is used and treatment effects vary, as this work shows, using the conventional estimator is akin to assuming that the experiment really is a random sample from the population, even when evidence points to the contrary. Without random sampling, all generalizations require modeling; the virtue of this method is that it makes this model and the requisite assumptions explicit instead of hidden.

Second, this article provides new results for the amount of bias reduction and variance inflation that may be expected when using a subclassification estimator to estimate the average treatment effect for a population, under particular ignorability assumptions. These results indicate that in the generalization context, an estimator with five strata is likely to be 56–98% less biased than the conventional estimator, though typically these values will be between 78% and 96%. Similarly, these results suggest that with five strata the variances are often increased between 4% and 104%, though in many cases (such as the example), these increases are moderate (in the 18–53% range). When fewer strata are used, the bias reduction is smaller, but so too is the variance inflation.

Finally, this article highlights an important problem that arises in generalization: In many situations, the population of interest is not well represented by the experimental sample. We propose two ways to deal with this problem. The first is to find the subpopulation that is well represented (i.e., that meets the ignorability conditions), to estimate an average treatment effect for this subpopulation, and to carefully describe membership in this subpopulation. This solution provides an answer by changing the estimand of interest. In contrast, the second approach proposed is to estimate the average treatment effect for the full population using a subclassification estimator, even though the sampling ignorability condition does not hold. We show that in comparison to the conventional estimator commonly used, a subclassification estimator with five strata: requires a careful discussion of assumptions; is as much as 53–90% less biased; and has sampling variance from 6 to 1200% larger (with typical values between 338% and 359%). We note, however, that it is often the case that bias will be reduced more in some covariates than others and that the decision to use the subclassification or conventional estimator for the full population will generally require substantive knowledge regarding the importance of particular covariates.

Appendix A

Proof of Proposition 2.2: Expected Variance Inflation via Subclassification in Generalization

From the assumptions given note that we can write,

$$\sigma^2 = \sigma_t^2 = \beta_T^2\sigma_s^2 + \sigma_{st}^2 = \beta_T^2\sigma_s^2 + (1-\rho_{st}^2)\sigma^2, \text{ and}$$

$$\sigma^2 = \sigma_c^2 = \beta_C^2\sigma_s^2 + \sigma_{sc}^2 = \beta_C^2\sigma_s^2 + (1-\rho_{sc}^2)\sigma^2,$$

where for each case $\beta^2 = \rho^2\sigma^2/\sigma_s{}^2$, σ^2 is the total variation in the outcomes, β is the linear coefficient relating s and the outcomes, $\sigma_s{}^2$ is the variation in $s(\mathbf{X})$ in the experiment, and ρ is the correlation between the outcomes Y and the variable s. In the experiment it can be shown that,

$$V(T) = V(\bar{Y}_T - \bar{Y}_C) = 4\sigma^2/n,$$

where as the variation of the subclassification estimator can be written,

$$V(T_{\text{sub}}) = \sum_{j=1}^{k} w_{p_j}^2 V\left(\bar{Y}_{T_j} - \bar{Y}_{C_j}\right) = \sum_{j=1}^{k} w_{p_j}^2 \left(\frac{\sigma_{T_j}^2}{n_{T_j}} + \frac{\sigma_{C_j}^2}{n_{C_j}}\right).$$

By Assumption 2, $E(\sigma_{Cj}{}^2) = E(\sigma_{Tj}{}^2) = \sigma_j{}^2$ and $E(n_{Cj}) = E(n_{Tj}) = n_j/2$, and using a first-order Taylor expansion,

$$E(V(T_{\text{sub}})) \approx 2\sum_{j=1}^{k} \frac{w_{p_j}}{n_j}\left\{\left(\beta_C^2 + \beta_T^2\right)\sigma_j^2 + [2 - \rho_{sc}^2 - \rho_{st}^2]\sigma^2\right\}.$$

Next let $E(n_j) = n\, w_{sj}$, where n_j is the observed number of sample units in stratum j and w_{sj} is the proportion expected when $s(\mathbf{X})$ follows a particular distribution. Then using a first-order Taylor expansion,

$$\text{EE}(V(T_{\text{sub}})) \approx \frac{2}{n}\sum_{j=1}^{k} \frac{w_{p_j}^2}{w_{s_j}}\left\{\left(\beta_C^2 + \beta_T^2\right)\sigma_j^2 + [2 - \rho_{sc}^2 - \rho_{st}^2]\sigma^2\right\} = \frac{4\sigma^2}{n}[\text{EVIF}].$$

Acknowledgments

The author thanks Jeremy Roschelle for sharing the SimCalc data used in the example.

Declaration of Conflicting Interests

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

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by grants 0437861 and 1118978 from the National Science Foundation.

References

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*. New York, NY: Russell Sage Foundation.

Bohrnstedt, G. W., & Stecher, B. M. (Eds.). (1999). *Class size reduction in California: Early evaluation findings, 1996-1998*. Palo Alto, CA: CSR Research Consortium, Year 1 Evaluation Report, American Institutes for Research.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295–313.

Cochran, W. G., & Cox, G. M. (1992). *Experimental designs. Wiley classics library*. New York, NY: Wiley.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. *American Journal of Epidemiology*, *172*, 107–115.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them: New directions for program evaluation* (Vol. 57, pp. 39–82). San Francisco, CA: Jossey-Bass.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750. doi:10.1002/pam.20375

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*, 907–949.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of American Statistical Association*, *94*, 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84*, 151–161.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., & Singer, E. (2009). *Survey methodology*. New York, NY: John Wiley.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Hedges, L. V., & O'Muircheartaigh, C. A. (2011). *Improving generalizations from designed experiments. Northwestern University* (Manuscript submitted for publication).

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin. *Statistics in Medicine*, *27*, 2055–2061.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1–28. Retrieved from http://www.jstatsoft.org/v42/i08/

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, *101*, 901–910.

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.

Imai, K., King, G., & Lau, O. (2006). Zelig: Everyone's statistical software. R package version 3.5.3. Retrieved from http://CRAN.R-project.org/package=Zelig

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.

Kish, L. (1987). *Statistical design for research. Wiley series in probability and mathematical statistics*. New York, NY: John Wiley.

Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography*, *1*, 296–315.

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

Nadarajah, S., & Kotz, S. (2006). R programs for computing truncated distributions. *Journal of Statistical Software*, *16*, 1–8.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. Technical Report R-372-A. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, August 7–11, 2011, San Francisco, CA (pp. 247–254). Menlo Park, CA: AAAI Press.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Roschelle, J., Schechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47,* 833–878.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi:10.1093/biomet/70.1.41

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20199225

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. *The Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1980). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472–480.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279–292.

Schneider, B., & McDonald, S. K. (2007). *Scale-up in education: Ideas in principle* (Vol. I). Lanham, MD: Rowman & Littlefield.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological methods*, *15*, 3–17. doi:10.1037/a0015916

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

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: 369–386. doi: 10.1111/j.1467-985X.2010.00673.x.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, *4*, 67–91.

Author

ELIZABETH TIPTON is an Assistant Professor of Applied Statistics in the Department of Human Development at Teachers College, Columbia University, 425 W 120th Street, New York, NY 10027; email: tipton@tc.columbia.edu. Her research interests are in the design and analysis of large-scale randomized experiments and meta-analysis.

Manuscript received May 10, 2011
First revision received September 2, 2011
Second revision received December 5, 2011
Third revision received Jan 18, 2012
Accepted January 30, 2012